AABSTRACT  Stance detection on tweets aims at classifying the attitude of tweets towards given targets. Existing work leverage attention-based models to learn target-aware stance representations. While those methods achieve substantial success, most of them usually train a model for each target separately despite the scarcity of annotated data for each target. To alleviate limitation of annotated data, some methods turn to external linguistic resources, additional sentiment annotations or target-aware data augmentation techniques for better detection results. We argue that the sharedness of stance-related features across targets in the existing stance detection dataset is not fully exploited. However, directly training on mixed examples of all targets may confuse the model in learning best features for each target. To this end, we borrow the idea from transfer learning and multi-task learning, and formulate stance detection on tweets as a multi-domain multi-task classification problem. We apply the target adversarial learning to capture stance-related features shared by all targets and target descriptors for learning stance-informative features correlating to specific targets. Experimental results on the benchmark SemEval 2016 dataset demonstrate the effectiveness of our model, which outperforms BERT model by over 2% on macro average F1 and achieves superior performance than many recent methods utilizing external resources. We further provide detailed analyses to illustrate the superiority of fully utilizing features shared by different targets.

INDEX TERMS artificial intelligence, natural language processing, sentiment analysis

I. INTRODUCTION
Stance detection [1, 2] is the task of identifying the attitude conveyed in the text towards a given target. It is useful for various opinion mining tasks, especially important for understanding peoples’ opinions expressed towards targets of interest on social media platforms. In this work, we focus on detecting stances on tweets. Unlike conventional aspect-level sentiment classification, the target may not appear in the text. In consequence, it is crucial to decide the relationship between the opinionated entity and the given target to accurately understanding the stance. For instance, in Figure 1, none of the words in the target appears in the tweet and we need to infer that the tweet is in favor of the target as it uses the phrase like have rights. Moreover, people tend to use different expressions towards different targets, which makes it difficult for a single model to learn diverse stance features.

Various models have been proposed for stance detection to tackle those challenges. Some [3, 4] have leverage models with intense target-text interactions to learn target-aware representations. Some [5, 6] utilized external knowledge bases with rich commonsense information to build connections between targets and mentioned entities in tweets. While those models greatly advance the performance of stance detection, most of them usually train separate models for each target and intensify the scarcity of annotated data available for each

FIGURE 1: An example of stance detection on tweets. Here, the target does not appear in the tweet. And the tweet expresses a favor attitude towards the target.
target. Recent work from [27] proposed a target-aware data augmentation method to alleviate this issue. Others [8, 9] tried to leverage training data from external stance detection datasets. However, most methods ignore the sharedness of stance features across different targets and such features in existing datasets are not fully exploited.

To remedy this, we train a single model for all targets in the hope of capturing stance-related patterns shared by different targets. Since people may use different expressions and opinionated entities when referring to different targets, a good model should be capable of disentangling such distinctions among targets. To this end, we propose S-MDMT (Stance Detection as Multi-Domain Multi-Task Classification), which treats different targets as different domains for the first time and employs sophisticated multi-domain text classification models with a target-aware design. Given a <tweet, target> pair, S-MDMT first represents the pair with a BERT encoder. Then it applies a shared encoder with a target-adversarial layer to learn stance features shared across targets. Meanwhile, it adopts target-descriptors and a target-aware attention structure to learn stance features specific to different targets. Then the target-invariant and target-specific stance features are aggregated for final stance classification.

We conduct extensive experiments on the SemEval 2016 task 6 dataset [1]. S-MDMT outperforms BERT by 2% in terms of average macro-f1 of Favor and Against class. Visualization of shared and private features from S-MDMT proves that our model captures stance features shared by different targets. Ablation study further verifies that these shared features help improve the performance.

In summary, the contributions of this work are three-fold:

- We novelly treat the task of stance detection on tweets as a multi-domain multi-task learning problem to fully utilize stance features shared by different targets in the existing dataset.

- The proposed S-MDMT model achieves new state-of-the-art results, outperforming many state-of-the-art methods that use any external resources, multiple stance detection datasets or data augmentation methods.

- We provide in-depth analysis of shared and private features of different targets on the stance detection task.

II. RELATED WORK

A. STANCE DETECTION

Recently, there is a growing interest in detecting stances of texts on microblog platforms [11, 10, 11, 12]. Unlike traditional sentiment analysis tasks, stance detection is more challenging as the target may not appear in the text and the relations between a target and an opinionated entity in the tweet is harder to infer. Existing work have studied various settings of stance detection on tweets. [11] proposed the single target stance detection task, where for a given tweet, only one target is considered at one time for stance prediction. As multi-target stance detection is quite useful for scenarios including political election and brand comparisons, some researchers [13, 14] introduced new datasets for multi-target stance detection, where multiple targets are given to a tweet. Researches in [15, 16, 17] further propelled this task by modeling dependency among different targets. Besides, some [18, 19, 20] studied the problem of cross-target stance detection with external resources like knowledge bases and lexicons, and neural models like graph networks. Recently, Allaway and McKeown [21] proposed the task of zero-shot stance detection and [22] further studied this problem with adversarial learning. Meanwhile, many [8, 9] tried to leverage training data from multiple stance detection datasets simultaneously and achieved better results than single dataset learning. While most datasets only contain sentence-level annotations, Ye et al. [23] proposed a token-level stance detection dataset with 2025 labeled tweets for fine-grained analysis of stance detection models.

In this work, we focus on the single target stance detection task and adopt the dataset from [11] as the testbed to evaluate our method. Under this setting, various target-aware approaches [24, 25] were proposed to integrate target semantics into final text representations. For example, BiCond [23] initialized the cell state of LSTM that modeled tweets with cell states of the LSTM that modeled targets. TAN [24] applied target-specific attention to each word in tweets and learned a weighted sentence representation concerning the given target. Research in [4] adopted an attention-based ensemble model to learn long-term dependency in the stance detection task. Ghosh et al. [26] improved existing models with better tweet-specific pre-processing.

Recent research work in [5] and [6] utilized external knowledge bases like ConceptNet and Wikipedia respectively to tackle the issue of absence of targets in tweets and acquired better performance. HAN [3] employed a hierarchical attention network that incorporated extra linguistic features. As tweets expressing stance are unlikely to sentiment-neutral, AT-JSS-Lex [27] incorporated additional sentiment annotations to facilitate stance detection using a multi-task learning manner. Since annotated data for each target are limited, Li et al. [27] also proposed to use target-aware data augmentation methods to enrich the training set. MeLT [28] leveraged additional user history messages by introducing a hierarchical message-transformer. Different from those work, to remedy the scarcity of labeled data, we fully exploit the existing datasets by mining shared features across targets instead of using additional annotations or augmented data. We employ the strong BERT [29] model to learn target-aware tweet representation by taking advantage of the BERT-Pair input format.

B. MULTI-DOMAIN AND MULTI-TASK LEARNING

Domain adaptation techniques have been extensively studied in sentiment analysis tasks including sentence-level cross-domain and multi-domain sentiment classification, aspect-level sentiment classification [30, 31, 32, 33]. These work usually deal with scenarios where different domains have different sizes of labeled data (sometimes even no labeled data for training). Various models had applied shared-private
architectures where shared part learns features invariant across domains and private part learns features belonging to specific domains. Gradient adversarial layer [34] had been employed to learn domain-agnostic representations by reversing the gradients of domain classification and the learned representations had been shown to have less domain-private features. Besides, [33] had applied mixture-of-experts to learn features only shared by some domains and [35] allowed instance-wise domain-similarity adaptation. We novelly treat stance detection on tweets as a multi-domain fine-grained text classification problem and solve it in a multi-task learning manner. In this way, we could make full use of existing techniques for domain adaptations in sentiment analysis. We incorporate these structures into a BERT-based model and applied to the stance detection task, which has not been studied by previous work.

III. MODEL

In the task of stance detection on tweets, people tend to use different opinion expressions and refer to distinct entities when expressing stance towards specific targets. Therefore, we could virtually treat tweets talking about different targets as tweets that come from different domains and borrow the idea of multi-domain and multi-task learning to tackle this task.

Figure 2 portrays an overview of the proposed model S-MDMT. It contains two parts: the shared part for learning target-agnostic stance features and the private part for capturing target-specific stance features. In the shared part, we only extract features from tweets and apply target adversarial learning to acquire target-invariant stance features. In the private part, we combine target and tweet to learn target-aware stance representations through multiple layers of self-attention. The stance-related features from shared and private parts are then concatenated for deciding final stance orientation.

A. PROBLEM DEFINITION AND NOTATIONS

Given a target $TA$ and a tweet $TW$, the task of stance detection aims to identify the attitude $ys$ of tweet towards the target as Favor, Against or None. Specifically, we have a total of $K$ targets and for each target we only have limited training data. The goal of this work is to overcome the scarcity of annotated data for each target by leveraging shared stance features across various targets.

B. BERT AS TEXT ENCODER

Our model use BERT [29] to acquire contextualized representations of target and tweet for further stance modeling. BERT build on Transformer networks [36] is widely adopted in sentiment analysis tasks. BERT is pre-trained by predicting masked tokens in the input (MLM task) and discriminating whether two sentences are consecutive (NSP task). The MLM task allows BERT to learn semantics for a token with both left and right context. And the NSP task enables BERT to infer the relationship between an input sentence pair. Downstream tasks usually exploit BERT by fine-tuning pre-trained models on corresponding datasets. Given an text snippet $X = \{x_1, x_2, \ldots, x_T\}$ of $T$ tokens, BERT encodes its semantics with multiple-layers of self-attention interactions. The hidden vector of [CLS] from the last layer is considered
as the representation of the whole text. Existing work [22] [37] utilized the input format of NSP task and extended it into tasks including machine reading comprehension and sentiment analysis. For example, [37] transformed the input format of aspect-based sentiment classification to be compatible with that of NSP. Following [37], in Section III-D we also transform the input into [CLS]target_SEP_tweet_SEP to take advantage of BERT’s ability in modeling relationship between input text pairs. In our model, we use BERT to encode tweet, target, and target-tweet pair.

C. SHARED FEATURE EXTRACTOR

The shared part of S-MDMT is used to capture stance features shared by all targets. Stance detection is essentially a target-dependent task, the final stance depends on the relationship between the target and opinioned entity in a tweet. Nevertheless, we argue that in many cases, the attitude a tweet expressed is deterministic whatever the opinioned entity in the tweet is. And the expressions used in those tweets are shared across targets. To better facilitate this idea, we only feed tweets into the shared part. Given a tweet \(X^t\) from the input pair, we first encode the tweet with BERT, and the hidden output of [CLS] (denoted as \(h_s\)) from the last transformer layer is used as the tweet representation. It is intuitive that if a strong classifier cannot tell the target involved in the tweet based on the extracted features, then those features are more likely to be target-agnostic. We enforce this by adding a gradient reversal layer (GRL) [34] before the target discriminator.

\[
\bar{h}_s = GRL(h_s)
\]

(1)

where \(h_s\) is the hidden output from the shared encoder, \(\bar{h}_s\) has the same value as \(h_s\) while the gradient of \(GRL\) is reversed by GRL.

In the forward pass, stance features go directly into the target classifier. In the backward pass, gradients of target classification are updated in the reverse direction for shared features and target-invariant stance features are learnt. The loss for target classification in the shared part is defined as:

\[
y_{tc1} = \text{softmax}(W_s \sigma(W_s \bar{h}_s))
\]

(2)

\[
L_{tc1} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{K} \hat{y}_{tc1}^i(j) \log \hat{y}_{tc1}^i(j)
\]

(3)

where \(W_s \in \mathbb{R}^{d_m \times d_h}\), \(W_{tc1} \in \mathbb{R}^{K \times d_m}\), \(d_h\) is the size of hidden dimension of BERT output, \(d_m\) is the dimension of intermediate hidden features, \(N_s\) is the number of training instances, \(K\) is the number of different targets, \(\hat{y}_{tc1}^i(j)\) is the predicted probability of \(j\)-th target of the \(i\)-th training instance and \(\hat{y}_{tc1}^i(j)\) is the corresponding gold probability.

D. PRIVATE FEATURE EXTRACTOR

As mentioned earlier, stance detection requires understanding the attitude that a tweet expressed towards a target. If the target is not directly mentioned or the opinioned entity has nothing to do with the target, it would be difficult to deliberate the actual stance of the tweet towards the target. Hence, learning target-aware representations is crucial for the above circumstances. To this end, the private part of S-MDMT is designed to capture such target-specific stance features. We achieve this through the following process: 1) We first encode the target and target-tweet pair using a BERT encoder separately. The target-tweet pair is used to capture deep interactions between a target and tweet pair. The target text is used to better represent semantics of a target. We refer to representations of target and target-tweet pair from BERT encoder as \(h^t_{[CLS]}\) and \(h^\text{pair}_{[CLS]}\) respectively. 2) We then combine \(h^t_{[CLS]}\) and \(h^\text{pair}_{[CLS]}\) as the enhanced target-aware tweet representation \(h^tt\).

1) Target Descriptors

Motivated by [32], we introduce target descriptors to explicitly capture stance features relating to specific targets. Each target descriptor is specific to one target. Those target descriptors are useful in aggregating target-aware stance features for stance predictions. The target descriptors are parameterized by a matrix \(D \in \mathbb{R}^{K \times d_n}\), where \(K\) is the number of targets and \(d_n\) is the dimension for each descriptor. Target descriptors are randomly initialized and learned jointly with other parameters during training.

Given the enhanced target-aware tweet representation \(h^tt\), we add a simple target classifier here to enforce its target-awareness:

\[
y_{tc2} = \text{softmax}(W_{p2} \sigma(W_{p1} h^tt))
\]

(4)

\[
L_{tc2} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{K} \hat{y}_{tc2}^i(j) \log \hat{y}_{tc2}^i(j)
\]

(5)

where \(W_{p1} \in \mathbb{R}^{d_n \times d_h}\), \(W_{p2} \in \mathbb{R}^{K \times d_n}\), \(d_h\) is the size of hidden dimension of BERT output, \(d_m\) is the dimension of intermediate hidden features, \(N_s\) is the number of training instances, \(K\) is the number of different targets, \(\hat{y}_{tc2}^i(j)\) is the predicted probability of \(j\)-th target of the \(i\)-th training instance and \(\hat{y}_{tc2}^i(j)\) is the corresponding gold probability.

We then evaluate its similarity to target descriptors and hope that the gold target descriptor gets higher similarity scores. We perform a simple bi-linear transformation between \(h^tt\) and target descriptor matrix \(D\) as follows:

\[
s = h^tt W_{sim} D^\top
\]

(6)

where \(W_{sim} \in \mathbb{R}^{2dh \times d_n}\), \(D \in \mathbb{R}^{K \times d_n}\), \(d_h\) is the size of hidden dimension of BERT output, \(d_n\) is the dimension of a target descriptor, and \(K\) is the number of targets.

We normalize the similarity scores and calculate a weighted sum of target descriptors as the virtual target \(\bar{h}_{ta}\) instead of the original target:

\[
a_i = \frac{s_i}{\sum_k^K s_i}
\]

(7)

\[
\bar{h}_{ta} = \sum_i a_i \times D_i
\]

(8)
where \( s_i \) is the \( i \)-th value in \( s \) (the similarity between the \( i \)-th target descriptor and \( h^{(i)} \)). \( D_i \) is the vector of \( i \)-th target descriptor and \( K \) is the number of targets.

We use this mixture of target descriptors as the target representation for the following reasons: 1) the original target usually gets the highest attention score and comprises the major semantics of the new target. 2) some features are shared and consistent in stance polarity across certain targets. For instance, a tweet expressing opinions in favor of ‘Feminist Movement’ and not tuned throughout the experiments.

### 2) Target-Aware Stance Representation

After acquiring a more comprehensive representation of the target, we learn a target-aware tweet representation based on BERT’s hidden outputs of target-tweet pair. As different tokens don’t contribute equally to stance detection, we employ a target-aware attention network to discriminate each token’s contribution to the final stance and relation to the target.

\[
\begin{align*}
    s_j &= h_j W_{att} h_{ta} \\
    a_j &= \frac{s_j}{\sum_{j=1}^{T} s_j} \\
    h_p &= \sum_{j=1}^{T} a_j \times h_j
\end{align*}
\]

where \( h_j \) is the feature vector of the \( j \)-th token in the target-tweet pair, \( W_{att} \in \mathbb{R}^{d_h \times d_m} \), \( s_j \) is the relatedness of \( h_j \) to the current target features \( h_{ta} \), \( T \) is the number of tokens in the target-tweet pair. \( h_p \) is considered as the target-aware tweet representation that encodes stance features towards the given target.

### E. STANCE CLASSIFIER

We then combine stance features from the shared and private parts of S-MDMT and feed the concatenated vector \( h_{sc} \) into a simple network with sigmoid non-linearity activation as follows:

\[
\begin{align*}
    h_{sc} &= h_s \oplus h_p \\
    y_s &= \text{softmax}(W_{sc1} \sigma(W_{sc2} h_{sc}))
\end{align*}
\]

where \( \oplus \) refers to vector concatenation, \( W_{sc1} \in \mathbb{R}^{d_m \times 2d_h} \), \( W_{sc2} \in \mathbb{R}^{N_o \times d_m} \), \( d_h \) is the size of hidden dimension of BERT output, \( d_m \) is the dimension of intermediate hidden features, \( N_o \) is the number of different stance labels, and \( y_s \) is probabilities of features through a softmax layer.

The classifier is trained with the following cross-entropy loss of stance classification:

\[
L_{sc} = -\frac{1}{N_s} \sum_{i=1}^{N_s} \sum_{j=1}^{N_o} \hat{y}_s^i(j) \log y_s^i(j)
\]

where \( N_s \) is the number of training instances and \( N_o \) is the number of different stance labels.

### F. MULTI-TASK LEARNING

The stance classifier and target discriminators are jointly trained with back-propagation. The final loss function for our model can be written as:

\[
L = \lambda_1 L_{tec1} + \lambda_2 L_{tec2} + L_{sc}
\]

(15)

where \( \lambda_1 \) and \( \lambda_2 \) are hyper-parameters balancing the influence of target classification in the shared part and the private part respectively.

### IV. EXPERIMENTAL SETUPS

In this section, we describe the dataset used for training and evaluation, implementation details of the proposed model, and strong baselines adopted for performance comparison.

### A. DATASET

We conduct experiments on SemEval 2016 Task 6 Sub-task A. The dataset is comprised of 4163 English tweets crawled from Twitter and each is assigned with a target and a manually annotated stance label towards that target. There are a total of five targets, which are ‘Atheism’ (CC), ‘Climate Change is a real Concern’ (CC), ‘Feminist Movement’ (FM), ‘Hillary Clinton’ (HC), and ‘Legislation of Abortion’ (LA). The detailed statistics of this dataset are shown in Table 1. We use the official train/test split. While the task didn’t provide an official validation set, we run a 5-fold cross-validation on the training set and report averaged results on the test set.

### B. IMPLEMENTATION DETAILS

We adopt uncased BERT\(_{base}\) for all our experiments, which has 12 transformer encoder layers. We fine-tune BERT\(_{base}\) model with Adam optimizer. The dropout rate is set to 0.5 for all parameters. The learning rate is chosen from \( \{1, 2, 3, 4, 5\} \times 10^{-5} \) and batch size for training is set to 8. We choose \( \lambda_1 \) and \( \lambda_2 \) from \([0.1, 1.0]\) with a step size of 0.1. Final choices of all hyper-parameters are selected according to performance on the validation set. \( \lambda_1 \) and \( \lambda_2 \) are set to 0.1 and 0.2 respectively. The learning rate is set to \( 4 \times 10^{-5} \). \( d_m \) is set to 200. The dimension of target descriptors (\( d_n \)) is empirically set to 200 and not tuned throughout the experiments.

Similar to previous work, we adopt the macro-average of F1-score across targets as the evaluation metric, which is calculated as:

\[
F_{Favor} = \frac{2 P_{Favor} R_{Favor}}{P_{Favor} + R_{Favor}}
\]

(16)

\[
F_{Against} = \frac{2 P_{Against} R_{Against}}{P_{Against} + R_{Against}}
\]

(17)

\[
F_{macro} = \frac{2(F_{Against} + F_{Favor})}{2}
\]

(18)

where \( P \) and \( R \) are precision and recall. Then the average of \( F_{Favor} \) and \( F_{Against} \) is calculated as the final metrics \( F_{macro} \).

1. https://www.twitter.com
TABLE 1: Statistics of SemEval2016 Task 6 Subtask A.

| Target                  | #total | #Train | % of instances in Train | % of instances in Test |
|-------------------------|--------|--------|-------------------------|------------------------|
|                         |        |        | Favor | Against | None | Favor | Against | None |
| Atheism                 | 733    | 513    | 17.9 | 59.3    | 22.8 | 22.0  | 72.7    | 12.7 |
| Climate Change          | 564    | 395    | 53.7 | 3.8     | 42.5 | 169   | 72.8    | 6.5  |
| Feminist Movement       | 949    | 664    | 31.6 | 49.4    | 19.0 | 285   | 20.4    | 15.4 |
| Hillary Clinton         | 984    | 689    | 17.1 | 54.4    | 27.1 | 283   | 16.5    | 26.4 |
| Abortion                | 1033   | 283    | 18.5 | 5.4     | 76.1 | 14.5  | 72.7    | 12.7 |
| All                     | 4163   | 2914   | 25.8 | 47.9    | 26.3 | 1249  | 24.3    | 57.3 |

TABLE 2: Results on the SemEval dataset. The average of $F_{Favor}$ and $F_{Against}$ is adopted as the evaluation metric.

Note that the final metrics do not disregard the None class. By taking the average F-score for only Favor and Against classes, we treat None as a class that is not of interest.

C. BASELINES

We compare our model with the following methods:
- SVM-ngrams [2]: Five SVM classifiers (one per target) trained on the corresponding training set for the target using word and character n-grams features.
- MITRE [38]: The best system in SemEval-2016 subtask A is MITRE. This model uses two RNNs: the first one is trained to predict the task-relevant hashtags on a very large unlabeled Twitter corpus. This network is used to initialize the second RNN classifier, which was trained with the provided subtask A data.
- BiCond [25]: A method that employs conditional LSTM to learn a representation of the tweet considering the target.
- TAN [24]: An LSTM-attention based model that incorporates target-specific information into stance identification.
- PNEM [4]: An ensemble model that uses two densely connected and nested BLSTM networks with attention to capture the long-term dependency in stance detection.
- Disc-STS [29]: A model that jointly models interactions among sentiment-stance-target information. An external sentiment classifier for tweets is used.
- HAN [3]: A hierarchical attention-based method that utilizes additional linguistic features like dependency information.
- AT-JSS-Lex [27]: A multi-task learning-based method that adopts sentiment classification as an auxiliary task and uses manual constructed stance and sentiment lexicons for attention.
- CKEMN [5]: A commonsense knowledge enhanced memory network for stance detection.
- BERTbase [29]: BERT model that takes <target, tweet> as input and uses a linear transformation with a softmax layer for final stance classification.
- Stancy [40]: A BERT based method that combines an additional cosine similarity between the target representation and the target-tweet pair representation.
- MoLE [8]: A RoBERTa [41] based model that leverages features from multiple stance detection datasets and learns from heterogeneous labels using label embeddings.
- ASDA [7]: A target-aware data augmentation method that generates new examples with conditional masked language modeling and auxiliary sentences.
- MeLT [28]: A hierarchical message-encoder which utilizes additional features from user history.

For Stancy [40], we run the experiments using the code 2 provided by original paper.

V. RESULTS AND ANALYSIS

2https://github.com/kashpop/stancy
with attention to capture dependency between targets and tweet with multiple layers and achieves better results. Pre-trained language model based methods (Stancy \cite{40} and MT-DNN \cite{9}) generally work better than methods using traditional non-pre-trained methods (BiCond \cite{25} and TAN \cite{24}). Thus, the BERT model serves as a better backbone model for learning target-aware stance representations compared to previously used BLSTM networks. Furthermore, recent models \cite{7,8,28} that combine both pre-trained language models and external resources further boost the performances.

The proposed S-MDMT model further outperforms strong baselines like BERT, MoLE, and MT-DNN by 2.99%, 0.62%, and 0.89% respectively, which demonstrates the effectiveness of exploiting shared features among different targets without using any external resources. This verifies that shared and private stance features of different targets are under-explored by previous methods. We leave it as future work to see whether these external resources are complementary to information captured by our model S-MDMT.

### B. ABLATION STUDY

We have shown the effectiveness of leveraging shared stance features among different targets in \cite{V-A}. Here, we present further analyses on the effect of different components of S-MDMT with some ablation experiments.

We first compare our model with two variants: one that doesn’t have the shared part and the other one that doesn’t have the private part of the full model. We can see that performances drop by over 1.9% when removing either shared part or private part, which reveals that both shared and private features contribute to the improvements over the BERT base model. The private part capturing target-aware stance features has a larger impact than the shared part, which again shown that target information plays an important role in stance detection.

Then we explore three extra variants to study the influences of other modules in S-MDMT. It can be observed that when removing the gradient reversal layer (GRL), the shared part degrades to learn target-specific stance features instead of shared stance-related features and its performance is similar to that of BERT base model as expected. Besides, we also experiment with removing the target classification module in the private part. Instead of using a mixture of target descriptors, this variant only use the descriptor of the corresponding target. It can be seen that the performance drop by 2.08% when neglecting mining stance features of related targets. This variant also illustrates the effectiveness of multi-task learning in stance detection task. Moreover, we testify the effect of utilizing target descriptors. This variant performs similarly with variant (4) as in both cases, private features specific to related targets are not fully exploited.

To intuitively assess sharedness of stance features in both shared and private parts, we visualize the features from the shared and private parts with t-SNE on the test set. As shown in Figure \ref{fig:tsne} there are more entangled data points in the shared part than in the private part, which verifies that the shared part learns more target-general stance features. Moreover, most data points are well separated in according to their corresponding targets, showing that most stance features are specific to corresponding targets. Besides, we can observe from Figure \ref{fig:tsne} that some data points corresponding...
to ‘Feminist Movement’ and ‘Legalization of Abortion’ are closer than to other data points. This is consistent with our intuition that some tweets expressing stance towards these two targets may be talking about similar topics with overlapped expressions.

C. VISUALIZATION OF FEATURES

We also demonstrate similar phenomenon from another perspective in Figure 4. We draw Figure 4 based on the normalized confusion matrix of the target classification in the private part. In the confusion matrix, the row is the right target and column is the predicted target. Darker color means more examples are predicted in the cell.

We can observe that:
- Cells in the diagonal have the darkest colors which reveal that most texts are correctly classified to their true targets. Thus, the feature extractor can learn good representations for target classification.
- Besides, there exists some dark cells in non-diagonal positions, which means that some cases from those targets are misclassified. This shows that some texts are hard to tell their corresponding target, as they may contain features shared by different targets. For instance, some texts from the ‘Legalization of Abortion’ are misclassified into ‘Atheism’ and ‘Feminist Movement’, and some cases from ‘Feminist Movement’ are misclassified into ‘Legalization of Abortion’. If those targets don’t have any topics in common, it would be easy for a classifier to distinguish between those targets. However, many cases from those targets are misclassified, which
proves that some features are shared across these targets. Contrary to those targets, as ‘Climate Change’ has little overlap of topics with other targets, most of its texts are correctly classified.

• By comparing between results of the BERT model and our model in Table 2, the improvement of macro averaged F1 score of ‘Legalization of Abortion’ is greater than that of ‘Atheism’. While the performance on ‘Climate change is a real concern’ also increases by a large margin, it is mainly caused by the class imbalance of in this target. As our model extracts better features than the BERT baseline and improves the prediction on Against class, which as a result improves the overall macro-average F1 of Favor and Against.

We also present 5 top important words of stance classification for each target using LIME [42] in Table 4. LIME is a probing method that weights the importance of each input token by corrupting the input text and calculating the changes in the confidence of predictions. Intuitively, removing important words from the text usually causes the decrease of the confidence of correct class, sometimes even resulting in the flipping of predictions. Those words chosen by LIME are key features used by our model to make stance predictions, which should be highly informative of stance towards their corresponding target. Extracted important words in Table 4 verify the above assumption, e.g. murder is frequently used to express an ‘against’ stance towards ‘Legalization of Abortion’. rights is commonly used in ‘Legalization of Abortion’ to support the birthright of women. Those words prove that our model learns proper stance features for each target to perform stance classification.

VI. CONCLUSION

In this paper, we propose the S-MDMT model for stance detection on tweets. We novelly formulate the task as a multi-domain multi-task learning problem. And we employ the shared-private structure in stance detection for the first time to fully exploit the shared stance features across different targets. Experimental results on the SemEval 2016 dataset demonstrate the effectiveness of the proposed S-MDMT model, which outperforms existing target-aware models that using external resources and additional sentiment annotations. We provide a detailed analysis of the shared features across different targets.

The proposed S-MDMT method could also be combined with existing methods that utilizing external resources like linguistic analyzers and commonsense knowledge bases. We leave it as future work to explore whether those resources are complementary to the multi-domain multi-task formulation.

REFERENCES

[1] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, “Semeval-2016 task 6: Detecting stance in tweets,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 31–41.

[2] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, “Stance and sentiment in tweets,” ACM Transactions on Internet Technology (TOIT), vol. 17, no. 3, pp. 1–23, 2017.

[3] Q. Sun, Z. Wang, Q. Zhu, and G. Zhou, “Stance detection with hierarchical attention network,” in Proceedings of the 27th international conference on computational linguistics, 2018, pp. 2399–2409.
[4] U. A. Siddiqua, A. N. Chy, and M. Aono, “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). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1868–1873. [Online]. Available: https://aclanthology.org/N19-1185

[5] J. Du, L. Gui, R. Xu, Y. Xia, and X. Wang, “Common-sense knowledge enhanced memory network for stance classification,” IEEE Intelligent Systems, vol. 35, no. 4, pp. 102–109, 2020.

[6] K. Hanawa, A. Sasaki, N. Okazaki, and K. Inui, “Stance detection attending external knowledge from wikipedia,” Journal of Information Processing, vol. 27, pp. 499–506, 2019.

[7] Y. Li and C. Caragea, “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. Online: Association for Computational Linguistics, Jun. 2021, pp. 1850–1860. [Online]. Available: https://aclanthology.org/2021.naacl-main.148

[8] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, “Cross-domain label-adaptive stance detection,” CoRR, vol. abs/2104.07467, 2021. [Online]. Available: https://arxiv.org/abs/2104.07467

[9] B. Schiller, J. Daxenberger, and I. Gurevych, “Stance detection benchmark: How robust is your stance detection?” CoRR, vol. abs/2001.01565, 2020. [Online]. Available: http://arxiv.org/abs/2001.01565

[10] D. Küçük and F. Can, “Stance detection: Concepts, approaches, resources, and outstanding issues,” in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2673–2676. [Online]. Available: https://doi.org/10.1145/3404835.3462815

[11] A. AlDayel and W. Magdy, “Stance detection on social media: State of the art and trends,” CoRR, vol. abs/2006.03644, 2020. [Online]. Available: https://arxiv.org/abs/2006.03644

[12] D. Küçük and F. Can, “Stance detection: A survey,” vol. 53, no. 1, Feb. 2020. [Online]. Available: https://doi.org/10.1145/3369026

[13] P. Sobhani, D. Inkpen, and X. Zhu, “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, 2017, pp. 551–557.

[14] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, “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. Online: Association for Computational Linguistics, Jul. 2020, pp. 1715–1724. [Online]. Available: https://aclanthology.org/2020.acl-main.157

[15] P. Sobhani, D. Inkpen, and X. Zhu, “Exploring deep neural networks for multitarget stance detection,” Computational Intelligence, vol. 35, no. 1, pp. 82–97, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/coi.12189

[16] Y. Li and C. Caragea, “A multi-task learning framework for multi-target stance detection,” in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, Aug. 2021, pp. 2320–2326. [Online]. Available: https://aclanthology.org/2021.findings-acl.204

[17] K. Kawintiranon and L. Singh, “Knowledge enhanced masked language model for stance detection,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, Jun. 2021, pp. 4725–4735. [Online]. Available: https://aclanthology.org/2021.naacl-main.376

[18] B. Zhang, M. Yang, X. Li, Y. Ye, X. Xu, and K. Dai, “Enhancing cross-target stance detection with transferable semantic-emotion knowledge,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, Jul. 2020, pp. 3188–3197. [Online]. Available: https://aclanthology.org/2020.acl-main.291

[19] B. Liang, Y. Fu, L. Gui, M. Yang, J. Du, Y. He, and R. Xu, “Target-adaptive graph for cross-target stance detection,” in Proceedings of the Web Conference 2021, ser. WWW ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 3453–3464. [Online]. Available: https://doi.org/10.1145/3442381.3449790

[20] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, “Synthetic examples improve cross-target generalization: A study on stance detection on a Twitter corpus,” in Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Online: Association for Computational Linguistics, Apr. 2021, pp. 181–187. [Online]. Available: https://aclanthology.org/2021.wassa-1.19

[21] E. Allaway and K. McKeown, “Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020, pp. 8913–8931. [Online]. Available: https://aclanthology.org/2020.emnlp-main.717
[22] E. Allaway, M. Srikanth, and K. McKeown, “Adversarial learning for zero-shot stance detection on social media,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, Jun. 2021, pp. 4756–4767. [Online]. Available: https://aclanthology.org/2021.naacl-main.379

[23] T. Ye, Y. Ma, X. Lin, and J. Li, “A token level multi-target stance detection dataset,” in 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), 2020, pp. 253–256.

[24] J. Du, R. Xu, Y. He, and L. Gui, “Stance classification with target-specific neural attention networks.” International Joint Conferences on Artificial Intelligence, 2017.

[25] I. Augenstein, T. Rocktäschel, A. Vlachos, and K. Bontcheva, “Stance detection with bidirectional conditional encoding,” arXiv preprint arXiv:1606.05464, 2016.

[26] S. Ghosh, P. Singhania, S. Singh, K. Rudra, and S. Ghosh, “Stance detection in web and social media: A comparative study,” in Experimental IR Meets Multilinguality, Multimodality, and Interaction, F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, Eds. Cham: Springer International Publishing, 2019, pp. 75–87.

[27] Y. Li and C. Caragea, “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). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6299–6305. [Online]. Available: https://aclanthology.org/D19-1657

[28] M. Matero, N. Soni, N. Balasubramanian, and H. A. Schwartz, “Melt: Message-level transformer with masked document representations as pre-training for stance detection,” CoRR, vol. abs/2109.08113, 2021. [Online]. Available: https://arxiv.org/abs/2109.08113

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[30] P. Liu, X. Qiu, and X. Huang, “Adversarial multi-task learning for text classification,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/P17-1001

[31] X. Chen and C. Cardie, “Multinomial adversarial networks for multi-domain text classification,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1226–1240. [Online]. Available: https://www.aclweb.org/anthology/N18-1111

[32] Q. Liu, Y. Zhang, and J. Liu, “Learning domain representation for multi-domain sentiment classification,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 541–550. [Online]. Available: https://aclanthology.org/N18-1050

[33] J. Guo, D. Shah, and R. Barzilay, “Multi-source domain adaptation with mixture of experts,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 4694–4703. [Online]. Available: https://www.aclweb.org/anthology/D18-1498

[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” The Journal of Machine Learning Research, vol. 17, no. 1, pp. 2096–2030, 2016.

[35] X. Chen, A. H. Awadallah, H. Hassan, W. Wang, and C. Cardie, “Multi-source cross-lingual model transfer: Learning what to share,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3098–3112. [Online]. Available: https://www.aclweb.org/anthology/P19-1299

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547deec9fbb053c1e4a845aa-Paper.pdf

[37] C. Sun, L. Huang, and X. Qiu, “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). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 380–385. [Online]. Available: https://aclanthology.org/N19-1035

[38] G. Zaremba and A. Marsh, “MITRE at SemEval-2016 task 6: Transfer learning for stance detection,” in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 458–463. [Online]. Available: https:
[39] J. Ebrahimi, D. Dou, and D. Lowd, “A joint sentiment-target-stance model for stance classification in tweets,” in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2656–2665. [Online]. Available: https://aclanthology.org/C16-1250

[40] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, “STANCY: Stance classification based on consistency cues,” 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). Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6413–6418. [Online]. Available: https://aclanthology.org/D19-1675

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” CoRR, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[42] M. T. Ribeiro, S. Singh, and C. Guestrin, “why should i trust you?: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

**LIMIN WANG** Limin Wang is currently a Ph.D candidate in Harbin Institute of Technology. Her research interest is mainly in the area of cyber-sociology.

**DEXIN WANG** Dexin Wang is received the Ph.D degree from Harbin Institute of Technology. He currently serves as a lecturer in Chengdu University of Technology. His research interest includes network research and new media research.

***