# Improving Stance Detection with Multi-Dataset Learning and Knowledge Distillation

**Yingjie Li** **Chenye Zhao** **Cornelia Caragea**

University of Illinois at Chicago
{yli300,czhao43,cornelia}@uic.edu

## Abstract

Stance detection determines whether the author of a text is in favor of, against or neutral to a specific target and provides valuable insights into important events such as legalization of abortion. Despite significant progress on this task, one of the remaining challenges is the scarcity of annotations. Besides, most previous works focused on a hard-label training in which meaningful similarities among categories are discarded during training. To address these challenges, first, we evaluate a multi-target and a multi-dataset training settings by training one model on each dataset and datasets of different domains, respectively. We show that models can learn more universal representations with respect to targets in these settings. Second, we investigate the knowledge distillation in stance detection and observe that transferring knowledge from a teacher model to a student model can be beneficial in our proposed training settings. Moreover, we propose an Adaptive Knowledge Distillation (AKD) method that applies instance-specific temperature scaling to the teacher and student predictions. Results show that the multi-dataset model performs best on all datasets and it can be further improved by the proposed AKD, outperforming the state-of-the-art by a large margin. We publicly release our code.[1]

## 1 Introduction

People often express their stances toward specific targets (e.g., political figures, or abortion) on social media. These opinions can provide valuable insights into important events, e.g., legalization of abortion. The goal of stance detection is to determine whether the author of a text is in favor of, against or neutral toward a specific target (Mohammad et al., 2016b; Küçük and Can, 2020; AlDayel and Magdy, 2020). For example, for the tweet in Table 1, we can infer that the author is against to the legalization of abortion implied by the presence of the words "protect the sanctity of life".

| | |
|---|---|
| **Tweet:** | We all have a duty to protect the sanctity of life...from the first cell division, to the last. #ProtectLife #pjnet #ctot #ccot #SemST |
| **Target:** | Legalization of Abortion |
| **Stance:** | Against |

Table 1: An example of stance detection.

Even though stance detection has received a lot of attention, one of the biggest challenges for the stance detection tasks is the scarcity of annotated data. Even worse, previous studies (Mohammad et al., 2017; Du et al., 2017; Wei et al., 2018; Li and Caragea, 2019; Siddiqua et al., 2019) focused on a per-target training strategy, which aims to train one model for each target and evaluate it on the test data corresponding to that target (which we call ad-hoc training). In this case, the model is more likely to make predictions based on specific words without fully considering the target information, and hence, to overfit in the ad-hoc training setting. Motivated by these observations, we aim to investigate the following Research Question (RQ):

***RQ1:*** *Can we improve the performance of a stance detection model by training one model on all targets of each dataset and can we improve the performance further by training one model on all datasets?*

Toward this question, we evaluate two training settings: multi-target training and multi-dataset training, by training one model on each dataset and five datasets of different domains, respectively. We expect the model to learn more universal representations on the combined dataset and alleviate overfitting. On the other hand, compared to having many single-target models, a multi-target model or a multi-dataset model is simpler to deploy and more scientifically meaningful from the perspec-

[1] https://github.com/chuchun8/MDL-Stance-Distillation



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345
November 7–11, 2021. 

tive of building general natural language processing systems.

Besides the limited training data, models might also overfit to the ground truth labels (one-hot labels) that the meaningful rankings are destroyed, i.e., models fail to consider the similarity among different categories during training. Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher model to a student model by training the student model to imitate the teacher's prediction logits (which we call soft labels). It is commonly believed that the soft labels of the teacher model can benefit the student model by providing more training signals than one-hot labels. However, less attention has been paid to applying knowledge distillation to the stance detection. This naturally gives rise to another research question:

***RQ2:*** *Can knowledge distillation benefit the stance detection task in different training settings?*

Regarding RQ2, we apply various knowledge distillation methods in both multi-target and multi-dataset training settings. We train a teacher model and a student model for each dataset and all datasets for multi-target learning and multi-dataset learning, respectively. Moreover, we propose an Adaptive Knowledge Distillation (AKD) method that applies instance-specific temperature scaling to the predictions of teacher and student models. Experimental results show that knowledge distillation contributes to the performance improvement of stance detection models.

Even though we show that knowledge distillation can be beneficial to the stance detection task, how to most effectively transfer knowledge to the student remains an open question. Therefore, our third research question investigates:

***RQ3:*** *Which knowledge distillation setting benefit the stance detection task the most?*

In this paper, we perform empirical comparisons of three knowledge distillation settings: Single→Single, Multiple→Multiple and Multiple→Single. More specifically, we train only one teacher model and student model on all datasets for Single→Single and use Multiple→Multiple to indicate distilling single-dataset teacher models into single-dataset student models, i.e., both teacher and student models are trained on one dataset. Multiple→Single indicates distilling multiple teacher models individually trained on each dataset into one student model trained on all datasets.

In order to answer these questions, we fine-tune a pre-trained BERTweet (Nguyen et al., 2020) model for stance detection and perform the self-distillation (Furlanello et al., 2018) in both multi-target and multi-dataset training settings, i.e., both teacher and student models have the same model architecture. Our contributions include the following: 1) We evaluate three training settings (ad-hoc, multi-target and multi-dataset settings) for stance detection and observe that models trained in multi-target and multi-dataset settings show substantially better performance than models trained in ad-hoc setting. The model that is trained on all datasets performs best, outperforming the state-of-the-art by a large margin. 2) We explore the effectiveness of knowledge distillation on the stance detection and experimental results show that knowledge distillation can help improve the performance of stance detection models. We further propose an instance-specific temperature scaling method, which achieves superior performance on five stance detection datasets. 3) We show that Single→Single consistently outperforms other distillation settings, indicating that transferring the knowledge from a well-trained teacher that learns more universal representations is more beneficial to the stance detection.

## 2 Related Work

Stance detection task aims to identify the stance toward a specific target (Mohammad et al., 2016b; Küçük and Can, 2020; AlDayel and Magdy, 2020). The target is usually a political figure (Sobhani et al., 2017; Darwish et al., 2017; Grimminger and Klinger, 2021; Li and Caragea, 2021a; Li et al., 2021), a controversial topic such as marijuana legalization (Hasan and Ng, 2014; Mohammad et al., 2016a; Xu et al., 2016; Taulé et al., 2017; Swami et al., 2018; Stab et al., 2018; Zotova et al., 2020; Conforti et al., 2020a; Lai et al., 2020; Vamvas and Sennrich, 2020; Conforti et al., 2020b; Miao et al., 2020; Glandt et al., 2021) or a claim that could be a rumor's post (Qazvinian et al., 2011; Derczynski et al., 2015; Ferreira and Vlachos, 2016; Bar-Haim et al., 2017; Rao and Pomerleau, 2017; Derczynski et al., 2017; Gorrell et al., 2019). Besides the in-target stance detection where the test topic has always been seen in the training stage, cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020) and zero-shot stance detection (Allaway and McKeown, 2020) have also

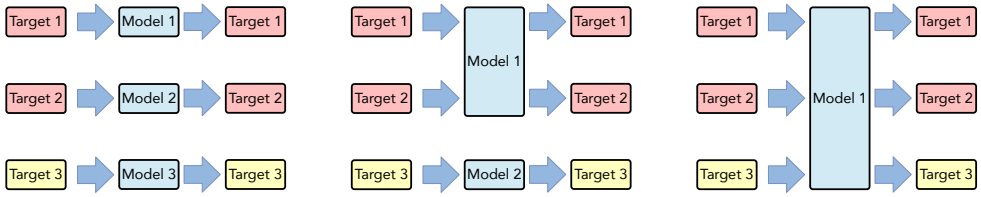


Figure 1: Overview of three training settings. The left figure represents the **ad-hoc** training setting in which a model is trained only on one target. The center figure represents the **multi-target** training setting in which a model is trained on all targets of each dataset. The right figure represents the **multi-dataset** training setting and a unified model is trained on all available targets. Assuming that dataset 1 consists of target 1 and target 2 and dataset 2 only contains target 3.

attracted a lot of attention in recent years. In this paper, we focus on the in-target stance detection.

Interestingly, despite significant progress on stance detection, the large-scale annotated datasets are limited and the number of training samples varies drastically between datasets. To make matters worse, previous studies (Mohammad et al., 2017; Du et al., 2017; Sun et al., 2018; Wei et al., 2018; Li and Caragea, 2019; Siddiqua et al., 2019; Sobhani et al., 2019; Li and Caragea, 2021b) adopted an ad-hoc training strategy, which means that the number of models that need to be trained is proportional to the number of targets. To address these issues, Schiller et al. (2021) explored multi-task learning for various stance detection tasks by fine-tuning the pre-trained BERT (Devlin et al., 2019) on multiple datasets. Different from Schiller et al. (2021), in this paper, we evaluate three different training settings on the datasets of diverse domains, showing the improvement brought by the joint training step by step. Moreover, we investigate whether knowledge distillation can help further improve the performance of stance detection models.

Knowledge distillation (Ba and Caruana, 2014; Hinton et al., 2015) aims to distill the knowledge from a teacher model into a student model and has been widely adopted and modified in computer vision (Romero et al., 2015; Gupta et al., 2016; Zagoruyko and Komodakis, 2017; Wang et al., 2019; Mirzadeh et al., 2020; Yuan et al., 2020) and natural language processing (Kim and Rush, 2016; Sun et al., 2019; Liu et al., 2019a; Aguilar et al., 2020; Sun et al., 2020; Tong et al., 2020; Currey et al., 2020; Jiao et al., 2020). Furlanello et al. (2018) proposed self-distillation in which the teacher and student models have identical architectures. Clark et al. (2019) further extended self-distillation to the multi-task setting to achieve superior performance than standard multi-task training. Zhang and Sabuncu (2020) attributed the success of self-distillation to the increasing uncertainty and diversity in teacher predictions.

Despite recent progress in knowledge distillation, less attention has been paid to combining knowledge distillation with stance detection. Miao et al. (2020) distilled knowledge in a semi-supervised manner for COVID-19 stance detection. However, experiments have only been conducted on a small dataset and the test set only contains one target. Motivated by recent works, we comprehensively investigate self-distillation in the stance detection under multi-target and multi-dataset training settings and evaluate the model performance on five datasets of different domains. Moreover, we propose an instance-specific temperature scaling method to further improve the self-distillation and explore how to effectively distill knowledge to the student model in a holistic way.

## 3 Methods

### 3.1 Model

BERTweet (Nguyen et al., 2020) is used as our base model, which is a pre-trained language model following the training procedure of RoBERTa (Liu et al., 2019c). We fine-tune the pre-trained BERTweet to predict the stance by appending a linear classification layer to the hidden representation of the $[CLS]$ token. The input is formulated as: *[CLS] target [SEP] sentence*.

### 3.2 Joint Training

Most previous work focused on an ad-hoc training setting (Figure 1(a)) that aims to train one model for each target, which fails to explore the potential of all the training data and is unable to learn universal representations of targets. Therefore, in order to explore the benefits of incorporating more training

data, we compare the performance of ad-hoc setting with two training settings: multi-target training and multi-dataset training, by training models on all targets of each dataset and on all targets of all datasets, respectively. More specifically, as shown in Figure 1(b), the multi-target model is trained and validated on data of all targets of each dataset, and tested on single target separately to be compared with the results of ad-hoc models. Different from the multi-target training, as shown in Figure 1(c), the multi-dataset model is trained and validated on the combination of all datasets in which training data come from different domains. Multi-target and multi-dataset training can be considered as one kind of multi-task learning approaches that help the pre-trained language models learn more generalized text representations by sharing the domain-specific information across the related targets.

### 3.3 Knowledge Distillation

In this subsection, we first introduce a vanilla knowledge distillation method, and then present our proposed Adaptive Knowledge Distillation (AKD) in details.

**Vanilla Knowledge Distillation** We assume that the training dataset $D^{tr}$ is composed of $m$ different datasets in multi-dataset training:

$$D^{tr} = D_1^{tr} + D_2^{tr} + ... + D_m^{tr}$$

$$D^{tr} = \{(x_i, t_i, y_i)\}_{i=1}^{n}$$

where $x_i$ is a sequence of words, $t_i$ is the corresponding target and $y_i$ is the hard label. The goal is to train a fixed-capacity model that performs well on targets of all $m$ datasets.

Standard supervised learning aims to minimize the cross-entropy loss $L_{CE}(p, y)$ of training data where $p$ denotes softmax outputs. In knowledge distillation, a teacher-student learning mechanism is used to improve the performance of the student model. Let $p_\tau^t$ denote the softmax outputs of the teacher model with temperature scaling and

$$p_\tau^t(k) = exp(z_k^t/\tau)/\sum_{j=1}^{K} exp(z_j^t/\tau)$$

where $\tau$ is the temperature used to scale the model predictions and $z^t$ is the output logits of the teacher model. The idea behind knowledge distillation (Hinton et al., 2015) is to transfer knowledge from the teacher model to the student model by minimizing the sum of cross-entropy loss between the predictions of student and hard labels and the distance loss between the predictions of student and teacher:

$$L_{KD} = (1-\alpha)L_{CE}(p, y) + \alpha L_{KL}(p_\tau, p_\tau^t)$$

where $L_{KL}$ is Kullback-Leibler (KL) divergence loss, $\alpha$ is the hyper-parameter that balances the importance of the cross-entropy loss and the KL divergence loss.

**Adaptive Knowledge Distillation** Previous works usually apply the same amount of temperature scaling to all teacher and student predictions. However, given a training set, we would expect some samples to be more representative of the label class than others and we hope to classify the typical examples with much greater confidence than the ambiguous ones. In this way, samples with larger confidence obtained from the teacher predictions should receive less amount of temperature scaling and vice versa. Therefore, we propose an Adaptive Knowledge Distillation (AKD) approach that applies instance-specific temperature scaling to the predictions of the teacher and student models. Formally, given a teacher output distribution $z_i^t$ of sample $i$, the temperature can be written as:

$$\tau_{z_i^t} = \begin{cases} T_1 & if \;\; 0 \leq max(softmax(z_i^t)) < a_1, \\ T_2 & if \;\; a_1 \leq max(softmax(z_i^t)) < a_2, \\ 1 & if \;\; a_2 \leq max(softmax(z_i^t)) \leq 1, \end{cases}$$

where $max(softmax(z_i^t))$ is the maximum probability of softmax output distribution, $a_1$ and $a_2$ are hyper-parameters to control the range of scaling, $T_1$ and $T_2$ are random variables that follow the uniform distributions, taking values in $(b_1, b_2)$, $(1, b_1)$, respectively, $b_1$ and $b_2$ are hyper-parameters that control the amount of scaling. By doing so, the amount of temperature scaling applied to a sample will be proportional to the amount of confidence the teacher model shows in that sample's prediction. Examples that are more challenging to classify will receive more temperature scaling applied to their soft labels. More specifically, we use higher temperature to soften the teacher prediction of a sample if the teacher shows lower confidence in that sample's prediction and vice versa.

We perform the self-distillation (Furlanello et al., 2018; Zhang and Sabuncu, 2020) in both multi-target and multi-dataset training settings, i.e., both teacher and student models have the same network

architecture. Self-distillation can be repeated iteratively to further improve the performance: the trained student model can be treated as the new teacher model and the knowledge can be further distilled to another student model. However, to better demonstrate the benefits of applying our proposed AKD approach to stance detection, we train teacher and student models only once for all distillation methods.

## 4 Experimental Settings

In this section, we first describe stance detection datasets used for evaluation and introduce the evaluation metrics. Then, we describe several baseline methods of knowledge distillation.

### 4.1 Datasets

Stance detection datasets of diverse domains are used to evaluate the performance of the proposed models. We train and validate the multi-dataset model on the combined dataset of SemEval, MT, AM, WT-WT and COVID-19. We then test the generalization abilities of stance detection models on unseen datasets WT-WT-E and Election-2020. Summary statistics of these datasets are shown in Tables 2, 3, 4, 5, 6, 7 and examples of these datasets are shown in Table 8. Datasets used for training a multi-dataset model are described as follows.

**SemEval** SemEval-2016 (Mohammad et al., 2016a) is a benchmark stance dataset and contains five different targets: "Atheism", "Climate Change", "Feminist Movement", "Hillary Clinton" and "Legalization of Abortion". The dataset is annotated for detecting whether the author is against to, neutral or in favor of a given target. We split the train set in a 5:1 ratio into train and validation sets and removed the target "Climate Change" due to the limited and highly skewed data. The test set of each target is the same as provided by the authors.

**MT** Multi-Target stance dataset (Sobhani et al., 2017) is a political dataset containing presidential candidates of 2016 US election. It contains three sets of tweets corresponding to target pairs: "Donald Trump and Hillary Clinton", "Donald Trump and Ted Cruz", "Hillary Clinton and Bernie Sanders". The task aims at detecting the stances (against, none or favor) toward two targets for each data. Train, validation and test sets are as provided by the authors.

**AM** AM (Stab et al., 2018) is an argument mining dataset containing eight different topics: "Abortion", "Cloning", "Death Penalty", "Gun Control", "Marijuana Legalization", "Minimum Wage", "Nuclear Energy" and "School Uniforms". The dataset is annotated for detecting whether an argument is in support of, neutral or opposed to a given topic. Train, validation and test sets are as provided by the authors.

**WT-WT** WT-WT (Conforti et al., 2020b) is a financial dataset and the task aims at detecting the stance toward mergers and acquisition operations between companies. This dataset consists of four targets in the healthcare domain and one target in the entertainment domain. We train the model on the four target pairs of healthcare domain. Each tweet of WT-WT is annotated with one of four labels (refute, comment, support and unrelated). We split the dataset in a 10:2:3 ratio into train, validation and test sets and removed the data of label "unrelated" to be consistent with other datasets.

**COVID-19** COVID-19 (Miao et al., 2020) is a stance detection dataset collected during COVID-19 pandemic, which contains one target "Lockdown in New York State". The dataset is annotated for detecting whether the author is in support of, neutral or against to the lockdown policy in New York State of United States. We split the train set in a 5:1 ratio into train and validation sets and used the test set as provided by the authors.

Two additional datasets are used to test the generalization abilities of stance detection models (no sample from these datasets is used for training).

**WT-WT-E** Target "Fox and Disney" of WT-WT (Conforti et al., 2020b) in the entertainment domain is used to test the generalization ability of stance detection models.

**Election-2020** Election-2020 (Grimminger and Klinger, 2021) is a political dataset containing two presidential candidates: "Donald Trump" and "Joe Biden" of 2020 US elections. The task aims at detecting the stance (favor, against, neutral, neither or mixed) toward a given target. We test the generalization ability of the model on this dataset and removed the data of label "neither" and "mixed" to be consistent with other datasets.

### 4.2 Evaluation Metrics

Similar to Mohammad et al. (2016a) and Sobhani et al. (2017), $F_{avg}$ and macro-average of F1-score

| Target | #Train | %Favor | %Against | %None | #Test | %Favor | %Against | %None |
|---|---|---|---|---|---|---|---|---|
| **Atheism** | 513 | 17.93 | 59.26 | 22.81 | 220 | 14.54 | 72.73 | 12.73 |
| **Climate** | 395 | 53.67 | 3.80 | 42.53 | 169 | 72.78 | 6.51 | 20.71 |
| **Feminism** | 664 | 31.63 | 49.40 | 18.97 | 285 | 20.35 | 64.21 | 15.44 |
| **Hillary** | 689 | 17.13 | 57.04 | 25.83 | 295 | 15.25 | 58.31 | 26.44 |
| **Abortion** | 653 | 18.53 | 54.36 | 27.11 | 280 | 16.43 | 67.50 | 16.07 |
| **Total** | 2,914 | 25.84 | 47.87 | 26.29 | 1,249 | 24.34 | 57.25 | 18.41 |

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

| Topic | #Total | %Support | %Oppose | %None |
|---|---|---|---|---|
| **Abortion** | 3,929 | 17.31 | 20.92 | 61.77 |
| **Cloning** | 3,039 | 23.23 | 27.61 | 49.16 |
| **Death Penalty** | 3,651 | 12.52 | 30.43 | 57.05 |
| **Gun Control** | 3,341 | 23.56 | 19.90 | 56.54 |
| **Marijuana Legalization** | 2,475 | 23.72 | 25.29 | 50.99 |
| **Minimum Wage** | 2,473 | 23.29 | 22.28 | 54.43 |
| **Nuclear Energy** | 3,576 | 16.95 | 23.82 | 59.23 |
| **School Uniforms** | 3,008 | 18.12 | 24.23 | 57.65 |
| **Total** | 25,492 | 19.40 | 24.30 | 56.30 |

Table 3: Data distribution of AM dataset (Stab et al., 2018).

| Target | #Total | %Refute | %Comment | %Support | %Unrelated |
|---|---|---|---|---|---|
| **Cigna and Express Scripts** | 2,527 | 10.01 | 37.47 | 30.58 | 21.92 |
| **Aetna and Humana** | 7,897 | 14.00 | 35.50 | 13.14 | 37.34 |
| **CVS Health and Aetna** | 11,622 | 4.45 | 47.49 | 21.24 | 26.80 |
| **Anthem and Cigna** | 11,044 | 17.82 | 28.05 | 8.78 | 45.33 |
| **21st Century Fox and Disney** | 18,044 | 2.09 | 46.92 | 7.73 | 43.26 |
| **Total** | 51,134 | 8.25 | 40.75 | 13.00 | 38.00 |

Table 4: Data distribution of WT-WT dataset (Conforti et al., 2020b).

| Target | #Total | %Favor | %Against | %Neutral | %Mixed | %Neither |
|---|---|---|---|---|---|---|
| **Donald Trump** | 3,000 | 26.00 | 28.07 | 11.37 | 0.66 | 33.90 |
| **Joe Biden** | 3,000 | 41.20 | 13.47 | 10.87 | 1.56 | 32.90 |

Table 5: Data distribution of Election-2020 dataset (Grimminger and Klinger, 2021).

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

Table 6: Distribution of instances in Multi-Target dataset (Sobhani et al., 2017).

| Target | #Total | #Train | #Test |
|---|---|---|---|
| **Lockdown in New York** | 1,097 | 733 | 364 |

Table 7: Distribution of instances in COVID-19 dataset (Miao et al., 2020).

($F_{macro}$) are adopted to evaluate the performance of our baseline models. $F_{avg}$ is calculated by averaging the F1-scores of label "Favor" and "Against". We calculate the $F_{avg}$ for each target and $F_{macro}$ is calculated by averaging the $F_{avg}$ across all targets for each dataset. Further, we can obtain $avgF_m$ by averaging the $F_{macro}$ across all datasets.

### 4.3 Baseline Methods

We run experiments with the following baseline methods:

**Base**: The pre-trained BERTweet (Nguyen et al., 2020) is fine-tuned under the PyTorch framework for 5 epochs. The maximum length is set to 128 and the batch size is 32. We use AdamW optimizer (Loshchilov and Hutter, 2019) and the learning rate is 2e-5. Each experiment is conducted on a single NVIDIA V100 GPU.

**KD**: A vanilla knowledge distillation method with temperature scaling. The student has the same model architecture as the teacher.

**LSR** (Szegedy et al., 2016): A label smoothing regularization technique used to encourage the base model to be less confident in making predictions.

**TFKD** (Yuan et al., 2020): A teacher-free knowledge distillation method that regularizes the model with manually designed label distribution.

The proposed methods are listed as follows:

**AKD**: The proposed adaptive knowledge distillation method that improves vanilla knowledge distillation by classifying the typical examples with greater confidence than an ambiguous example. ($a_1$, $a_2$) are chosen from {(0.6, 0.8),(0.7, 0.9)}.

| Dataset | Target | Tweet | Stance |
|---|---|---|---|
| SemEval | Feminist Movement | We celebrate the company of 50 great people & organisations. From #fiji-lovers #perfumers #ecowarrior #vegan #humanrights #healers #SemST | None |
| AM | Nuclear Energy | It has been determined that the amount of greenhouse gases have decreased by almost half because of the prevalence in the utilization of nuclear power . | Favor |
| WT-WT | Aetna and Humana | $AET $HUM Hearing Aetna Humana deal blocked by Federal Judge as anticompetitive | Against |
| MT | Bernie Sanders | #Hillary > #Bernie That word salad couldn't have been better if it came from Sarah Palin. @OnlyTruthReign @yoloswagnamstyl @coltonjbauer | Against |
| COVID-19 | Lockdown in New York | @MichaelSholler2 The pandemic doesn't care about plans. They'd have to lock down until a vaccine to avoid a repeat. NY might minimize it next time but it's unavoidable. People are morons and can't help but screw things up. And yes, rest of us are screwed the rest of 2020 | Favor |
| WT-WT-E | Fox and Disney | 21st Century Fox president believes Disney is a better fit for the company than Comcast | Favor |
| Election-2020 | Joe Biden | @alienfound @Panther7112 Nope Joe Biden is soft on china second he and his VP has bad thing to say about India relating to removal of 370 it will for sure benefit Pakistan in future that's for sure so it's going to be tough international pressure is going to be huge because of Kamala harish | Against |

Table 8: Example from each stance detection dataset.

$b_1$ and $b_2$ are set to 2 and 3, respectively[2].

**AKD-plus**: A variation of AKD with oversampling. First, we find the target with the largest number of training samples. Let $T_{max}$ denote this number. Second, for each of the remaining targets, we oversample its sentences until we obtain $T_{max}$ training samples for that target.

## 5 Results

In this section, we thoroughly discuss the experimental results to answer our research questions presented in §1. First, we study the performance of models in three training settings in §5.1. Then we explore the effectiveness of different distillation models on stance detection datasets and test the generalization ability of knowledge distillation models in §5.2. We finally compare the distillation models in different settings in §5.3.

### 5.1 Multi-Target Training and Multi-Dataset Training (RQ1)

We use the Base$_{Multiple}$ and Base$_{Single}$ to indicate the base models trained in multi-target and multi-dataset settings, respectively[3]. Table 9 shows performance comparisons of ad-hoc, multi-target and multi-dataset settings. Each result is the average of six runs with different initializations. First, we can observe that Base$_{Multiple}$ and Base$_{Single}$ significantly outperform the model trained in the ad-hoc setting. The performance of Base$_{Multiple}$ is the same with Base$_{Ad-hoc}$ on the COVID-19 dataset because there is only one target in this dataset.

Second, Base$_{Single}$ shows promising improvements over Base$_{Multiple}$, which demonstrates that Base$_{Single}$ learns more universal representations with respect to targets by leveraging the data from datasets of diverse domains. Note that Base$_{Single}$ achieves a substantial improvement over Base$_{Multiple}$ on SemEval and COVID-19 datasets. One possible reason is that Base$_{Multiple}$ still overfits the training data heavily and training on all datasets can alleviate overfitting. Last, we can also observe that Base$_{Single}$ outperforms the current state-of-the-art models on the MT and SemEval stance datasets, demonstrating its effectiveness.

[2]We expect to see further improvements by tuning the hyper-parameters in wider range. However, we only tune the hyper-parameters within a small range due to limited computational resources.

[3]The subscripts "Multiple" and "Single" mean that we need to train multiple models for multi-target training (i.e., one model for one dataset) and train only single model for multi-dataset training (i.e., one model for all datasets).

| Model | WT-WT | MT | SemEval | COVID-19 | AM | avgF$_m$ |
|---|---|---|---|---|---|---|
| State-of-the-art | - | 58.72* | 64.75* | - | - | - |
| Base$_{Ad-hoc}$ | 70.88 | 59.45 | 65.96 | 51.67 | 60.08 | 61.61 |
| Base$_{Multiple}$ | 74.04† | 68.57† | 65.25 | 51.67 | 64.11† | 64.73 |
| Base$_{Single}$ | **74.60**† | **69.56**† | **67.39**† | **58.02**† | **64.91**† | **66.90** |

Table 9: Performance comparisons of different training settings on stance detection datasets. Bold scores are best overall. $*$: the result is retrieved from Siddiqua et al. (2019). †: the proposed model improves the ad-hoc model at $p < 0.05$ with paired t-test. Model performance on each target of each dataset is shown in Appendix A.1.

| Model | WT-WT | MT | SemEval | COVID-19 | AM | avgF$_m$ |
|---|---|---|---|---|---|---|
| Base$_{Multiple}$ | 74.04 | 68.57 | 65.25 | 51.67 | 64.11 | 64.73 |
| +KD$_{Multiple\rightarrow Multiple}$ | 74.60 | 69.94 | 66.47 | 52.02 | 64.43 | 65.49 |
| +LSR$_{Multiple\rightarrow Multiple}$ | 74.61 | 69.49 | 65.60 | 52.65 | 64.03 | 65.28 |
| +TFKD$_{Multiple\rightarrow Multiple}$ | 74.50 | 69.37 | 66.06 | 52.06 | 63.65 | 65.13 |
| +**AKD**$_{Multiple\rightarrow Multiple}$ | **75.12** | 69.95 | 67.31 | 53.61 | 64.86 | 66.17 |
| Base$_{Single}$ | 74.60 | 69.56 | 67.39 | 58.02 | 64.91 | 66.90 |
| +KD$_{Single\rightarrow Single}$ | 74.92 | 70.02 | 67.34 | 57.66 | 64.75 | 66.94 |
| +LSR$_{Single\rightarrow Single}$ | 74.56 | 70.24 | 67.49 | 58.68 | 63.88 | 66.97 |
| +TFKD$_{Single\rightarrow Single}$ | 74.50 | 69.07 | 68.79 | 58.14 | 64.54 | 67.01 |
| +**AKD**$_{Single\rightarrow Single}$ | 75.00 | **70.45** | **69.18** | **61.27** | **64.93** | **68.17** |

Table 10: Performance comparisons of different distillation models on stance detection datasets. Underlined scores are best within groups of models with same teachers; bold scores are best overall. Model performance on each target of each dataset is shown in Appendix A.2.

| Model | WT-WT-E | Election-2020 |
|---|---|---|
| Base$_{Single}$ | 46.94 | 73.42 |
| +KD$_{Single\rightarrow Single}$ | 46.76 | 73.53 |
| +LSR$_{Single\rightarrow Single}$ | 47.00 | 73.87 |
| +TFKD$_{Single\rightarrow Single}$ | 45.79 | 72.17 |
| +**AKD**$_{Single\rightarrow Single}$ | **48.28** | **74.52** |

Table 11: Performance comparisons of distillation models on the unseen datasets. Best results are marked in bold.

## 5.2 Stance Detection with Knowledge Distillation (RQ2)

Table 10 shows performance comparisons of different distillation models on five stance detection datasets. We observe that all distillation models show improvements over their base models in $avgF_m$, which demonstrates that knowledge distillation can benefit the stance detection. Moreover, our proposed model AKD that performs instance-specific temperature scaling outperforms knowledge distillation with fixed temperature for each instance in both settings. Specifically, AKD$_{Multiple\rightarrow Multiple}$ and AKD$_{Single\rightarrow Single}$ outperform the vanilla knowledge distillation models by 0.68% and 1.23% in $avgF_m$ in multi-target and multi-dataset training settings, respectively, which indicates that training with instance-specific temperature scaling leads to better performance. Note that distillation models show less improvements in multi-dataset learning. One explanation is that knowledge distillation can be viewed as the instance-specific regularization on the softmax outputs of neural networks and the effect of knowledge distillation diminishes with increasing the size of train set (Zhang and Sabuncu, 2020).

We test the generalization ability of knowledge distillation models on the unseen WT-WT-E dataset and Election-2020 dataset. Even though target "Donald Trump" of Election-2020 dataset has been seen in training data, the task is still challenging since the target-related topics in election 2016 are quite different from the ones in 2020[4]. Table 11 shows performance comparisons of various distillation models in multi-dataset training setting. We can observe that our proposed model achieves the best performance on both datasets, showing better generalization abilities.

## 5.3 Different Distillation Settings (RQ3)

We further compare Single→Single distillation with several variants (Multiple→Single and Single→Single with oversampled data). Experimental results of training models in different distillation settings are shown in Table 12. First, we can observe that AKD$_{Single\rightarrow Single}$ performs best overall. Specifically, AKD$_{Single\rightarrow Single}$

[4]The tweets of Donald Trump in 2016 are different from the ones in 2020 because the target-related events vary a lot. People may support Donald Trump by attacking the email scandal of Hillary Clinton in 2016, and support Donald Trump by attacking the corruption of Joe Biden in 2020.

| Model | WT-WT | MT | SemEval | COVID-19 | AM | avgF$_m$ |
|---|---|---|---|---|---|---|
| Base$_{Multiple}$ | 74.04 | 68.57 | 65.25 | 51.67 | 64.11 | 64.73 |
| +AKD$_{Multiple \rightarrow Multiple}$ | **75.12** | 69.95 | 67.31 | 53.61 | 64.86 | 66.17 |
| +AKD$_{Multiple \rightarrow Single}$ | 74.54 | 70.16 | 68.73 | 57.31 | 64.91 | 67.13 |
| Base$_{Single}$ | 74.60 | 69.56 | 67.39 | 58.02 | 64.91 | 66.90 |
| +AKD-plus$_{Single \rightarrow Single}$ | 74.04 | **70.56** | 68.88 | 58.44 | 64.00 | 67.18 |
| +AKD$_{Single \rightarrow Single}$ | 75.00 | 70.45 | **69.18** | **61.27** | **64.93** | **68.17** |

Table 12: Performance comparisons of different distillation settings on stance detection datasets. Underlined scores are best within groups of models with same teachers; bold scores are best overall.

| Model | avgF$_m$ |
|---|---|
| Base$_{Single}$ | 66.90 |
| +Single-Dataset Fine-Tuning | +0.49 |
| Base$_{Single}$+AKD$_{Single \rightarrow Single}$ | 68.17 |
| +Single-Dataset Fine-Tuning | +0.22 |

Table 13: Performance of single-dataset fine-tuning after multi-dataset training on the combined dataset (SemEval, MT, AM, WT-WT and COVID-19).

leads to significant performance gains than AKD$_{Multiple \rightarrow Multiple}$ on SemEval and COVID-19 datasets, reinforcing the claim that multi-dataset training helps models learn more generalized text representations. Moreover, AKD$_{Single \rightarrow Single}$ consistently outperforms AKD$_{Multiple \rightarrow Single}$, indicating that transferring the knowledge from a well-trained teacher model is more beneficial to the stance detection task.

Second, we can observe that AKD$_{Single \rightarrow Single}$ shows improvements over AKD-plus$_{Single \rightarrow Single}$. This might be due to the difference in size between the target with the largest train set (CVS and Aetna, 5,040 sentences) and the target with the smallest train set (Atheism, 439 sentences).

### 5.4 Single-Dataset Fine-Tuning

Multi-task models such as MT-DNN (Liu et al., 2019b) achieve further improvements by continuing training the model on individual tasks after the multi-task training. However, we do not fine-tune the model on each dataset after multi-dataset training because our goal is training one model for all datasets instead of training one model for each dataset. Moreover, one multi-dataset model is much easier to deploy, and thus has more practical value.

We evaluate the effectiveness of single-dataset fine-tuning on the base model and distillation model in Table 13. We first train a multi-dataset model and then fine-tune the model on each dataset. It is unsurprising to observe that single-dataset fine-tuning further improves the performance of both models.

## 6 Conclusion

In this paper, we formulated three research questions for which evidence-based answers were unknown. We conducted extensive experiments on stance detection datasets and answer the questions as follows: 1) The performance of a stance detection model can be significantly improved by training on all targets of each dataset and on multiple datasets. Moreover, the model trained on datasets of diverse domains shows superior performance than the model trained on each dataset, indicating that the multi-dataset model benefits from learning with more training data and the multi-target model might still overfit the training data. 2) Self-distillation can further improve the stance detection in both training settings. Our proposed AKD benefits stance detection the most and shows better generalization abilities over other knowledge distillation methods. 3) We explore different distillation settings and observe that Single→Single achieves the best performance overall, which indicates that distilling knowledge from a well-trained teacher is more beneficial to the stance detection.

Future work includes further strengthening the multi-dataset model by incorporating more stance detection datasets. It would be also interesting to extend the knowledge distillation to more stance detection tasks such as rumour detection and multilingual stance detection.

## Acknowledgments

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

## References


Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. 2020. Knowledge distillation from internal representations. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7350–7357.

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

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems*, volume 27.

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

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. Bam! Born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020a. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101.

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

Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. Distilling multiple domains for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What went viral during the 2016 US presidential election. In *Social Informatics*, pages 143–161.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Tomas Pariente Lobo, Anna Kolliakou, Robert Stewart, Sara-Jayne Terp, Geraldine Wong, Christian Burger, Arkaitz Zubiaga, Rob Procter, and Maria Liakata. 2015. Pheme: Computing veracity — the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

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

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1607–1616.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2827–2836.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

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

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

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

Yingjie Li and Cornelia Caragea. 2021a. A multi-task learning framework for multi-target stance detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2320–2326.

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

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Linqing Liu, Huan Wang, Jimmy Lin, Richard Socher, and Caiming Xiong. 2019a. Attentive student meets multi-task teacher: Improved knowledge distillation for pretrained models. *arXiv preprint arXiv:1911.03588*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019c. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on COVID-19 intervention measures. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher. *AAAI 2020*.

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

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

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

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

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Delip Rao and Dean Pomerleau. 2017. Fake news challenge.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*.

Umme Aymun Siddiqua, Abu Nowshed Chy, and Masaki Aono. 2019. Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1868–1873.

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

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

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674.

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

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4323–4332.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

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

Meihan Tong, Bin Xu, Shuai Wang, Yixin Cao, Lei Hou, Juanzi Li, and Jun Xie. 2020. Improving event detection via open-domain trigger knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5887–5897.

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

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4933–4942.

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

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

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

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. 2020. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911.

Sergey Zagoruyko and Nikos Komodakis. 2017. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197.

Zhilu Zhang and Mert Sabuncu. 2020. Self-distillation as instance-specific label smoothing. In *Advances in Neural Information Processing Systems*, volume 33, pages 2184–2195.

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

# A More Experimental Results

## A.1 RQ1

We use the $\text{Base}_{Multiple}$ and $\text{Base}_{Single}$ to indicate the base models trained in multi-target setting and multi-dataset setting, respectively. Table 14, 15, 16, 20 show performance comparisons of ad-hoc, multi-target and multi-dataset models on each target of each dataset. Multi-target and multi-dataset models are first trained and validated on all targets of each dataset and all targets of all datasets, respectively. Then, the well-trained models are tested on single target separately to be compared with the results of ad-hoc models. Note that we do not report the results of COVID-19 dataset here because COVID-19 dataset only consists of one target and we have reported the results in the paper. Experimental results show that $\text{Base}_{Single}$ consistently outperforms $\text{Base}_{Multiple}$ and $\text{Base}_{Ad-hoc}$ on all stance datasets, indicating that $\text{Base}_{Single}$ learns more universal representations with respect to targets by leveraging the data from datasets of diverse domains.

## A.2 RQ2

Table 17, 18, 19, 21 show performance comparisons of different distillation models on stance detection datasets. $\text{KD}_{M\rightarrow M}$ and $\text{KD}_{S\rightarrow S}$ are short for $\text{KD}_{Multiple\rightarrow Multiple}$ and $\text{KD}_{Single\rightarrow Single}$, respectively. We can observe that our proposed AKD outperforms the vanilla knowledge distillation model on 16 and 15 (out of 20) targets in multi-target and multi-dataset training settings, respectively, which reinforces our claim that training with instance-specific temperature scaling leads to better performance.

| **Targets** | CI_ES | AET_HUM | CVS_AET | AN_CI |
|---|---|---|---|---|
| $\text{Base}_{Ad-hoc}$ | 57.28 | 75.83 | 74.50 | 75.89 |
| $\text{Base}_{Multiple}$ | 67.04 | **77.53** | 75.55 | 76.05 |
| $\text{Base}_{Single}$ | **69.27** | 77.44 | **75.57** | **76.12** |

Table 14: Performance comparisons of different training settings on the **WT-WT** dataset. Bold scores are best overall.

| **Targets** | Trump-Clinton | Trump-Cruz | Clinton-Sanders |
|---|---|---|---|
| $\text{Base}_{Ad-hoc}$ | 63.41 | 57.84 | 57.10 |
| $\text{Base}_{Multiple}$ | 68.04 | 70.02 | 67.64 |
| $\text{Base}_{Single}$ | **69.40** | **70.60** | **68.68** |

Table 15: Performance comparisons of different training settings on the **MT** dataset. Bold scores are best overall.

| **Targets** | Atheism | Feminism | Clinton | Abortion |
|---|---|---|---|---|
| $\text{Base}_{Ad-hoc}$ | **70.75** | 59.55 | 70.45 | 63.09 |
| $\text{Base}_{Multiple}$ | 68.54 | 59.35 | 66.73 | 66.36 |
| $\text{Base}_{Single}$ | 67.09 | **62.48** | **72.55** | **67.42** |

Table 16: Performance comparisons of different training settings on the **SemEval** dataset. Bold scores are best overall.

| **Targets** | CI_ES | AET_HUM | CVS_AET | AN_CI |
|---|---|---|---|---|
| $\text{Base}_{Multiple}$ | 67.04 | 77.53 | 75.55 | 76.05 |
| $+\text{KD}_{M\rightarrow M}$ | 68.62 | 77.57 | 76.11 | 76.11 |
| $+\textbf{AKD}_{M\rightarrow M}$ | 69.21 | **77.97** | **77.10** | 76.20 |
| $\text{Base}_{Single}$ | 69.27 | 77.44 | 75.57 | 76.12 |
| $+\text{KD}_{S\rightarrow S}$ | **70.21** | 77.67 | 75.12 | 76.66 |
| $+\textbf{AKD}_{S\rightarrow S}$ | 69.58 | 77.72 | 75.80 | **76.91** |

Table 17: Performance comparisons of different distillation models on the **WT-WT** dataset. Underlined scores are best within groups of models with same teachers; bold scores are best overall.

| **Targets** | Trump-Clinton | Trump-Cruz | Clinton-Sanders |
|---|---|---|---|
| $\text{Base}_{Multiple}$ | 68.04 | 70.02 | 67.64 |
| $+\text{KD}_{M\rightarrow M}$ | 69.42 | 70.78 | 69.61 |
| $+\textbf{AKD}_{M\rightarrow M}$ | 69.16 | 71.01 | 69.69 |
| $\text{Base}_{Single}$ | 69.40 | 70.60 | 68.68 |
| $+\text{KD}_{S\rightarrow S}$ | 69.58 | 70.79 | 69.70 |
| $+\textbf{AKD}_{S\rightarrow S}$ | **70.16** | **71.38** | **69.80** |

Table 18: Performance comparisons of different distillation models on the **MT** dataset. Underlined scores are best within groups of models with same teachers; bold scores are best overall.

| **Targets** | Atheism | Feminism | Clinton | Abortion |
|---|---|---|---|---|
| $\text{Base}_{Multiple}$ | 68.54 | 59.35 | 66.73 | 66.36 |
| $+\text{KD}_{M\rightarrow M}$ | 71.19 | 59.38 | 69.09 | 66.20 |
| $+\textbf{AKD}_{M\rightarrow M}$ | 70.99 | 59.83 | 69.29 | 69.13 |
| $\text{Base}_{Single}$ | 67.09 | 62.48 | 72.55 | 67.42 |
| $+\text{KD}_{S\rightarrow S}$ | 65.14 | 61.56 | **73.40** | 69.27 |
| $+\textbf{AKD}_{S\rightarrow S}$ | **71.59** | **62.95** | 71.73 | **70.46** |

Table 19: Performance comparisons of different distillation models on the **SemEval** dataset. Underlined scores are best within groups of models with same teachers; bold scores are best overall.

| **Targets** | Abortion | Cloning | Death Penalty | Gun Control | Marijuana | Minimum Wage | Nuclear Energy | Uniform |
|---|---|---|---|---|---|---|---|---|
| Base$_{Ad-hoc}$ | 52.87 | 69.90 | 53.57 | 49.63 | 65.14 | 65.66 | 60.01 | 63.91 |
| Base$_{Multiple}$ | 57.02 | 71.44 | 58.23 | 55.23 | 67.84 | 67.94 | **64.31** | 70.86 |
| Base$_{Single}$ | **59.49** | **72.17** | **60.18** | **56.09** | **67.87** | **68.23** | 63.83 | **71.38** |

Table 20: Performance comparisons of different training settings on the **AM** dataset. Bold scores are best overall.

| **Targets** | Abortion | Cloning | Death Penalty | Gun Control | Marijuana | Minimum Wage | Nuclear Energy | Uniform |
|---|---|---|---|---|---|---|---|---|
| Base$_{Multiple}$ | 57.02 | 71.44 | 58.23 | 55.23 | 67.84 | 67.94 | 64.31 | 70.86 |
| +KD$_{M\rightarrow M}$ | 56.62 | 71.01 | 59.65 | 55.54 | **68.28** | 68.56 | 64.35 | 71.39 |
| +**AKD**$_{M\rightarrow M}$ | 57.53 | 72.21 | 59.20 | 55.81 | 68.05 | **69.06** | 64.72 | 72.31 |
| Base$_{Single}$ | 59.49 | 72.17 | 60.18 | **56.09** | 67.87 | 68.23 | 63.83 | 71.38 |
| +KD$_{S\rightarrow S}$ | 58.84 | **72.80** | **60.58** | 54.32 | 66.63 | 67.20 | **65.06** | 72.60 |
| +**AKD**$_{S\rightarrow S}$ | **59.83** | 72.37 | 59.10 | 54.56 | 68.03 | 67.44 | 65.00 | **73.10** |

Table 21: Performance comparisons of different distillation models on the **AM** dataset. Underlined scores are best within groups of models with same teachers; bold scores are best overall.