Target-Aware Data Augmentation for Stance Detection

Yingjie Li
University of Illinois at Chicago
yli300@uic.edu

Cornelia Caragea
University of Illinois at Chicago
cornelia@uic.edu

Abstract

The goal of stance detection is to identify whether the author of a text is in favor of, neutral or against a specific target. Despite substantial progress on this task, one of the remaining challenges is the scarcity of annotations. Data augmentation is commonly used to address annotation scarcity by generating more training samples. However, the augmented sentences that are generated by existing methods are either less diversified or inconsistent with the given target and stance label. In this paper, we formulate the data augmentation of stance detection as a conditional masked language modeling task and augment the dataset by predicting the masked word conditioned on both its context and the auxiliary sentence that contains target and label information. Moreover, we propose another simple yet effective method that generates target-aware sentence by replacing a target mention with the other. Experimental results show that our proposed methods significantly outperforms previous augmentation methods on 11 targets.

1 Introduction

Nowadays, people often take to social media to express their stances toward specific targets (e.g., political figures or abortion). These stances in an aggregate can provide valuable information for obtaining insight into some important events such as presidential elections. The goal of the stance detection task is to determine from a piece of text whether the author of the text is in favor of, neutral or against toward a specific target (Mohammad et al., 2016; Lin et al., 2019), which indicates that all elements, the sentence, the target, and the label, are used to train a stance detection model. We can further classify the task as single-target stance detection and multi-target stance detection (Küçük and Can, 2020; AlDayel and Magdy, 2020) where we need to detect the stances toward two different targets simultaneously.

Orig.	We all have a duty to protect the sanctity of life.	
G1	We all have a life to protect the sanctity of duty.	✗
G2	We all have a duty to protect the sanctitude of life.	✗
G3	We all have a responsibility to protect the unborn lives.	✓

Table 1: Examples of data augmentation on the target "Legalization of Abortion".

One of the biggest challenges for stance detection tasks is the scarcity of annotated data. Data augmentation (DA) is an effective strategy for handling scarce data situations. However, we face three main obstacles when applying the existing augmentation methods to the stance detection tasks. First, existing augmentation methods do not generalize well, which means some methods are tailored to specific tasks and models, and thus difficult to be extended to the stance detection tasks. Second, consider an original sample that is against to the target "Legalization of Abortion" in Table 1. Using previous augmentation methods we may end up with the first generation example (G_1) that deviates from its original meaning due to the unawareness of target and label information during augmentation. Third, previous augmentation methods could generate the sentence (G_2) with less diversified patterns. To address these issues, we propose an augmentation method that can generate more diversified sentence (G_3) that is consistent with target and label information. Moreover, we expect the proposed method to generalize well to other tasks.

A common data augmentation strategy is based on word replacement. Zhang et al. (2015) augmented a sentence by substituting the replaceable words with synonyms from WordNet (Miller, 1995). However, synonym replacement can only generate limited diversified patterns. Wu et al. (2019) formulated the text data augmentation as a Conditional Masked Language Modeling (C-MLM) task and proposed a Conditional BERT

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1850–1860
June 6–11, 2021. ©2021 Association for Computational Linguistics

(CBERT) where segmentation embeddings of BERT (Devlin et al., 2019) are replaced with the annotated label during augmentation. This method seems to be able to generate label-compatible sentences, yet it does not consider the target information for stance detection. Moreover, CBERT can hardly be extended to other pre-trained language models that do not use segmentation embeddings in inputs, and cannot be applied to the multi-target stance detection due to the inability to encode two stance labels in segmentation embeddings. Wei and Zou (2019) proposed a simple effective method that uses operations such as random deletion or swap to help train more robust model. However, similar to the above methods, it fails to take target information into considerations. Another commonly used strategy for augmentation is back-translation (Yu et al., 2018), however, it is less controllable and may change the target information unpredictably.

Inspired by the recent advances of applying auxiliary sentence to aspect-based sentiment analysis (Sun et al., 2019) and the task of recognising agreement and disagreement between stances (Xu et al., 2019), in this paper, we propose an Auxiliary Sentence based Data Augmentation (ASDA) method that generates target-relevant and label-consistent data samples based on the C-MLM task. Specifically, we fine-tune a pre-trained BERTweet (Nguyen et al., 2020) model through C-MLM task in which the masked word is conditioned on both its context and the prepended auxiliary sentence that contains target and label information. The same task is also adopted in the augmentation stage to generate data samples. Besides, we propose a simple Target Replacement (TR) method that generates target-aware sentence by replacing a target mention in a sentence with the other.

Our contributions include the following: 1) In this paper, we propose a novel data augmentation method called ASDA. As far as we know, this is the first attempt to explore the conditional data augmentation of stance detection. Our proposed ASDA significantly outperforms strong baselines on three different stance detection datasets with 11 targets in total, demonstrating its effectiveness. Experimental results show that prepending auxiliary sentence contributes to the performance gain; 2) We further propose a simple yet effective method called Target Replacement (TR) that achieves highly competitive performance even without fine-tuning during the augmentation; 3) Our proposed ASDA can be also employed on other baseline to help improve the performance, which indicates that ASDA is not tailored to specific model.

2 Related Work

2.1 Stance Detection

Most previous studies on stance detection focused on the detection of stance from text that contains expressions of stance towards one single target, i.e., single-target stance detection. Mohammad et al. (2016) presented the SemEval-2016 dataset that contains 5 independent targets, e.g., Legalization of Abortion and Hillary Clinton. Conforti et al. (2020) constructed WT-WT, a financial dataset on which the task is to detect whether two companies (e.g., Cigna and Express Scripts) will merge or not. Inspired by the attention mechanism (Bahdanau et al., 2015), various target-specific attention-based approaches (Du et al., 2017; Sun et al., 2018; Wei et al., 2018b; Li and Caragea, 2019; Siddiqua et al., 2019; Sobhani et al., 2019) were proposed to connect the target with the sentence representation. Moreover, gated mechanism (Dauphin et al., 2017) and BERT (Devlin et al., 2019) have drawn a lot attention these years and achieved promising performance on aspect-based sentiment analysis (Xue and Li, 2018; Huang and Carley, 2018). We used the models from Du et al. (2017), Huang and Carley (2018) and Devlin et al. (2019) as strong base classifiers for our evaluation.

Sobhani et al. (2017) introduced the multi-target stance detection task and presented the Multi-Target stance dataset. The task is to detect the stances toward two presidential candidates (e.g., Donald Trump and Ted Cruz) simultaneously. They also proposed an attention-based encoder-decoder (Seq2Seq) model that predicts stance labels by focusing on different parts of a tweet. Wei et al. (2018a) proposed a dynamic memory network for detecting stance. We used the above three datasets (Mohammad et al., 2016; Sobhani et al., 2017; Conforti et al., 2020) for our evaluation.

2.2 Text Data Augmentation

One of the main challenges for stance detection tasks is the scarcity of annotated training data, which is costly to obtain. Therefore, data augmentation becomes appealing, particularly when the training models become increasingly large. Generative models are commonly used for data augmentation in previous studies, including variational autoencoders (VAE) (Kingma and Welling, 2014), generative adversarial networks (GAN) (Goodfellow et al., 2014) and pre-trained language generation models (Anaby-Tavor et al., 2020; Li et al., 2020; Kumar et al., 2020). Besides, Sennrich et al. (2016) and Yu et al. (2018) generated the data by using back-translation, which first translates the English sentence into another language (e.g., French) and then translates it back to English.

Another commonly used way for data augmentation is to substitute local words. Zhang et al. (2015) and Wang and Yang (2015) substituted the replaceable words with synonyms from WordNet (Miller, 1995) and Word2Vec (Mikolov et al., 2013), respectively. Kobayashi (2018) proposed a contextual data augmentation method. A bidirectional language model is used to predict the word given the context surrounding the original word. Wu et al. (2019) formulated the text data augmentation as a C-MLM task, retrofitting BERT (Devlin et al., 2019) to predict the masked word based on its context and annotated label. Wei and Zou (2019) boosted the performance on text classification by using simple operations such as random deletion or insertion, and received substantial attention from the research community recently.

However, the augmentation methods mentioned above mostly focus on the sentence-level natural language processing tasks and the resulting augmented sentence can either change the stance toward the given target unexpectedly or generate only limited diverse patterns for stance detection tasks.

3 Methodology

3.1 Problem Formulation

Suppose a given training dataset of size n is D_{train} = $\{(x_i, t_i, y_i)\}_{i=1}^{n}$ where $x_i = [x_i^1, x_i^2, ..., x_i^l]$ is a sequence of l words, t_i is the corresponding target and $y_i \in \{1, ..., c\}$ is the label. The objective of our data augmentation task is to generate an augmented sentence $\hat{x}_i$ that is consistent with the target t_i and label y_i. Note that $t_i = [t_i^1, t_i^2]$ and $y_i = [y_i^1, y_i^2]$ for multi-target stance detection, which makes the augmentation task more challenging.

3.2 Auxiliary Sentence based Data Augmentation

Previous conditional data augmentation methods such as (Wu et al., 2019) could generate target-unaware data samples and cannot be applied to the multi-target stance detection task. In this paper, we propose an Auxiliary Sentence based Data Augmentation (ASDA) method that can generate target-relevant and label-consistent data samples based on the C-MLM task.

3.2.1 Construction of the Auxiliary Sentence

ASDA generates augmented sentence by predicting the masked word that is conditioned on both its context and the auxiliary sentence. We propose the following method to construct the auxiliary sentence.

ASDA: Given a training sample E_1, we prepend both another training sample E_2 with the same target and label as E_1 and the description sentence that contains target and label information to E_1. The complete sentence is: *The authors of the following tweets are both [Label] [Target]. The first tweet is:* E_2*. The second tweet is:* E_1.

The sentences before E_1 are the auxiliary sentences we construct. “Target” and “Label” are the target name and stance label with regard to the given training sample. E_2 that contains the same target and stance label with E_1 is sampled from the training dataset. Specifically, suppose we are given a training example in the SemEval-2016 dataset: *We all have a duty to protect the sanctity of life.* Target: *Legalization of Abortion*; Label: *Against*. We can have the following masked words and auxiliary sentences in fine-tuning or augmentation stage: *The authors of the following tweets are both against to legalization of abortion. The first tweet is: Every human life is worth the same, and worth saving. The second tweet is: We all have a [MASK] to protect the [MASK] of life.* Target: *Legalization of Abortion*; Label: *Against*. With the auxiliary sentence, the masked word is not only conditioned on its context in the second tweet, but also conditioned on the first tweet of same target “Legalization of Abortion” and label “Against”.

We expect the agreement between stances to benefit the data augmentation by adding a reference sentence E_2. The introduction of the E_2 not only generates more diversified samples for fine-tuning the pre-trained language model, but also provides a strong guideline to help generate target-relevant and label-compatible sentences in the augmentation stage. Moreover, ASDA is not tailored to specific model because it does not rely on the model architecture, and thus can be easily extended to different language models.

3.2.2 Conditional DA using BERTweet

BERTweet (Nguyen et al., 2020) is a large-scale language model pre-trained on 850M English tweets. BERTweet follows the training procedure of RoBERTa (Liu et al., 2019) and uses the same model configuration with BERT-base (Devlin et al., 2019). We fine-tune the pre-trained BERTweet via C-MLM on stance detection tasks. The fine-tuning step is summarized in Algorithm 1. Note that words of auxiliary sentence A are never masked (see Algorithm 1 lines 4-6) because we want to preserve all target and label information.

Algorithm 1: Fine-tuning

```
   Input : Training dataset $D_{train}$
           Total training steps $S$
           Auxiliary sentence $A$
           Batch size $B$
           Language model $M$
           Proportion of sentence to mask $p$
1  for each i = 1, 2, ..., S do
2      $Batch_i = \emptyset$
3      for each j = 1, 2, ..., B do
4          Randomly sample a sentence $s$ from $D_{train}$
5          Randomly mask words of $s$ with probability $p$ to obtain $s_m$
6          Prepend the auxiliary sentence $A$ that contains corresponding target
           and label information to the $s_m$ to obtain $\hat{s}$
7          $Batch_i = Batch_i \cup \{\hat{s}\}$
8      end
9      Fine-tune the language model $M$ with $Batch_i$
10 end
11 return M
```

After fine-tuning the BERTweet on the training dataset for a few epochs, we use the well-trained model for augmentation. Similar to the fine-tuning procedure as shown in Algorithm 1, for a training sentence s from D_{train}, we randomly mask words in s and prepend the corresponding auxiliary sentence A to obtain the masked sentence $\hat{s}$. Then, the BERTweet model is used to predict the masked words and we repeat these steps over all training data to get $\hat{D}_{train}$. The above steps can be implemented multiple times with different masked positions, and hence, different augmented samples can be generated from the original training dataset. Finally, we merge the D_{train} with $\hat{D}_{train}$ and perform classification task on this combined dataset.

3.3 Target Replacement Method

Besides ASDA, we propose a Target Replacement (TR) method to increase the size of the training set by replacing a target mention in a sentence with the other, which improves model robustness so that meaningful lexical patterns are learned by the model instead of learning undesirable correlation between a target and its contexts. In case a target is mentioned more than once, we continue to replace the target until all targets are replaced. Hashtags and mentions that contain target information (e.g., #Cigna) are also considered for replacement. Consider the following example in single-target stance detection: *#CI Shareholders vote to approve merger Cigna and Express Scripts.* Target: *Cigna and Express Scripts*; Label: *Support*. After applying TR, we have: *#ESRX Shareholders vote to approve merger Express Scripts and Cigna.* Target: *Cigna and Express Scripts*; Label: *Support*. CI and ESRX represent Cigna and Express Scripts, respectively.

TR can be also applied to the multi-target stance detection with minor changes. Consider the following example: *#Cruz supporters want people to think his words alone are good enough. #DonaldTrump has created jobs and businesses we need in this country.* Target1: *Donald Trump*; Target2: *Ted Cruz*; Label1: *Favor*; Label2: *Against*. TR could potentially generate contradictory content with the labels if we only replace the target mentions since the task is to detect the stances toward two different targets simultaneously. Therefore, we replace the target mentions and swap the stance labels for multi-target stance detection. Consider the same example as above after applying the target replacement and label swap: *#DonaldTrump supporters want people to think his words alone are good enough. #Cruz has created jobs and businesses we need in this country.* Target1: *Donald Trump*; Target2: *Ted Cruz*; Label1: *Against*; Label2: *Favor*.

4 Experiments

In this section, we first describe three stance detection datasets used for evaluation and several baseline methods of data augmentation and stance detection. Then, we introduce the evaluation metrics and report the experimental results.

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
Total	2,914	25.84	47.87	26.29	1,249	24.34	57.25	18.41

Table 2: Data distribution of SemEval-2016 dataset (Mohammad et al., 2016).

Target	#Total	%Refute	%Comment	%Support	%Unrelated
Cigna and Express Scripts	2,527	10.01	37.47	30.58	21.92
Aetna and Humana	7,897	14.00	35.50	13.14	37.34
CVS Health and Aetna	11,622	4.45	47.49	21.24	26.80
Anthem and Cigna	11,044	17.82	28.05	8.78	45.33
Total	33,090	11.62	37.38	15.87	35.13

Table 3: Data distribution of WT-WT dataset (Conforti et al., 2020).

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 4: Distribution of instances in Multi-Target stance dataset (Sobhani et al., 2017).

4.1 Datasets

Three stance detection datasets, the SemEval-2016 dataset (Mohammad et al., 2016), the WT-WT financial dataset (Conforti et al., 2020) and the Multi-Target election dataset (Sobhani et al., 2017), are used to evaluate the performance of augmentation methods. The SemEval-2016 dataset and WT-WT dataset are both single-target stance datasets and the third dataset is a multi-target stance dataset, which contains stances toward two targets in each tweet. Summary statistics of three datasets are shown in Tables 2, 3, 4, respectively.

SemEval-2016 SemEval-2016 is a benchmark dataset containing five different targets: “Atheism”, “Climate Change is a Real Concern”, “Feminist Movement”, “Hillary Clinton” and “Legalization of Abortion”. The dataset is annotated for detecting whether the author is against to, neutral or in favor of a given target. We split the train set in a 5:1 ratio into train and validation sets and removed the target “Climate Change” because of the limited and highly skewed data. The test set of each target is the same as provided by the authors.

WT-WT WT-WT is a financial dataset and the task aims at detecting the stance toward mergers and acquisition operations between companies. This dataset consists of four target pairs in the healthcare domain and each data is annotated with four labels (refute, comment, support and unrelated). We split the dataset in a 10:2:3 ratio into train, validation and test sets.

Multi-Target Multi-Target stance dataset consists of three sets of tweets corresponding to target pairs: Donald Trump and Hillary Clinton, Donald Trump and Ted Cruz, Hillary Clinton and Bernie Sanders. The task aims at detecting the stances (against, none or favor) toward two targets for each data. We used the train, validation and test sets as provided by the authors.

4.2 Baseline Methods

We compare the proposed augmentation methods with the following baselines:

- Synonym Replacement (SR): A data augmentation method that randomly replaces words with their synonyms from WordNet.
- EDA (Wei and Zou, 2019): A simple data augmentation method that consists of four operations: synonym replacement, random deletion, random swap and random insertion.
- BT (Yu et al., 2018): A back-translation method that first translates the English sentence into French and then translates back to English.
- CBERT (Wu et al., 2019): A C-MLM method that generates label-compatible words by replacing the segmentation embeddings of BERT with label embeddings.

Three base classifiers are used to evaluate the performance of different augmentation methods:

- PGCNN (Huang and Carley, 2018): A parameterized convolutional neural network that uses target-sensitive filters and gated mechanism to incorporate the target information.
- TAN (Du et al., 2017): An attention-based LSTM model that extracts target specific features.

Method	Atheism	Feminist	Hillary	Abortion	avgF$_1$
PGCNN	68.08	55.31	61.51	67.26	63.04
+SR	66.67	56.07	61.12	67.14	62.75
+EDA	66.43	55.26	62.06	66.07	62.46
+BT	67.52	56.01	61.70	65.18	62.60
+CBERT	66.36	57.24	61.97	66.19	62.94
+TR	-	-	-	-	-
+CBERT-ASDA	**68.70**‡	56.74	**64.97**‡	**68.77***‡	**64.80**
+ASDA-base	65.61	**58.11**	62.10	66.75	63.14
+ASDA	**69.78***†	**59.03***	**63.62**	**68.44***†	**65.22**
TAN	59.27	56.45	56.58	59.29	57.90
+SR	63.49	55.76	56.61	59.52	58.85
+EDA	64.86	57.23	55.35	59.69	59.28
+BT	65.36	57.37	58.95	58.45	60.03
+CBERT	65.43	57.27	59.39	60.18	60.57
+TR	-	-	-	-	-
+CBERT-ASDA	**66.60**	**58.59**‡	**61.02**	**62.72***‡	**62.23**
+ASDA-base	63.40	57.55	56.03	60.96	59.49
+ASDA	**68.47***†	**58.73**†	**60.09**†	**63.66***†	**62.74**
BERT	70.69	54.57	66.06	57.80	62.28
+SR	**72.78**	54.85	66.20	59.42	63.31
+EDA	71.69	53.82	65.51	60.18	62.80
+BT	72.49	54.98	66.74	61.91	64.03
+CBERT	71.88	55.79	64.29	**62.36**	63.58
+TR	-	-	-	-	-
+CBERT-ASDA	**74.93***‡	**56.43**	**67.01**‡	61.66	**65.01**
+ASDA-base	70.67	54.18	64.67	61.34	62.72
+ASDA	72.30†	**55.97**†	**68.09**†	**63.02**†	**64.85**

Table 5: Performance comparisons of applying different augmentation methods to the base model on the SemEval-2016 stance dataset. $*$: the proposed methods improve the best baseline at p $<$ 0.05 with paired t-test. †: ASDA improves the ASDA-base at p $<$ 0.05 with paired t-test. ‡: CBERT-ASDA improves the CBERT at p $<$ 0.05 with paired t-test. $avgF_1$ is the average of all target pairs.

- 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 various stance detection tasks.

The proposed methods are listed as follows:

- Target Replacement (TR): A method that replaces target words with the other.
- CBERT-ASDA: The CBERT that uses our proposed auxiliary sentences during fine-tuning and augmentation.
- ASDA-base: A variation of ASDA that only prepends the description sentence to the given training sample. The complete sentence is: *The author of the following tweet is [Label] [Target].* E_1.
- ASDA: The full method that uses both description and reference sentences as auxiliary sentences during fine-tuning and augmentation.

4.3 Evaluation Metric and Hyperparameters

F_{avg} is adopted to evaluate the performance of the proposed model. First, the F1-score of label "Favor" and "Against" is calculated as follows:

$$F_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor} + R_{favor}} \tag{1}$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}} \tag{2}$$

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

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \tag{3}$$

We calculate the F_{avg} for each target. The same evaluation metric was used in SemEval-2016 dataset and Multi-Target stance datasets. To be consistent with the previous work, we evaluate the performance of augmentation methods on WT-WT dataset by using the same evaluation metric F_{avg}, which is calculated by averaging the F1-scores of label "Support" and "Refute". Moreover, we get $avgF_1$ by calculating the average of F_{avg} across all targets for each dataset.

We use the pre-trained uncased BERTweet model for fine-tuning and augmentation under the PyTorch framework. When fine-tuning, the batch

Method	CI_ESRX	AET_HUM	CVS_AET	ANTM_CI	avgF$_1$
PGCNN	71.34	77.43	73.73	71.70	73.55
+SR	71.96	77.25	73.54	71.78	73.63
+EDA	70.97	77.28	73.85	71.90	73.50
+BT	71.57	77.59	74.17	71.56	73.72
+CBERT	71.57	77.31	73.53	71.59	73.50
+TR	**73.51**	77.85	**75.42***	**72.57***	**74.84**
+CBERT-ASDA	73.02‡	**78.65***‡	74.30*‡	72.19	74.54
+ASDA-base	71.61	77.97	73.77	72.14	73.87
+ASDA	**74.25***†	**78.36***	**74.63**†	**72.63***	**74.97**
TAN	68.39	76.06	69.83	68.72	70.75
+SR	67.88	75.46	70.37	69.02	70.68
+EDA	68.02	75.40	69.86	69.06	70.59
+BT	67.69	75.19	70.57	67.99	70.36
+CBERT	68.55	75.75	70.89	68.88	71.02
+TR	68.02	75.85	69.66	69.10	70.66
+CBERT-ASDA	**70.40***‡	**76.35**	**71.50**	**69.87***‡	**72.03**
+ASDA-base	67.19	76.29	70.83	69.32	70.91
+ASDA	**70.13***†	**77.53***	**71.73***†	**70.18***	**72.39**
BERT	71.12	78.47	75.28	74.11	74.75
+SR	73.24	78.57	75.65	73.80	75.32
+EDA	73.44	78.51	75.85	73.98	75.45
+BT	72.30	77.47	75.97	73.80	74.89
+CBERT	72.83	77.99	75.49	73.66	74.99
+TR	74.17	78.80	76.30	74.24	75.88
+CBERT-ASDA	**74.58**‡	**78.95**‡	**76.46**‡	**74.46**‡	**76.11**
+ASDA-base	72.49	78.76	76.09	74.01	75.34
+ASDA	**75.45***†	**78.99**	**76.41***	**74.48***	**76.33**

Table 6: Performance comparisons of applying different augmentation methods to the base model on the WT-WT stance dataset. $*$: the proposed methods improve the best baseline at p $<$ 0.05 with paired t-test. †: ASDA improves the ASDA-base at p $<$ 0.05 with paired t-test. ‡: CBERT-ASDA improves the CBERT at p $<$ 0.05 with paired t-test. $avgF_1$ is the average of all target pairs.

size is 32, maximum sequence length is 128, learning rate is 2e-5 and proportion of sentence to mask is 15%. For classification, we train our PGCNN and TAN models using a mini-batch of 128 and the learning rate of Adam optimizer (Kingma and Ba, 2015) is 1e-3. Maximum sequence length is 50 and word vectors are initialized using fastText embeddings (Bojanowski et al., 2017) with dimension 300. For BERT classifier, we fine-tune the pre-trained BERT to predict the stance by appending a linear classification layer to the hidden representation of the $[CLS]$ token. The maximum sequence length is set to 128 and the learning rate is 2e-5.

4.4 Experimental Results

We generate one augmented sentence for each training data, doubling the original train set in size for fair comparison. Experimental results on SemEval-2016, WT-WT and Multi-Target datasets are shown in Tables 5, 6 and 7, respectively. Bold scores are best two results for each classifier. Each result is the average of ten runs with different initializations. Since CBERT and TR cannot be applied to the Multi-Target and SemEval-2016 datasets, respectively, we didn't report the results of these methods.

First, we can observe that our proposed ASDA performs the best in avgF$_1$ on almost all datasets. Moreover, ASDA has better performance than ASDA-base on all targets, demonstrating the effectiveness of adding reference sentences. Second, CBERT can be only used in single-target stance detection tasks due to the segmentation embeddings. In contrast, ASDA-base that achieves similar performance with CBERT can be applied to all datasets, which indicates that constructing auxiliary sentence contributes to the C-MLM task. Third, Tables 5 and 6 show that constructing the auxiliary sentence can not only perform well on the BERTweet model, but also help improve the baseline CBERT, indicating that our proposed method is not tailored to specific masked language model. Fourth, TR achieves promising improvements on WT-WT and Multi-Target datasets, outperforming the EDA in the average of avgF$_1$ on three classifiers by 0.61% and 1.54%, respectively. Further comparison between TR and Random Swap of EDA is discussed later in this section. At last, we can observe that improvements brought by the baselines are limited on three datasets, verifying that target-based stance detection tasks are more challenging.

Method	Tr-Cl	Tr-Cr	Cl-Sa	avgF$_1$
PGCNN	57.38	52.73	50.00	53.37
+SR	57.95	53.42	50.97	54.11
+EDA	57.57	51.99	50.56	53.37
+BT	57.94	52.33	**51.29**	53.85
+CBERT	-	-	-	-
+TR	**59.86***	**56.06***	50.32	**55.41**
+ASDA-base	57.97	53.65	50.62	54.08
+ASDA	**59.98***†	**54.67***	**52.66***†	**55.77**
TAN	56.26	53.25	50.13	53.21
+SR	56.62	53.73	50.17	53.51
+EDA	55.55	54.12	50.11	53.26
+BT	56.31	52.61	50.83	53.25
+CBERT	-	-	-	-
+TR	**58.07***	**56.57***	48.38	54.34
+ASDA-base	57.45*	55.45	**50.99**	**54.63**
+ASDA	**57.76***	**56.44***†	**52.92***†	**55.71**
BERT	54.87	54.32	**53.49**	54.23
+SR	53.99	52.97	52.72	53.23
+EDA	55.46	52.69	53.21	53.79
+BT	54.81	53.95	53.30	54.02
+CBERT	-	-	-	-
+TR	**61.83***	**55.23**	48.85	**55.30**
+ASDA-base	52.98	53.27	51.80	52.68
+ASDA	**56.31**†	**55.48**	**53.72**†	**55.17**

Table 7: Performance comparisons of applying different augmentation methods to the base model on the Multi-Target stance dataset. $*$: the proposed methods improve the best baseline at p $<$ 0.05 with paired t-test. $\dagger$: ASDA improves the ASDA-base at p $<$ 0.05 with paired t-test. $avgF_1$ is the average of all target pairs.

We further explore the effect of the auxiliary sentence by comparing the proposed ASDA with other Prepending based Data Augmentation (PDA) (Schick and Schütze, 2020; Kumar et al., 2020) in which no description sentence is constructed and the complete sentence is: *[Label] [Target]* E_1. Moreover, we consider the reference sample E_2 as mentioned in Section 3.2.1 for PDA and the complete sentence is *[Label] [Target]* E_2 E_1. Comparison results on SemEval-2016 dataset are shown in Table 8. We can observe that both ASDA and PDA-ASDA show better performance over their base models, which indicates that the reference sentence contributes to the performance improvement and our proposed method is not tailored to specific auxiliary sentence.

We compare the proposed methods with other augmentation methods in Table 9. We can observe that both ASDA and TR consider the target information during augmentation. However, TR cannot be applied to SemEval-2016 dataset because unlike WT-WT dataset that corresponds to the merger of two target companies, only single target is available in SemEval-2016 dataset.

Method	Atheism	Feminist	Hillary	Abortion
PGCNN	68.08	55.31	61.51	67.26
+PDA	65.82	56.36	61.66	65.80
+PDA-ASDA	66.77	57.34	**65.50**	66.36
+ASDA-base	65.61	58.11	62.10	66.75
+ASDA	**69.78**	**59.03**	63.62	**68.44**
TAN	59.27	56.45	56.58	59.29
+PDA	65.16	57.09	58.76	58.21
+PDA-ASDA	67.36	**58.85**	**60.98**	60.80
+ASDA-base	63.40	57.55	56.03	60.96
+ASDA	**68.47**	58.73	60.09	**63.66**
BERT	70.69	54.57	66.06	57.80
+PDA	70.35	55.27	66.50	61.11
+PDA-ASDA	**72.93**	**56.04**	66.54	59.37
+ASDA-base	70.67	54.18	64.67	61.34
+ASDA	72.30	55.97	**68.09**	**63.02**

Table 8: Performance comparisons of applying augmentation methods with different auxiliary sentences to the base model on the SemEval-2016 dataset.

Method	Target-aware	All datasets	Require FT
SR	✗	✓	✗
EDA	✗	✓	✗
BT	✗	✓	✗
CBERT	✗	✗	✓
TR	✓	✗	✗
ASDA	✓	✓	✓

Table 9: Overall method comparisons on the stance detection. "Target aware" means the method is aware of target information during augmentation. "All datasets" means the augmentation method can be applied to all three stance detection datasets. "Require FT" means the method requires fine-tuning before augmentation.

Random Swap is an augmentation method that randomly chooses two words in the sentence and swaps their positions. However, Random Swap can potentially generate augmented sentences that contain contradictory content with the labels. Since TR shares similar features with Random Swap by swapping the target mentions in some cases, we compare our proposed TR with Random Swap on WT-WT and Multi-Target datasets in Table 10. The results show that TR achieves better performance on 6, 5 and 4 targets for PGCNN, TAN and BERT, respectively, demonstrating the effectiveness of this method. Note that TR does not perform well on the target pair Clinton-Sanders; one possible reason is that there is more target-related information in this target pair. Since only target words (e.g., "Hillary Clinton") are swapped in TR, target-related words like "feminism" and "Benghazi" still appear in the same position in the generated sentence, which may lead to the inconsistency of target information.

5 Case Study

In this section, we present several augmented examples in Table 11 to show the effectiveness of our

Method	**Tr-Cl**	**Tr-Cr**	**Cl-Sa**	**avgF$_1$**	**CI_ESRX**	**AET_HUM**	**CVS_AET**	**ANTM_CI**	**avgF$_1$**
PGCNN									
+RS	57.98	51.32	**51.45**	53.58	71.05	77.37	74.33	71.12	73.47
+TR	**59.86**	**56.06**	50.32	**55.41**	**73.51**	**77.85**	**75.42**	**72.57**	**74.84**
TAN									
+RS	56.45	52.89	**49.30**	52.88	67.19	74.20	**70.87**	68.22	70.12
+TR	**58.07**	**56.57**	48.38	**54.34**	**68.02**	**75.85**	69.66	**69.10**	**70.66**
BERT									
+RS	55.14	53.57	**53.46**	54.06	73.07	**78.80**	**76.45**	73.98	75.58
+TR	**61.83**	**55.23**	48.85	**55.30**	**74.17**	**78.80**	76.30	**74.24**	**75.88**

Table 10: Performance comparison between EDA Random Swap (RS) and the proposed Target Replacement (TR) on the Multi-Target stance dataset and WT-WT stance dataset. $avgF_1$ is the average of all target pairs.

Target:	Feminist Movement.
Source:	What do feminists want: all humans, male and female, should have equal political, economic and social rights.
EDA:	What do **libber** want: all humans, male and female, should have equal political, economic and social rights.
BT:	What **the** feminists want: all humans, **men** and **women** should have **the same** political, economic and social.
CBERT:	What do feminists want: all humans, male and female, **will** have **had what.** economic and social**.**.
ASDA:	What **real** feminists want: all humans, male and female, **to** have equal political **rights and equal** social rights.
Target:	Cigna and Express Scripts.
Source:	Cigna stockholders greenlight merger with Express Scripts.
EDA:	Cigna stockholders greenlight ~~**merger**~~ with Express Scripts.
BT:	Cigna **merger to shareholders with GreenLight** Express Scripts.
CBERT:	Cigna stockholders **relight** merger with Express Scripts.
TR:	**Express Scripts** stockholders greenlight merger with **Cigna**.
ASDA:	Cigna stockholders **vote for** merger with Express Scripts.
Target:	Donald Trump and Ted Cruz.
Source:	Make america great again!! No socialist/liberals. Principals that made this country great can make it great again! Trump Cruz
EDA:	Make america great again!! No **Cruz**/liberals. Principals that made this country great can make it great again! Trump **socialist**
BT:	**Do it** again !! **Not great america liberal socialist /**. **Managers who have made this great country can do much more**! Trump Cruz
TR:	Make america great again!! No socialist/liberals. Principals that made this country great can make it great again! **Cruz Trump**
ASDA:	Make america great again!! No **democrats**/liberals. **The people who** made this country great can make **america** great again! Trump Cruz

Table 11: Examples generated by the augmentation methods. Texts in bold represent generated words.

proposed methods. Synonym Replacement, Random Deletion and Random Swap of EDA are applied to the targets “Feminist Movement”, “Cigna and Express Scripts” and “Donald Trump and Ted Cruz”, respectively. We can observe that the generated words of ASDA and TR are more consistent with the target and label information. In contrast, the augmented words of baseline methods especially EDA could be incompatible with the labels of the original sentences.

6 Conclusion

In this paper, we presented two data augmentation methods, called ASDA and TR, for stance detection. Different from the existing augmentation methods that are either unaware of target information or hard to be applied to different stance detection tasks, ASDA performs better in generating target-relevant and label-compatible sentences and can be easily applied to various tasks. Results show that ASDA can not only achieve best performance on BERTweet model but also help improve the existing augmentation method such as CBERT. Unlike other rule-based word replacement methods that may produce undesirable correlation between a target and its contexts, TR replaces a target mention with the other, generating qualified sentences with meaningful lexical patterns. In addition, both ASDA and TR will be applicable if we need to detect the stances toward more than two targets simultaneously in the future.

Future work includes extending the proposed methods to various directions, e.g., argument mining, aspect-based sentiment analysis and hate-speech detection, and generating more diversified samples through conditional generation.

Acknowledgments

This work is partially supported by the NSF Grants IIS-1912887 and IIS-1903963. We thank our reviewers for their insightful comments.

References

Abeer AlDayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *arXiv preprint arXiv:2006.03644*.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7383–7390.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

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

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 933–941.

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.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, page 2672–2680.

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.

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

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

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

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066.

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.

Junjie Lin, Qingchao Kong, Wenji Mao, and Lei Wang. 2019. A topic enhanced approach to detecting multiple standpoints in web texts. *Information Sciences*, 501:483–494.

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*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, page 3111–3119.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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

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.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few-shot text classification and natural language inference. *Computing Research Repository*, arXiv:2001.07676.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. 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 1868–1873.

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.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2019. Exploring deep neural networks for multitarget stance detection. *Computational Intelligence*, 35(1):82–97.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. 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 380–385.

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.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388.

Penghui Wei, Junjie Lin, and Wenji Mao. 2018a. Multi-target stance detection via a dynamic memory-augmented network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1229–1232.

Penghui Wei, Wenji Mao, and Daniel Zeng. 2018b. 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.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising agreement and disagreement between stances with reason comparing networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4665–4671.

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.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 649–657.