Calibrating Student Models for Emotion-related Tasks

Mahshid Hosseini **Cornelia Caragea**
Computer Science
University of Illinois at Chicago
mhosse4@uic.edu cornelia@uic.edu

Abstract

Knowledge Distillation (KD) is an effective method to transfer knowledge from one network (a.k.a. teacher) to another (a.k.a. student). In this paper, we study KD on the emotion-related tasks from a new perspective: *calibration*. We further explore the impact of the mixup data augmentation technique on the distillation objective and propose to use a simple yet effective mixup method informed by training dynamics for calibrating the student models. Underpinned by the regularization impact of the mixup process by providing better training signals to the student models using training dynamics, our proposed mixup strategy gradually enhances the student model's calibration while effectively improving its performance. We evaluate the calibration of pre-trained language models through knowledge distillation over three tasks of emotion detection, sentiment analysis, and empathy detection. By conducting extensive experiments on different datasets, with both in-domain and out-of-domain test sets, we demonstrate that student models distilled from teacher models trained using our proposed mixup method obtained the lowest Expected Calibration Errors (ECEs) and best performance on both in-domain and out-of-domain test sets.

1 Introduction

It has been shown that transferring knowledge from a teacher model with desired high performance to a student model, through knowledge distillation, can lead to better performance of student models distilled (Furlanello et al., 2018; Yim et al., 2017). However, little is known about the impact of the distillation process on the calibration of the student model. Evaluating the uncertainty of a model's predictions is crucial, specifically in applications where the cost of an error is high. For instance, in a computer-assisted therapy session, an accurate and calibrated emotion or empathy detection model can inform the doctor when a model's predictions should (or should not) be trusted, which is helpful for them in deciding the preferred treatment for patients. In this work, we aim to shed light on the impact of knowledge distillation on the calibration of the student models on emotion-related tasks. Calibration measures the discrepancy between the correctness of the prediction (i.e., accuracy) and the (empirical) probability that a model assigns to a prediction (i.e., confidence). A well-calibrated model *knows* how often it is correct or wrong; predicting an event with p confidence shall empirically be true p of the time (Guo et al., 2017).

Recently, a large body of work has investigated why neural networks have become miscalibrated (Platt et al., 1999; Niculescu-Mizil and Caruana, 2005; Nguyen and O'Connor, 2015a; Kuleshov and Liang, 2015; Kuleshov and Ermon, 2016; Guo et al., 2017; Desai and Durrett, 2020). More recent attention, however, has focused on methods to alleviate this problem. Specifically on natural language processing tasks, Guo et al. (2017) proposed a simple extension of Platt scaling (Platt et al., 1999) that softens the softmax by a learned scalar parameter which can effectively calibrate probabilistic models. Pereyra et al. (2017a); Müller et al. (2019); Desai and Durrett (2020) also showed that regularization techniques such as label smoothing could prevent over-confident predictions and result in better model calibration.

Along these lines, in this paper, we empirically examine the impact of the mixed sample data augmentation technique, Mixup (Zhang et al., 2018), on the performance and calibration of the student models in a distillation setup and propose a simple yet effective mixup strategy to attain more accurate and better-calibrated models. Mixup (Zhang et al., 2018) is a popular data augmentation and regularization technique that generates a weighted combination of random input pairs from the training data. It has been empirically shown that mixup can hone the accuracy and calibration of the pre-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9266 - 9278

diction emanating from the desirable regularization effects it induces (Carratino et al., 2020; Zhang et al., 2018; Thulasidasan et al., 2019). By employing mixup, the goal is to provide the teacher models with more useful information and impart the student models with better supervision signals during the distillation of the emotion-related models. In addition, mixup may introduce some noise to the training data (as real-world emotion-related datasets), which enables us to gain additional information about relatively similar data. This, in turn, makes teacher models more robust, helping the student models to be more accurate and produce better-calibrated predictions.

While mixup is making significant inroads in a broad range of tasks ranging from computer vision (Zhang et al., 2018; Thulasidasan et al., 2019; Carratino et al., 2020; Wang et al., 2020a) to natural language processing (Guo et al., 2019; Guo, 2020; Chen et al., 2020; Yin et al., 2021; Kong et al., 2020; Liang et al., 2021), there has hitherto been a limited number of works focusing on its effectiveness on model calibration specifically in NLP (Kong et al., 2020; Park and Caragea, 2022). With that caveats, what is not yet studied is using mixup for calibrating the student model predictions on the knowledge distillation setting; that is what this paper focuses on.

In this paper, we study, for the first time to our knowledge, the impact of the mixup data augmentation technique on the distillation objective and propose a simple yet effective mixup strategy that is informed by training dynamics (Swayamdipta et al., 2020) for calibrating the student models. To this end, we first characterize data instances based on their contributions to the model's learning, which yields distinct regions in the data, presenting easy-to-learn, ambiguous, or hard-to-learn instances. Then, we generate mixup samples by interpolating easy-to-learn with ambiguous samples as a regularization technique to promote generalization to both in-domain (ID) and out-of-domain (OOD) test sets and improve the student model calibration. While ambiguous/hard-to-learn instances are intuitively the most challenging yet informative for learning, easy-to-learn instances are essential for convergence (Swayamdipta et al., 2020). Therefore, interpolating samples from different regions (e.g., easy-to-learn with ambiguous) in the teacher model can potentially result in a better-calibrated student model with improved ID and OOD performance. To contextualize examples in our datasets based on training dynamics, we utilize data maps (Swayamdipta et al., 2020). Data maps is a model-based tool that characterizes datasets based on the model's behavior on each of the instances. By leveraging training dynamics, data maps estimates two measures, i.e., confidence and variability, the mean and standard deviation of the ground-truth probabilities, predicted for each instance across training epochs.

We further experimentally explore the effect of popular regularization techniques like temperature scaling (Guo et al., 2017) and label smoothing (Pereyra et al., 2017a) along with our informed mixup on the calibration of the student models in a teacher-student training set-up. We carried out extensive experiments to evaluate the proposed informed mixup data augmentation technique by creating teacher networks on two pre-trained models, BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019). Student networks will then be distilled and evaluated against several data sets on three different text classification tasks, including emotion detection (Demszky et al., 2020), sentiment analysis (Zhang et al., 2015), and empathy detection (Sharma et al., 2020). Our contributions are thus summarized as follows:

- We show that the dark knowledge of a pre-trained language teacher model can act as a regularization process, helping to calibrate the student model's confidence in its predictions.
- We demonstrate that using training dynamics to inform the interpolation process in the mixup data augmentation on a teacher model can effectively improve the calibration of the student model in a distillation setting. Based on the confidence and variability of each example, we divide training samples into distinct categories where we propose to mix easy-to-learn and ambiguous samples in the teacher model for the student model calibration.
- We also examine the performance of the distilled student models under distributional shift, and show the effectiveness of the informed mixup method to coax the student model into generating more calibrated predictions.
- Through extensive experiments, we show that student models distilled from teacher models trained using our proposed mixup are not only more accurate but also better-calibrated on

both in-domain and out-of-domain test sets than strong baselines on different text classification tasks of emotion detection, sentiment analysis, and empathy detection.

2 Related Work

Knowledge Distillation: Knowledge distillation (KD) is an efficient method broadly used for transferring knowledge from a teacher network to a student network. In the knowledge distillation setting, a student model is trained to obtain the knowledge of a deeper or more complex teacher model and can therefore estimate the capacity of the powerful teacher model by incorporating the extra knowledge. KD was first introduced as an approach to compress large networks into smaller networks (Ba and Caruana, 2014; Buciluǎ et al., 2006) for computational efficiency. The advances of KD, however, go beyond model compression. Zhang and Sabuncu (2020) empirically explained the reason behind the enhanced performance of self-distillation and proposed a framework that employs instance-specific regularization for teacher predictions. Phuong and Lampert (2019) examined the impact of distillation on student models by analyzing linear and deep linear classifiers. Unlike previous works, we are interested in analyzing the impact of knowledge distillation on the calibration of the models. Thus, we examine the calibration of large-scale pre-trained models through knowledge distillation. We further analyze the impact of dataset shift on calibration for all these settings. We evaluate the predictive uncertainty on both in-domain and out-of-domain test sets from known and unknown distributions on emotion-related datasets.

Mixup: Mixup (Zhang et al., 2018) was first proposed to improve the generalization of deep neural networks in computer vision. Since then, many studies have explored mixup in natural language processing tasks (Guo et al., 2019; Guo, 2020; Chen et al., 2020; Yin et al., 2021; Kong et al., 2020; Liang et al., 2021). Liang et al. (2021) proposed a data-agnostic distillation framework that leverages mixup to confer the student model with better generalization ability. Kong et al. (2020) examined BERT calibration using mixup by generating augmented samples based on a cosine distance of extracted features. Park and Caragea (2022) also improved pre-trained language models calibration by leveraging Area Under the Margins (Pleiss et al., 2020) along with saliency maps (Simonyan et al., 2014) to generate mixup samples. In contrast to these works, we study the impact of mixup data augmentation technique on the distillation objective.

Calibration: Calibration and uncertainty of the models have been investigated on several natural language processing tasks, including question answering (Zhang et al., 2021), neural machine translation (Lu et al., 2021; Müller et al., 2019; Kumar and Sarawagi, 2019; Wang et al., 2020b), language understanding (Desai and Durrett, 2020), estimating proportions from annotations (Card and Smith, 2018), and coreference resolution (Nguyen and O'Connor, 2015b). Ovadia et al. (2019) provided a benchmark of models on image and text classification tasks and explored the influence of distributional shift on accuracy and calibration. Focusing on the pre-trained models, Desai and Durrett (2020) examined calibration over three tasks of paraphrase detection, natural language inference, and commonsense reasoning. Unlike these works, we study the calibration of emotion-related tasks through knowledge distillation and propose a mixup strategy to enhance the performance and calibration of the Transformer-based student models.

3 Methods

3.1 Teacher-Student Training

Given a k-class classification task and a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{n}$ consisting of sentence-label pairs, standard supervised learning is optimized based on the one-hot labels by minimizing the cross-entropy loss $\mathcal{L}_{ce}$ of training data which is defined as:

$$\mathcal{L}_{ce}(p, y) = -\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} log p(y = j|x_i) \quad (1)$$

where p indicates the softmax outputs. In knowledge distillation, a teacher-student training method is employed to enhance the performance of the student model where the softmax outputs of the teacher model, p^t, is computed as:

$$p_i^t = \frac{exp(z_i/\tau)}{\sum_j exp(z_j/\tau)} \quad (2)$$

where τ is the softmax temperature, and z is the logits from the teacher model. In general, the knowledge distillation framework (Hinton et al., 2015) incorporates the knowledge obtained from the logits of a teacher model and transfers the knowledge

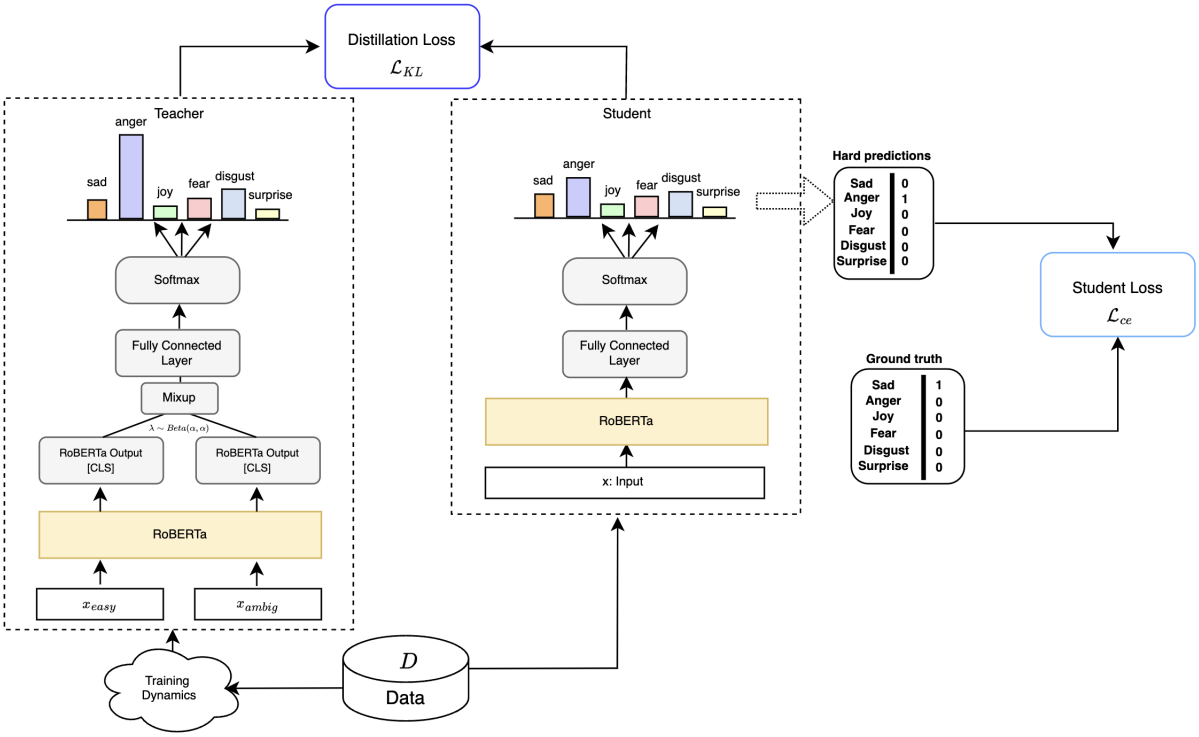

Figure 1: Our proposed mixup for teacher–student framework in the self-distillation setting. We first train the pre-trained language models (i.e., BERT or RoBERTa) using our informed mixup on each task's training dataset. Then, the student networks are built from the BERT or RoBERTa with no data augmentation or regularization techniques added.

to a small student model. In this way, better training signals can be retrieved from the data using a teacher-student framework. This is done by minimizing the sum of cross-entropy loss between hard labels and student' predictions and the difference loss between the student's and teacher's predictions:

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{ce}(p, y) + (1 - \alpha)\mathcal{L}_{KL}(p, p^t) \quad (3)$$

where $\mathcal{L}_{KL}$ is Kullback-Leibler (KL) divergence loss, and $\alpha \in [0, 1]$ is the hyper-parameter that controls the impact of cross-entropy loss and the KL divergence loss.

Self-distillation is a particular case of teacher-student training where both the teacher and student models have the same architecture. For example, in Figure 1 we have both teacher and student models based on RoBERTa.

3.2 Mixup Training

In an attempt to provide the teacher models with more useful information and impart the student models with better supervision signals during the distillation of the emotion-related models, we empirically examine standard mixup and propose a simple yet effective strategy to hone the performance and calibration of the student model in a distillation setup.

Given a training dataset of sentence-label pairs $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{n}$ and a language model f, standard mixup creates the vicinal dataset by calculating a weighted average of training points based on the following simple rule by (Zhang et al., 2018):

$$(\widetilde{x}_{ij}, \widetilde{y}_{ij}) := (\lambda x_i + (1 - \lambda)x_j, \lambda y_i + (1 - \lambda)y_j) \quad (4)$$

where (x_i, y_i) and (x_j, y_j) are two input examples that are randomly drawn from the training set, and weight λ is sampled from a beta distribution, $\beta(\alpha, \alpha)$ with parameter $\alpha > 0$, generally taken to be relatively small, so that the weighted averages do not stray too far from the original data points. Mixup augments the training data by linearly interpolating training samples and their corresponding labels in the input space.

We propose to use a novel mixup data augmentation technique on the teacher models that is informed by training dynamics to improve the student model calibration on the distillation objective. Our proposed mixup creates vicinal distribution steered by the data maps (Swayamdipta et al., 2020) as described below.

Mixup with Training Dynamics. Figure 1 depicts our proposed mixup strategy for teacher–student framework in the self-distillation setting (where both the teacher and student models have the same architecture, e.g., RoBERTa). We first contextualize each training instance of our $\mathcal{D}_{train}$ into three categories, namely easy-to-learn, ambiguous, and hard-to-learn, based on training dynamics (statistics deriving from the behavior of the model across time). The training dynamics of instance (x_i, y_i) are defined as statistics, i.e., confidence and variability computed across the E epochs. Confidence is calculated as the mean model probability of the true label y_i across epochs:

$$\hat{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} p_{\theta_e}(y_i|x_i) \tag{5}$$

where θ denotes our model parameters and $p_{\theta e}$ indicates the model's probability at the end of the e_{th} epoch. Intuitively, a high-confidence instance is easier for the given learner.

Variability is defined as the standard deviation of the ground-truth probabilities $p_{\theta_e}(y_i|x_i)$ across different epochs:

$$\hat{\sigma}_i = \sqrt{\frac{\sum_{e=1}^{E}(p_{\theta_e}(y_i|x_i) - \hat{\mu}_i)}{E}} \tag{6}$$

Intuitively, samples to which the model confidently assigns the true label (i.e., high confidence) and constantly the same label (i.e., low variability) corresponds to easy-to-learn examples (for the model). The samples with low confidence and low variability resemble hard-to-learn examples (for the model), and examples with high variability that the model is uncertain about during training are ambiguous (to the model).

Using the model's confidence and variability of the instances, our informed mixup method first splits $\mathcal{D}_{train}$ into three distinct categories, i.e., $\mathcal{D}_{easy}$, $\mathcal{D}_{hard}$, and $\mathcal{D}_{ambig}$ (Figure 2), each containing 33% of train set. Then, it generates mixup samples by randomly selecting and interpolating samples from our $\mathcal{D}_{easy}$ and $\mathcal{D}_{ambig}$ as a regularization technique to improve the calibration of the student model (with all original examples, including hard-to-learn examples be used during the training process). We mix samples from two distinct groups of easy-to-learn and ambiguous[1], as easy samples play an important role in model optimization, and

[1]Mixing samples from these two data categories yields the best results in our experiments, so we only report the mixup results in this setting.

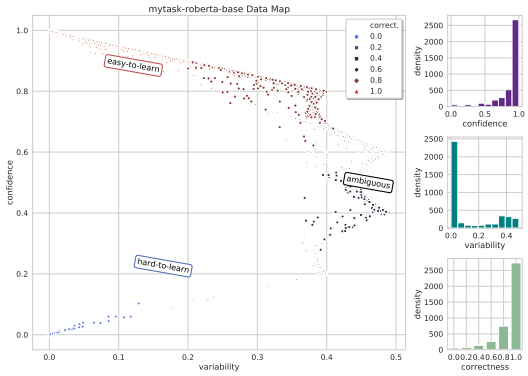

Figure 2: Data map for the EPITOME train set, based on a `RoBERTa-base` classifier.

ambiguous samples are essential for learning. In this way, we confer the augmented samples in the teacher model to embrace useful information from both easy-to-learn and ambiguous samples, adjusting the difficulty of samples and hence perturbing the student model's predictions to be better calibrated. Our informed mixup approach interpolates samples on the final hidden state corresponding to the [CLS] token generated by the task-specific layer on top of our teacher model. We conduct our mixup procedure using mini-batch SGD to update the model weights in our experiments.

3.3 Calibration

A probabilistic model is considered calibrated if its predicted probabilities of classes are equivalent to the actual probabilities of those classes. Intuitively, if a model allots 80% posterior probability to a class, that class should appear 80% of the time. Considering class predictions, suppose a model assigns probability q to a class y, formally the model is perfectly *calibrated* if $\forall p \in [0, 1], \mathcal{P}[Y = y|q = p] = p$ (i.e., the model is calibrated when q is always the true probability p). To evaluate the calibration, following Guo et al. (2017), we use the expected calibration error (ECE) (Naeini et al., 2015). ECE measures model miscalibration by binning the predicted probabilities and measuring the gap between them and the average accuracies of these bins: $\sum_{s=1}^{S} \frac{b_s}{N}|acc(s) - conf(s)|$, where S is the overall number of bins, and b_s represents the number of predictions in the s-th bin. N denotes the total number of data points, and $acc(s)$ and $conf(s)$ represent the accuracy and confidence of the s-th bin, respectively. We use $S = 10$ for the experiments in this paper.

3.4 Post-processing and Regularization

We additionally experiment with post-processing methods[2] employed to tune a model's calibration.

Temperature Scaling. In temperature scaling (Guo et al., 2017), before the softmax operation, a single scalar hyperparameter T divides logits (which then go through softmax). $T \longrightarrow \infty$ yields maximum uncertainty with uniform probabilities, and as $T \longrightarrow 0$, the probability drops to a point mass. $T = 1$ obtains the original probabilities, i.e., $T = 1$ corresponds to no temperature scaling. This process is shown to make the re-calibrated probabilities in over-confident models smaller than the main probabilities and helps the models to be slightly less confident.

Label Smoothing. Label smoothing (LS) (Szegedy et al., 2016; Pereyra et al., 2017b) is a regularization technique that preserves a reasonable proportion between the logits of the incorrect classes by keeping uncertainty across the label space throughout training (Szegedy et al., 2016). Therefore, without changing the model architecture, LS prevents overconfident predictions and could results in better model calibration (Müller et al., 2019). Having a hyperparameter[3] $\alpha \in (0, 1)$, we use label smoothing to regularize a model with k output values by converting the hard 0 and 1 targets with targets of $\frac{\alpha}{k-1}$ and $1 - \alpha$, respectively. The case of $\alpha = 0$ corresponds to learning from one-hot labels.

4 Experiments

4.1 Tasks and Datasets

We perform evaluations on three text classification tasks of emotion detection, sentiment analysis, and empathy detection. We analyze tasks with challenging domain shifts where out-of-domain performance is considerably lower. We explain our in-domain and out-of-domain datasets below.

Emotion Detection. GoEmotions corpus is a large-scale emotion detection dataset from Reddit comments labeled with 27 emotion categories[4] or neutral (Demszky et al., 2020). We use the 6 basic emotions (joy, anger, fear, sadness, disgust, and surprise) and neutral, proposed by Ekman (1992) and conduct an Ekman-style grouping into six coarse categories. Meld (Poria et al., 2019) contains dialogues from the popular Friends TV series annotated with Ekman's six universal emotions and two additional emotion labels of neutral and non-neutral, which we use as unseen test domains. For consistency, in Meld, we use Ekman-6 emotions and neutral similar to GoEmotions.

Sentiment Analysis. Yelp is a dataset for binary sentiment classification, which consists of reviews from Yelp (Zhang et al., 2015). Our out-of-domain setting is Stanford Sentiment Treebank (SST-2) (Socher et al., 2013), which is composed of sentences of movie reviews and their sentiment.

Empathy Detection. EPITOME is a corpus annotated with three levels of empathy communication (Sharma et al., 2020). We consider weak and strong communications as our positive class, and no communication as negative class. Buechel et al. (2018) dataset on empathy (which we refer to as NewsEmpathy) is our out-of-domain empathy dataset consisting of empathic reactions to news stories. For consistency, we use binary empathy labels to model empathy in NewsEmpathy in a binary setting.

4.2 Models

To evaluate calibration of large-scale pre-trained models, we fine-tune BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) . We additionally experiment with self-distillation (a particular case of the teacher-student training), where both teacher and student models have the same architecture (Zhang and Sabuncu, 2020). In this setting, we first train a pre-trained language model (i.e., BERT or RoBERTa) as a teacher model and then train a student model (with the same architecture) to mimic the output of the teacher model. We also experiment with knowledge distillation, where we distill knowledge from a pre-trained language model with a different architecture than the student model. We present the result of the knowledge distillation setting in Appendix B. We further compare the performance and calibration of the standalone and distilled models using standard Mixup (Zhang et al., 2018) and our proposed mixup method. The experimental setting are discussed in Appendix A.

4.3 Results

Test accuracies and expected calibration errors (ECE) are summarized in Tables 1 and 2, respectively. First, we train the model on the in-domain

[2]We do not use these approaches independently but in combination guided by prior work.

[3]For instance, when $\alpha = 0.1$, the one-hot label vector $[1, 0, 0]$ is converted to [0.9, 0.05, 0.05] smoothed label vector.

[4]For samples with more than one label, we randomly chose one label to do a multi-class classification.

Model	In-Domain			Out-of-Domain		
	GoEmotions	Yelp	EPITOME	Meld	SST	NewsEmpathy
BERT	$68.10_{0.2}$	$95.87_{0.4}$	$67.42_{0.3}$	$41.64_{0.2}$	$79.83_{0.5}$	$57.10_{0.4}$
BERT+(LS+TS)	$68.03_{0.6}$	$95.42_{0.2}$	$66.25_{0.5}$	$41.86_{1.2}$	$80.41_{0.8}$	$56.75_{0.9}$
BERT+Mixup	$67.83_{0.6}$	$95.81_{0.4}$	$67.35_{0.7}$	$41.48_{0.3}$	$79.70_{0.5}$	$57.08_{0.2}$
BERT+Mixup+(LS+TS)	$69.58_{0.4}$	$96.52_{0.3}$	$67.83_{0.5}$	$42.71_{0.5}$	$80.65_{0.8}$	$57.81_{0.7}$
BERT+Ours	$70.52_{0.3}$	$96.90_{0.2}$	$68.93_{0.5}$	$44.12_{0.6}$	$81.10_{0.5}$	$58.45_{0.5}$
BERT+Ours+(LS+TS)	$\underline{70.84}_{0.6}$	$\underline{96.99}_{0.3}$	$\underline{69.83}_{0.4}$	$\underline{45.58}_{0.7}$	$\underline{81.65}_{0.7}$	$\underline{59.20}_{0.9}$
SD_{BERT}	$68.53_{0.3}$	$96.22_{0.2}$	$68.57_{0.2}$	$42.33_{0.1}$	$80.69_{0.4}$	$58.82_{0.2}$
SD_{BERT}+(LS+TS)	$68.32_{0.5}$	$96.14_{0.4}$	$68.36_{0.7}$	$42.72_{0.5}$	$80.88_{0.6}$	$58.98_{0.7}$
SD_{BERT}+Mixup	$68.22_{0.2}$	$96.10_{0.7}$	$68.16_{0.3}$	$42.58_{0.3}$	$80.20_{0.8}$	$58.01_{0.4}$
SD_{BERT}+Mixup+(LS+TS)	$68.10_{0.5}$	$96.24_{0.8}$	$69.11_{0.7}$	$44.16_{0.4}$	$79.67_{0.3}$	$58.19_{0.5}$
SD_{BERT}+Ours	$\underline{71.82}_{0.2}$	$\underline{97.85}_{0.4}$	$70.60_{0.4}$	$49.82_{0.3}$	$82.65_{0.7}$	$60.40_{0.6}$
SD_{BERT}+Ours+(LS+TS)	$71.50_{0.2}$	$97.71_{0.6}$	$\underline{71.32}_{0.7}$	$\underline{49.98}_{0.8}$	$\underline{82.76}_{0.4}$	$\mathbf{61.47_{0.2}}$
RoBERTa	$68.25_{0.5}$	$96.16_{0.7}$	$68.38_{1.2}$	$42.17_{0.8}$	$82.84_{0.6}$	$56.88_{0.4}$
RoBERTa+(LS+TS)	$68.17_{0.4}$	$96.05_{0.6}$	$67.87_{0.2}$	$42.94_{0.3}$	$82.96_{0.7}$	$55.73_{0.5}$
RoBERTa+Mixup	$68.20_{0.8}$	$96.07_{0.7}$	$68.24_{0.5}$	$43.12_{0.4}$	$82.77_{0.2}$	$56.81_{0.3}$
RoBERTa+Mixup+(LS+TS)	$68.47_{0.7}$	$96.79_{0.2}$	$68.65_{0.5}$	$44.80_{1.2}$	$83.24_{0.8}$	$\underline{58.96}_{0.2}$
RoBERTa+Ours	$70.57_{0.4}$	$97.25_{0.3}$	$69.60_{0.5}$	$48.32_{0.4}$	$85.24_{0.6}$	$58.67_{0.6}$
RoBERTa+Ours+(LS+TS)	$\underline{70.82}_{0.2}$	$\underline{97.40}_{0.5}$	$\underline{70.21}_{0.5}$	$\underline{49.10}_{0.3}$	$\underline{85.44}_{0.4}$	$58.92_{0.4}$
$SD_{RoBERTa}$	$68.73_{0.6}$	$96.74_{0.3}$	$68.68_{0.2}$	$42.64_{0.5}$	$84.73_{0.4}$	$57.51_{0.7}$
$SD_{RoBERTa}$+(LS+TS)	$68.70_{0.7}$	$96.40_{0.6}$	$68.31_{1.1}$	$43.10_{0.4}$	$84.80_{0.9}$	$57.65_{0.4}$
$SD_{RoBERTa}$+Mixup	$68.59_{0.8}$	$96.43_{0.7}$	$68.16_{0.4}$	$42.17_{0.3}$	$84.35_{0.8}$	$57.33_{0.3}$
$SD_{RoBERTa}$+Mixup+(LS+TS)	$\underline{68.24}_{0.4}$	$\underline{96.17}_{0.8}$	$\underline{68.36}_{1.0}$	$\underline{43.23}_{0.6}$	$\underline{84.61}_{0.5}$	$\underline{58.15}_{0.2}$
$SD_{RoBERTa}$+Ours	$72.61_{0.4}$	$97.93_{0.4}$	$71.45_{0.3}$	$50.04_{0.7}$	$86.54_{0.9}$	$60.83_{0.6}$
$SD_{RoBERTa}$+Ours+(LS+TS)	$\mathbf{73.21_{0.3}}$	$\mathbf{97.95_{0.4}}$	$\mathbf{71.80_{0.5}}$	$\mathbf{50.29_{0.6}}$	$\mathbf{87.10_{0.4}}$	$\underline{61.29}_{0.4}$

Table 1: Accuracy in percentage (%) for in-domain (GoEmotions, Yelp, EPITOME) and out-of-domain (Meld, SST, NewsEmpathy) datasets. LS, TS, and SD refer to the label smoothing, temperature scaling, and self-distillation, respectively. All reported values for the methods are $\text{mean}_{\pm \text{ std}}$ of three repetitions. Best results are **bolded**.

Model	In-Domain			Out-of-Domain		
	GoEmotions	Yelp	EPITOME	Meld	SST	NewsEmpathy
BERT	$3.50_{0.6}$	$1.18_{0.3}$	$5.53_{0.8}$	$6.23_{0.4}$	$3.69_{0.3}$	$6.63_{0.7}$
BERT+(LS+TS)	$3.98_{0.4}$	$1.87_{0.6}$	$6.49_{0.3}$	$5.33_{0.6}$	$3.29_{0.5}$	$5.46_{0.6}$
BERT+Mixup	$3.57_{0.5}$	$1.21_{0.2}$	$5.98_{0.7}$	$6.35_{0.4}$	$3.77_{0.3}$	$6.72_{0.4}$
BERT+Mixup+(LS+TS)	$2.27_{0.3}$	$0.73_{0.6}$	$4.40_{0.6}$	$5.58_{0.7}$	$2.71_{0.2}$	$5.24_{0.8}$
BERT+Ours	$2.21_{0.5}$	$0.49_{0.3}$	$4.53_{0.2}$	$5.18_{0.3}$	$2.56_{0.2}$	$4.89_{0.5}$
BERT+Ours+(LS+TS)	$\underline{1.96}_{0.6}$	$\underline{0.45}_{0.6}$	$\underline{3.29}_{0.7}$	$\underline{4.86}_{0.5}$	$\underline{2.08}_{0.6}$	$\underline{4.60}_{0.4}$
SD_{BERT}	$2.53_{0.5}$	$0.65_{0.4}$	$3.29_{0.6}$	$5.49_{0.7}$	$2.96_{0.5}$	$4.95_{0.3}$
SD_{BERT}+(LS+TS)	$2.79_{1.3}$	$0.78_{0.6}$	$3.52_{0.5}$	$5.22_{0.2}$	$2.70_{0.7}$	$4.86_{0.7}$
SD_{BERT}+Mixup	$2.76_{0.8}$	$0.81_{0.6}$	$3.47_{0.7}$	$5.57_{0.4}$	$3.10_{0.6}$	$5.05_{0.9}$
SD_{BERT}+Mixup+(LS+TS)	$2.81_{0.4}$	$0.94_{0.8}$	$3.66_{0.2}$	$5.72_{0.5}$	$3.08_{0.3}$	$5.26_{0.6}$
SD_{BERT}+Ours	$\underline{2.03}_{0.3}$	$0.47_{0.5}$	$2.88_{0.6}$	$4.64_{0.7}$	$2.12_{0.6}$	$4.42_{0.6}$
SD_{BERT}+Ours+(LS+TS)	$2.10_{0.5}$	$\underline{0.39}_{0.2}$	$\underline{2.75}_{0.3}$	$\underline{4.24}_{0.4}$	$\underline{2.04}_{0.5}$	$\underline{3.59}_{0.4}$
RoBERTa	$4.15_{0.6}$	$1.23_{0.8}$	$4.08_{0.3}$	$5.52_{0.6}$	$3.05_{0.2}$	$6.71_{0.5}$
RoBERTa+(LS+TS)	$4.35_{1.2}$	$1.67_{0.6}$	$4.92_{0.8}$	$5.21_{0.5}$	$2.69_{0.6}$	$6.38_{0.9}$
RoBERTa+Mixup	$2.31_{0.4}$	$1.28_{0.3}$	$4.19_{0.7}$	$5.65_{0.2}$	$3.13_{0.2}$	$6.82_{0.3}$
RoBERTa+Mixup+(LS+TS)	$1.96_{0.6}$	$0.74_{0.7}$	$3.15_{0.4}$	$3.58_{0.3}$	$2.84_{0.5}$	$5.60_{0.2}$
RoBERTa+Ours	$1.57_{0.3}$	$\underline{0.32}_{0.2}$	$2.61_{0.4}$	$2.20_{0.6}$	$1.92_{0.6}$	$3.12_{0.5}$
RoBERTa+Ours+(LS+TS)	$\underline{1.52}_{0.7}$	$0.33_{0.6}$	$\underline{2.18}_{0.8}$	$\underline{2.14}_{0.4}$	$\underline{1.90}_{0.5}$	$\underline{2.87}_{0.6}$
$SD_{RoBERTa}$	$1.78_{0.8}$	$0.55_{0.4}$	$2.93_{0.9}$	$5.34_{0.6}$	$2.51_{0.5}$	$4.32_{0.2}$
$SD_{RoBERTa}$+(LS+TS)	$1.85_{0.9}$	$0.67_{0.7}$	$3.10_{0.5}$	$5.06_{0.6}$	$2.34_{0.3}$	$4.16_{0.5}$
$SD_{RoBERTa}$+Mixup	$1.85_{1.2}$	$0.75_{0.8}$	$3.46_{0.5}$	$5.62_{0.6}$	$2.66_{0.3}$	$4.69_{0.6}$
$SD_{RoBERTa}$+Mixup+(LS+TS)	$1.97_{0.6}$	$0.88_{0.8}$	$3.70_{0.6}$	$5.62_{0.4}$	$2.93_{0.7}$	$4.41_{0.8}$
$SD_{RoBERTa}$+Ours	$1.30_{0.4}$	$0.21_{0.6}$	$0.98_{0.8}$	$2.03_{0.7}$	$1.72_{0.6}$	$2.29_{0.5}$
$SD_{RoBERTa}$+Ours+(LS+TS)	$\mathbf{1.13_{0.3}}$	$\mathbf{0.18_{0.5}}$	$\mathbf{0.83_{0.4}}$	$\mathbf{1.86_{0.7}}$	$\mathbf{1.38_{0.4}}$	$\mathbf{2.08_{0.4}}$

Table 2: Expected calibration errors (ECE) in percentage (%) for in-domain (GoEmotions, Yelp, EPITOME) and out-of-domain (Meld, SST, NewsEmpathy) datasets. LS, TS, and SD refer to the label smoothing, temperature scaling, and self-distillation, respectively. All reported values for the methods are $\text{mean}_{\pm \text{ std}}$ of three repetitions. Best results are **bolded**.

training set for each task. Then, we evaluate its performance on both the in-domain and out-of-domain test sets. We make a few remarks below.

First, distillation leads to improved accuracy and model calibration compared to the standalone models (BERT or RoBERTa), both in the in-domain and out-of-domain settings. We can see from Table 2 that SD_* yield better-calibrated models with lower ECE in all of the experiments with our setup. As shown in Table 2, the errors obtained with self-distillation are much smaller in general compared to the standalone models. For example, on Yelp, SD_{BERT} reduces ECE by a factor of 2 compared to the vanilla pre-trained BERT. The results indicate that the dark knowledge of a teacher model can act as a regularization process, helping to calibrate the student model's confidence in its predictions. This phenomenon is visually shown in Figure 3 in Appendix C.

Second, compared to the vanilla pre-trained models (with or without distillation), label smoothing with temperature scaling[5] (+(LS + T S)) does not always improve the calibration or accuracy of the models, specifically in the in-domain setting. For example, on EPITOME, the accuracy of the BERT model is decreased by 1.17%, and the ECE is increased from 5.53 to 6.49. On the other hand, in the out-of-domain setting, employing label smoothing with temperature scaling (+(LS + TS)) results in a decrease in the calibration error in most settings. We also observe that in the distillation settings with (+(LS + TS)), the increase in calibration, in most cases, coincides with the stagnation of student test accuracy, which in turn shows the inefficacy of such regularization techniques (especially for the in-domain setting). The results indicate that label smoothing with temperature scaling may not *always* be effective in calibrating the pre-trained language model's predictions as they do not show a consistent behavior on the calibration and performance. Consequently, we conclude that stronger regularization strategies are required to temper the miscalibration of the pre-trained language models.

Third, as shown in Tables 1 and 2, no significant improvements in calibration or performance are observed by solely incorporating plain mixup (i.e., Mixup) on the standalone models (i.e., BERT or RoBERTa). Similarly, we observe that if a teacher model is trained using plain mixup, the student model distilled from it is impaired in calibration and its generalization capabilities in most cases. Such an aggravation of miscalibration may be due to the quality of the generated augmented samples in the mixup process that afflicts the models to capture the intricacies of the data. We hypothesize that this adversarial impact leads to a loss in the quality of the supervision signal during training or distillation. In contrast, incorporating (+(LS + TS)) on plain mixup leads to lower ECEs on some cases. Nevertheless, solely incorporating plain mixup without other regularization strategies (in our case (+(LS + TS))) is not effective in calibrating the model's predictions.

Finally, it is worth noting that we obtain encouraging results with our proposed informed mixup. From the Table 2, we see that the errors obtained with our mixup method are much smaller in general compared to the other settings (Figure 4 in Appendix C). Interestingly, we observe that the student models distilled from a teacher trained using our mixup strategy yield the best-calibrated models on both the in-domain and out-of-domain data (see self-distillation + Ours ECE compared with other settings). Moreover, we find that incorporating (+(LS + TS)) generally helps further to improve the calibration and performance of the pre-trained language models. The results suggest that, unlike the baseline mixup method that focuses more on the class-specific features, by incorporating training dynamics into the mixup process, we focus more on the instance-specific major features that lead to the more calibrated models. In that capacity, we boost the amount of information encoded in all the latent features encoded by the teacher model, which spurs the student model to generate more generalized and calibrated predictions.

5 Conclusion

In this paper, we showed that the dark knowledge of a pre-trained language teacher model could act as a regularization process, helping to calibrate the student model's confidence in its predictions. We further proposed an informed mixup process and demonstrated that using training dynamics to guide the interpolation process in the mixup data augmentation on a teacher model can effectively improve the calibration of the student model in a

[5]Since prior work (Desai and Durrett, 2020; Kong et al., 2020; Park and Caragea, 2022) showed that solely incorporating label smoothing or temperature scaling cannot consistently improve the calibration of the pre-trained langue models (in- or out-of-distribution), we only report the effect of these two techniques together in our (+(LS + TS)) setting.

distillation setting. We showed that student models distilled from such teacher models trained using our proposed mixup method not only achieved the best performance but also obtained the lowest expected calibration errors (ECEs) on both in-domain and out-of-domain test sets on emotion-related tasks.

6 Limitations

Our proposed approach shows that using training dynamics to generate mixup samples along with the dark knowledge of a pre-trained language teacher model can act as a regularization process, which helps to calibrate the student model's confidence in its predictions. It would be interesting to analyze the impact of adding mixed data augmentation techniques to the student networks on the calibration of the pre-trained language models. One potential limitation of our approach is using a small additional overhead for calculating statistics with the data maps tool. However, this is a common limitation for all approaches that use this data maps.

Acknowledgements

This research is supported in part by NSF Convergence Accelerator award #2137846, NSF IIS award #2107487, and NSF BigData award #1912887. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF. We thank AWS for computational resources that supported this work. We also thank our anonymous reviewers for their constructive feedback.

References

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2654–2662.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *CoRR*, abs/1808.10399.

Dallas Card and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1636–1646.

Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. 2020. On mixup regularization. *CoRR*, abs/2006.06049.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2147–2157. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics.

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 2020, Online, November 16-20, 2020*. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions?

Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1602–1611. PMLR.

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

Hongyu Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 4044–4051. AAAI Press.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *CoRR*, abs/1905.08941.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. 2020. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1326–1340. Association for Computational Linguistics.

Volodymyr Kuleshov and Stefano Ermon. 2016. Reliable confidence estimation via online learning. *CoRR*, abs/1607.03594.

Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated structured prediction. *Advances in Neural Information Processing Systems*, 28.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *CoRR*, abs/1903.00802.

Kevin J. Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2021. Mixkd: Towards efficient distillation of large-scale language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1288–1298. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Khanh Nguyen and Brendan O'Connor. 2015a. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1587–1598. The Association for Computational Linguistics.

Khanh Nguyen and Brendan O'Connor. 2015b. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. The Association for Computational Linguistics.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *CoRR*, abs/1906.02530.

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. *arXiv preprint arXiv:2203.07559*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017a. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017b. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5142–5151. PMLR.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 EMNLP*, pages 1631–1642.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP 2020, Online, November 16-20, 2020*, pages 9275–9293. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13888–13899.

Dongdong Wang, Yandong Li, Liqiang Wang, and Boqing Gong. 2020a. Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model. In *CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1495–1504. Computer Vision Foundation / IEEE.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020b. On the inference calibration of neural machine translation. In *Proceedings of ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Junho Yim, Donggyu Joo, Ji-Hoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7130–7138. IEEE Computer Society.

Wenpeng Yin, Huan Wang, Jin Qu, and Caiming Xiong. 2021. Batchmixup: Improving training by interpolating hidden states of the entire mini-batch. In *Findings of ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4908–4912. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1958–1970. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

Zhilu Zhang and Mert R. Sabuncu. 2020. Self-distillation as instance-specific label smoothing. In *NeurIPS 2020, December 6-12, 2020, virtual*.

A Implementation Details

For creating the data maps, we use `RoBERTa-base` (Liu et al., 2019) with the same set of hyper-parameters as (Swayamdipta et al., 2020). For the experiments, we fine-tune BERT `bert-base-uncased` (Devlin et al., 2019), and RoBERTa `roberta-base` from the HuggingFace Transformers library (Wolf et al., 2019). All the models are trained with a maximum of 3 epochs. BERT is fine-tuned with a batch size of 16, learning rate of $2e-5$, gradient clip of 1.0, and no weight decay. RoBERTa is fine-tuned with a batch size of 32, learning rate of $1e-5$, gradient clip of 1.0, and weight decay of 0.1. Models are optimized with AdamW (Loshchilov and Hutter, 2019). We train our models with subsets containing top $[50\%, 33\%, 25\%]$ easy-to-learn and ambiguous samples and find 33% (i.e., total 66% train set) results in the best in-domain and out-of-domain performance. For mixup, the α parameter 0.4 in the beta distribution works best in our settings amongst $[0.3, 0.4, 0.5]$. For label smoothing, we change the smoothing hyper-parameter α in the range $[0.05, 0.1, 0.2, 0.3, 0.4]$, and find 0.1 to work best for our settings. We utilize the in-domain development set for temperature scaling to obtain an optimum temperature T in the range of $[0.01, 5.0]$ with a granularity of 0.01. For distillation with data augmentation, we first train the pre-trained language models (i.e., BERT or RoBERTa) using the data augmentation techniques discussed above (i.e., standard mixup or our informed mixup) on each task's training dataset. Then, the student networks are built from the BERT or RoBERTa with no data augmentation techniques added. In the knowledge distillation setting, all the regularization methods (i.e., mixup methods, label smoothing, and temperature scaling) are only applied to the teacher model before using the model as a teacher for the student model. For all the results, we report the mean performance over 3 random seeds. Finally, fine-tuning all the models took in total less than one day on our NVIDIA GP100 16GB GPU.

B Knowledge Distillation Experiments

Tables 3 and 4 present the comparison results of the knowledge distillation setting with baselines (the baseline results are borrowed from Tables 1 and 2). Unlike the self-distillation setting, in this setting, we first train a pre-trained language model (i.e., BERT or RoBERTa) as a teacher model and then train a student model (with a different architecture) to mimic the output of the teacher model. For example, if we choose to train RoBERTa as the teacher model, then we train BERT as the student model to learn from the output of the RoBERTa teacher model, and vice versa.

From the Tables 3 and 4, we can observe that in most cases teacher-student training in the knowledge distillation setting (i.e., KD_*) results in better-calibrated student models compared to the standalone models. For example, on NewsEmpathy, $\text{KD}_{\text{RoBERTa}}$ reduces ECE by 3.29% compared to the vanilla pre-trained RoBERTa.

C Visualisations of Comparison of the Knowledge and Self-distillation with Vanilla Models

Figure 3 compares knowledge and self-distillation with vanilla models and illustrates that distillation settings yield better-calibrated models with lower ECE in all of the experiments with our setup (with the self-distillation being the best setting). Figure 4 shows that the errors obtained with our mixup method are much smaller in general compared to the other settings and yields best calibrated models.

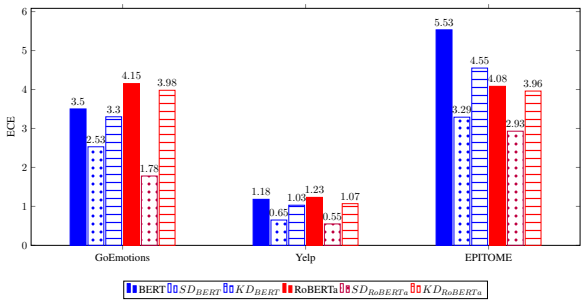

Figure 3: In-domain expected calibration errors (ECE) of vanilla BERT, SD_{BERT}, KD_{BERT}, RoBERTa, $\text{SD}_{\text{RoBERTa}}$, and $\text{KD}_{\text{RoBERTa}}$.

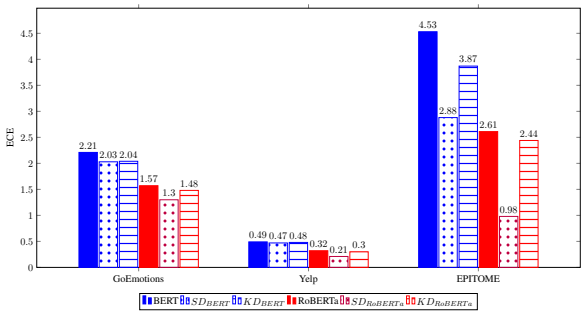

Figure 4: In-domain expected calibration errors (ECE) of BERT, SD_{BERT}, KD_{BERT}, RoBERTa, $\text{SD}_{\text{RoBERTa}}$, and $\text{KD}_{\text{RoBERTa}}$ with our proposed mixup.

	In-Domain			Out-of-Domain		
Model	GoEmotions	Yelp	EPITOME	Meld	SST	NewsEmpathy
BERT	$68.10_{0.2}$	$95.87_{0.4}$	$67.42_{0.3}$	$41.64_{0.2}$	$79.83_{0.5}$	$57.10_{0.4}$
BERT+(LS+TS)	$68.03_{0.6}$	$95.42_{0.2}$	$66.25_{0.5}$	$41.86_{1.2}$	$80.41_{0.8}$	$56.75_{0.9}$
BERT+Mixup	$67.83_{0.6}$	$95.81_{0.4}$	$67.35_{0.7}$	$41.48_{0.3}$	$79.70_{0.5}$	$57.08_{0.2}$
BERT+Mixup+(LS+TS)	$69.58_{0.4}$	$96.52_{0.3}$	$67.83_{0.5}$	$42.71_{0.5}$	$80.65_{0.8}$	$57.81_{0.7}$
BERT+Ours	$70.52_{0.3}$	$96.90_{0.2}$	$68.93_{0.5}$	$44.12_{0.6}$	$81.10_{0.5}$	$58.45_{0.5}$
BERT+Ours+(LS+TS)	$\underline{70.84}_{0.6}$	$\underline{96.99}_{0.3}$	$\underline{69.83}_{0.4}$	$\underline{45.58}_{0.7}$	$\underline{81.65}_{0.7}$	$\underline{59.20}_{0.9}$
KD_{BERT}	$68.15_{0.2}$	$95.93_{0.3}$	$67.85_{0.5}$	$42.37_{0.4}$	$80.33_{0.2}$	$57.84_{0.6}$
KD_{BERT}+(LS+TS)	$68.03_{0.7}$	$95.46_{0.5}$	$67.28_{0.5}$	$42.50_{0.2}$	$80.56_{0.6}$	$58.16_{0.3}$
KD_{BERT}+Mixup	$67.90_{1.2}$	$95.61_{0.8}$	$67.44_{0.6}$	$41.98_{0.7}$	$80.11_{0.6}$	$57.34_{0.4}$
KD_{BERT}+Mixup+(LS+TS)	$67.83_{0.8}$	$95.71_{0.6}$	$67.23_{0.8}$	$41.52_{0.3}$	$80.29_{0.2}$	$57.84_{0.6}$
KD_{BERT}+Ours	$70.63_{0.3}$	$96.98_{0.5}$	$69.51_{0.2}$	$42.65_{0.4}$	$81.38_{0.5}$	$59.09_{0.8}$
KD_{BERT}+Ours+(LS+TS)	$\mathbf{71.24_{0.3}}$	$\underline{97.10}_{0.6}$	$\underline{69.86}_{0.2}$	$\underline{42.80}_{0.5}$	$\underline{81.63}_{0.6}$	$\underline{59.79}_{0.3}$
RoBERTa	$68.25_{0.5}$	$96.16_{0.7}$	$68.38_{1.2}$	$42.17_{0.8}$	$82.84_{0.6}$	$56.88_{0.4}$
RoBERTa+(LS+TS)	$68.17_{0.4}$	$96.05_{0.6}$	$67.87_{0.2}$	$42.94_{0.3}$	$82.96_{0.7}$	$55.73_{0.5}$
RoBERTa+Mixup	$68.20_{0.8}$	$96.07_{0.7}$	$68.24_{0.5}$	$43.12_{0.4}$	$82.77_{0.2}$	$56.81_{0.3}$
RoBERTa+Mixup+(LS+TS)	$68.47_{0.7}$	$96.79_{0.2}$	$68.65_{0.5}$	$44.80_{1.2}$	$83.24_{0.8}$	$58.86_{0.2}$
RoBERTa+Ours	$70.57_{0.4}$	$97.25_{0.3}$	$69.60_{0.5}$	$48.32_{0.4}$	$85.24_{0.6}$	$58.67_{0.6}$
RoBERTa+Ours+(LS+TS)	$\underline{70.82}_{0.2}$	$\underline{97.40}_{0.5}$	$\mathbf{70.21_{0.5}}$	$\underline{49.10}_{0.3}$	$\underline{85.44}_{0.4}$	$\underline{58.92}_{0.4}$
$KD_{RoBERTa}$	$68.33_{0.8}$	$96.18_{0.5}$	$68.40_{0.6}$	$49.35_{1.2}$	$82.90_{0.5}$	$56.89_{0.8}$
$KD_{RoBERTa}$+(LS+TS)	$68.47_{0.5}$	$96.10_{0.8}$	$68.17_{0.4}$	$49.77_{0.7}$	$83.20_{0.3}$	$57.36_{0.7}$
$KD_{RoBERTa}$+Mixup	$68.21_{0.6}$	$96.37_{0.8}$	$67.75_{0.3}$	$49.42_{0.5}$	$82.63_{0.4}$	$56.64_{0.4}$
$KD_{RoBERTa}$+Mixup+(LS+TS)	$68.55_{0.3}$	$96.21_{0.6}$	$67.42_{0.7}$	$49.50_{0.2}$	$82.78_{0.8}$	$57.13_{0.7}$
$KD_{RoBERTa}$+Ours	$70.71_{0.4}$	$\mathbf{97.90_{0.5}}$	$69.84_{0.4}$	$\mathbf{49.84_{0.6}}$	$85.25_{0.3}$	$59.83_{0.5}$
$KD_{RoBERTa}$+Ours+(LS+TS)	$\underline{70.93}_{0.3}$	$97.83_{0.2}$	$\underline{70.09}_{0.7}$	$49.61_{0.3}$	$\mathbf{85.60_{0.8}}$	$\mathbf{60.26_{0.6}}$

Table 3: Accuracy in percentage (%) for in-domain (GoEmotions, Yelp, EPITOME) and out-of-domain (Meld, SST, NewsEmpathy) datasets. LS, and TS refer to the label smoothing, and temperature scaling, respectively. KD refers to knowledge distillation from a teacher model that is specified at the subscript and a student model with a different architecture (for example, in $KD_{RoBERTa}$, RoBERTa is used as the teacher in the knowledge distillation setting and BERT is used a the student model). For space restrictions, we do not include the student model. All reported values for the methods are $mean_{\pm\ std}$ of three repetitions.

	In-Domain			Out-of-Domain		
Model	GoEmotions	Yelp	EPITOME	Meld	SST	NewsEmpathy
BERT	$3.50_{0.6}$	$1.18_{0.3}$	$5.53_{0.8}$	$6.23_{0.4}$	$3.69_{0.3}$	$6.63_{0.7}$
BERT+(LS+TS)	$3.98_{0.4}$	$1.87_{0.6}$	$6.49_{0.3}$	$5.33_{0.6}$	$3.29_{0.5}$	$5.46_{0.6}$
BERT+Mixup	$3.57_{0.5}$	$1.21_{0.2}$	$5.98_{0.7}$	$6.35_{0.4}$	$3.77_{0.3}$	$6.72_{0.4}$
BERT+Mixup+(LS+TS)	$2.27_{0.3}$	$0.73_{0.6}$	$4.40_{0.6}$	$5.58_{0.7}$	$2.71_{0.2}$	$5.24_{0.8}$
BERT+Ours	$2.21_{0.5}$	$0.49_{0.3}$	$4.53_{0.2}$	$5.18_{0.3}$	$2.56_{0.2}$	$4.89_{0.5}$
BERT+Ours+(LS+TS)	$\underline{1.96}_{0.6}$	$\underline{0.45}_{0.6}$	$\underline{3.29}_{0.7}$	$\underline{4.86}_{0.5}$	$\underline{2.08}_{0.6}$	$\underline{4.60}_{0.4}$
KD_{BERT}	$3.30_{0.6}$	$1.03_{0.3}$	$4.55_{0.7}$	$6.17_{0.4}$	$3.65_{0.6}$	$4.57_{0.5}$
KD_{BERT}+(LS+TS)	$3.18_{0.8}$	$1.58_{0.6}$	$4.96_{0.4}$	$5.52_{0.7}$	$3.27_{0.3}$	$4.08_{0.4}$
KD_{BERT}+Mixup	$3.47_{0.3}$	$1.38_{0.3}$	$4.96_{0.7}$	$6.21_{0.5}$	$4.02_{0.2}$	$4.70_{0.6}$
KD_{BERT}+Mixup+(LS+TS)	$3.21_{0.3}$	$1.16_{0.5}$	$4.29_{0.4}$	$5.67_{0.2}$	$3.67_{0.8}$	$4.42_{0.4}$
KD_{BERT}+Ours	$2.04_{0.2}$	$0.48_{0.4}$	$3.87_{0.6}$	$4.26_{0.4}$	$2.47_{0.3}$	$3.05_{0.2}$
KD_{BERT}+Ours+(LS+TS)	$\underline{2.00}_{0.4}$	$\underline{0.36}_{0.3}$	$\underline{2.64}_{0.3}$	$\underline{3.10}_{0.4}$	$\underline{2.23}_{0.5}$	$\underline{2.82}_{0.2}$
RoBERTa	$4.15_{0.6}$	$1.23_{0.8}$	$4.08_{0.3}$	$5.52_{0.6}$	$3.05_{0.2}$	$6.71_{0.5}$
RoBERTa+(LS+TS)	$4.35_{1.2}$	$1.67_{0.6}$	$4.92_{0.8}$	$5.21_{0.5}$	$2.69_{0.6}$	$6.38_{0.9}$
RoBERTa+Mixup	$2.31_{0.4}$	$1.28_{0.3}$	$4.19_{0.7}$	$5.65_{0.2}$	$3.13_{0.2}$	$6.82_{0.3}$
RoBERTa+Mixup+(LS+TS)	$1.96_{0.6}$	$0.74_{0.7}$	$3.15_{0.4}$	$3.58_{0.3}$	$2.84_{0.5}$	$5.60_{0.2}$
RoBERTa+Ours	$1.57_{0.3}$	$0.32_{0.2}$	$2.61_{0.4}$	$2.20_{0.6}$	$1.92_{0.6}$	$3.12_{0.5}$
RoBERTa+Ours+(LS+TS)	$\underline{1.52}_{0.7}$	$\underline{0.33}_{0.6}$	$\mathbf{2.18_{0.8}}$	$\underline{2.14}_{0.4}$	$\underline{1.90}_{0.5}$	$\underline{2.87}_{0.6}$
$KD_{RoBERTa}$	$3.98_{0.4}$	$1.07_{0.5}$	$3.96_{0.4}$	$5.49_{0.3}$	$3.05_{0.2}$	$3.42_{0.6}$
$KD_{RoBERTa}$+(LS+TS)	$4.26_{0.8}$	$0.93_{0.6}$	$4.12_{0.5}$	$5.20_{0.7}$	$2.84_{0.6}$	$3.35_{0.3}$
$KD_{RoBERTa}$+Mixup	$3.43_{0.7}$	$1.58_{0.6}$	$3.90_{0.9}$	$4.74_{0.8}$	$3.06_{0.4}$	$3.24_{0.4}$
$KD_{RoBERTa}$+Mixup+(LS+TS)	$3.36_{0.5}$	$1.77_{0.5}$	$3.98_{0.6}$	$4.69_{1.1}$	$2.90_{0.8}$	$3.18_{0.7}$
$KD_{RoBERTa}$+Ours	$1.48_{0.4}$	$\mathbf{0.30_{0.3}}$	$2.44_{0.5}$	$\mathbf{2.08_{0.4}}$	$1.88_{0.6}$	$2.94_{0.2}$
$KD_{RoBERTa}$+Ours+(LS+TS)	$\mathbf{1.12_{0.5}}$	$0.37_{0.7}$	$\underline{2.27}_{0.3}$	$2.17_{0.6}$	$\mathbf{1.32_{0.8}}$	$\mathbf{2.55_{0.5}}$

Table 4: Expected calibration errors (ECE) in percentage (%) for in-domain (GoEmotions, Yelp, EPITOME) and out-of-domain (Meld, SST, NewsEmpathy) datasets. For the definitions refer to Table 3.