24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

DAN: Dual-View Representation Learning for Adapting Stance Classifiers to New Domains

Chang Xu[1] and **Cécile Paris**[1] and **Surya Nepal**[1] and **Ross Sparks**[1]
Chong Long[2] and **Yafang Wang**[2]

Abstract. We address the issue of having a limited number of annotations for stance classification in a new domain, by adapting out-of-domain classifiers with domain adaptation. Existing approaches often align different domains in a single, global feature space (or view), which may fail to fully capture the richness of the languages used for expressing stances, leading to reduced adaptability on stance data. In this paper, we identify two major types of stance expressions that are linguistically distinct, and we propose a tailored **d**ual-view **a**daptation **n**etwork (DAN) to adapt these expressions across domains. The proposed model first learns a separate view for domain transfer in each expression channel and then selects the best adapted parts of both views for optimal transfer. We find that the learned view features can be more easily aligned and more stance-discriminative in either or both views, leading to more transferable overall features after combining the views. Results from extensive experiments show that our method can enhance the state-of-the-art single-view methods in matching stance data across different domains, and that it consistently improves those methods on various adaptation tasks.

1 Introduction

There has been a growing interest in the relatively new task of stance classification in opinion mining, which aims at automatically recognising one's attitude or position (e.g., *favour* or *against*) towards a given controversial topic (e.g., feminist movement) [32, 12, 21, 4]. Recently, deep neural networks (DNNs) have been used to learn representations for stance expressions, resulting in state-of-the-art performance on multiple stance corpora [3, 17, 9, 30]. However, DNNs are notorious for relying on abundant labelled data for training, which could be hardly available for a new trending topic, as obtaining quality stance annotations is often costly [22].

To address this issue, domain adaptation [7] enables adapting what has been learned from a *source* domain to a *target* domain, usually by aligning the source and target data distributions in a shared feature space. This process makes the learned features invariant to the domain shift and thus become generalisable across the domains. Recently, due to their effectiveness and seamless integration with DNNs, *adversarial* adaptation methods [10, 26] have gained popularity among various NLP tasks [15, 37, 1, 19]. In these approaches, a *domain examiner* (also called *domain classifier* [10] or *domain critic* [2]) is introduced to assess the discrepancy between the domains, and, by confusing it with an adversarial loss, one obtains domain-invariant features.

[1] CSIRO Data61, Australia, email: {Chang.Xu, Cecile.Paris, Surya.Nepal, Ross.Sparks}@data61.csiro.au
[2] Ant Financial Services Group, email: {huangxuan.lc, yafang.wyf}@antfin.com

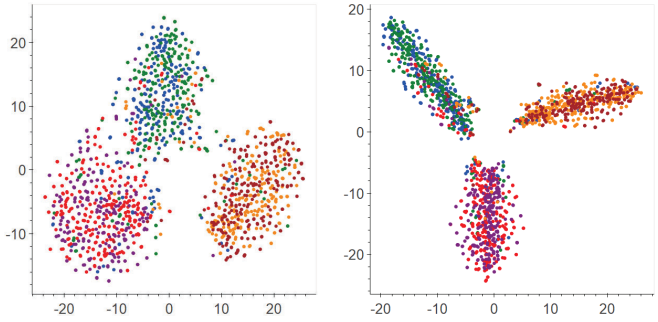
However, as the domain examiner solely induces a global feature space (view) for aligning the domains, it might not fully capture the various ways stances are expressed in real-world scenarios. For example, Table 1 shows examples of commonly observed stance-bearing utterances where stances are expressed in two distinct ways: *explicitly* via *subjective* expressions carrying opinions (e.g., "really incredible") and/or *implicitly* via *objective* expressions that provide facts to support the stance. With only a single feature space, the distributions of different expression types could be intertwined in the space, which could hinder the domains from being optimally aligned, leading to inferior adaptation performance.

Table 1: Examples of stances conveyed explicitly (with opinions) or implicitly (with facts).

Utterance	Topic	Stance
(1) Its really incredible how much this world needs the work of missionaries! (**Opinion**)	Atheism	Against
(2) Women who aborted were 154% more likely to commit suicide than women who carried to term. (**Fact**)	Legalisation of Abortion	Against

Source: SemEval-2016 Task 6

In this paper, to cope with the heterogeneity in stance-expressing languages and adapt stance classifiers to the shift of domains, we first identify the aforementioned types of stance expressions, i.e., *subjective* and *objective* stance expressions, as the major elements for a better characterisation of a stance-bearing utterance. In particular, we propose a hypothesised *dual-view stance model* which regards a stance-bearing utterance as a *mixture* of the subjective and objective stance expressions. Under this hypothesis, the characterisation of a stance-bearing utterance is then reduced to modelling the more fine-grained subjective and/or objective expressions appearing in the utterance, each of which can receive a different, finer treatment. Moreover, to implement this dual-view stance model, we propose DAN, the **d**ual-view **a**daptation **n**etwork, for adapting stance classifiers with signals from both the subjective and objective stance expressions. Specifically, DAN aims at learning transferable subjective and objective features that are both **domain-invariant and stance-discriminative in their respective views (i.e., feature spaces)**. To achieve this, DAN is designed to perform three view-centric subtasks: 1) *view formation* that creates the subjective/objective views for learning the view-specific, stance-discriminative features; 2) *view adaptation* that employs a view-specific domain examiner to make each of the view features domain-invariant; and 3) *view fusion* where the view features are made more transferable after being fused in an optimal manner. All these subtasks are trained jointly in DAN with standard back-propagation.



(a) A single-view method (DANN) (b) Our method

Figure 1: Feature spaces learned by a state-of-the-art single-view method (a) and our method (b) on a common adaptation task. Source samples are coloured by orange (*favour*), blue (*against*), and red (*neutral*), while target samples by brown (*favour*), green (*against*), and purple (*neutral*). The features produced by our method are more domain-invariant and stance-discriminative than those by the single-view method. The feature dimension is reduced to 2 by using t-SNE for better visualisation.

We evaluate our method extensively via both quantitative and qualitative analyses on various adaptation tasks. We find that DAN can enhance the single-view adaptation methods by delivering more transferable features. As an example, Figure 1 shows the features learned by a state-of-the-art single-view approach (DANN [11]) and its DAN-enhanced version (our method) on a common adaptation task. As shown, DANN can produce features with good but limited transferability (Figure 1.a): in terms of domain invariance, features of the source and target samples are aligned, but they are scattered around a relatively wide area, indicating non-trivial distances between the source and target features; in terms of stance discrimination, samples from different classes are separated, but the boundary between the *against* and *neutral* classes is still not fully clear. In contrast, after being enhanced by the proposed DAN to separately adapt the subjective and objective views, the learned features (after view fusion) in the enhanced feature space of DAN (Figure 1.b) exhibit much better transferability: not only do the source/target features become more concentrated, but they are also well separated over the stance classes[3]. This result suggests that our dual-view treatment of the stance-bearing utterances can yield a more fine-grained adaptation strategy that enables better characterising and transferring heterogeneous stance expressions.

2 Dissecting Stance Expressions

To better characterise stance-bearing utterances for domain transfer, we identify two major types of stance expressions to facilitate fine-grained modelling of stance-bearing utterances, which are the opinion-carrying *subjective* stance expressions and the fact-stating *objective* stance expressions.

Subjective stance expressions: This type of expressions is common in stance-bearing utterances. When stating a stance, people may also express certain feelings, sentiments, or beliefs towards the discussed topic, which are the various forms of *subjective* expressions [33]. The role of sentiment information in stance classification has recently been examined [27, 22], and the identification of sentiment-bearing words in an utterance has been shown to help recognise its stance. For instance, the underlined sentiment words in the following utterances reveal various stances towards the topic of *Feminist Movement*,

- *Women are strong, women are smart, women are bold*. (**Favour**)
- *My feminist heart is so angry right now, wish I could scream my hate for inequality right now*. (**Favour**)
- *The concept of #RapeCulture is a puerile, intellectually dishonest glorification of a crime by feminists*. (**Against**)

Based on this observation, we seek to find such stance-indicative subjective expressions in an utterance for stance determination.

Objective stance expressions: A stance can also be expressed objectively, usually by presenting some facts for backing up the stance. For example, all the following utterances mention particular evidence for supporting their stances towards *Legalisation of Abortion*,

- *Life Fact: At just 9 weeks into a pregnancy, a baby begins to sigh and yawn*. (**Against**)
- *There are 3000 babies killed DAILY by abortion in the USA*. (**Against**)
- *Over the last 25 years, more than 50 countries have changed their laws to allow for greater access to abortion*. (**Favour**)

In such case, no explicit subjective signals (e.g., sentiment or emotional words) can be observed; a stance is instead implied in text providing some facts related to the stance. Usually, such factual information would serve as the reasons for supporting the stances [13, 8], thus it may also be stance-specific and become stance-indicative. Therefore, a different treatment from the one for characterising the subjectivity is needed for capturing such (implicit) objectivity.

A dual-view stance model: Motivated by the observations made above, we propose a hypothesised *dual-view stance model* to characterise the subjective and objective aspects of stance-bearing utterances, aiming at learning more transferable representations. Specifically, in this model, we regard any stance-bearing utterance as *a mixture of the subjective and objective stance expressions*. Formally, given a stance-bearing utterance $\mathbf{x}$, we use the following transformation U for such dual-view characterisation,

$$\mathbf{f}_{\text{dual}} = U(F_{\text{subj}}(\mathbf{x}), F_{\text{obj}}(\mathbf{x}); \theta_U) \quad (1)$$

where F_{subj} and F_{obj} are the two *view feature functions* (or *view functions* for short) for extracting the subjective and objective features of $\mathbf{x}$, respectively. $\mathbf{f}_{\text{dual}}$ is the *dual-view stance feature* of $\mathbf{x}$, resulting from applying U to unify both view features provided by the view functions F_{subj} and F_{obj}. θ_U denotes the parameters of U, characterising how much contribution from each view to the overall adaptation. Based on this dual-view stance model, we formulate **our task of dual-view adaptation of stance classifiers** as follows[4]: given a set of *labelled* samples $S = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{|S|}$ from a source domain $\mathcal{D}^s$ and a set of *unlabelled* samples $T = \{\mathbf{x}_i^t\}_{i=1}^{|T|}$ from a target domain $\mathcal{D}^t$, where $\mathbf{x}$ is a stance-bearing utterance and $\mathbf{y}$ its stance label, the goal is to learn a transferable (i.e., domain-invariant and stance-discriminative) dual-view stance feature $\mathbf{f}_{\text{dual}}^t$ for any target sample $\mathbf{x}^t$, and train a classifier C_{stance} to predict its stance label $\mathbf{y}^t$ with $\mathbf{f}_{\text{dual}}^t$.

3 DAN: Dual-View Adaptation Network

In this section, we introduce the dual-view adaptation network (DAN) for our task. Figure 2 shows a sketch of DAN, which involves three view-centric subtasks designed for realising the above dual-view stance model, i.e., *view formation*, *adaptation*, and *fusion*. In *view formation*, DAN learns the view functions F_{subj} and F_{obj} which create independent feature spaces for accommodating the subjective and objective features of input utterances, respectively. In particular, it leverages *multi-task learning* for obtaining the subjective/objective features that are also stance-discriminative. Then, DAN performs *view adaptation* to make each view feature invariant to the shift of

[3] We also visualise the features in the intermediate subjective and objective views of DAN in our experiments (Figure 3).

[4] We confine ourselves to the *unsupervised* domain adaptation setting in this work, where we lack annotations in the target domain.

domains. This is done by solving a *confusion maximisation* problem in each view which encourages the view features of both source and target samples to confuse the model about their origin domains (thus making the features domain-invariant). Finally, DAN realises the transformation U (i.e., Eq. 1) in *view fusion* that unifies both views to form the dual-view stance feature $\mathbf{f}_{\text{dual}}$, which is used to produce the ultimate stance predictions. Next, we elaborate each of these subtasks and show how they can be jointly trained to make the view features both stance-discriminative and domain-invariant.

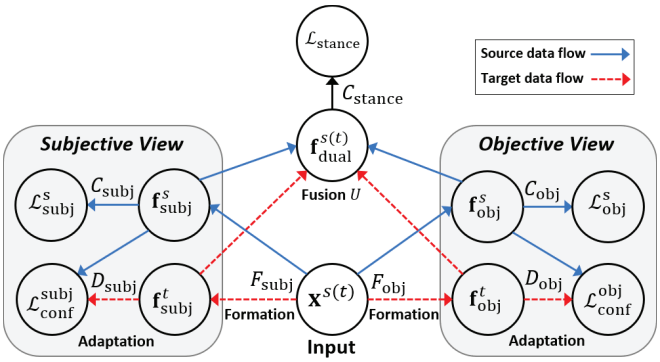


Figure 2: Model scheme of DAN. Both source and target utterances $\mathbf{x}^{s(t)}$ are fed into the view functions F_{subj} and F_{obj} to produce the subjective and objective features $\mathbf{f}_{\text{subj}}^{s(t)}$ and $\mathbf{f}_{\text{obj}}^{s(t)}$, respectively. These features are then adapted separately and finally fused to form $\mathbf{f}_{\text{dual}}^{s(t)}$ for the stance classification task.

3.1 View Formation

The first and foremost subtask in DAN is to create the split subjective/objective views for the input utterances. The challenge of this, however, lies in how to effectively extract the subjective and objective signals from an utterance $\mathbf{x}$ which also reveal its stance polarity. A straightforward solution is to utilise specialised lexica to identify words in $\mathbf{x}$ that carry subjective and/or objective information. For example, in [22], several sentiment lexica were used to derive sentiment features for stance classification. However, while there are many off-the-shelf lexica for finding the sentiment words, the ones for spotting the objective words/expressions are rarely available in practice. Moreover, this approach does not guarantee finding stance-discriminative subjective/objective expressions. Instead of searching the subjective/objective signals at the word level, we focus on learning the **stance-discriminative subjective and objective features** for the entire utterance. In particular, we resort to *multi-task learning* and formulate this view formation problem as learning stance-discriminative subjective/objective features with supervision from multiple related tasks.

Concretely, to learn the stance-discriminative subjective feature, we perform the stance classification task (main) together with a *subjectivity classification* task (auxiliary), which predicts whether $\mathbf{x}$ contains any subjective information. Similarly, to learn the stance-discriminative objective feature, we perform the stance classification task (main) together with an *objectivity classification* task (auxiliary), which predicts whether $\mathbf{x}$ contains any objective information. Formally, the above tasks can be expressed as follows,

$$\begin{aligned}
\textit{View Learning:}\quad & \mathbf{f}_{\text{subj}} = F_{\text{subj}}(\mathbf{x};\theta_{F_{\text{subj}}}),\ \mathbf{f}_{\text{obj}} = F_{\text{obj}}(\mathbf{x};\theta_{F_{\text{obj}}}) \\
\textit{Subj-Auxiliary Task:}\quad & \hat{\mathbf{y}}_{\text{subj}} = C_{\text{subj}}(\mathbf{f}_{\text{subj}};\theta_{C_{\text{subj}}}) \\
\textit{Obj-Auxiliary Task:}\quad & \hat{\mathbf{y}}_{\text{obj}} = C_{\text{obj}}(\mathbf{f}_{\text{obj}};\theta_{C_{\text{obj}}}) \\
\textit{Main Task:}\quad & \hat{\mathbf{y}}_{\text{stance}} = C_{\text{stance}}(\mathbf{f}_{\text{dual}};\theta_{C_{\text{stance}}})
\end{aligned} \tag{2}$$

where a view function F maps the input $\mathbf{x}$ into its d-dimensional view feature $\mathbf{f}$ with parameter θ_F; C denotes a classifier parameterised by θ_C, and $\hat{\mathbf{y}}$ a prediction. To jointly learn these tasks, we minimise the negative log-likelihood of the ground truth class for each *source* sample (as target samples are assumed to be unlabelled),

$$\begin{aligned}
&\mathcal{L}_{\text{stance}} + \alpha\mathcal{L}_{\text{subj}} + \beta\mathcal{L}_{\text{obj}} = \\
&-\sum_{i=1}^{|S|} \mathbf{y}_{\text{stance}}^{(i)} \ln \hat{\mathbf{y}}_{\text{stance}}^{(i)} - \alpha \sum_{i=1}^{|S|} \mathbf{y}_{\text{subj}}^{(i)} \ln \hat{\mathbf{y}}_{\text{subj}}^{(i)} - \beta \sum_{i=1}^{|S|} \mathbf{y}_{\text{obj}}^{(i)} \ln \hat{\mathbf{y}}_{\text{obj}}^{(i)}
\end{aligned} \tag{3}$$

where $\mathbf{y}$s denote the true classes, and α, β the balancing coefficients. Notice that both tasks C_{subj} and C_{stance} (C_{obj} and C_{stance}) share the same underlying feature $\mathbf{f}_{\text{subj}}$ ($\mathbf{f}_{\text{obj}}$) for making predictions, and that $\mathbf{f}_{\text{dual}}$ is a function of both $\mathbf{f}_{\text{subj}}$ and $\mathbf{f}_{\text{obj}}$ (Eq.1); minimising Eq.3 thus encourages $\mathbf{f}_{\text{subj}}$ ($\mathbf{f}_{\text{obj}}$) to be stance-discriminative.

The ground-truth subjectivity and objectivity labels $\mathbf{y}_{\text{subj}}$s and $\mathbf{y}_{\text{obj}}$s are essential for computing the losses $\mathcal{L}_{\text{subj}}$ and $\mathcal{L}_{\text{obj}}$, respectively. One can obtain gold standard labels from human annotations, which, however, is often a costly process. To seek a cost-effective solution, we pre-train a subjective and an objective classifier with a publicly available subjectivity vs. objectivity corpus, and then use the pre-trained models to assign a subjective and an objectivity label to each utterance in our data as the *silver* standard labels. The benefits of this practice are two-fold. First, it automates the label acquisition process; Second, although these silver labels may be less reliable than the human-annotated gold standard ones, we find that the silver labels produced by the pre-trained models trained with large amounts of subjectivity/objectivity data are adequately informative in practice to indicate useful subjective/objective signals. More details on obtaining such silver labels are discussed later in the experiments.

3.2 View Adaptation

With both view features learned, we then perform feature alignment to match the distributions of source and target features $\{\mathbf{f}_i^s\}$ and $\{\mathbf{f}_i^t\}$ in each view, so that they both become invariant to the domain shift. To achieve this, we introduce a *confusion* loss, which *adversarially* encourages the model (*domain examiner* in particular) to be confused with the origin of a sample, i.e., which domain it comes from. Then, by *maximising* the confusion loss, the source and target features $\mathbf{f}^s$ and $\mathbf{f}^t$ would become asymptotically similar so as to confuse the model. The more alike $\mathbf{f}^s$ and $\mathbf{f}^t$ are, the more likely that the stance classifier C_{stance} trained on $\mathbf{f}^s$ would perform similarly well on $\mathbf{f}^t$. In this work, we experimented with two implementations of the confusion loss, both of which assess the level of confusion by measuring the discrepancy between domains in different ways.

The first one measures the $\mathcal{H}$-divergence between the domains, approximated as the classification loss incurred by the domain examiner to distinguish the source/target samples [10]. Specifically, the domain examiner learns a function $D^{\mathcal{H}}$, with parameter $\theta_{D^{\mathcal{H}}}$, that maps a feature $\mathbf{f} = F(\mathbf{x})$ to a binary class label indicating its domain. Then, by maximising the following binary domain prediction loss with respect to $\theta_{D^{\mathcal{H}}}$, while minimising it with respect to θ_F, one obtains the domain-invariant $\mathbf{f}$,

$$\mathcal{L}_{\text{conf}}^{\mathcal{H}} = \sum_{i=1}^{|S|+|T|} \mathbb{1}_{[\mathbf{x}_i \in S]} \ln D^{\mathcal{H}}(\mathbf{f}_i) + \mathbb{1}_{[\mathbf{x}_i \in T]} \ln(1 - D^{\mathcal{H}}(\mathbf{f}_i)) \tag{4}$$

The other one measures the Wasserstein distance between domains for the purpose of stabilising adversarial training [26]. Specifically, the domain examiner learns a function $D^{\mathcal{W}}$, with parameter $\theta_{D^{\mathcal{W}}}$, that maps $\mathbf{f} = F(\mathbf{x})$ to a real number. Then one can approximate the

empirical Wasserstein distance by maximising the following domain critic loss with respect to $\theta_{D^{\mathcal{W}}}$,

$$\mathcal{L}_{\text{conf}}^{\mathcal{W}} = \frac{1}{|S|}\sum_{\mathbf{x}_i \in S} D^{\mathcal{W}}(\mathbf{f}_i) - \frac{1}{|T|}\sum_{\mathbf{x}_i \in T} D^{\mathcal{W}}(\mathbf{f}_i) \tag{5}$$

To obtain the domain-invariant features, one minimises $\mathcal{L}_{\text{conf}}^{\mathcal{W}}$ with respect to θ_F when the domain examiner is trained to optimality [26].

We create a different domain examiner for each of the subjective and objective views to achieve separate adaptation in each view.

3.3 View Fusion

Finally, we combine the two adapted view features $\mathbf{f}_{\text{subj}}$ and $\mathbf{f}_{\text{obj}}$ to produce the dual-view stance feature $\mathbf{f}_{\text{dual}}$. The key is to select the best aligned dimensions of $\mathbf{f}_{\text{subj}}$ and $\mathbf{f}_{\text{obj}}$ to attain optimal combination. For this, we merge $\mathbf{f}_{\text{subj}}$ and $\mathbf{f}_{\text{obj}}$ by learning a *fusion function* $U(\cdot;\theta_U)$, which serves as a realisation of the transformation U in Eq. 1, to weigh each dimension of $\mathbf{f}_{\text{subj}}$ and $\mathbf{f}_{\text{obj}}$ in two steps: 1) It learns a fusion vector $\mathbf{g}$ via a feed-forward network with sigmoid activation: $\mathbf{g} = \text{sigmoid}(\mathbf{W}_u[\mathbf{f}_{\text{subj}};\mathbf{f}_{\text{obj}}] + \mathbf{b}_u)$, where $[;]$ denotes vector concatenation, and $\theta_U = \{\mathbf{W}_u, \mathbf{b}_u\}$ the trainable parameters; 2) It merges the views using $\mathbf{g}$ to deliver the dual-view stance feature $\mathbf{f}_{\text{dual}}$,

$$\mathbf{f}_{\text{dual}} = \mathbf{g} \odot \mathbf{f}_{\text{subj}} + (\mathbf{1} - \mathbf{g}) \odot \mathbf{f}_{\text{obj}} \tag{6}$$

where $\odot$ is the element-wise product. Note that during training, the fusion is applied to the source data only, as the target domain is assumed to be unlabelled. After fusion, the dual-view stance feature $\mathbf{f}_{\text{dual}}$ is used by the stance classifier C_{stance} to produce a predicted stance label $\hat{\mathbf{y}}_{\text{stance}}$ for the stance classification task (Eq.2).

3.4 Training and Prediction

To train DAN, we jointly optimise all the losses introduced above by using both the labelled source and unlabelled target data, so that the view features are learned to be both domain-invariant and stance-discriminative. In this process, we need to minimise the classification losses ($\mathcal{L}_{\text{stance}}$, $\mathcal{L}_{\text{subj}}$, $\mathcal{L}_{\text{obj}}$) with respect to the classifiers (C_{stance}, C_{subj}, C_{obj}), the fusion function U, and the view functions (F_{subj}, F_{obj}) for obtaining stance-discriminative features, while *adversarially* maximising the confusion losses ($\mathcal{L}_{\text{conf}}^{\text{subj}}$, $\mathcal{L}_{\text{conf}}^{\text{obj}}$) with respect to the domain examiners (D_{subj}, D_{obj}) and view functions (F_{subj}, F_{obj}) to make those features domain-invariant. We thus formulate the training as a min-max game between the above two groups of losses, which involves alternating the following min and max steps until convergence:

Min step: Update the parameters of the view functions $\{\theta_{F_{\text{subj}}}, \theta_{F_{\text{obj}}}\}$, the fusion function θ_U, and the classifiers $\{\theta_{C_{\text{stance}}}, \theta_{C_{\text{subj}}}, \theta_{C_{\text{obj}}}\}$ with the following minimisation task,

$$\min_{\substack{\theta_{F_{\text{subj}}},\theta_{F_{\text{obj}}},\theta_U \\ \theta_{C_{\text{stance}}},\theta_{C_{\text{subj}}},\theta_{C_{\text{obj}}}}} \mathcal{L}_{\text{stance}} + \alpha\mathcal{L}_{\text{subj}} + \beta\mathcal{L}_{\text{obj}} + \gamma(\mathcal{L}_{\text{conf}}^{\text{subj}} + \mathcal{L}_{\text{conf}}^{\text{obj}}) \tag{7}$$

where α, β, γ are the balancing coefficients.

Max step: Train the domain examiners $\{D_{\text{subj}}, D_{\text{obj}}\}$ (could be $D^{\mathcal{H}}$ or $D^{\mathcal{W}}$) to optimality by maximising the confusion losses,

$$\max_{\theta_{D_{\text{subj}}},\theta_{D_{\text{obj}}}} \mathcal{L}_{\text{conf}}^{\text{subj}} + \mathcal{L}_{\text{conf}}^{\text{obj}} \tag{8}$$

The above training process can be implemented with the standard back-propagation, the algorithm for which is summarised in Algorithm 1. Once all parameters converge, the view feature $\mathbf{f}_{\text{subj}}$ ($\mathbf{f}_{\text{obj}}$) would become both domain-invariant and stance-discriminative, as the view function F_{subj} (F_{obj}) has received gradients from both the confusion loss $\mathcal{L}_{\text{conf}}^{\text{subj}}$ ($\mathcal{L}_{\text{conf}}^{\text{obj}}$) and stance classification loss $\mathcal{L}_{\text{stance}}$ during back-propagation (lines 15∼16 in Algorithm 1).

Algorithm 1: Adversarial Training of DAN

Input: source data S; target data T; batch size m; domain examiner training step n; balancing coefficient α, β, γ; learning rate λ_1, λ_2

Output: $\theta_{F_{\text{subj}}}, \theta_{F_{\text{obj}}}, \theta_{D_{\text{subj}}}, \theta_{D_{\text{obj}}}, \theta_{C_{\text{stance}}}, \theta_{C_{\text{subj}}}, \theta_{C_{\text{obj}}}, \theta_U$

1 Initialise the parameters of view functions, domain examiners, classifiers, and fusion function with random weights
2 **repeat**
3 Sample batch $\{\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^m$, $\{\mathbf{x}_i^t\}_{i=1}^m$ from S and T
4 **for** $k = 1, ..., n$ **do**
   `// Maximisation Step`
5 $\mathbf{f}_{\text{subj}}^s = F_{\text{subj}}(\mathbf{x}^s)$, $\mathbf{f}_{\text{obj}}^s = F_{\text{obj}}(\mathbf{x}^s)$
6 $\mathbf{f}_{\text{subj}}^t = F_{\text{subj}}(\mathbf{x}^t)$, $\mathbf{f}_{\text{obj}}^t = F_{\text{obj}}(\mathbf{x}^t)$
7 $\theta_{D_{\text{subj}}} \mathrel{+}= \lambda_1 \nabla_{\theta_{D_{\text{subj}}}} \mathcal{L}_{\text{conf}}^{\text{subj}}(\mathbf{f}_{\text{subj}}^s, \mathbf{f}_{\text{subj}}^t)$
8 $\theta_{D_{\text{obj}}} \mathrel{+}= \lambda_1 \nabla_{\theta_{D_{\text{obj}}}} \mathcal{L}_{\text{conf}}^{\text{obj}}(\mathbf{f}_{\text{obj}}^s, \mathbf{f}_{\text{obj}}^t)$
9 **end**
   `// Minimisation Step`
10 $\mathbf{f}_{\text{dual}}^s \leftarrow U(\mathbf{f}_{\text{subj}}^s, \mathbf{f}_{\text{obj}}^s)$
11 $\theta_U \mathrel{-}= \lambda_2 \nabla_{\theta_U} \mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s)$
12 $\theta_{C_{\text{subj}}} \mathrel{-}= \lambda_2 \nabla_{\theta_{C_{\text{subj}}}} [\mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s) + \alpha\mathcal{L}_{\text{subj}}(\mathbf{f}_{\text{subj}}^s, \mathbf{y}_{\text{subj}}^s)]$
13 $\theta_{C_{\text{obj}}} \mathrel{-}= \lambda_2 \nabla_{\theta_{C_{\text{obj}}}} [\mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s) + \beta\mathcal{L}_{\text{obj}}(\mathbf{f}_{\text{obj}}^s, \mathbf{y}_{\text{obj}}^s)]$
14 $\theta_{C_{\text{stance}}} \mathrel{-}= \lambda_2 \nabla_{\theta_{C_{\text{stance}}}} \mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s)$
15 $\theta_{F_{\text{subj}}} \mathrel{-}= \lambda_2 \nabla_{\theta_{F_{\text{subj}}}} [\mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s) + \alpha\mathcal{L}_{\text{subj}}(\mathbf{f}_{\text{subj}}^s, \mathbf{y}_{\text{subj}}^s) + \gamma\mathcal{L}_{\text{conf}}^{\text{subj}}(\mathbf{f}_{\text{subj}}^s, \mathbf{f}_{\text{subj}}^t)]$
16 $\theta_{F_{\text{obj}}} \mathrel{-}= \lambda_2 \nabla_{\theta_{F_{\text{obj}}}} [\mathcal{L}_{\text{stance}}(\mathbf{f}_{\text{dual}}^s, \mathbf{y}_{\text{stance}}^s) + \beta\mathcal{L}_{\text{obj}}(\mathbf{f}_{\text{obj}}^s, \mathbf{y}_{\text{obj}}^s) + \gamma\mathcal{L}_{\text{conf}}^{\text{obj}}(\mathbf{f}_{\text{obj}}^s, \mathbf{f}_{\text{obj}}^t)]$
17 **until** $\theta_{F_{subj}}, \theta_{F_{obj}}, \theta_{D_{subj}}, \theta_{D_{obj}}, \theta_{C_{stance}}, \theta_{C_{subj}}, \theta_{C_{obj}}, \theta_U$ *converge*

Once the training finishes, we are ready to make stance predictions on the target domain $\mathcal{D}^t$. The prediction phase is more straightforward compared to the training, as it only involves chaining together the learned view functions F_{subj} and F_{obj}, the fusion function U, and the stance classifier C_{stance} to transform a target utterance $\mathbf{x}^t \sim \mathcal{D}^t$ into a stance prediction: $\hat{\mathbf{y}}^t = C_{\text{stance}}(U(F_{\text{subj}}(\mathbf{x}^t), F_{\text{obj}}(\mathbf{x}^t)))$.

4 Experiments

In this section, we evaluate the performance of DAN on a wide range of adaptation tasks. We first conduct a quantitative study on the overall cross-domain classification performance of DAN on all the tasks. Then, a series of qualitative experiments is performed to further examine the various properties of DAN.

4.1 Experimental Setup

Dataset: To evaluate DAN, we utilise the dataset publicised by SemEval-2016 Task 6[5] on tweet stance classification, which has been widely used for benchmarking stance classifiers. It contains stance-bearing tweets on five different domains/topics: Climate Change is a Real Concern (CC: 564), Feminist Movement (FM: 949), Hillary Clinton (HC: 984), Legalisation of Abortion (LA: 933), and Atheism (AT: 733)[6]. Each tweet in the dataset is associated with one of the three stance labels: *favour*, *against*, and *neutral*. On these domains, we construct the complete set of adaptation tasks over all 20 (S)ource→(T)arget domain pairs. For each $S \rightarrow T$ pair, we use 90% tweets from S and all from T (without labels) as the training data,

[5] `http://alt.qcri.org/semeval2016/task6/`
[6] The number of tweets in each domain is shown in the parentheses.

Table 2: Performance (macro-$F1$ %) of the baselines and their DAN-enhanced versions on various adaptation tasks. The highest performance on each task is underlined. The improvements of a DAN-enhanced method over its original version are shown in the parentheses, with the largest one on each task in bold.

S	T	SO	CORAL	DANN	WDGRL	D-SO	D-CORAL	D-DANN	D-WDGRL	TO
AT	CC	31.96	48.66	49.45	48.39	35.43 (3.46)**	54.97 (6.31)***	55.78 (**6.32**)***	53.36 (4.96)***	59.18
FM	CC	36.63	58.83	49.89	56.06	39.25 (2.61)***	62.04 (2.21)***	52.25 (**2.35**)*	58.35 (2.28)*	
HC	CC	37.47	64.76	59.48	65.54	40.21 (2.74)**	71.84 (**7.07**)***	63.54 (4.06)***	70.76 (5.21)***	
LA	CC	36.91	59.31	48.88	56.42	40.79 (3.88)***	62.17 (2.85)*	53.51 (**4.62**)*	59.91 (3.48)***	
AT	HC	35.11	63.12	66.90	66.08	38.05 (2.93)*	67.87 (**4.75**)*	70.05 (3.14)*	69.34 (3.26)**	68.33
CC	HC	38.73	66.53	64.32	66.75	41.17 (2.44)*	68.86 (2.32)*	67.53 (**3.21**)***	69.49 (2.73)***	
FM	HC	44.65	68.10	69.75	75.08	46.97 (2.32)*	73.12 (5.02)**	76.77 (**7.01**)***	77.72 (2.64)**	
LA	HC	38.05	59.21	63.31	55.61	40.64 (2.59)*	63.78 (**4.57**)***	66.90 (3.59)**	59.33 (3.71)**	
AT	LA	35.58	71.13	69.84	75.36	38.14 (2.56)***	75.96 (4.82)***	77.58 (**7.73**)***	78.25 (2.89)***	68.41
CC	LA	42.47	62.75	74.04	69.26	45.27 (2.80)*	68.99 (**6.23**)**	76.60 (2.55)***	72.88 (3.61)***	
FM	LA	43.15	68.60	67.47	69.37	46.14 (2.99)***	73.75 (**5.15**)***	72.41 (4.93)***	73.41 (4.03)***	
HC	LA	40.16	52.11	53.42	71.16	43.25 (3.09)*	61.70 (**9.59**)***	61.05 (7.62)**	74.93 (3.77)**	
AT	FM	34.37	65.10	52.77	62.91	37.91 (3.53)	70.71 (**5.60**)*	56.43 (3.65)*	66.22 (3.31)*	61.49
CC	FM	40.57	66.42	60.17	52.23	44.18 (3.60)***	69.29 (2.87)***	65.33 (5.15)***	58.08 (**5.85**)***	
HC	FM	41.82	72.47	63.02	66.96	45.73 (3.90)***	74.59 (2.12)***	71.77 (**8.74**)***	69.74 (2.78)***	
LA	FM	42.71	59.92	55.80	57.36	45.51 (2.79)**	62.52 (2.60)**	58.67 (**2.87**)**	60.12 (2.75)**	
CC	AT	31.29	73.14	64.09	64.95	35.15 (3.85)***	75.34 (2.20)***	69.43 (**5.34**)***	67.05 (2.09)***	70.81
FM	AT	32.19	70.08	70.70	77.46	37.32 (5.13)***	73.91 (3.82)***	76.13 (**5.42**)***	80.05 (2.59)***	
HC	AT	34.87	76.31	72.27	67.28	38.21 (3.34)***	81.16 (**4.84**)***	74.37 (2.10)***	71.22 (3.93)***	
LA	AT	42.43	62.89	74.04	71.39	45.09 (2.66)**	69.37 (**6.48**)***	79.44 (5.39)***	74.05 (2.65)**	
Average		38.06	64.47	62.48	64.78	41.22 (3.15)	69.09 (4.62)	67.27 (**4.78**)	68.21 (3.42)	65.64

(Two-tailed t-test: *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$)

the rest 10% from S as the validation data, and all labelled tweets from T for testing.

Baselines: We consider the following approaches as our baselines: 1) **SO**: a source-only stance classification model based on a Bidirectional LSTM [3]; it is trained on the source data only, without using any adaptation; 2) **CORAL** [29]: it performs correlation alignment for minimising domain discrepancy by aligning the second-order statistics of the source/target distributions; 3) **DANN** [11]: an adversarial adaptation network that approximates the $\mathcal{H}$-divergence between domains; and 4) **WDGRL** [26]: an adversarial adaptation network that approximates the Wasserstein distance between domains.

All the above methods are single-view based and can be enhanced by DAN, i.e., they can be extended by specific components of DAN to learn the subjective and objective views, forming their respective dual-view variants: 5) **D-SO**: SO trained with the two view functions $\{F_{\text{subj}}, F_{\text{obj}}\}$ and classifiers $\{C_{\text{subj}}, C_{\text{obj}}\}$, and the fusion function U for combining the two views; 6) **D-CORAL**: besides adding $\{F_{\text{subj}}, F_{\text{obj}}, C_{\text{subj}}, C_{\text{obj}}, U\}$ to the model, CORAL is also extended with two CORAL losses [29] for the objective and subjective views, respectively; 7) **D-DANN**: in *view adaptation*, two $\mathcal{L}^{\mathcal{H}}_{\text{conf}}$-focused domain examiners $\{D^{\mathcal{H}}_{\text{subj}}, D^{\mathcal{H}}_{\text{obj}}\}$ are used to align the source/target data in the respective views; 8) **D-WDGRL**: two $\mathcal{L}^{\mathcal{W}}_{\text{conf}}$-focused domain examiners $\{D^{\mathcal{W}}_{\text{subj}}, D^{\mathcal{W}}_{\text{obj}}\}$ are used for *view adaptation*; and 9) **TO** [18]: we finally include as a reference the in-domain results from a state-of-the-art target-only method on the same dataset used here[7].

4.2 Implementation Details

Model: Each view function F is implemented as a RNN-based encoder to convert an utterance $\mathbf{x}$ into its feature $\mathbf{f}$. In this encoder, each word $x_j \in \mathbf{x}$ is first embedded into a d_e-dimensional word vector $\mathbf{w}_j = W[x_j]$, where $W \in \mathbb{R}^{d_e \times V}$ is the word embedding matrix, V is the vocabulary size, and $W[x]$ represents the x-th column of W. Then, a bi-directional LSTM with hidden size d_h is used to encode the word sequence $\mathbf{w} = \{w_j\}$ as $\mathbf{h}_j = \text{BiLSTM}(\mathbf{h}_{j-1}, \mathbf{h}_{j+1}, \mathbf{w}_j)$, where $\mathbf{h}_j \in \mathbb{R}^{2d_h}$ is the output for the jth time step. Finally, a linear mapping is used to project each $\mathbf{h}_j$ back to dimension d_h: $\mathbf{h}_j = \mathbf{W}_l\mathbf{h}_j + \mathbf{b}_l$, with $\mathbf{W}_l \in \mathbb{R}^{d_h \times 2d_h}, \mathbf{b}_l \in \mathbb{R}^{d_h}$ the trainable parameters. Each of the classifiers $\{C_{\text{subj}}, C_{\text{obj}}, C_{\text{stance}}\}$ and domain examiners $\{D^{\mathcal{H}}_{\text{subj}}, D^{\mathcal{H}}_{\text{obj}}, D^{\mathcal{W}}_{\text{subj}}, D^{\mathcal{W}}_{\text{obj}}\}$ is realised with a unique two-layer feed-forward network with hidden size d_f and ReLU activation.

Training: The pre-trained GloVe word vectors (d_e=200, glove.twitter.27B) are used to initialise the word embeddings, which are fixed during training. Batches of 8 samples from each domain are used in each training step. All models are optimised using Adam [16], with a varied learning rate based on the schedule: $lr = 10^{-3} \cdot \min(\frac{1}{\sqrt{step}}, \frac{step}{warmup})$. α, β, γ are set equally to 0.1 for balancing the corresponding loss terms. Each compared method is ran ten times with random initialisations under different seeds. The mean value of the evaluation metric is reported. The hidden sizes of LSTM (d_h) and feed-forward network (d_f) are randomly sampled from the interval $[100, 300]$ upon each run. A light dropout (0.1) is used. We pre-train a subjectivity classifier and an objectivity one to obtain the silver standard subjectivity/objectivity labels for our data. A widely-used subjectivity vs. objectivity dataset [24] is used for the pre-training, which consists of 5000 subjective sentences (movie reviews) and 5000 objective sentences (plot summaries). The pre-training is implemented with the FastText library [14].

4.3 Quantitative Results

We report the overall classification performance of DAN and the baselines on all adaptation tasks in Table 2.

First, all the DAN-enhanced methods (D-X) are shown to improve over their original versions on all adaptation tasks, with the improvements ranging from 2.12% to 9.59% at different significant levels. This shows that it is empirically superior to apply DAN for domain adaptation on stance data, and that the separate feature alignment for the subjective and objective stance expressions is effective in alleviating the domain drift issue. Among these methods, D-CORAL obtains

[7] The in-domain result on a target domain from [18] is measured on the official test set of that domain, while in this work the whole dataset (both official training and test sets) is used for testing. The results are partially comparable due to the shared test set used in both cases.

generally better performance than others on half of all tasks (10/20), suggesting that most domains of this dataset can be much aligned by the second-order statistics (feature covariances) of the samples.

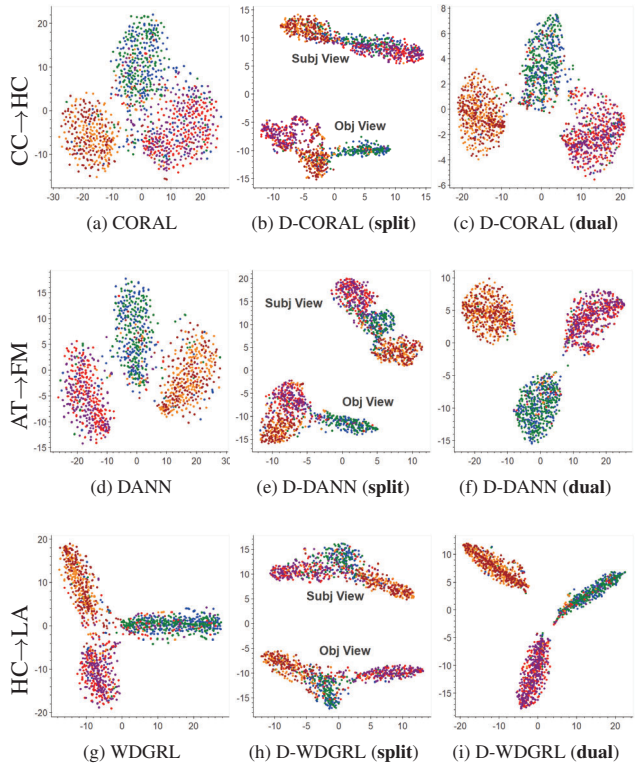
Second, the improvements achieved by the adaptative models (D-CORAL, D-DANN, and D-WDGRL) are generally higher than those by the non-adaptative model (D-SO). This suggests that the feature alignment between the source and target data (adaptation) can benefit more from the dual-view modelling of DAN than the source-only learning (no adaptation). This points out the key benefit of DAN that the features could be more easily aligned in the split views.

Finally, we notice that the results on certain target domains (e.g., tasks with CC as the target domain) are generally worse than those on others. One possible reason for this could be related to the inherent complexity of the data distributions of different domains; for example, the in-domain performance on CC is reportedly the poorest among all the five domains [21], thus probably making it the most challenging to transfer knowledge from other domains to it.

4.4 Qualitative Analysis

To gain a better understanding of what leads to DAN's superiority over the single-view methods, we conduct further experiments to qualitatively evaluate the properties of DAN.

(1) Visualising view features in DAN's enhanced feature space. We first derive insights into what has been learned in the feature space of DAN that makes its view features more transferable. For this, we plot the feature distributions of the source and target samples learned by all compared (adaptative) methods in Figure 3.



(a) CORAL (b) D-CORAL (**split**) (c) D-CORAL (**dual**)

(d) DANN (e) D-DANN (**split**) (f) D-DANN (**dual**)

(g) WDGRL (h) D-WDGRL (**split**) (i) D-WDGRL (**dual**)

Figure 3: The t-SNE plots of the feature distributions of the source and target samples learned by the three adaptative methods, {CORAL, DANN, WDGRL}, and their DAN-enhanced versions {D-CORAL, D-DANN, D-WDGRL}. Source samples are coloured by orange (*favour*), blue (*against*), and red (*neutral*), while target samples by brown (*favour*), green (*against*), and purple (*neutral*). Each row compares the plots of a baseline method (X) and its DAN-enhanced version (D-X) on a randomly selected task. For each of the DAN-enhanced methods (D-X), we plot both the intermediate subjective/objective features $\{\mathbf{f}_{\text{subj}}, \mathbf{f}_{\text{subj}}\}$ (Eq. 2) in its **split** view and the ultimately fused dual-view stance features $\mathbf{f}_{\text{dual}}$ (Eq. 6) in its **dual** view.

The first row shows the case of CORAL and D-CORAL, where we observe how DAN improves features' stance discriminating power. First, the plot of CORAL (Figure 3.a) exhibits a good separation between each of the three stance classes except that between *against* (blue/green) and *neutral* (red/purple), as the boundary between these two is rather blurred by many source *against* features (blue) invading the *neutral* region. The reason behind this seems to be revealed in the split view plot of D-CORAL (Figure 3.b), where a similar pattern (*against* and *neutral* classes overlap) is also seen in the **subj**ective view of D-CORAL. Fortunately, its **obj**ective view remedies this issue, by yielding features that better separate the problematic classes (i.e., *against* vs. *neutral*). As a result, the overall fused features in the dual view plot of D-CORAL (Figure 3.c) become more stance-discriminative, leading to a much better separation of all classes.

The second row demonstrates the case of how DAN improves features' domain-invariance with better feature alignment. In this case, the features learned by DANN already shows good discrimination in stance classes (Figure 3.d), but the alignment of the source and target features seems less effective, as in each class they tend to scatter over a relatively large area, making the distance longer between the two feature distributions[8]. In contrast, both the subjective and objective views of D-DANN produce more compact feature distributions within each class (Figure 3.e), suggesting smaller distances obtained between the source and target features. Consequently, we observe a much stronger feature alignment achieved within each class in the ultimate dual view of D-DANN (Figure 3.f).

Finally, the last row shows a more successful case of DAN than before, where it improves both the domain invariance and stance-discriminative power of the source and target features. As shown, although WDGRL already achieves superior feature learning than the previous single-view cases, D-WDGRL manages to produce even better dual-view features (Figure 3.i), which are shown to be more compact within the same classes and separable over different classes.

Overall, the above results suggest that, compared to the indiscriminate treatment applied in the single-view methods, the separate characterisation of the subjective and objective stance expressions in the split views of DAN could learn more transferable features for better adapting stance classifier across domains.

(2) Ablation analysis: subjective view vs. objective view. As the two views of DAN have shown to sometimes learn features with distinct transferability in the previous experiment, it is necessary to further examine how differently each view contributes to the overall adaptation. To this end, we build two ablation variants of DAN, namely D-DANN-SUBJ and D-DANN-OBJ (DANN is used here as an exemplified base for DAN), with each working in a specific view only. Specifically, for D-DANN-SUBJ (D-DANN-OBJ), we keep the subjectivity (objectivity) classifier for learning the view feature and the subjectivity (objectivity) domain examiner for making the feature domain-invariant. Figure 4 shows the results of all ablated D-DANN variants over the 20 adaptation tasks. As we can see, D-DANN-OBJ surpasses D-DANN-SUBJ on most of the tasks (16/20), with an averaged improvement of 3.38% on its dominating tasks. This indicates that the objective information in stance expressions is relatively more transferable than the subjective one. Since much of the subjectivity expressed in this dataset appears to be sentiments/emotions, this

[8] Note that it happens to be the case that each stance class from one domain matches its counterpart in the other domain in this experiment (i.e., source *favour* vs. target *favour*, etc.). The feature alignment in DAN is class-agnostic; it does not assume any kind of match between particular class labels a priori, nor does it impose any such constraint during training.

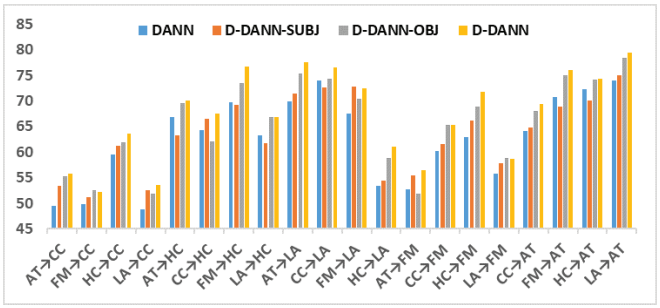


Figure 4: Performance of the ablated variants of DAN.

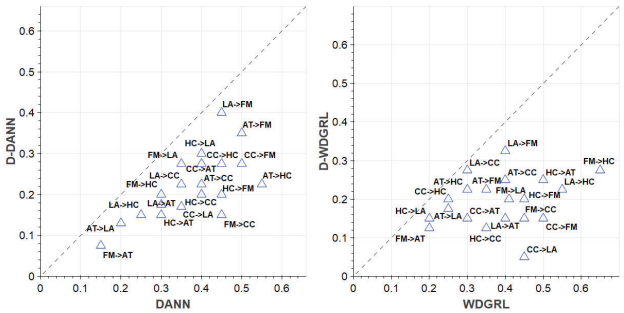


Figure 5: Proxy $\mathcal{A}$-distance between every pair of the evaluated domains.

finding is somewhat consistent with previous studies on stance and sentiment analysis [27, 22], where the utility of sentiments to stance recognition is sometimes limited. This may occur when the sentiments and stances of the same utterances do not correlate well [22], potentially causing the model to collapse in the case of disagreed subjectivity and stance signals. As a result, the subjective features alone could sometimes degrade the overall performance (D-DANN-SUBJ underperforms DANN on 6/20 tasks). Indeed, the objective information such as facts tend to be more stance-indicative, since the reasons (usually stating some facts) that support a stance are often found to be more specific to that stance [13, 35]. Finally, we observe that D-DANN, the full model combining both views, gives the best performance on almost all tasks (18/20), suggesting that *view fusion* in DAN is essential for obtaining the best of both views.

(3) Reduced domain discrepancy. Achieving good feature alignment is the key to successful domain adaptation. The visualisation in Figure 3 already shows some evidence for the superiority of DAN in matching the source and target features. Here we provide further analysis to quantify how good such feature matching is. In particular, we use the Proxy $\mathcal{A}$-distance (PAD) [5] as the metric, which is a common measure for estimating the discrepancy between a pair of domains. It is defined by $2(1 - 2\epsilon)$, where ϵ is the generalisation error on the problem of distinguishing between the source and target samples. Following the same setup in [10], we compute the PAD value by training and testing a linear SVM using a data set created from both source and target features of the training examples. For the DAN-enhanced methods, the fused dual-view features $\mathbf{f}_{\text{dual}}$ are used.

Figure 5 displays the results of comparing the PAD values of the two DAN-enhanced methods, D-DANN and D-WDGRL, with their original versions on the 20 adaptation tasks. As shown, both D-DANN and D-WDGRL achieve lower PAD values than their original counterparts across all the tasks. This validates the previous observations in Figure 3 that the source/target dual-view features learned by D-DANN and D-WDGRL are better aligned than the respective cases of DANN and WDGRL. Therefore, both the quantitative (Figure 5) and qualitative (Figure 3) results manifest the benefit of the proposed dual-view adaptation in matching stance data of different domains.

5 Related Work

In-domain stance classification has been studied in the context of online debate forums [28, 32, 12] and social media [3, 9, 22]. Recently, deep neural networks have been applied to learn rich representations of the stance-bearing utterances [3, 9, 22, 30], which could be further enhanced by incorporating the subjective and objective supervision into the representation learning.

While domain adaption has been successful in the related task of cross-domain sentiment classification [23, 19, 25], its efficacy in the stance context is much less explored. Along this line, pre-trained stance classifiers are fine-tuned for predicting stances on new domains [36]. Attention-based models are built to extract salient information from the source data, which is expected to also work well on a related target domain [34]. In contrast, we take a different approach to exploiting existing knowledge, by making features invariant to the shift of domains.

Adversarial approaches have recently gained popularity in domain adaptation for aligning feature distributions [20, 11, 31, 26]. In these approaches, a global feature space is solely induced to coarsely match the data from different domains. By contrast, we explore the potential for finding a more fine-grained alignment of the domains by creating split feature spaces (views) to fully characterise the subjective and objective stance expressions.

The interaction between stance and subjectivity in stance classification has been studied recently [27, 22], where the sentiment subjectivity shows its potential for predicting stances, although it is not as useful for stance classification as it is for sentiment classification. There are few efforts on exploring the utility of objective information in stance expressions. Some research examines reasons mentioned in the stance-bearing utterances [13, 8], which often contain factual information for supporting the stances. Different from all the above efforts, we leverage both subjective and objective information for better capturing the variety of stance expressions.

6 Conclusions and Future Work

In this paper, we propose the dual-view adaptation network, DAN, to adapt stance classifiers to new domains, which learns a subjective view and an objective view for every input utterance. We show that DAN allows existing single-view methods to acquire the dual-view transferability, and that it empirically improves those methods on various adaptation tasks. A series of qualitative analyses on the properties of DAN shows that more fine-grained adaptation for stance data can lead to more reduced domain discrepancies and finer stance discrimination, and that a proper view fusion is necessary for obtaining better overall features by leveraging the best of both views.

In the future, our work could be extended in several ways. First, we plan to evaluate our method on more diverse stance datasets with different linguistic properties. For instance, the utterances in posts of online debate forums [28] are typically longer, which may pose new challenges such as capturing dependencies across multiple sentences as well as richer subjective/objective expressions. Second, as DAN is input-agnostic (as long as the input is feature vectors), it would be interesting to apply it to other scenarios suitable for dual-view modelling. One example is modelling the social network user behaviours [6] where the networks of users together with their interactions provide a dual-view of their behaviours. Finally, it is possible that DAN could be extended for multi-view adaptation, e.g., by adding more view functions and domain examiners.

ACKNOWLEDGEMENTS

We would like to thank all anonymous referees for their constructive comments. We would also like to thank Xiang Dai for his helpful comments on drafts of this paper.

REFERENCES

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran, 'Domain adaptation with adversarial training and graph embeddings', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1077–1087, (2018).

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, 'Wasserstein generative adversarial networks', in *International conference on machine learning*, pp. 214–223, (2017).

[3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva, 'Stance detection with bidirectional conditional encoding', *arXiv preprint arXiv:1606.05464*, (2016).

[4] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim, 'Stance classification of context-dependent claims', in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 251–261, (2017).

[5] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, 'Analysis of representations for domain adaptation', in *Advances in neural information processing systems*, pp. 137–144, (2007).

[6] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida, 'Characterizing user behavior in online social networks', in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pp. 49–62. ACM, (2009).

[7] John Blitzer, Ryan McDonald, and Fernando Pereira, 'Domain adaptation with structural correspondence learning', in *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128. Association for Computational Linguistics, (2006).

[8] Filip Boltužić and Jan Šnajder, 'Back up your stance: Recognizing arguments in online discussions', in *Proceedings of the First Workshop on Argumentation Mining*, pp. 49–58, (2014).

[9] Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui, 'Stance classification with target-specific neural attention networks', in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, (2017).

[10] Yaroslav Ganin and Victor Lempitsky, 'Unsupervised domain adaptation by backpropagation', in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, pp. 1180–1189, (2015).

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, 'Domain-adversarial training of neural networks', *The Journal of Machine Learning Research*, **17**(1), 2096–2030, (2016).

[12] Kazi Saidul Hasan and Vincent Ng, 'Stance classification of ideological debates: Data, models, features, and constraints', in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1348–1356, (2013).

[13] Kazi Saidul Hasan and Vincent Ng, 'Why are you taking this stance? identifying and classifying reasons in ideological debates', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 751–762, (2014).

[14] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, 'Bag of tricks for efficient text classification', *arXiv preprint arXiv:1607.01759*, (2016).

[15] Young-Bum Kim, Karl Stratos, and Dongchan Kim, 'Adversarial adaptation of synthetic or stale data', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1297–1307, (2017).

[16] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[17] Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei, 'Deep memory networks for attitude identification', in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 671–680. ACM, (2017).

[18] Yingjie Li and Cornelia 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)*, pp. 6299–6305, Hong Kong, China, (November 2019). Association for Computational Linguistics.

[19] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang, 'Hierarchical attention transfer network for cross-domain sentiment classification', in *Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI)*, (2018).

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan, 'Learning transferable features with deep adaptation networks', *International Conference on Machine Learning (ICML)*, (2015).

[21] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry, 'Semeval-2016 task 6: Detecting stance in tweets', in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 31–41, (2016).

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

[23] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen, 'Cross-domain sentiment classification via spectral feature alignment', in *Proceedings of the 19th international conference on World wide web*, pp. 751–760. ACM, (2010).

[24] Bo Pang and Lillian Lee, 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271. Association for Computational Linguistics, (2004).

[25] Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang, 'Cross-domain sentiment classification with target domain specific information', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2505–2513, (2018).

[26] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu, 'Wasserstein distance guided representation learning for domain adaptation', in *Thirty-Second AAAI Conference on Artificial Intelligence*, (2018).

[27] Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko, 'Detecting stance in tweets and analyzing its interaction with sentiment', in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pp. 159–169, (2016).

[28] Swapna Somasundaran and Janyce Wiebe, 'Recognizing stances in ideological on-line debates', in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 116–124. Association for Computational Linguistics, (2010).

[29] Baochen Sun and Kate Saenko, 'Deep coral: Correlation alignment for deep domain adaptation', in *European Conference on Computer Vision*, pp. 443–450. Springer, (2016).

[30] Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou, 'Stance detection with hierarchical attention network', in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2399–2409, (2018).

[31] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, 'Adversarial discriminative domain adaptation', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7167–7176, (July 2017).

[32] Marilyn A Walker, Pranav Anand, Robert Abbott, and Ricky Grant, 'Stance classification using dialogic properties of persuasion', in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 592–596. Association for Computational Linguistics, (2012).

[33] Janyce Wiebe et al., 'Learning subjective adjectives from corpora', *Aaai/iaai*, **20**(0), 0, (2000).

[34] Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks, 'Cross-target stance classification with self-attention networks', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 778–783. Association for Computational Linguistics, (2018).

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

[36] Guido Zarrella and Amy Marsh, 'Mitre at semeval-2016 task 6: Transfer learning for stance detection', in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 458–463, (2016).

[37] Yuan Zhang, Regina Barzilay, and Tommi Jaakkola, 'Aspect-augmented adversarial networks for domain adaptation', *Transactions of the Association for Computational Linguistics*, **5**, 515–528, (2017).