Model Calibration for Emotion Detection

Mihaela Petre-Vlad[◇] **Cornelia Caragea**[♣] **Florentina Hristea**[◇]
[◇]Computer Science, University of Bucharest, Romania
[♣]Computer Science, University of Illinois Chicago, USA
mihaela.petre-vlad@my.fmi.unibuc.ro; cornelia@uic.edu; fhristea@fmi.unibuc.ro

Abstract

In this paper, we propose a unified approach to model calibration for emotion detection that exploits the complementary strengths of knowledge distillation and the MixUp data augmentation technique to enhance the trustworthiness of emotion detection models. Specifically, we use a MixUp method informed by training dynamics that generates augmented data by interpolating easy-to-learn with ambiguous samples based on their similarity and dissimilarity provided by saliency maps. We use this MixUp method to calibrate the teacher model in the first generation of the knowledge distillation process. To further calibrate the teacher models in each generation, we employ dynamic temperature scaling to update the temperature used for scaling the teacher predictions. We find that calibrating the teachers with our method also improves the calibration of the student models. We test our proposed method both in-distribution (ID) and out-of-distribution (OOD). To obtain better OOD performance, we further fine-tune our models with a simple MixUp method that interpolates a small number of OOD samples with ambiguous ID samples.

1 Introduction

Emotion detection in written content (Kusal et al., 2022; Plaza-del Arco et al., 2024) is a text classification task that entails the analysis of textual data in order to discern the emotional state expressed in it. Text-based emotion detection has applications in many fields, ranging from customer service (Agarwal et al., 2021; Brun et al., 2025) to mental health support (Sosea and Caragea, 2020; Nijhawan et al., 2022), where it is important not only to correctly predict an emotion but also to know how trustworthy the prediction is. This is where model calibration comes in, which measures the discrepancy between the confidence of the model in a prediction (that is, the probability output assigned to the prediction) and the correctness of that prediction (i.e., accuracy). A well-calibrated model can inform us through its confidence when it is uncertain about a prediction (i.e., its confidence is low) and there is a chance that the prediction is wrong. However, modern neural networks tend to suffer from miscalibration (Guo et al., 2017; Desai and Durrett, 2020).

Recently, Hosseini and Caragea (2022) studied the calibration of the pre-trained models BERT and RoBERTa across three emotion-related tasks: emotion detection, sentiment analysis, and empathy detection. They obtained better calibrated models both in and out of domain by training a model with MixUp (Zhang et al., 2018), which has been found to improve model calibration across diverse tasks (Thulasidasan et al., 2019), and then applied knowledge distillation (Hinton et al., 2015), which involves transferring knowledge from a teacher model to a student model. The specific MixUp method they used interpolates the top 33% easy-to-learn samples with the top 33% most ambiguous samples. They also experimented with two popular regularization techniques widely used to reduce miscalibration: temperature scaling (Guo et al., 2017) and label smoothing (Pereyra et al., 2017).

In contrast to the above study, which focused on calibrating the student models in a single generation of knowledge distillation, our study aims to calibrate both the teacher and student models in multiple generations using MixUp and dynamic temperature scaling. We also used a different MixUp method to train the teacher in the first generation. Our MixUp method uses 100% of the data to generate more informative samples based on training dynamics and saliency maps (Simonyan et al., 2014). Specifically, we first leveraged the training dynamics to separate the training data into two equally sized sets: easy-to-learn and ambiguous samples (which include hard-to-learn samples). We then used MixUp to combine samples from the two sets with the most similar and dissimilar samples from

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20442–20457
November 4-9, 2025

the other set according to their saliency signals.

Our contributions are summarized as follows:

- We used the MixUp method based on training dynamics and saliency to train the initial teacher model in the first generation in order to obtain a better calibrated and more robust teacher. We found that doing knowledge distillation on calibrated teachers helps improve the performance and calibration of the student models.
- In order to calibrate the teacher model in each generation, we used dynamic temperature scaling to estimate the temperature that best calibrates the teacher model on the validation set.
- By combining knowledge distillation with dynamic temperature scaling and the previously described MixUp method, we effectively improved the calibration of the student models.
- To increase the performance in the OOD setting, we further trained our models on a small percentage of the OOD training data together with additional samples created by mixing the OOD samples with ambiguous ID samples in the feature space.

2 Related Work

2.1 Model Calibration

Miscalibration is a common problem in modern neural networks (Guo et al., 2017; Desai and Durrett, 2020), which implies that the models' confidence is not reliable and, consequently, their predictions are not trustworthy. Methods to improve the calibration of neural network models have been studied on both image classification and natural language processing tasks. Guo et al. (2017) found that using temperature scaling before the softmax operation is one of the most effective methods to reduce calibration errors for both image and document classification tasks. Thulasidasan et al. (2019) showed that deep neural networks trained with MixUp are better calibrated on both image classification and NLP tasks. Desai and Durrett (2020) used temperature scaling and label smoothing to obtain better calibration for the pre-trained language models BERT and RoBERTa on three language understanding tasks: natural language inference, paraphrase detection, and commonsense reasoning, both in-domain and out-of-domain. Kong et al. (2020) studied the calibration of BERT with MixUp for both in-domain and out-of-domain settings. New samples were generated using the cosine distance between samples in the feature space. Jung et al. (2020) proposed an end-to-end training procedure called posterior calibrated (PosCal) training that reduced the calibration error on two benchmarks for NLP classification tasks: GLUE and xSLUE. They fine-tuned the BERT model by jointly optimizing a cross-entropy loss and a calibration loss while dynamically minimizing the difference between the predicted and the true posterior probabilities during training. Park and Caragea proposed two new MixUp methods for calibrating the BERT model, both in-domain and out-of-domain: TD-MixUp (2022a) and MixUp guided by AUM and saliency (2022b). The MixUp methods were evaluated on three language understanding tasks: natural language inference, paraphrase detection, and commonsense reasoning. Li and Caragea (2023) explored the effect of knowledge distillation over multiple generations and dynamic temperature scaling on calibration for stance detection.

2.2 Emotion Detection

Suresh and Ong (2021) proposed Label-Aware Contrastive Loss (LCL), a method for fine-grained emotion classification tasks that incorporates relationships between labels into contrastive learning. LCL was proven to help the model differentiate between easily confusable classes, but its effect on calibration was not studied. The method was also not evaluated in the OOD setting.

Zanwar et al. (2022) addressed the challenge of out-of-domain generalization in emotion detection. They proposed hybrid models that combine transformer-based architectures (BERT and RoBERTa) with Bidirectional Long Short-Term Memory (BiLSTM) networks trained on psycholinguistic features. Their approach improved generalizability across various text-based emotion classification datasets.

3 Proposed Approach

3.1 Overview

A model is well-calibrated if its confidence reflects the likelihood of the predicted outcome. In other words, an event predicted by a well-calibrated model with confidence p should empirically be true p of the time (Guo et al., 2017). The metric we use to evaluate calibration is the **Expected Calibration Error (ECE)** (Naeini et al., 2015). Lower ECE

values indicate better calibrated models.

In this section we introduce our proposed approach for emotion detection calibration that unifies ideas for combating miscalibration. Specifically, we combined two techniques: a MixUp method that creates more challenging examples and a knowledge distillation method that employs dynamic temperature scaling. This holistic approach produces better calibrated teachers during the knowledge distillation process, which, in turn, results in better calibrated students. We explain our key components in the following subsections.

3.2 Calibrated Knowledge Distillation

Knowledge distillation (KD) introduced by Hinton et al. (2015) refers to the process of transferring knowledge from a **teacher** model to a **student** model by training the student model to mimic the teacher's output probabilities. In our study we used **born-again networks (BANs)** (Furlanello et al., 2018), a particular case of knowledge distillation in which the teacher and student models have the same architecture (also called self-distillation). Specifically, we used the RoBERTa base architecture. Furlanello et al. (2018) showed that BAN student models can achieve better performance than their teachers over multiple generations. In each new generation, the student model from the previous generation takes the role of the teacher to transfer its knowledge to a new fresh student model with identical architecture.

Given a k-class text classification task and the training data $D^{tr} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is an input sentence and y_i is the corresponding one-hot hard label, standard supervised learning trains a model by minimizing the cross-entropy loss L_{CE} of the training data, described below.

Let z_i be the model's prediction logits (i.e., the model's unnormalized output) for input x_i and let $p_i = \frac{exp(z_i)}{\sum_{j=1}^{k} exp(z_{ij})}$ be the model's softmax output for z_i, then the cross-entropy loss is defined as:

$$L_{CE} = - \sum_{(x_i, y_i) \in D^{tr}} l_{CE}(p_i, y_i)$$

where $l_{CE}(p_i, y_i)$ is the cross-entropy loss for a single training example:

$$l_{CE}(p_i, y_i) = \sum_{j=1}^{k} y_{ij} log(p_{ij})$$

In the self-distillation setting, a student model is trained using the hard labels and the teacher prediction logits (i.e., the soft labels) by minimizing L_{KD}, the weighted sum of the cross-entropy loss between the hard labels and the student's predictions and the difference loss between the teacher's and student's predictions:

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

where L_{KL} is the Kullback-Leibler (KL) divergence loss:

$$L_{KL} = \sum_{x_i \in D^{tr}} l_{KL}(p^s(x_i), p^t(x_i))$$

where $l_{KL}(p^s(x_i), p^t(x_i))$ is the KL divergence loss for the training example x_i and $p^t(x_i) = \sigma(z^t(x_i))$ and $p^s(x_i) = \sigma(z^s(x_i))$ denote the softmax predictions of the teacher and student models, respectively, where σ is the softmax function and $z^t(x_i)$ and $z^s(x_i)$ denote the output logits of the teacher and student models for input x_i.

The KL divergence loss for a single input is defined as:

$$l_{KL}(p^s(x_i), p^t(x_i)) = \sum_{j=1}^{k} p_j^t(x_i) log \frac{p_j^t(x_i)}{p_j^s(x_i)}$$

In the L_{KD} loss, the coefficient $\lambda \in (0, 1)$ is used to weight the importance of the two loss functions. We use **teacher annealing**, introduced by Clark et al. (2019) to optimize λ. This method gradually transitions the student model from self-distillation to supervised learning as the training progresses.

To effectively calibrate the prediction probabilities of the teacher models during the knowledge distillation process, we use temperature scaling (TS) (Guo et al., 2017). For an instance x, temperature scaling divides the logit vector $z(x)$ by the scalar T before the softmax operation. Therefore, the new confidence prediction for x obtained with TS is $\sigma(z(x)/T)$, where σ denotes the softmax operation.

When temperature scaling is combined with knowledge distillation, a slightly modified version of the KL divergence loss is used. Typically, the new formula for L_{KL} is the following:

$$L_{KL} = T^2 \sum_{x_i \in D^{tr}} l_{KL}(\sigma(z^s(x_i)/T), \sigma(z^t(x_i)/T)$$

where $z^t(x_i)/T$ and $z^s(x_i)/T$ are the scaled output logits of the teacher and student models, respectively, and T is usually fixed. However, we

Algorithm 1 : Sim-Mixup

Require: $\mathcal{D}_{train} = \{(x_i, y_i)\}_{i=1}^{n}$; model f

Compute the data map of $\mathcal{D}_{train}$ and sort the samples in descending order by confidence. Split $\mathcal{D}_{train}$ into the easy-to-learn subset $\mathcal{D}_{easy}$ and the ambiguous subset (that includes hard-to-learn samples) $\mathcal{D}_{ambig}$ of equal size.
for e := 0 to E **do**
    $L_{total} \leftarrow 0$
    for i := 0 to $|\mathcal{D}_{train}|$ **do**
        $L_{CE} \leftarrow CrossEntropy(f(x_i), y_i)$
        Construct a saliency map S by computing the gradient of L_{CE} with respect to the logit vector z_i.
        if $(x_i, y_i) \in \mathcal{D}_{easy}$ **then**:
            Find the most similar and dissimilar samples from $\mathcal{D}_{ambig}$
        else if $(x_i, y_i) \in \mathcal{D}_{ambig}$ **then**:
            Find the most similar and dissimilar samples from $\mathcal{D}_{easy}$
        end if
        Generate two MixUp samples by interpolating (x_i, y_i) with its most similar and dissimilar sample.
        Compute the cross-entropy losses L' and L'' of the two MixUp samples.
        $L_{total} = L_{total} + 0.8L_{CE} + 0.1L' + 0.1L''$
    end for
    Update the model weights
end for

obtained generally better results when the student logits were not scaled (see ***Scaling vs not scaling the student*** in the Appendix), so we did not use temperature scaling on the student models in the final experiments.

We dynamically choose the temperature that minimizes the ECE of the teacher model on the ID validation set in each generation. This method is called Calibration-based Knowledge Distillation (CKD) (Li and Caragea, 2023). The choice of the optimal temperature is time-efficient since it only involves dividing the teacher's softmax outputs by potential temperature values and then computing the corresponding ECE values.

3.3 MixUp

The original **MixUp**, introduced by Zhang et al. (2018), is a data augmentation technique that generates new samples during training by combining

Algorithm 2 : Our Proposed Method

Require: training set $\mathcal{D}_{train}$, validation set $\mathcal{D}_{val}$, number generations G

$gen \leftarrow 1$
Train the first teacher model for E epochs using the Sim-MixUp method in Algorithm 1.
Do temperature scaling on the teacher's output probabilities with temperature T that minimizes the ECE of the teacher on $\mathcal{D}_{val}$. Stop the algorithm if gen is equal to G, otherwise continue to step 4.
Train on $\mathcal{D}_{train}$ a student model with the same architecture as the teacher by minimizing the weighted sum of the cross-entropy loss and the KL-divergence loss:

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

where λ is optimized using teacher annealing.
Take the student as a new teacher, increase gen by one and return to step 2.

random pairs of training samples and their labels in the input space. The new samples are created using the following rule:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j$$

where x_i and x_j are two randomly selected input instances with their associated one-hot encoded labels y_i and y_j, and λ is a mixing ratio sampled from a Beta(α, α) distribution with a hyper-parameter α.

Verma et al. (2019) showed that, rather than combining input-level features, interpolating **hidden representations** in the feature space leads to better regularization effects because the model is encouraged to focus on the representations of the training examples in a low dimensional subspace.

Inspired by prior work (Hosseini and Caragea, 2022; Park and Caragea, 2022a) that used MixUp guided by training dynamics (Swayamdipta et al., 2020) for model calibration, we split the training data evenly into easy-to-learn and ambiguous samples using the median confidence as a threshold. According to Swayamdipta et al. (2020), easy-to-learn samples are useful for model optimization and help the model converge, while the ambiguous samples are the most challenging for the model and are beneficial for learning since they push the

model to become more robust.

We utilized MixUp to interpolate in the feature space the most similar and dissimilar samples across the two data categories. Specifically, for a training sample (x_i, y_i) we leverage its saliency map S based on gradients to find the most similar sample (x_i', y_i') using Eq. 1 and the most dissimilar sample (x_i'', y_i'') using Eq. 2:

$$(x_i', y_i') = \underset{(x_j,y_j)\in\mathcal{D}_{other}}{\text{argmax}} \; CosSim(S, S^{(x_j,y_j)}) \quad (1)$$

$$(x_i'', y_i'') = \underset{(x_j,y_j)\in\mathcal{D}_{other}}{\text{argmin}} \; CosSim(S, S^{(x_j,y_j)}) \quad (2)$$

where $\mathcal{D}_{other}$ is the subset that (x_i, y_i) does not belong to.

We denote the MixUp strategy used in our proposed method by Sim-MixUp. We use Sim-MixUp to generate additional training examples that are more useful for learning since they share the characteristics of both easy-to-learn and ambiguous samples. The models are trained on the generated data together with the easy-to-learn and ambiguous samples using cross-entropy loss.

In this study we combined the MixUp method described above with calibration-based self-distillation over multiple generations by training the first teacher model with the MixUp method instead of standard supervised learning. The steps of our approach are detailed in Algorithm 2.

4 Experimental Setup

4.1 Datasets

We perform emotion detection both in-distribution (ID) and out-of-distribution (OOD). ID refers to the data distribution used to train a model, while OOD data is significantly different from the training data and is used to evaluate how well models adapt beyond their training distribution. We also mapped the emotion labels from one ID dataset to Ekman's emotions and from another ID dataset to Plutchik's emotions to assess how well our approach generalizes across different emotion schemes. We chose the mapping according to the alignment between the original emotions (in a dataset) and each scheme. Specifically, Empathetic Dialogues includes labels corresponding to the two Plutchik emotions (anticipation and trust) that are not part of Ekman's set (see Table 6 in the Appendix).

4.1.1 In-Distribution Data

(1) **GoEmotions** (Demszky et al., 2020) is a dataset consisting of 58k Reddit comments labeled with 27 emotions or neutral. The dataset is multilabeled, i.e, some samples have more than one label. In this study we first removed the neutral label because the other two ID datasets do not have that label and because we wanted to be consistent with prior work that also did not use the neutral label (Suresh and Ong, 2021). We then replaced each emotion label with one of Ekman's six basic emotions (Ekman, 1992) as shown in Table 5 from the Appendix. Finally, we removed the samples that still had more than one label, as well as duplicate instances (i.e., repeated comments with identical labels).

(2) **Empathetic Dialogues** (Rashkin et al., 2019) consists of 25k two-way conversations between a speaker and listener grounded in emotional situations. Each conversation has an associated prompt (i.e., a description of an emotional situation) which is labeled with one of 32 emotions. In this work, we only used the prompts to train and evaluate the models. The emotion labels were grouped into Plutchik's eight fundamental emotions (Plutchik, 2001) as shown in Table 6 in Appendix, then the duplicate instances were removed.

(3) **ISEAR** (short for "International Survey on Emotion Antecedents and Reactions") (Scherer and Wallbott, 1994) contains reports describing emotional events and labeled with one of seven emotions: joy, fear, anger, sadness, disgust, shame, or guilt, each emotion having approximately 1070 examples. Note that 154 reports with no descriptions (e.g., 'No response', 'Never experienced', 'Does not apply') were removed. Situations labeled with the 'guilt' emotion were also removed since that emotion is not present in the corresponding OOD dataset (Emotion-Stimulus).

4.1.2 Out-of-Distribution Data

(1) The **DailyDialog** (Li et al., 2017) is the OOD dataset corresponding to the GoEmotions dataset. It consists of 13K multi-turn dialogues about daily life. Each utterance is labeled with one of Ekman's 6 emotions or with 'no emotion'. We only used the utterances with emotions to evaluate the models.

(2) **The Stance Sentiment Emotion Corpus (SSEC)** (Schuff et al., 2017) is the OOD

dataset corresponding to the Empathetic Dialogues dataset. SSEC consists of 4,870 tweets annotated with multiple emotion labels per tweet following Plutchik's 8 emotions. The dataset has a train and test subset that have 5 different label assignments available based on the fraction t of annotators that agree on any given emotion. We used the train and test sets with $t = 0.66$, from which we removed the tweets with multiple emotions. For validation we used 5% of the train set (the rest was used for OOD finetuning).

(3) The **Emotion-Stimulus** dataset (Ghazi et al., 2015) corresponds to the ID dataset ISEAR. It contains 2413 sentences annotated with one of Ekman's 6 emotions or *shame*. We removed the sentences labeled with the emotion *surprise*, since the ISEAR dataset does not contain that emotion. We evenly split 95% of the dataset into two to obtain a validation and a test subset with which to evaluate the models.

4.2 Baseline Methods

We compare our results with the following baselines for text classification:

- **Label Aware Contrastive Loss (LCL)** (Suresh and Ong, 2021): a method that adapts contrastive learning for fine-grained emotion classification. A dual-model approach is used: one model learns the inter-label relationships that are used in the main model's contrastive objective. In the paper both models were initialized with HuggingFace' $\text{ELECTRA}_{\text{base}}$.
- **RoBERTa** (short for "Robustly Optimized BERT Approach"): a variant of the BERT model proposed by Liu et al. (2019) that has been shown to outperform BERT and other state-of-the-art models on a variety of NLP tasks.
- **Manifold MixUp (M-MixUp)** (Verma et al., 2019): an extension of the original MixUp (Zhang et al., 2018) with better regularization that interpolates training samples in the feature space. The interpolation is performed on the features obtained from the task-specific layer on top of the RoBERTa model. We report the results for random interpolation (**Rand-M-MixUp**) and the interpolation of easy and ambiguous samples using 66% of the training data (**66% EA-M-MixUp**), which is the method used by Hosseini and Caragea (2022).
- **Self-distillation (SD)** (Furlanello et al., 2018): a knowledge distillation method over multiple generations without temperature scaling in which the student has the same architecture as the teacher.
- **Calibration-based Self-Distillation (CSD)** (Li and Caragea, 2023): an SD method that dynamically updates the temperature used to scale the teacher predictions in each generation. Note that the student predictions are not scaled.
- Hosseini and Caragea (2022): their method combines one generation self-distillation (**1-G SD**) with **66% EA-M-MixUp**. We also report the result of adding temperature scaling to this method, which is equivalent with one generation CSD (**1-G CSD**) with **66% EA-M-MixUp** for a fair comparison with our approach.

We also compare our proposed method with the following instruction-tuned large language models (LLMs):

- **Llama 3** from Meta with 8b parameters;
- **Gemma 2** from Google with 9b parameters.

4.3 Evaluation Metrics

For the emotion classification task, the macro F1 score is a good metric to evaluate a model's ability to distinguish between different emotions. However, if the dataset used to train the model is imbalanced, the weighted F1 score considers the contributions of each class more appropriately. Since two of the datasets used in this study are imbalanced, we report both scores in our results. As mentioned in Subsection 3.1, we use the **Expected Calibration Error (ECE)** (Naeini et al., 2015) to evaluate the calibration of the models.

4.4 Experimetal Details

We performed knowledge distillation over 4 generations. We utilized the in-domain development set for dynamic temperature scaling to determine an optimum temperature T in the range of [0.01, 5.0] with a granularity of 0.01. For MixUp, we used 5.0 as the value for the parameter α. To obtain the easy-to-learn subset, we took the top 50% samples with the highest confidence. The rest of the samples were placed in the ambiguous subset.

To improve the OOD performance, we further trained the models on 1%, 3%, and 5% of the OOD training data, together with new samples created by mixing the OOD samples with random ambiguous ones from the ID training data. We found that

	GoEmotions		Empathetic Dialogues		Isear	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
llama3-8b(zs)	49.68(64.58)	17.03	60.69(62.53)	20.83	69.81(69.88)	14.56
gemma2-9b-it(zs)	49.95(64.85)	22.03	60.53(62.80)	25.75	74.39(74.42)	13.50
llama3-8b(fs-3prompts)	$46.87_{0.2}(61.83_{0.2})$	$23.49_{0.5}$	$58.53_{0.7}(60.89_{0.6})$	$24.59_{0.2}$	$69.74_{0.6}(69.81_{0.6})$	$17.27_{0.8}$
gemma2-9b(fs-3prompts)	$50.19_{0.5}(65.50_{0.4})$	$23.73_{0.8}$	$61.23_{0.6}(62.94_{0.5})$	$28.15_{0.2}$	$74.41_{0.4}(74.44_{0.4})$	$15.66_{0.4}$
Electra+LCL	$72.13_{0.6}(82.26_{0.2})$	$4.27_{0.3}$	$77.46_{0.2}(79.73_{0.3})$	$6.06_{0.2}$	$76.29_{0.5}(76.31_{0.5})$	$8.71_{0.5}$
RoBERTa	$72.71_{0.5}(82.20_{0.2})$	$5.65_{0.3}$	$77.56_{0.3}(79.86_{0.3})$	$5.52_{0.5}$	$75.60_{0.4}(75.62_{0.4})$	$9.12_{0.3}$
RoBERTa + TS	same as above	$1.70_{0.3}$	same as above	$1.96_{0.3}$	same as above	$4.00_{0.3}$
Rand-M-MixUp	$71.46_{0.6}(81.59_{0.2})$	$5.52_{0.4}$	$77.10_{0.5}(79.44_{0.5})$	$3.94_{0.7}$	$74.43_{0.7}(74.46_{0.7})$	$7.75_{0.8}$
Rand-M-MixUp + TS	same as above	$3.16_{0.5}$	same as above	$2.60_{0.5}$	same as above	$3.74_{0.8}$
66% EA-M-MixUp	$68.95_{0.4}(80.52_{0.3})$	$4.89_{0.3}$	$75.33_{0.9}(77.86_{0.8})$	$4.15_{1.3}$	$72.10_{0.9}(72.12_{0.9})$	$6.79_{0.6}$
66% EA-M-MixUp+TS	same as above	$4.60_{0.5}$	same as above	$2.56_{0.5}$	same as above	$3.13_{0.9}$
1G-SD + 66%EA-MixUp	$71.28_{0.5}(81.70_{0.2})$	$5.08_{0.2}$	$77.09_{0.2}(79.48_{0.2})$	$4.54_{0.3}$	$75.96_{0.3}(75.97_{0.3})$	$6.64_{0.3}$
1G-CSD + 66%EA-MixUp	$71.65_{0.3}(81.79_{0.1})$	$2.50_{0.4}$	$77.72_{0.4}(80.07_{0.3})$	$2.44_{0.6}$	$75.69_{0.5}(75.71_{0.5})$	$3.83_{0.6}$
SD	$73.12_{0.4}(82.20_{0.2})$	$5.39_{0.2}$	$\mathbf{77.91_{0.3}(80.23_{0.3})}$	$4.83_{0.3}$	$76.63_{0.4}(76.65_{0.4})$	$8.22_{0.4}$
CSD	$73.22_{0.3}(82.33_{0.2})$	$1.50_{0.2}$	$77.84_{0.3}(79.98_{0.2})$	$1.91_{0.4}$	$76.29_{0.3}(76.31_{0.3})$	$3.57_{0.7}$
Sim-MixUp + CSD (ours)	$\mathbf{74.01_{0.3}(82.68_{0.1})}$	$\mathbf{1.44_{0.1}}$	$77.90_{0.4}(80.02_{0.3})$	$\mathbf{1.81_{0.4}}$	$\mathbf{76.79_{0.3}(76.81_{0.3})}$	$\mathbf{3.13_{0.3}}$

Table 1: Results for the ID datasets. The F1 score columns show the $F_{macro}(F_{weighted})$ values in percentage.

	DailyDialog		SSEC		Emotion-Stimulus	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
RoBERTa (w/o finetuning)	$47.63_{0.7}(76.03_{0.5})$	$14.60_{0.4}$	$38.87_{0.6}(49.06._{0.5})$	$16.33_{1.7}$	$63.58_{1.8}(70.50_{1.7})$	$8.20_{0.9}$
llama3-8b(fs)	$54.71_{0.5}(77.25_{0.3})$	$10.04_{0.4}$	$41.33_{1.1}(49.44_{0.5})$	$36.36_{0.5}$	$71.59_{0.5}(75.12_{0.6})$	$11.40_{0.8}$
gemma2-9b(fs)	$\mathbf{58.86_{0.6}}(80.85_{0.8})$	$\mathbf{9.83_{1.0}}$	$43.30_{1.0}(53.03_{0.6})$	$37.80_{0.6}$	$69.72_{0.4}(75.80_{0.2})$	$12.39_{0.3}$
RoBERTa	$56.41_{2.4}(82.21_{0.8})$	$11.10_{0.8}$	$42.59_{1.4}(53.89_{1.7})$	$16.28_{2.4}$	$79.68_{3.0}(87.49_{1.1})$	$5.29_{1.8}$
Rand-M-MixUp	$56.51_{1.0}(82.17_{0.4})$	$11.08_{0.7}$	$41.23_{2.1}(53.48_{1.5})$	$13.92_{1.2}$	$80.90_{2.7}(88.20_{1.2})$	$3.66_{0.8}$
66% data EA-M-MixUp	$56.71_{1.5}(82.50_{0.6})$	$10.09_{0.9}$	$39.89_{1.9}(52.05_{1.8})$	$15.93_{2.5}$	$79.28_{3.1}(87.54_{1.2})$	$\mathbf{3.53_{0.9}}$
1G-SD + 66%EA-MixUp	$58.56_{1.3}(\mathbf{82.90_{0.6}})$	$10.28_{1.3}$	$41.60_{0.7}(54.03_{0.8})$	$14.53_{0.9}$	$78.91_{1.7}(86.99_{1.2})$	$4.65_{0.6}$
1G-CSD + 66%EA-MixUp	$58.18_{1.4}(82.65_{0.4})$	$10.30_{1.0}$	$41.58_{0.6}(53.31_{1.1})$	$14.56_{2.1}$	$80.74_{2.2}(87.89_{1.2})$	$4.12_{0.6}$
SD	$56.29_{1.5}(82.08_{0.7})$	$11.26_{1.0}$	$41.68_{0.9}(53.82_{1.8})$	$15.42_{2.7}$	$79.53_{2.4}(87.32_{0.9})$	$5.57_{0.8}$
CSD	$56.09_{1.9}(82.23_{0.6})$	$11.21_{0.8}$	$42.55_{0.8}(53.67_{1.1})$	$14.59_{2.8}$	$79.89_{2.4}(86.94_{1.3})$	$5.86_{0.9}$
Sim-MixUp + CSD (ours)	$57.23_{1.4}(82.54_{0.6})$	$10.30_{1.2}$	$\mathbf{43.64_{0.8}(54.38_{0.4})}$	$\mathbf{11.88_{2.0}}$	$\mathbf{82.76_{1.2}(88.59_{0.8})}$	$4.08_{1.0}$

Table 2: OOD results after finetuning with 3% of the OOD training data (bottom block) compared with the results of the LLMs and RoBERTa without finetuning (top block).

finetuning using 3% of the OOD data and a small batch size for 4 epochs lead to the best balance between performance and calibration. The batch size used was 16 for Empathetic Dialogues / SSEC and 8 for the other datasets.

For the LLMs we used a zero-shot prompt and 3 few-shot prompts with 2 examples per class, for which we report the mean results. The prompts used are shown in the Appendix.

5 Results

The test performance and expected calibration errors of the trained models are summarized in Tables 1 (ID results) and 2 (OOD results). We report the mean values across 5 training runs with the standard deviation shown in subscript. The best results are shown in bold. For the models trained using KD we report the values from the best generation. For the results of all generations see Tables 11 and 12 in Appendix.

Main Results: Firstly, we can see in Table 1 that in the ID setting the performance of the LLMs on the imbalanced datasets (GoEmotions and Empathetic Dialogues) is very low compared to the other models. One possible reason for this is that the emotions in the two datasets were originally more granular and therefore more complex and difficult to understand. The LLMs also have the highest ECE in the ID setting, which indicates that they are not well calibrated. On the other hand, the F1 scores of the LLMs in the OOD setting are much higher than those of the RoBERTa model without

	GoEmotions		Empathetic Dialogues		Isear	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
Rand-M-MixUp	$71.46_{0.6}(81.59_{0.2})$	$5.52_{0.4}$	$77.10_{0.5}(79.44_{0.5})$	$3.94_{0.7}$	$74.43_{0.7}(74.46_{0.7})$	$7.75_{0.8}$
EA-M-MixUp	$71.37_{0.5}(81.54_{0.2})$	$5.43_{0.3}$	$77.24_{0.4}(79.57_{0.3})$	$4.25_{0.4}$	$74.32_{0.6}(74.34_{0.6})$	$8.19_{0.6}$
Sim-M-MixUp (ours)	$71.59_{0.6}(81.70_{0.3})$	$6.13_{0.3}$	$77.11_{0.4}(79.42_{0.5})$	$3.65_{0.7}$	$74.60_{0.7}(74.62_{0.7})$	$7.87_{0.8}$
CSD	$73.22_{0.3}(82.33_{0.2})$	$1.50_{0.2}$	$77.84_{0.3}(79.98_{0.2})$	$1.91_{0.4}$	$76.29_{0.3}(76.31_{0.3})$	$3.57_{0.7}$
Rand-M-MixUp + CSD	$73.23_{0.2}(82.24_{0.1})$	$1.56_{0.2}$	$\mathbf{78.18_{0.5}(80.38_{0.4})}$	$1.90_{0.5}$	$76.51_{0.4}(76.52_{0.4})$	$3.13_{0.4}$
EA-M-MixUp + CSD	$73.32_{0.3}(82.24_{0.1})$	$1.51_{0.2}$	$77.79_{0.2}(80.08_{0.1})$	$2.05_{0.3}$	$76.74_{0.2}(76.76_{0.2})$	$3.29_{0.7}$
Sim-MixUp + CSD (ours)	$\mathbf{74.01_{0.3}(82.68_{0.1})}$	$\mathbf{1.44_{0.1}}$	$77.90_{0.4}(80.02_{0.3})$	$\mathbf{1.81_{0.4}}$	$\mathbf{76.79_{0.3}(76.81_{0.3})}$	$\mathbf{3.13_{0.3}}$

Table 3: ID Ablation Study

	DailyDialog		SSEC		Emotion-Stimulus	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
Rand-M-MixUp	$56.51_{1.0}(82.17_{0.4})$	$11.08_{0.7}$	$41.23_{2.1}(53.48_{1.5})$	$13.92_{1.2}$	$80.90_{2.7}(88.20_{1.2})$	$3.66_{0.8}$
EA-M-MixUp	$57.04_{1.4}(82.47_{0.4})$	$10.48_{1.0}$	$42.78_{1.7}(54.42_{1.4})$	$13.07_{2.4}$	$79.19_{3.2}(87.59_{1.5})$	$4.07_{0.9}$
Sim-MixUp	$56.57_{1.6}(82.48_{0.5})$	$10.58_{1.2}$	$41.59_{1.6}(53.78_{1.6})$	$13.76_{2.5}$	$80.85_{1.6}(88.29_{0.7})$	$\mathbf{3.59_{0.7}}$
CSD	$56.09_{1.9}(82.23_{0.6})$	$11.21_{0.8}$	$42.55_{0.8}(53.67_{1.1})$	$14.59_{2.8}$	$79.89_{2.4}(86.94_{1.3})$	$5.86_{0.9}$
Rand-M-MixUp + CSD	$\mathbf{57.72_{1.4}}(82.48_{0.7})$	$10.79_{1.4}$	$41.81_{0.9}(53.37_{1.0})$	$15.78_{1.8}$	$81.26_{3.5}(88.22_{2.0})$	$4.22_{1.2}$
EA-M-MixUp + CSD	$56.91_{1.3}(82.35_{0.5})$	$10.63_{1.1}$	$41.43_{0.8}(54.35_{1.2})$	$14.09_{2.0}$	$82.73_{1.5}(88.57_{1.0})$	$3.74_{0.7}$
Sim-MixUp + CSD (ours)	$57.23_{1.4}(\mathbf{82.54_{0.6}})$	$\mathbf{10.30_{1.2}}$	$\mathbf{43.64_{0.8}(54.38_{0.4})}$	$\mathbf{11.88_{2.0}}$	$\mathbf{82.76_{1.2}(88.59_{0.8})}$	$4.08_{1.0}$

Table 4: Finetuned OOD Ablation Study

finetuning (first line in Table 2).

Secondly, we can observe in Table 1 that, in the ID setting, self-distillation over multiple generations (SD) improves the performance of RoBERTa the most out of all baselines while simultaneously decreasing the ECE. Adding dynamic temperature scaling leads to better calibration such that CSD has the lowest ID ECE among the baselines. On the other hand, the overall performance and calibration of MixUp methods without self-distillation are suboptimal. This suggests that, while MixUp can produce useful augmentations, it does not lead to better calibration or performance by itself.

We can see that our proposed method that combines Sim-MixUp with CSD brings further improvement to the calibration in the ID setting, as well as similar or better performance to the baselines. In the OOD setting, after finetuning the models on 3% of the OOD training data, the performance of our model comes close to or, in the case of the SSEC and Emotion-Stimulus datasets, exceeds the performance of the LLMs, without a significant increase in the calibration error.

Ablation Study: Our approach shows the best ID macro/weighted F1 scores on two out of three datasets (GoEmotions and ISEAR) and the best ECE across all three datasets. In the OOD setting, our model achieves the best ECE on two datasets (DailyDialog and SSEC) and is only slightly worse than Sim-MixUp on Emotion-Stimulus. It also obtains the best OOD weighted F1 score on all datasets and the best OOD macro F1 score on SSEC and Emotion-Stimulus. Thus, combining CSD with Sim-MixUp provides clear gains over both random and EA-MixUp variants, in both ID and OOD settings, confirming the benefit of mixing the most similar and dissimilar examples. This suggests that MixUp guided by training dynamics and saliency yields more informative synthetic examples than arbitrary or partially informed pairings.

6 Error Analysis

In this section we analyze the emotions that are commonly misclassified in all 5 training runs. Note that all mean confidences reported have the standard deviation in subscript.

6.1 GoEmotions

As we can observe in Figure 1, the model trained on the GoEmotions dataset is skewed towards the most prevalent emotions in the training data (see Table 13), specifically joy, anger and surprise. The most notable error is the tendency to mistake surprise for joy, likely because the model associates comments that express positive surprise with joy.

Another notable finding is the strong and reciprocal confusion between anger and joy, two very different emotions. This suggests that the model struggles to distinguish between these two high-arousal emotions. One explanation might be that

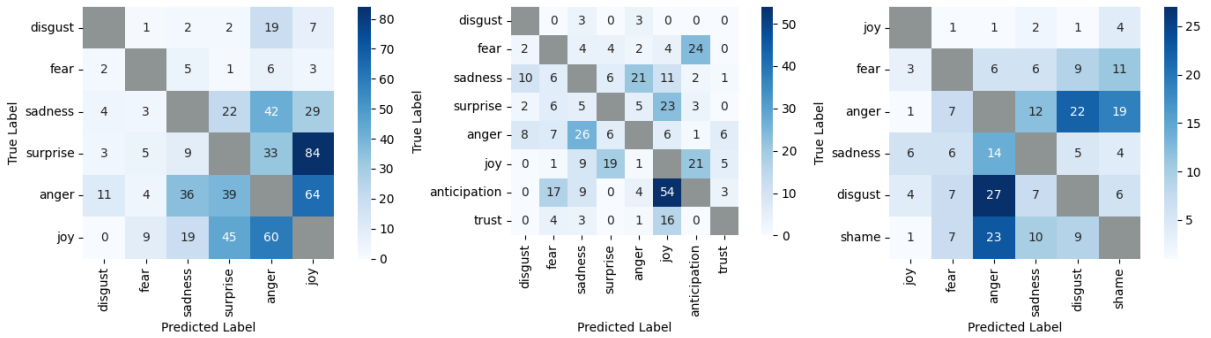

Figure 1: Confusion matrix (errors) for the test set of GoEmotions (left), Empathetic Dialogues (middle) and ISEAR (right).

sarcasm or satire in the Reddit comments confuse the model, when the words used suggest joy, but the underlying emotion is anger and vice versa.

Several examples of misclassified test samples that illustrate these trends can be seen in Table 10 from the Appendix.

6.2 Empathetic Dialogues

As we can see in Figure 1, the model trained on Empathetic Dialogues most commonly mistakes anticipation for joy. One possible explanation is that the emotions mapped to anticipation (like hopeful and prepared) and joy (like excited and confident) are very closely related. For example, the following statements in the test set about a future event are inherently positive and can be easily confused with an expression of joy:

- I have a certification exam coming up and I think I'll do well! (**mean confidence:** $\mathbf{81.36_{1.3}}$)
- I just got an email from property management confirming that I've got the new apartment I wanted. I can't wait to move in! (**mean confidence:** $\mathbf{80.37_{2.8}}$)

Another likely source of errors is that many complex emotions from the original dataset that could be a blend of feelings, such as nostalgic (sadness and joy) and anxious (anticipation and fear), were mapped to a single primary emotion (see Table 6). This overlap creates inherent ambiguity in the data. Notably, the label 'disgust' is the only one which does not incorporate multiple emotions and has the fewest errors.

6.3 ISEAR

The most prominent errors in ISEAR's confusion matrix in Figure 1 are the bidirectional anger-disgust and anger-shame confusions. These pairs of emotions share a high intensity and negative sentiment. The model likely recognizes the similarities between anger and disgust/shame but struggles to differentiate between them. Anger and disgust are especially similar, as seen in the examples below:

- For "Some people were unfairly treated, because of their nationality/color." our model predicted anger instead of disgust with $69.20_{2.5}$ confidence.
- For "When a man, a stranger to me, personally insulted a close woman friend of mine in public." our model predicted disgust instead of anger with $66.34_{0.8}$ confidence.

On the other hand, 'joy' has the fewest instances of being misclassified, likely because it is the only purely positive emotion in the labels set.

7 Conclusion

In this work, we proposed a novel calibration method for pre-trained language models that combines MixUp with knowledge distillation in order to calibrate teacher models for the emotion detection task, which in turn helps train better students. Our method calibrates the first teacher using an informed MixUp method that interpolates easy-to-learn with ambiguous samples guided by saliency signals. This way we provide the teacher models with more useful information that is then imparted to the student models during the distillation of the emotion-detection models. During the distillation process, we calibrate the teacher predictions in each generation by dynamically updating the temperature used for scaling. We empirically validated that our method achieves competitive performance and calibrates the pre-trained model RoBERTa on various emotion-detection datasets, both in-domain and out-of-domain.

Acknowledgement

This research is supported in part by the NSF IIS award 2107487. We thank our anonymous reviewers for their constructive feedback, which helped improve the quality of our paper.

Limitations

One limitation of our method is that it runs for multiple generations, therefore it requires longer training time and extra memory to store the teacher model. However, this is a common limitation for all methods that use knowledge distillation over multiple generations, not just specific to our approach.

References

Akshay Agarwal, Shashank Maiya, and Sonu Aggarwal. 2021. Evaluating empathetic chatbots in customer service settings. *Preprint*, arXiv:2101.01334.

Antonin Brun, Ruying Liu, Aryan Shukla, Frances Watson, and Jonathan Gratch. 2025. Exploring emotion-sensitive llm-based conversational ai. *Preprint*, arXiv:2502.08920.

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. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan 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*, pages 4040–4054. 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)*, page 295–302.

Paul Ekman. 1992. Are there basic emotions? *Psychological review*, 99:550–553.

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. PMLR.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing*, page 152–165. Springer International Publishing.

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. PMLR.

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

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. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Hua Cheng, Lucas Mentch, and Thomas Schaaf. 2020. Posterior calibrated training on sentence classification tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 2723–2730. Association for Computational Linguistics.

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)*, pages 1326—-1340. Association for Computational Linguistics.

Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2022. A review on text-based emotion detection – techniques, applications, datasets, and future directions. *Preprint*, arXiv:2205.03235.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995. Asian Federation of Natural Language Processing.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6316–6329. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

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

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.

Tanya Nijhawan, Girija Attigeri, and Ananthakrishna Thalengala. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9(1):1–24.

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. 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. Association for Computational Linguistics.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *Preprint*, arXiv:1701.06548.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in nlp: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 5696–5710. ELRA and ICCL.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, page 13–23. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Preprint*, arXiv:1312.6034.

Tiberiu Sosea and Cornelia Caragea. 2020. Cancer-emo: A dataset for fine-grained emotion detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 8892–8904. Association for Computational Linguistics.

Varsha Suresh and Desmond Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 4381–4394. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275—-9293. Association for Computational Linguistics.

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*, volume 32.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6438—-6447. PMLR.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, page 1–13. Association for Computational Linguistics.

Hongyi Zhang, Moustapha Cisse, Yann N, Dauphin, , and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*.

A Appendix

A.1 Emotion Mapping and Distribution

Tables 5 and 6 show the original emotion labels in the GoEmotions and Empathetic Dialogues datasets and the corresponding Ekman and Plutchik emotions. Tables 13, 14, 15, 16, 17, 18 show the distribution of the emotions in all six datasets used in this study.

GoEmotions Label	Ekman Label
disgust	disgust
fear, nervousness	fear
anger, annoyance, disapproval	anger
surprise, realization, confusion, curiosity	surprise
sadness, disappointment, embarrassment, grief, remorse	sadness
joy, relief, gratitude, approval, admiration, pride, love, desire, caring, optimism, amusement, excitement	joy

Table 5: Ekman emotions mapping for the GoEmotions dataset

Empathetic Dialogues Label	Plutchik Label
disgusted	disgust
trusting, faithful	trust
surprised, impressed	surprise
afraid, terrified, apprehensive	fear
angry, annoyed, furious, jealous	anger
anticipating, hopeful, prepared, anxious	anticipation
joyful, content, proud, excited, grateful, confident, caring	joy
sad, devastated, lonely, disappointed, guilty, ashamed, embarrassed, nostalgic, sentimental	sadness

Table 6: Plutchik emotions mapping for the Empathetic Dialogues dataset

A.2 Implementation Details

For the experiments, we fine-tuned *roberta-base* from the HuggingFace Transformers library with 125M parameters, All the models were optimized with AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.001 and gradient clip of 3.0. The models were trained for a maximum of 3 epochs. We report the mean performance over 5 random seeds. The batch size and learning rate used to fine-tune RoBERTa for each dataset are shown in Table 7. The total time for training the models with our proposed approach was under 12 hours on a NVIDIA RTX A5000 24G GPU.

Dataset	Batch Size	Learning Rate
GoEmotions	64	2.5e-5
EmpDialogues	64	5e-5
ISEAR	32	2.5e-5

Table 7: Hyper-parameters for the RoBERTa models

A.3 Scaling vs not scaling the student

In the ID setting, as we can see in Table 8, not scaling the student during training results in better or equivalent macro and weighted F1 scores for the models that were trained using knowledge distillation combined with temperature scaling. Our model in particular achieves better ID performance without scaling the student on all tasks. In contrast, scaling the student consistently worsens the ID calibration of our model across all three datasets. For the "CSD" model, scaling the student improves ID calibration on Empathetic Dialogues and ISEAR, providing a better ID calibration than our model only for the ISEAR dataset.

In the OOD setting, scaling the student model is harmful to the performance and calibration of our model, as the F1 scores drop and the ECE increases across all three OOD datasets when the student is scaled during training. In the case of the 'CSD' model, the performance and calibration only improve on the DailyDialog dataset and only the weighted F1 score surpasses that of our model.

A.4 GoEmotions Error Examples

Table 10 provides concrete examples of the most common errors in the GoEmotions test set discussed in the Error Analysis section. The first three examples show that in cases of positive surprise, the model confidently defaults to "Joy", mistaking the tone of the expression for the core emotion itself. Similarly, in the cases of sarcastic comments, the model confidently misinterprets the underlying anger as joy. This indicates that strong positive keywords can mislead the model to a point where it is highly certain of its incorrect assessment, a behavior likely caused by the overwhelming amount of training samples for "Joy" compared to other emotions (see Table 13).

The final two examples show the model misclassifying "Joy" as "Anger" but with significantly

	GoEmotions		Empathetic Dialogues		Isear	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
CSD	$73.22_{0.3}(82.33_{0.2})$	$1.50_{0.2}$	$77.84_{0.3}(79.98_{0.2})$	$1.91_{0.4}$	$76.29_{0.3}(76.31_{0.3})$	$3.57_{0.7}$
CSD + scaled student	$73.19_{0.5}(82.24_{0.3})$	$1.68_{0.3}$	$77.83_{0.3}(80.17_{0.3})$	$1.84_{0.3}$	$76.29_{0.6}(76.31_{0.6})$	$\mathbf{2.98_{0.3}}$
Sim-MixUp + CSD (ours)	$\mathbf{74.01_{0.3}(82.68_{0.1})}$	$\mathbf{1.44_{0.1}}$	$\mathbf{77.90_{0.4}(80.02_{0.3})}$	$\mathbf{1.81_{0.4}}$	$\mathbf{76.79_{0.3}(76.81_{0.3})}$	$3.13_{0.3}$
Ours + scaled student	$73.78_{0.3}(82.51_{0.2})$	$1.61_{0.3}$	$77.79_{0.3}(79.93_{0.2})$	$1.86_{0.4}$	$76.52_{0.1}(76.54_{0.1})$	$3.27_{0.3}$

Table 8: ID Results for the methods that use KD with the student not scaled and scaled. The best values for each model are shown in bold. The F1 score columns show the $F_{macro}(F_{weighted})$ values in percentage.

	DailyDialog		SSEC		Emotion-Stimulus	
Model	F1 score	ECE	F1 score	ECE	F1 score	ECE
CSD	$56.09_{1.9}(82.23_{0.6})$	$11.21_{0.8}$	$42.55_{0.8}(53.67_{1.1})$	$14.59_{2.8}$	$79.89_{2.4}(86.94_{1.3})$	$5.86_{0.9}$
CSD + scaled student	$57.08_{2.0}(\mathbf{82.61_{0.7}})$	$10.87_{0.9}$	$42.04_{1.1}(53.56_{1.4})$	$15.61_{1.3}$	$79.91_{2.2}(86.90_{1.5})$	$6.02_{0.8}$
Sim-MixUp + CSD (ours)	$\mathbf{57.23_{1.4}}(82.54_{0.6})$	$\mathbf{10.30_{1.2}}$	$\mathbf{43.64_{0.8}(54.38_{0.4})}$	$\mathbf{11.88_{2.0}}$	$\mathbf{82.76_{1.2}(88.59_{0.8})}$	$4.08_{1.0}$
Ours + scaled student	$57.13_{0.9}(82.53_{0.3})$	$10.41_{1.0}$	$42.26_{1.2}(53.62_{1.4})$	$13.81_{2.6}$	$80.06_{2.5}(87.13_{1.6})$	$4.89_{1.2}$

Table 9: OOD Results after 3% finetuning for the methods that use KD with the student not scaled and scaled. The best values for each model are shown in bold.

Comment	True Emotion	Prediction	Mean Confidence
Omg didn't even think of that! So clever! Yes!	Surprise	Joy	$79.02_{1.4}$
Where can I get some? These are awesome!	Surprise	Joy	$94.16_{0.2}$
Only for a week? You are a great optimist!	Surprise	Joy	$95.56_{0.3}$
Yeah, it's pretty funny how incapable most of them are of actual satire and subtlety.	Anger	Joy	$96.45_{0.4}$
Nah man I prefer endless cause of the part where he goes WOMWOMWOMWOMWOMWOM WOOOOOO REEEEEEEEE like that's lyrical genius right there	Anger	Joy	$84.55_{0.8}$
Nothing against bartenders at all. Dumbass people like [NAME], I make fun of because I can.	Joy	Anger	$57.09_{2.7}$
This isn't even his highlight of his week. It's the highlight of the last few decades. Damn onions.	Joy	Anger	$55.51_{1.2}$

Table 10: Examples of the most common test errors in the GoEmotions dataset

lower confidence. In these cases, the model likely associates strong negative words ("Dumbass," "Damn") with "Anger", but the surrounding joyful context provides conflicting signals, reducing the model's overall confidence. This suggests that the model is less certain when the overall context does not align with powerful keywords.

A.5 LLM Prompts

The Zero-shot prompt we used to tell a LLM to classify a text and return a confidence score is the following:

You will receive a short text. Your task is to classify the emotion expressed in it into one of the following categories: {list_of_emotions}.

You must also provide a confidence score between 0 and 1 to indicate how sure you are in the categorization. Answer only with a JSON string that has the following format: {'prediction': emotion, 'confidence': float}

The text to classify is '{short_text}'.

The structure of the few-shot prompts we used to tell a LLM to classify a text and return a confidence score is the following:

Given the following text: (Instruction)

{input_text} (Input Text)

Classify the text into one of the following categories depending on the emotion expressed by the text: (Instruction)

1. emotion1; 2. emotion2; ...

Two example classifications for each emotion are shown below. (Instruction)

text1 (Few-shot example #1 text)

Emotion: emotion1 (Few-shot example #1 label)

text2 (Few-shot example #2 text)

Emotion: emotion2 (Few-shot example #2

label)

You must provide a confidence score between 0 and 1 to indicate how sure you are in the categorization. (Instruction)

Your response must be a JSON string that has the following format: {'prediction': 'emotion', 'confidence': float} (Instruction)"

A.6 Results per SD generation

Tables 11 and 12 show the macro F1 scores and calibration errors across 4 generations of self-distillation (SD) for the three emotion classification datasets in each setting, in-domain (ID) and out-of-domain (OOD). We can observe that in the ID setting, student models trained in later generations consistently outperform their predecessors and are better calibrated. Notably, in the ID setting most student models achieve the best macro F1 score and ECE in the third or fourth generation, demonstrating the benefits of multi-generation distillation. Moreover, our proposed approach that combines Sim-MixUp with CSD attains the best ID performance and calibration across all datasets. In the OOD setting, we can observe that student models generally also perform better in later generations and our model performs best on two out of three datasets (SSEC and Emotion-Stimulus).

	Gen 1		Gen 2		Gen 3		Gen 4	
	Macro F1	ECE	Macro F1	ECE	Macro F1	ECE	Macro F1	ECE
GoEmotions								
RoBERTa+SD	$72.71_{0.5}$	$5.65_{0.3}$	$\mathbf{73.12_{0.4}}$	$\mathbf{5.39_{0.2}}$	$73.06_{0.4}$	$5.49_{0.3}$	$72.80_{0.3}$	$5.54_{0.3}$
RoBERTa+CSD	$72.71_{0.5}$	$1.70_{0.3}$	$73.01_{0.6}$	$1.62_{0.4}$	$\mathbf{73.22_{0.3}}$	$\mathbf{1.50_{0.2}}$	$73.13_{0.4}$	$1.71_{0.4}$
RoBERTa+Rand-MixUp+CSD	$71.46_{0.6}$	$3.16_{0.5}$	$72.92_{0.4}$	$2.46_{0.2}$	$\mathbf{73.23_{0.2}}$	$\mathbf{1.56_{0.2}}$	$72.76_{0.2}$	$2.57_{0.2}$
RoBERTa+EA-MixUp+CSD	$71.37_{0.5}$	$3.10_{0.4}$	$72.48_{0.4}$	$2.80_{0.3}$	$72.98_{0.3}$	$1.75_{0.3}$	$\mathbf{73.32_{0.3}}$	$\mathbf{1.51_{0.2}}$
RoBERTa+Sim-MixUp+CSD	$71.59_{0.6}$	$2.96_{0.4}$	$72.62_{0.4}$	$2.11_{0.2}$	$73.17_{0.4}$	$1.81_{0.2}$	$\mathbf{74.01_{0.3}}$	$\mathbf{1.44_{0.1}}$
Empathetic Dialogues								
RoBERTa+SD	$77.56_{0.3}$	$5.52_{0.5}$	$77.46_{0.2}$	$5.12_{0.3}$	$77.88_{0.3}$	$4.86_{0.2}$	$\mathbf{77.91_{0.3}}$	$\mathbf{4.83_{0.3}}$
RoBERTa+CSD	$77.56_{0.3}$	$1.96_{0.3}$	$77.74_{0.2}$	$1.93_{0.5}$	$\mathbf{77.84_{0.3}}$	$\mathbf{1.91_{0.4}}$	$77.74_{0.4}$	$2.14_{0.4}$
RoBERTa+Rand-MixUp+CSD	$77.10_{0.5}$	$2.60_{0.5}$	$77.59_{0.3}$	$2.17_{0.3}$	$\mathbf{78.18_{0.5}}$	$\mathbf{1.90_{0.5}}$	$78.04_{0.3}$	$1.94_{0.3}$
RoBERTa+EA-MixUp+CSD	$77.24_{0.4}$	$2.63_{0.6}$	$77.63_{0.2}$	$2.35_{0.2}$	$77.66_{0.3}$	$2.18_{0.3}$	$\mathbf{77.79_{0.2}}$	$\mathbf{2.05_{0.3}}$
RoBERTa+Sim-MixUp+CSD	$77.11_{0.4}$	$2.12_{0.5}$	$77.85_{0.4}$	$1.89_{0.5}$	$\mathbf{77.90_{0.3}}$	$\mathbf{1.81_{0.4}}$	$77.48_{0.4}$	$2.00_{0.2}$
ISEAR								
RoBERTa+SD	$75.60_{0.4}$	$9.12_{0.3}$	$76.26_{0.3}$	$9.09_{0.8}$	$76.32_{0.7}$	$8.66_{0.6}$	$\mathbf{76.63_{0.4}}$	$\mathbf{8.22_{0.4}}$
RoBERTa+CSD	$75.60_{0.4}$	$4.00_{0.3}$	$75.76_{0.4}$	$4.15_{0.6}$	$\mathbf{76.29_{0.3}}$	$\mathbf{3.57_{0.7}}$	$75.93_{0.4}$	$4.03_{0.4}$
RoBERTa+Rand-MixUp+CSD	$74.43_{0.7}$	$3.74_{0.8}$	$75.21_{0.4}$	$4.16_{0.3}$	$\mathbf{76.51_{0.4}}$	$\mathbf{3.13_{0.4}}$	$76.48_{0.2}$	$3.21_{0.2}$
RoBERTa+EA-MixUp+CSD	$74.32_{0.6}$	$4.52_{0.7}$	$75.34_{0.3}$	$3.96_{0.3}$	$75.82_{0.3}$	$3.83_{0.4}$	$\mathbf{76.74_{0.2}}$	$\mathbf{3.29_{0.7}}$
RoBERTa+Sim-MixUp+CSD	$74.60_{0.7}$	$3.77_{0.5}$	$76.34_{0.3}$	$4.17_{0.6}$	$76.35_{0.4}$	$3.86_{1.0}$	$\mathbf{76.79_{0.3}}$	$\mathbf{3.13_{0.3}}$

Table 11: ID Results for each generation. The best values for each model are shown in bold.

	Gen 1		Gen 2		Gen 3		Gen 4	
	Macro F1	ECE	Macro F1	ECE	Macro F1	ECE	Macro F1	ECE
DailyDialog								
RoBERTa+SD	$56.41_{2.4}$	$11.10_{0.8}$	$56.29_{1.5}$	$11.26_{1.0}$	$\mathbf{57.14_{1.8}}$	$\mathbf{10.72_{0.8}}$	$56.80_{2.5}$	$11.28_{1.0}$
RoBERTa+CSD	$56.41_{2.4}$	$11.10_{0.8}$	$\mathbf{57.43_{1.4}}$	$\mathbf{10.66_{0.9}}$	$56.09_{1.9}$	$11.21_{0.8}$	$56.94_{2.5}$	$11.39_{1.2}$
RoBERTa+Rand-MixUp+CSD	$56.51_{1.0}$	$11.08_{0.7}$	$56.88_{1.0}$	$11.14_{1.2}$	$\mathbf{57.72_{1.4}}$	$10.79_{1.4}$	$57.25_{1.8}$	$\mathbf{9.96_{1.4}}$
RoBERTa+EA-MixUp+CSD	$\mathbf{57.04_{1.4}}$	$\mathbf{10.48_{1.0}}$	$56.78_{1.4}$	$10.69_{1.0}$	$56.48_{1.4}$	$10.50_{1.0}$	$56.91_{1.3}$	$10.63_{1.1}$
RoBERTa+Sim-MixUp+CSD	$56.57_{1.6}$	$10.58_{1.2}$	$56.89_{1.7}$	$10.50_{1.2}$	$\mathbf{57.38_{1.1}}$	$10.37_{1.3}$	$57.23_{1.4}$	$\mathbf{10.30_{1.2}}$
SSEC								
RoBERTa+SD	$\mathbf{42.59_{1.4}}$	$16.28_{2.4}$	$41.72_{1.4}$	$15.44_{2.3}$	$42.03_{1.8}$	$\mathbf{14.37_{3.0}}$	$41.68_{0.9}$	$15.42_{2.7}$
RoBERTa+CSD	$42.59_{1.4}$	$16.28_{2.4}$	$41.94_{1.2}$	$\mathbf{13.34_{2.3}}$	$42.55_{0.8}$	$14.59_{2.8}$	$\mathbf{43.18_{2.1}}$	$14.06_{2.9}$
RoBERTa+Rand-MixUp+CSD	$41.23_{2.1}$	$13.92_{1.2}$	$\mathbf{42.41_{1.1}}$	$\mathbf{12.95_{1.7}}$	$41.81_{0.9}$	$14.78_{1.8}$	$40.73_{2.1}$	$15.23_{0.9}$
RoBERTa+EA-MixUp+CSD	$\mathbf{42.78_{1.7}}$	$\mathbf{13.07_{2.4}}$	$42.01_{1.6}$	$14.51_{2.6}$	$42.37_{1.4}$	$13.60_{3.7}$	$41.43_{0.8}$	$14.09_{2.0}$
RoBERTa+Sim-MixUp+CSD	$41.59_{1.6}$	$13.76_{2.5}$	$42.51_{1.2}$	$12.59_{1.0}$	$\mathbf{43.64_{0.8}}$	$\mathbf{11.88_{2.0}}$	$42.08_{0.5}$	$11.92_{1.8}$
Emotion-Stimulus								
RoBERTa+SD	$\mathbf{79.68_{3.0}}$	$\mathbf{5.29_{1.8}}$	$78.23_{1.7}$	$5.42_{0.6}$	$78.71_{1.9}$	$5.48_{0.7}$	$79.53_{2.4}$	$5.57_{0.8}$
RoBERTa+CSD	$79.68_{3.0}$	$5.29_{1.8}$	$79.19_{3.4}$	$5.41_{1.0}$	$79.89_{2.4}$	$5.86_{0.9}$	$\mathbf{81.12_{2.3}}$	$\mathbf{5.14_{1.4}}$
RoBERTa+Rand-MixUp+CSD	$80.90_{2.7}$	$\mathbf{3.66_{0.8}}$	$81.39_{2.8}$	$4.45_{1.4}$	$81.26_{3.5}$	$4.22_{1.2}$	$\mathbf{82.57_{1.7}}$	$4.14_{0.6}$
RoBERTa+EA-MixUp+CSD	$79.19_{3.2}$	$4.07_{0.9}$	$82.21_{2.2}$	$\mathbf{3.42_{0.9}}$	$80.75_{1.3}$	$4.53_{1.0}$	$\mathbf{82.73_{1.5}}$	$3.74_{0.7}$
RoBERTa+Sim-MixUp+CSD	$80.85_{1.6}$	$\mathbf{3.59_{0.7}}$	$78.93_{2.3}$	$4.71_{1.1}$	$81.02_{2.9}$	$4.60_{1.0}$	$\mathbf{82.76_{1.2}}$	$4.08_{1.0}$

Table 12: OOD Results for each generation after 3% finetuning. The best values for each model are shown in bold.

Set	Disgust	Fear	Sadness	Surprise	Anger	Joy	Total
training	523	565	2448	4152	4584	15691	27963
validation	62	77	278	497	596	2014	3524
test	77	85	298	533	609	1931	3533

Table 13: GoEmotions dataset emotion distribution

Emotion	Training	Validation	Test
Disgust	612	81	85
Trust	867	123	114
Surprise	1616	249	207
Fear	1691	243	214
Anticipation	2408	330	309
Anger	2510	367	324
Joy	4318	596	596
Sadness	5199	764	693
Total	19221	2753	2542

Table 14: Empathetic Dialogues dataset emotion distribution

Set	Shame	Sadness	Disgust	Fear	Joy	Anger	Total
training	626	631	638	645	646	651	3837
validation	208	210	213	215	216	217	1279
test	209	211	213	215	215	216	1279

Table 15: ISEAR emotion distribution

Set	Disgust	Fear	Sadness	Surprise	Anger	Joy	Total
validation	3	11	79	102	74	604	873
test	47	17	102	113	114	980	1373

Table 16: DailyDialog dataset emotion distribution

Emotion	Test	Validation
Surprise	30	49
Fear	33	64
Disgust	39	34
Sadness	49	94
trust	74	127
Joy	122	285
Anticipation	158	159
Anger	286	462
Total	791	1274

Table 17: SSEC emotion distribution

Set	Disgust	Shame	Fear	Joy	Anger	Sadness	Total
validation	45	69	201	227	229	273	1044
test	45	69	201	228	229	273	1045

Table 18: Emotion-Stimulus dataset emotion distribution