Distilling Calibrated Knowledge for Stance Detection

Yingjie Li   Cornelia Caragea

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

Abstract

Stance detection aims to determine the position of an author toward a target and provides insights into people's views on controversial topics such as marijuana legalization. Despite recent progress in this task, most existing approaches use hard labels (one-hot vectors) during training, which ignores meaningful signals among categories offered by soft labels. In this work, we explore knowledge distillation for stance detection and present a comprehensive analysis. Our contributions are: 1) we propose to use knowledge distillation over multiple generations in which a student is taken as a new teacher to transfer knowledge to a new fresh student; 2) we propose a novel dynamic temperature scaling for knowledge distillation to calibrate teacher predictions in each generation step. Extensive results on three stance detection datasets show that knowledge distillation benefits stance detection and a teacher is able to transfer knowledge to a student more smoothly via calibrated guiding signals. We publicly release our code to facilitate future research.[1]

Sentence:	Illegal drugs such as marijuana are responsible for a disproportionate share of violence and social decline in America.
Target:	Marijuana Legalization
Stance:	Against

Table 1: An example of stance detection.

1 Introduction

The stance detection task aims to identify the position of a user toward a specific target (Mohammad et al., 2016b; Küçük and Can, 2020; AlDayel and Magdy, 2020). The target is usually a controversial topic (Stab et al., 2018; Glandt et al., 2021), a public figure (Sobhani et al., 2017; Li et al., 2021a) or a claim that could be a rumor's post (Derczynski et al., 2017; Gorrell et al., 2019). For example, for the sentence in Table 1, we can infer that the author is against the marijuana legalization implied by the presence of the words "illegal drugs" and "disproportionate share of violence". Even though impressive progress has been made in stance detection, most previous works rely on one-hot annotation labels in which meaningful signals among different categories are ignored during training. Knowledge distillation (KD) (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). Recent work has started to investigate knowledge distillation in the context of stance detection. Li et al. (2021b) evaluated knowledge distillation on stance detection datasets and proposed an adaptive knowledge distillation method (AKD) that applies less temperature scaling to the samples with larger confidence obtained from teacher predictions. However, the improvement brought by AKD could be limited if the teacher model is poorly calibrated (over-confident in its predictions).

Model miscalibration can be widely observed in modern neural networks (Guo et al., 2017; Yang and Song, 2021; Guo et al., 2021). If a model is well calibrated, then the probability associated with the predicted label should reflect its ground-truth correctness. According to our empirical observations, teacher models of AKD trained on stance detection datasets are not well-calibrated, producing peaked distributions of confidence in stance detection. Yang et al. (2019) showed that the teacher that provides less peaked training signals makes it possible for the student to learn better signals from interclass similarity and can potentially reduce overfitting. Therefore, we associate the performance of knowledge distillation with the calibration of teacher models in stance detection and propose to further improve the task performance by calibrating teacher predictions. Specifically, we recalibrate

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

Findings of the Association for Computational Linguistics: ACL 2023, pages 6316–6329
July 9-14, 2023 ©2023 Association for Computational Linguistics

teacher predictions using a post-processing method called temperature scaling (Platt, 1999; Guo et al., 2017) and the student model is trained based on both hard labels and calibrated teacher predictions.

Further, *born-again networks* (Furlanello et al., 2018), in which the teacher and student models have identical model architectures, have achieved additional improvements with multiple students generations. At each consecutive step, a student model is taken as a new teacher to transfer knowledge to a new fresh student model. In this paper, we explore the born-again networks for stance detection and propose to calibrate teacher predictions in each generation. Extensive experiments on stance detection datasets show that a teacher can transfer knowledge to a student more smoothly via calibrated soft labels generated by the teacher and training student models over multiple generations helps improve the task performance.

Our contributions are summarized as follows:

- We investigate knowledge distillation in generations for stance detection and observe performance gains over multiple generations.
- We explore the connection between knowledge distillation and calibration in stance detection and propose a Calibration-based Knowledge Distillation method (which we call *CKD*) that dynamically updates the temperature used in knowledge distillation in each generation.
- Our CKD consistently outperforms strong baselines of stance detection, indicating that transferring knowledge from a well-calibrated teacher is more beneficial to the stance detection task.

2 Related Work

Previous works for stance detection mainly focus on the in-target setting (Mohammad et al., 2016b; Du et al., 2017; Sobhani et al., 2017; Siddiqua et al., 2019; Li and Caragea, 2019, 2021a) where the test target has always been seen in the training stage. Recently, cross-target stance detection (Augenstein et al., 2016; Xu et al., 2018; Zhang et al., 2020; Liang et al., 2021) and zero-shot stance detection (Allaway and McKeown, 2020; Allaway et al., 2021; Liu et al., 2021a; Liang et al., 2022a,b; Li et al., 2023) have also attracted a lot of attention. In this paper, we mainly focus on the in-target stance detection.

An ad-hoc training strategy that trains one model for one target has been widely used in previous works (Mohammad et al., 2017; Du et al., 2017; Sun et al., 2018; Wei et al., 2018; Li and Caragea, 2019, 2021b). However, it only considers one target during training and thus it fails to exploit the potential of all training data. Recent works (Schiller et al., 2021; Li and Caragea, 2021a) have shown that multi-target training that trains one model on all targets of a dataset can benefit the stance detection task. In our work, we adopt this multi-target training strategy and conduct extensive experiments in the multi-target training setting on three stance detection datasets (Stab et al., 2018; Glandt et al., 2021; Li et al., 2021a).

Knowledge distillation, initially proposed by Hinton et al. (2015), has been widely used in natural language processing to distill external knowledge into a model (Kim and Rush, 2016; Sun et al., 2019; Aguilar et al., 2020; Tong et al., 2020; Currey et al., 2020; Jiao et al., 2020; Hosseini and Caragea, 2021; Zhao and Caragea, 2021). Interestingly, despite that recent works (Furlanello et al., 2018; Clark et al., 2019; Yang et al., 2019; Mobahi et al., 2020; Liu et al., 2021b) have made impressive progress in knowledge distillation in which the teacher and student have the same model architecture, not much attention has been paid to using knowledge distillation for stance detection. One exception is the work by Li et al. (2021b) who proposed an adaptive knowledge distillation method (AKD) that applies instance-specific temperature scaling to the teacher predictions in one generation. In contrast to this prior work which uses only a single generation step for knowledge distillation, we explore born-again networks on three stance datasets and observe further improvements in performance of stance detection models when they are trained over multiple generations.

Motivated by recent advances in calibration of neural networks (Guo et al., 2017; Desai and Durrett, 2020; Guo et al., 2021; Park and Caragea, 2022a,b; Hosseini and Caragea, 2022), we hypothesize that the miscalibration of teacher predictions has a negative impact on the student model and further test this hypothesis by calibrating teacher's predictions in born-again networks. In this work, we conclude that calibrating teacher's predictions in each generation benefits the student model by providing smoother supervision signals in stance detection task.

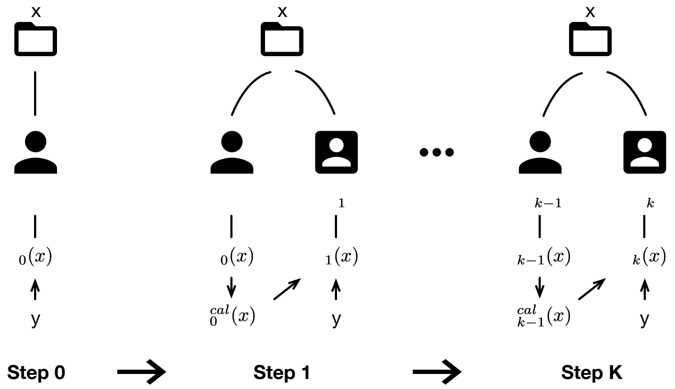

Figure 1: The Calibrated-Knowledge Distillation training procedure over multiple generations.

3 Approach

3.1 Knowledge Distillation

Suppose a given training stance dataset of size n is $D^{tr} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i includes an input sentence and a corresponding target and y_i is the hard label. The goal of stance detection is to predict the stance label given an input sentence and a target.

Standard supervised learning aims to minimize the cross-entropy loss of training data:

$$L_{CE} = \sum_{(x_i,y_i)\in D^{tr}} l_{CE}(p(x_i), y_i) \qquad (1)$$

where $l_{CE}(.)$ represents the cross-entropy, $p(x_i) = \sigma(z(x_i))$ denotes softmax predictions of the model. In knowledge distillation, a student model is trained based on two signals: the hard labels and the teacher estimates, which are known as soft labels. Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher model to a student model by minimizing L_{KD}, which is the sum of the cross-entropy loss between the student predictions and hard labels and the distance loss between the predictions of the student and those of the teacher:

$$L_{KD} = (1 - \lambda)L_{CE} + \lambda L_{KL} \qquad (2)$$

$$L_{KL} = \sum_{x_i \in D^{tr}} l_{KL}(p_s^{\tau}(x_i), p_t^{\tau}(x_i)) \qquad (3)$$

where L_{KL} is Kullback-Leibler (KL) divergence loss, and $p_t^{\tau}(x) = \sigma(z_t(x)/\tau)$ and $p_s^{\tau}(x) = \sigma(z_s(x)/\tau)$ denote the softmax outputs of the teacher and student models (respectively) with temperature scaling τ, where τ is the fixed temperature used to scale the model predictions, $\sigma(.)$ is the softmax function, $z_t(x)$ and $z_s(x)$ denote the output logits of the teacher and student models respectively; λ is a coefficient to balance the importance of the two loss functions and it could be optimized using teacher annealing (Clark et al., 2019) that dynamically mixes the teacher prediction with the ground-truth label during training. The student model learns more from teacher predictions at the early training stage, and mostly relies on the ground-truth labels at the end of training.

Model performance can be further improved with *multiple students generations* (Furlanello et al., 2018) in which a student is taken as a new teacher to transfer knowledge to a new student. In this paper, we explore the knowledge distillation over multiple generations on stance detection (see Figure 1 for training over multiple generations).

3.2 Calibration of Neural Networks

Miscalibration can be widely observed in modern neural networks (Guo et al., 2017; Desai and Durrett, 2020). The prediction confidence of a well-calibrated model is expected to reflect its classification accuracy. For example, given 100 model predictions, each with confidence of 0.7, we expect 70 of them to be correctly classified. One notion of miscalibration is Expected Calibration Error (ECE) (Naeini et al., 2015) that represents the difference in expectation between confidence and accuracy. ECE can be defined as:

$$ECE = \mathbb{E}_{\hat{P}}[|\mathbb{P}(\hat{Y} = Y|\hat{P} = p) - p|] \qquad (4)$$

where Y and $\hat{Y}$ denote the ground-truth label and the predicted label, respectively, $\hat{P}$ is the output probability associated with the predicted label and $p \in (0, 1)$ is a given confidence. As Eq. (4) cannot be estimated using a finite number of samples if $\hat{P}$ is a continuous random variable, empirical approximations (Guo et al., 2017) are usually adopted by partitioning predictions into M bins of equal size and computing the weighted average:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \qquad (5)$$

where B_m denotes the set of samples whose prediction confidence falls into the interval $(\frac{m-1}{M}, \frac{m}{M}]$, M is the bin size, $acc(B_m)$ and $conf(B_m)$ are accuracy and average confidence of B_m, respectively.

Temperature scaling (Guo et al., 2017; Desai and Durrett, 2020) is an effective post-processing calibration method that produces calibrated probabilities. Given the logit vector $z(x)$, the new confidence prediction is $\sigma(z(x)/\tau)$ after temperature scaling with temperature τ being usually fixed in knowledge distillation.

3.3 Calibrated Knowledge Distillation

In this paper, we propose an improved knowledge distillation method called CKD and show that transferring knowledge from a well-calibrated teacher improves the performance of students for stance detection.

The overall training procedure of CKD is shown in Algorithm 1. In the first step, a teacher model is trained from hard labels of stance detection dataset. In the second step, we do temperature scaling of teacher predictions but in our proposed approach, instead of using a fixed temperature τ that shows the best performance on the validation set, we dynamically choose the temperature τ that minimizes the ECE in each generation step (see Step 2 in Algorithm 1). Note that the selection of temperature in each generation is time-efficient as we simply need to divide the softmax teacher outputs by potential temperature values and then compute the ECE. Then, in each consecutive step, a new student model is initialized with an identical model architecture and trained based on both hard labels and calibrated teacher predictions. We expect the calibrated teacher predictions to help students learn better inter-class similarity and achieve better performance. The algorithm is also illustrated in Figure 1.

To summarize, CKD is different from previous adaptive knowledge distillation (Li et al., 2021b) of stance detection in the following aspects:

- We empirically connect the temperature of knowledge distillation with the calibration of distilled networks in CKD.
- The temperature scaling hyperparameter τ is dynamically updated in each generation step.

4 Experimental Settings

In this section, we first describe the stance detection datasets used for evaluation and introduce the evaluation metrics. Then, we describe several stance detection baselines that are used to be compared with our proposed method.

Algorithm 1: Calibrated-Knowledge Distillation with Dynamic Temp. Scaling

```
Require : Train set $D^{tr} = \{(x_i, y_i)\}_{i=1}^{n}$
          Val set $D^{val}$
1 Train the first teacher model by minimizing
    the cross-entropy loss of $D^{tr}$
        $L_{CE} = \sum_{x_i \in D^{tr}} l_{CE}(p(x_i), y_i)$
2 Do temperature scaling of teacher
    predictions with updated $\tau$ of temperature
    scaling that minimizes the ECE of teacher
    predictions on the $D^{val}$
3 Train a student model by minimizing the
    sum of cross-entropy loss and the
    KL-divergence loss of $D^{tr}$
        $L_{KD} = (1 - \lambda)L_{CE} + \lambda L_{KL}$
4 Iterative training: Use the student as a new
    teacher and go back to step 2.
```

4.1 Datasets

Three stance detection datasets of diverse domains are used to evaluate the performance of the proposed method. Train, validation and test sets are as provided by the authors. Examples from these datasets are shown in Table 2 and summary statistics of these datasets are shown in Tables 3, 4, 5. Details of these datasets are described as follows.

AM AM (Stab et al., 2018) is an argument mining dataset containing eight 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.

COVID-19 COVID-19 (Glandt et al., 2021) is a stance detection dataset collected during COVID-19 pandemic, which contains four targets “Face Mask”, “Fauci”, “Stay at Home Orders” and “School Closures”. The dataset is annotated for detecting whether the author is in favor of, neutral or against these topics.

P-Stance P-Stance (Li et al., 2021a) is a stance detection dataset collected during the 2020 U.S. presidential election, which contains three public figures “Donald Trump”, “Joe Biden” and “Bernie Sanders”. The dataset is annotated for detecting

Dataset	Target	Tweet	Stance
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
COVID-19	Face Masks	@SpeakerVos Masks in public places are a necessary prevention tool. You DON'T need to have symptoms to be infected. Many people get Covid and never experience symptoms- They're known as silent Carriers, who are indeed, Spreading the virus.	Favor
P-Stance	Joe Biden	Holy shat! Is @JoeBiden sleep walking and dreaming. Hey Joe, do think anyone believes Obummers Administration isnt guilty of treason and sedition. How about the Logan Act? Your team are going to be in jail in 2 years for long long sentences. #JusticeComing2019	Against

Table 2: Example from each stance detection dataset.

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	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	%Favor	%Against	%None
Face Mask	1,707	23.72	25.29	50.99
Fauci	1,864	23.29	22.28	54.43
Stay at Home	1,372	16.95	23.82	59.23
School Closure	1,190	18.12	24.23	57.65
Total	6,133	19.40	24.30	56.30

Table 4: Data distribution of COVID-19 dataset (Glandt et al., 2021).

Target	#Total	#Train	#Dev	#Test
Donald Trump	7,953	6,362	795	796
Joe Biden	7,296	5,806	745	745
Bernie Sanders	6,325	5,056	634	635
Total	21,574	17,224	2,174	2,176

Table 5: Distribution of instances in P-Stance dataset (Li et al., 2021a).

whether the author is in favor of or against these presidential candidates during the election.

4.2 Evaluation Metrics

Similar to Mohammad et al. (2016a), macro-average of F1-score (F_{macro}) and micro-average of F1-score (F_{micro}) are adopted to evaluate the performance of models. F_{avg} is first calculated by averaging the F1-scores of label "Favor" and "Against".[2] 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 obtain F_{micro} by averaging the F1-scores of "Favor" and "Against" across all targets for each dataset.

4.3 Baseline Methods

We run experiments with the following strong baselines of stance detection:

FNN: Feed-forward networks that take texts as inputs without considering the target information.

BiCE (Augenstein et al., 2016): A BiLSTM model that uses conditional encoding for stance detection. The topic is first encoded by a BiLSTM, whose hidden representations are then utilized to initialize a second BiLSTM with texts as inputs.

TAN (Du et al., 2017): TAN is an attention-based LSTM model that learns the correlation between target and text representations.

CrossNet (Xu et al., 2018): CrossNet improves the BiCE by introducing a self-attention layer.

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

TGA (Allaway and McKeown, 2020): A BERT-based model that exploits topic-grouped attention.

BERT-KE (Kawintiranon and Singh, 2021): A BERT model that is pre-trained with knowledge enhanced masked language modeling.

AKD (Li et al., 2021b): An adaptive knowledge distillation method that uses instance-specific temperature for stance detection.

In addition, we compare the CKD with the following knowledge distillation ablation methods:

KD-1: A vanilla knowledge distillation method without temperature scaling (τ = 1). The student has the same model architecture as the teacher (which is also known as self-distillation (Furlanello

[2] Note that we calculate the F_{avg} and F_{micro} by averaging the F1-scores of label "Favor", "Against" and "Neutral" for COVID-19 dataset to be consistent with their results.

et al., 2018; Zhang and Sabuncu, 2020)).

KD-T: A knowledge distillation method with temperature τ. τ is chosen from $\{2, 3.5, 5\}$ on the validation set. 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.

The proposed method is listed as follows:

CKD: A knowledge distillation method with temperature τ. Unlike KD-T where τ is fixed, in our proposed method τ is dynamically chosen from 1 to 5 with a step size of 0.01 to minimize the ECE in each generation (step 2 of Algorithm 1). The pretrained BERTweet (Nguyen et al., 2020) is used as teacher and student models for CKD and knowledge distillation methods. The student has the same model architecture as the teacher. We use the default bin size of 10 to compute ECE and we report the performance of using different bin sizes in the next section.

Note that unlike CKD, temperature τ of KD-T is chosen out of three options because it is impractical for KD-T to select the temperature in the range of 1 to 5 with step size 0.01, which requires to repeat the whole training procedure hundreds of times. Since temperature scaling will not change the prediction label, we can only tune the temperature according to the classification performance of student model on the validation set for KD-T, which is one of weaknesses of previous born-again networks. In this paper, we propose to select the temperature according to the ECE, which can be simply calculated with each teacher model alone (no training involved in temperature selection step) and thus it is much more time-efficient.

We adopt the teacher annealing (Clark et al., 2019) for all knowledge distillation methods in our experiments. We performed all experiments on a single NVIDIA A5000 GPU. We implement our model in PyTorch framework (Paszke et al., 2019) using HuggingFace Transformers library (Wolf et al., 2020). More details of the hyperparameters are described in Appendix A.

5 Results

In this section, we first compare the proposed CKD with strong baselines of stance detection. Then we present an ablation study to show that dynamically updating the temperature helps improve the performance in generations. In addition, we compare the performance of different bin sizes of our proposed CKD on stance detection datasets. Next, we compare our CKD with the best baseline AKD and visualize the relationship between the confidence and the evaluation metric (F_{micro}) in the form of reliability diagrams. At last, we discuss the performance of our CKD and other KD baselines in ECE on stance detection datasets.

Model	COVID-19	AM	P-Stance
FNN	61.83 (59.03)	47.21 (45.91)	72.08 (71.32)
BiCE	63.07 (60.34)	48.63 (47.59)	74.54 (73.66)
TAN	68.37 (66.45)	49.98 (49.55)	76.06 (75.49)
CrossNet	67.94 (66.16)	51.66 (51.20)	76.15 (75.48)
BERT	71.23 (68.71)	59.99 (59.51)	78.38 (77.96)
TGA	71.88 (69.69)	60.24 (59.72)	77.99 (77.66)
BERT-KE	72.78 (70.60)	58.27 (57.89)	79.74 (79.24)
BERTweet	74.76 (71.95)	63.70 (63.71)	81.97 (81.55)
AKD	75.07 (72.54)	64.82 (64.86)	81.98 (81.54)
CKD	**76.93*** **(74.53*****)**	**66.92*** **(67.02*****)**	**82.34 (81.97)**

Table 6: Performance comparisons of different stance detection models. We report F_{micro} (F_{macro}) over four runs. Bold scores are best results. $*$: CKD improves the best baseline at p $<$ 0.05 with paired t-test.

Main results Table 6 shows performance comparisons of CKD with the stance detection baselines on all three stance detection datasets. We can observe that our proposed CKD performs best in overall. Specifically, the best student model of CKD outperforms the best-performing baseline AKD by 1.86%, 2.10% and 0.36% in F_{micro} on COVID-19, AM, and P-Stance datasets, respectively. A more detailed comparison between our CKD and AKD is presented later in this section. Note that CKD shows less improvements on P-Stance dataset. One explanation is that P-Stance has much larger train set for targets and the effect of knowledge distillation diminishes with increasing the size of train set—a fact that was observed before (Zhang and Sabuncu, 2020).

Ablation study Table 7 shows the comparison results of our proposed method with the ablation methods of fixed temperature or no temperature ($\tau = 1$) mentioned above on all three stance detection datasets. We train born-again networks sequentially with multiple generations. Gen0 and Gen1 columns show the performance of the first teacher (i.e., the BERTweet model in Table 6) and student, respectively. Then the first student is taken as a new teacher to transfer knowledge to the second student and results of the second student are shown

Model	Gen 0	Gen 1	Gen 2	Gen 3
COVID-19				
KD-1	74.76 (71.95)	74.19 (71.61)	74.40 (72.70)	75.02 (72.90)
KD-T	74.76 (71.95)	74.85 (72.98)	75.29 (73.90)	74.49 (71.94)
LSR	74.34 (71.70)	–	–	–
CKD	74.76 (71.95)	74.90 (72.38)	**76.93*** **(74.53)**	75.70 (72.71)
AM				
KD-1	63.70 (63.71)	64.54 (64.56)	65.23 (65.28)	64.62 (64.68)
KD-T	63.70 (63.71)	64.66 (64.63)	65.12 (65.22)	64.20 (64.24)
LSR	64.46 (64.52)	–	–	–
CKD	63.70 (63.71)	65.19 (65.12)	**66.92*** **(67.02*****)**	65.55 (65.57)
P-Stance				
KD-1	81.97 (81.55)	81.74 (81.41)	81.66 (81.25)	81.71 (81.30)
KD-T	81.97 (81.55)	81.55 (81.11)	81.43 (80.98)	81.29 (80.82)
LSR	81.43 (81.04)	–	–	–
CKD	81.97 (81.55)	81.70 (81.31)	82.16 (81.71)	**82.34 (81.97)**

Table 7: Performance comparisons of different models on stance detection datasets over multiple generations. We report F_{micro} (F_{macro}) over four runs. $*$: CKD improves the best baseline at p $<$ 0.05 with paired t-test.

Model	Gen 1	Gen 2	Gen 3
COVID-19			
KD-T	74.85 (72.98)	75.29 (73.90)	74.49 (71.94)
Bin-5	75.28 (72.70)	75.18 (72.46)	75.34 (72.78)
Bin-10	74.90 (72.38)	**76.93 (74.53)**	75.70 (72.71)
Bin-15	75.20 (72.66)	75.92 (73.48)	75.85 (72.85)
AM			
KD-T	64.66 (64.63)	65.12 (65.22)	64.20 (64.24)
Bin-5	64.80 (64.65)	66.27 (66.32)	65.99 (65.93)
Bin-10	65.19 (65.12)	**66.92 (67.02)**	65.55 (65.57)
Bin-15	64.84 (64.78)	66.26 (66.43)	66.34 (66.31)
P-Stance			
KD-T	81.55 (81.11)	81.43 (80.98)	81.29 (80.82)
Bin-5	81.58 (81.11)	81.63 (81.21)	82.06 (81.63)
Bin-10	81.70 (81.31)	82.16 (81.71)	**82.34 (81.97)**
Bin-15	81.59 (81.16)	82.26 (81.83)	81.94 (81.47)

Table 8: Performance comparisons of our proposed models using different bin sizes in F_{micro} (F_{macro}). KD-T is the best-performing distillation baseline.

Model	Gen 1	Gen 2	Gen 3
COVID-19			
AKD	75.07 (72.54)	75.62 (73.00)	75.03 (72.67)
CKD	74.90 (72.38)	**76.93*** **(74.53)**	75.70 (72.71)
AM			
AKD	64.82 (64.86)	64.47 (64.46)	65.47 (65.54)
CKD	65.19 (65.12)	**66.92*** **(67.02*****)**	65.55 (65.57)
P-Stance			
AKD	81.98 (81.54)	81.92 (81.54)	81.48 (81.15)
CKD	81.70 (81.31)	82.16 (81.71)	**82.34 (81.97*****)**

Table 9: Performance comparisons of CKD and AKD in F_{micro} (F_{macro}). $*$: CKD improves the best AKD at p $<$ 0.05 with paired t-test.

in Gen2 column.

We can observe that both CKD and baselines achieve superior performance over the first teacher model (Gen0) on COVID-19 and AM, demonstrating the effectiveness of born-again networks on stance detection. Moreover, our proposed CKD consistently outperforms the distillation baselines on almost all stance detection datasets in different generations, which indicates that dynamically updating τ according to ECE as compared with a fixed τ yields better performance and a well-calibrated teacher can help train a better student.

Bin size We evaluate the proposed CKD using different bin sizes. Table 8 shows performance comparisons of CKD using different bin sizes on stance datasets. We can see that CKD models with different bin sizes outperform the best distillation baseline KD-T in most cases, demonstrating the effectiveness of the proposed method. Moreover, we observe that the default bin size of 10 achieves the best performance on all stance datasets in overall.

CKD vs. AKD First, we compare CKD with AKD in terms of F_{micro} and F_{macro} in Table 9. AKD (Li et al., 2021b) adopts an instance-specific temperature scaling strategy and is trained in only one generation. Here, we further extend AKD over multiple generations. First, we see that AKD achieves better performance with more generations, reinforcing the claim that training student models in multiple generations can improve the task performance. Second, our CKD shows superior performance over AKD on most generations for each dataset, which indicates that a well-calibrated teacher contributes more to stance detection task in generations.

Second, we compare reliability diagrams of our CKD and AKD on COVID-19, AM, and P-Stance datasets in Figures 2 and 3. Reliability diagrams are a visual representation of model calibration (Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). The prediction space is usually discretized

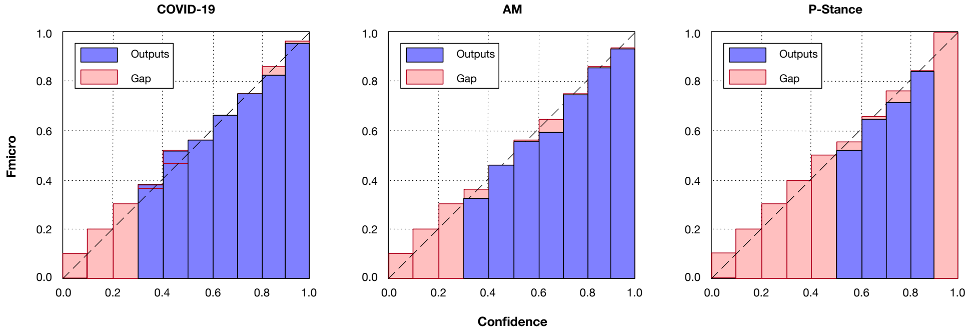

Figure 2: Reliability diagram for our CKD on COVID-19, AM and P-Stance datasets. Confidence values are retrieved from the teacher predictions of last generation.

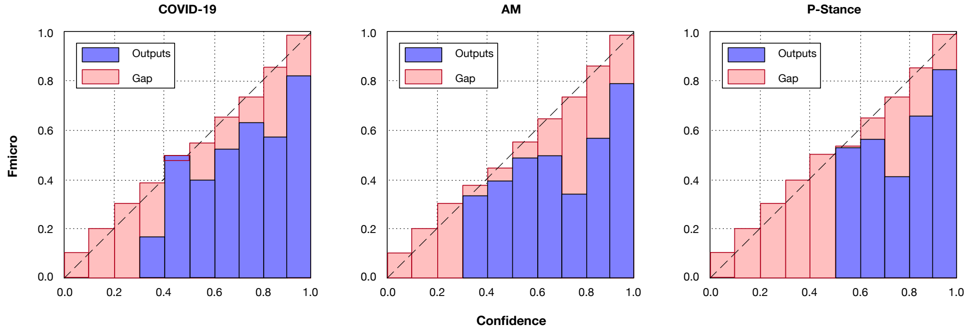

Figure 3: Reliability diagram for baseline AKD on COVID-19, AM and P-Stance datasets. Confidence values are retrieved from the teacher predictions of last generation.

into ten bins and the mean predicted value (i.e., confidence) is plotted against the expected sample accuracy in each bin. If a model is perfectly calibrated, then $acc(B_m)$ will be equal to $conf(B_m)$ for all $m \in \{1, ..., M\}$. We can see that AKD is over-confident in its predictions on all datasets, with large gap areas in Figure 3. We then can observe that our CKD produces more calibrated predictions on COVID-19, AM and P-Stance datasets, verifying the calibration effects of temperature scaling. Based on these observations, we conclude that a student learns better from calibrated teacher predictions that provide less peaked supervision signals in each generation.

ECE results In order to better understand the role of calibration in knowledge distillation in generations, we show the comparison of experimental results (ECE) on three stance detection datasets in Table 10. For each model, we report the performance of the teacher that helps train the best student on each dataset. We can observe that our proposed CKD shows the best performance, achieving the lowest ECE on each dataset. Moreover, the best student of our CKD achieves the highest micro-averaged F1 on all datasets, as shown in Table 7, which indicates that calibrating teacher predictions can benefit the stance detection task.

Model	COVID-19	AM	P-Stance
KD-1	11.75	20.25	8.21
KD-T	3.79	14.23	5.03
AKD	12.38	17.70	7.73
CKD	**2.48**	**3.79**	**3.58**

Table 10: Expected Calibration Error (ECE) in percentage (%) of different models on stance detection datasets. Bold scores are best ECE results. Lower ECE implies better-calibrated models.

6 Conclusion

In this paper, we study the problem of knowledge distillation in generations on stance detection. We show that knowledge distillation in multiple generations can be beneficial to stance detection. Moreover, based on the existing works, we provide a new perspective that a well-calibrated teacher can benefit the student by providing smoother training signals and make it possible for the student to learn from inter-class similarity. Our proposed CKD produces calibrated teacher predictions by dynamically updating the temperature used for scaling in each generation. Experimental results show that our proposed method consistently outperforms the best-performing baseline on different stance detection datasets. Future work includes extending our proposed method to a broader range of NLP tasks such as emotion classification.

7 Limitations

One limitation of our method is that it requires multiple generations to achieve the best performance on stance detection datasets. While the best student model significantly outperforms strong baselines, it takes longer training time and requires extra memory for the teacher model. This is a common limitation for knowledge distillation in generations. Another limitation of our method is that the improvements brought by knowledge distillation saturate after a few generations, which can be also observed in previous work. We will explore how to improve the performance saturation in the future.

8 Ethical Considerations

Beyond the proposed method that helps correctly identify the stance towards specific targets, it is very important to consider the ethical implications of stance detection systems. Since stance detection systems could automatically collect and aggregate the topical stance for a specific target, these systems may have significant impact on decision-making. Algorithms are not perfect, and thus a potential harm is that these systems may make incorrect predictions and further mislead the decision-making. Researchers should be aware of potential harms from the misuse of stance detection systems, and should respect people's privacy during the data collection.

Acknowledgments

We thank the National Science Foundation for support from grants IIS-1912887, IIS-2107487, and ITE-2137846 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.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767.

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.

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.

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.

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.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

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.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2021. An overview of uncertainty calibration for text classification and the role of distillation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 289–306.

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

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahshid Hosseini and Cornelia Caragea. 2022. Calibrating student models for emotion-related tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9266–9278, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735.

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.

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. 2021a. 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.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2021b. Improving stance detection with multi-dataset learning and knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6332–6345.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, page 1500–1509.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, page 2738–2747.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, page 3453–3464.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021a. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Yang Liu, Sheng Shen, and Mirella Lapata. 2021b. Noisy self-knowledge distillation for text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–703.

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

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. 2020. Self-distillation amplifies regularization in hilbert space. In *Advances in Neural Information Processing Systems*, volume 33, pages 3351–3361.

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.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907.

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.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, page 625–632.

Seo Yeon Park and Cornelia Caragea. 2022a. A data cartography based MixUp for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4244–4250, Seattle, United States. Association for Computational Linguistics.

Seo Yeon Park and Cornelia Caragea. 2022b. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74.

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.

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.

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.

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.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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.

Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L. Yuille. 2019. Training deep neural networks in generations: A more tolerant teacher educates better students. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5628–5635.

Lehan Yang and Jincen Song. 2021. Rethinking the knowledge distillation from the perspective of model calibration. *arXiv preprint arXiv:2111.01684*.

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.

Chenye Zhao and Cornelia Caragea. 2021. Knowledge distillation with BERT for image tag-based privacy prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1616–1625, Held Online. INCOMA Ltd.

A Hyperparameters

For each non-BERT model, AdamW optimizer (Loshchilov and Hutter, 2019) is used with a learning rate of 1e-3, and gradient clipping if the norm of the gradients exceeds 1. Each model is trained for 30 epochs, with a mini-batch size of 128 in each iteration. A dropout of 0.5 is used after the embedding layer. The dimension of feed-forward layers is 300 and the hidden dimension of LSTM is 300 for TAN, BiCE and CrossNet.

For each BERT-based model, AdamW optimizer is used with a learning rate of 2e-5, and gradient clipping if the norm of the gradients exceeds 1. Each model is fine-tuned for 5 epochs, with a mini-batch size of 32.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section after the Conclusion.

☑ A2. Did you discuss any potential risks of your work?
Ethical Considerations Section after the Conclusion.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We discussed our usage of baseline models and previous datasets in Section 4.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.

C ☑ Did you run computational experiments?

Section 5.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.