Transferable Post-hoc Calibration on Pretrained Transformers in Noisy Text Classification

Jun Zhang[1,2*], **Wen Yao**[2*], **Xiaoqian Chen**[2], **Ling Feng**[1†],

[1] Tsinghua University

[2] National Innovation Institute of Defense Technology, Chinese Academy of Military Science

jun-zhan19@mails.tsinghua.edu.cn, wendy0782@126.com,

chenxiaoqian@nudt.edu.cn, fengling@tsinghua.edu.cn

Abstract

Recent work has demonstrated that pretrained transformers are overconfident in text classification tasks, which can be calibrated by the famous post-hoc calibration method temperature scaling (TS). Character or word spelling mistakes are frequently encountered in real applications and greatly threaten transformer model safety. Research on calibration under noisy settings is rare, and we focus on this direction. Based on a toy experiment, we discover that TS performs poorly when the datasets are perturbed by slight noise, such as swapping the characters, which results in distribution shift. We further utilize two metrics, predictive uncertainty and maximum mean discrepancy (MMD), to measure the distribution shift between clean and noisy datasets, based on which we propose a simple yet effective transferable TS method for calibrating models dynamically. To evaluate the performance of the proposed methods under noisy settings, we construct a benchmark consisting of four noise types and five shift intensities based on the QNLI, AG-News, and Emotion tasks. Experimental results on the noisy benchmark show that (1) the metrics are effective in measuring distribution shift and (2) transferable TS can significantly decrease the expected calibration error (ECE) compared with the competitive baseline ensemble TS by approximately 46.09%.

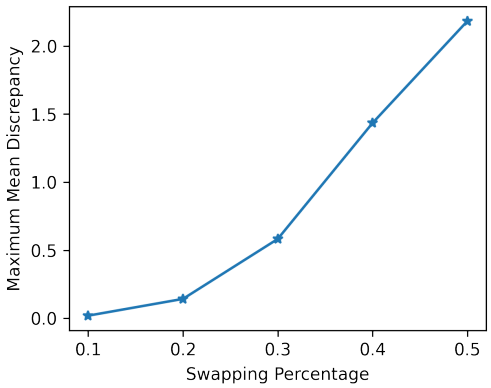


Figure 1: The distributional shift (measured by the maximum mean discrepancy between clean and noisy datasets) results in the poor calibration performance of temperature scaling (TS) under noisy settings, which increases with the noise proportion. The noisy texts are generated by swapping the adjacent characters based on the clean Emotion dataset.

Introduction

Recently, pretrained transformer-based models have been used widely in natural language processing (NLP), such as machine translation (Vaswani et al. 2017), question answering (Pearce et al. 2021), dialog generation (Oluwatobi and Mueller 2020), and text classification (Li et al. 2021). Recent studies (Guo et al. 2017) demonstrate that transformer-based models are seriously overconfident in their predictions, which leads to severe consequences, especially in safe-critical applications, such as suicidal ideation detection (Cao, Zhang, and Feng 2020). Calibrating a classifier means matching the probability of correct predictions at all levels of confidence (Guo et al. 2017). Numerous calibration methods have been proposed and can be organized into three aspects to deal with overconfidence issues.

The first category is multimodel methods, which alleviate overconfidence issues by averaging multiple model predictions. Bayesian neural networks (BNNs) (Welling and Teh 2011; Blundell et al. 2015) calibrate the prediction probability by sampling multiple models from the approximated posterior distribution of weights. MC dropout (MCD) (Gal and Ghahramani 2016) can be viewed as a special BNN case, which achieves multiple predictions by maintaining an activated dropout layer in the inference phase. Deep ensemble (DE) (Lakshminarayanan, Pritzel, and Blundell 2017) trains multiple individual models by different random initialization schemes. BNNs are generally difficult to train because the prior distribution of weights and initialization schemes are difficult to determine. DE is not practical for transformer-based models due to the considerable memory costs.

The second category regularizes the loss function with a calibration-related term. Some recent works incorporate differential calibration regularizers, such as MMCE (Kumar, Sarawagi, and Jain 2018), AvUC (Krishnan and Tickoo 2020) and SB-ECE (Karandikar et al. 2021), in the train-

*These authors contributed equally.

†Corresponding author.

	Selected Example	Accuracy(↑)	ECE(↓)	ECE-TS (↓)
Clean Dataset	i am feeling more confident that we will be able to take care of this baby	93.13	0.020	0.019
Noisy Dataset (10%)	i am **beeling** more confident that we will be able to take care of this baby	83.53	0.099	0.096
Noisy Dataset (20%)	i am **beeling** more confdient that we will be able to **tkae caFe** of this baby	67.78	0.220	0.215
Noisy Dataset (30%)	i am **beeling** more **confdiet** that we will be **Cble** to **tkea caFe** of this **bby**	51.60	0.291	0.284
Noisy Dataset (40%)	i am **beeKing** more **conDdiet** that we will be **Cbxle** to **tkae caF** of this **pbby**	37.25	0.321	0.313
Noisy Dataset (50%)	i am **beeKipng** more **clonDdiet** that we will be **Cbxel** to **tkea cuaF** of this **pYbby**	31.15	0.301	0.292

Table 1: Performance of BERT on Emotion dataset under different noise proportions implemented by the CharSwap method in TextAttack.

ing phase. This category of approaches needs to retrain the models, which consumes considerable time costs.

The third category, post-hoc methods, is more applicable for large-scale pretrained models. Temperature scaling (TS) (Guo et al. 2017) is the most widespread in this category due to its simplicity and accuracy preservation. TS calibrates the models by rescaling the outputs by dividing a scaling parameter T, which is learned based on the validation set. To improve the expressive power, ensemble TS (Zhang, Kailkhura, and Han 2020) calibrates a model via a weighted sum of three scaling parameters. TS calibrates well when the test set is independently and identically distributed (IID) from the training set but performs poorly under out-of-distribution settings (Desai and Durrett 2020), e.g., a model is trained by SNLI (Bowman et al. 2015) but tested by MNLI (Williams, Nangia, and Bowman 2018).

Text noise is frequently encountered in real applications due to character or word spelling mistakes, and grammatical errors. In this paper, we focus on the robustness of the post-hoc calibration approach under noisy settings in text classification tasks.

Firstly, we evaluate the pretrained BERT (Devlin et al. 2018) model on the original and noisy Emotion datasets. We generate five levels of noisy datasets based on the clean texts via increasing the noise proportion by CharSwap method in TextAttack[1] framework. As shown in Table 1, we discover that the model accuracy decreases with the noise proportion, and the calibration performance under noisy setting is not improved when using TS (see column **ECE** and **ECE-TS** in the table).

To understand the above results, we conduct a toy experiment by BERT on Emotion dataset, as shown in Fig 1, which reflects the relationship between the swapping percentage and maximum mean discrepancy (MMD) score, where MMD is used to measure the distance between the clean and noisy datasets. We control the noise proportion by swapping different amount of adjacent characters, which indicates different levels of data shift. Distributional shift from clean dataset results in the poor calibration performance of TS under noisy settings.

Based on the findings, we aim to improve the robustness of post-hoc calibration methods by measuring the distributional shift. How can the shift be measured effectively? We present two solutions. First, MMD is a natural choice. From the other perspective, when a classifier processes more difficult samples, such as noisy samples, it exhibits greater uncertainty (Hendrycks and Gimpel 2016), we adopt the predictive entropy ratio between noisy and clean datasets. Based on the measured distribution shift, we propose a simple yet effective post-hoc method named transferable TS, which readjusts the scaling parameter T learned on the clean set to adapt to noisy datasets.

The contributions of this paper lie in three aspects.

(1) To the best of our knowledge, this is the first work to study calibration under noisy settings in text classification tasks. We discover that TS performs poorly when the test sets are corrupted by noises, which lead to the distributional shift of dataset.

(2) We propose a simple yet effective post-hoc calibration method named transferable TS based on the measured distribution shift between clean and noisy datasets. In this paper, we apply two metrics to measure the shift.

(3) To evaluate the robustness of the calibration methods, we construct a noisy text classification benchmark using the TextAttack framework, which is famous in adversarial attacks and data augmentation in NLP. Empirical experiments demonstrate that our proposed method can improve calibration performance significantly under noisy settings compared with the competitive baselines.

The remainder of the paper is organized as follows. We review some closely related work in Section 2. We introduce the preliminaries of this paper in Section 3 and describe our transferable temperature scaling method in Section 4. We conduct empirical experiments to validate the effectiveness of the proposed method in Section 5. We conclude the paper with a brief discussion of future work in Section 6.

Related Work

This section reviews previous work in four parts: multimodel methods, calibration regularizers, post-hoc approach, and noisy text classification.

Multimodel Methods

Deep ensemble (DE) is a state-of-the-art calibration method in many vision benchmarks (Ovadia et al. 2019; Lakshminarayanan, Pritzel, and Blundell 2017). The individual DE models are trained by different initialization schemes. However, the DE memory cost is too expensive. The Bayesian neural network (BNN) (Welling and Teh 2011; Blundell et al. 2015) achieves calibration prompted by sampling multiple models from the approximated posterior distribution.

[1]https://github.com/QData/TextAttack

As a special BNN case, Monte Carlo-based methods, such as MC dropout (Gal and Ghahramani 2016), are easy to deploy. For the transformer-based models, DE and general BNNs are not practical due to the massive parameters, while MC dropout can be chosen as a baseline due to its ease of use in the inference phase.

Calibration Regularizers

Some research focuses on incorporating calibration-related regularizers into the training phase. (Kumar, Sarawagi, and Jain 2018) proposed a reproducing kernel Hilbert space-based trainable calibration metric named MMCE as a regularization term. (Krishnan and Tickoo 2020) introduced a differentiable accuracy versus uncertainty calibration loss function named AvUC. (Karandikar et al. 2021) proposed a differentiable loss named SB-ECE by softening the binning operation in calibration error estimators. In addition, some of the previously proposed loss functions, such as focal loss (Lin et al. 2020) and label smoothing (Pereyra et al. 2017), have been proven to improve calibration implicitly. All the methods in this category work in the training phase, but in this paper, we aim to study the calibration improvement in the inference phase.

Post-hoc Approach

Temperature scaling (TS) (Guo et al. 2017) is the most widespread method in this category due to its simplicity and accuracy preservation. The optimal scaling parameter T in TS is obtained by minimizing the negative log-likelihood based on the validation set. To improve the expressive power, ensemble TS (Zhang, Kailkhura, and Han 2020) calibrates a model via a weighted sum of three scaling parameters: a learnable T, a fixed $T = 1$ and $T = \infty$. TS methods perform poorly when the training and test set are not drawn from the same distribution (Ovadia et al. 2019). We choose TS and ETS as the baselines to compare our proposed method.

Noisy Text Classification

In real applications, noisy texts are commonly occurred due to misspelled words or phrases, which decrease the text classifier accuracy. Noisy texts can be used to validate the robustness of models in an adversarial style (Wang et al. 2023; Huq and Pervin 2020) and boost the performance by perturbing the clean texts in an augmented manner (Bayer, Kaufhold, and Reuter 2022). Recent works demonstrate that the generated adversarial examples easily fool DNN-based text classifiers at the char-level, word-level, sentence-level, and multilevel(Huq and Pervin 2020). Some works discover that synonym replacement, random insertion, random swap, and random deletion can boost performance (Wei and Zou 2019; Bayer, Kaufhold, and Reuter 2022). Most previous works aim to study the calibration under the IID setting. This paper focuses on the calibration improvement of post-hoc methods under noisy settings in text classification tasks. We use the abovementioned text augmentation methods to generate noisy datasets.

Preliminary

In this paper, we focus on the text classification task with k classes. Given an input sample x_i, a pretrained model outputs a class prediction $\hat{y}_i$ and confidence score $\hat{q}_i$. The logits z_i are k-dimensional vectors, where $\hat{y}_i = \arg\max_k z_i^k$, and $\hat{q}_i = \max_k \sigma_s(z_i)^{(k)}$, σ_s denotes the softmax function.

Temperature Scaling

Given the logit vector z_i, the calibrated confidence is:

$$\hat{q}_i = \max_k \sigma_s(z_i/T)^{(k)}. \tag{1}$$

where T is learned by minimizing by the negative log-likelihood (NLL) loss function based on the validation set D_{val}, which samples from the training dataset:

$$T = \underset{(\boldsymbol{x},y)\in D_{val}}{\arg\min} -\sum_{i=1}^{N} \log p(y|\boldsymbol{x_i}, \boldsymbol{\theta}). \tag{2}$$

where $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ denotes the probability of model output y given input $\boldsymbol{x}$. $\boldsymbol{\theta}$ denotes the pretrained model weights.

Expected Calibration Error

To evaluate DNN calibration performance, expected calibration error (ECE) is often used, which represents the difference between model accuracy and confidence (Naeini, Cooper, and Hauskrecht 2015). ECE can be computed as follows:

$$ECE = \sum_{l=1}^{L} \frac{|B_l|}{N} |acc(B_l) - conf(B_l)|. \tag{3}$$

where L denotes the number of equally spaced confidence bins, and $l = 1, \ldots, L$ denotes the bin number l. N is the total number of data samples. The accuracy of B_l can be defined as follows:

$$acc(B_l) = \frac{1}{|B_l|} \sum_{i\in B_l} \mathbf{1}(\hat{y}_i = y_i). \tag{4}$$

where $\mathbf{1}(\cdot)$ denotes an indicator function, which equals 1 when the condition is true and 0 otherwise. The average confidence of B_l can be defined as follows:

$$conf(B_l) = \frac{1}{|B_l|} \sum_{i\in B_l} \hat{q}_i. \tag{5}$$

Transferable Temperature Scaling Based on Distributional Shift

As shown in Fig 2, to improve the calibration performance of TS under noisy settings from the perspective of distributional shift, two issues should be solved: (1) how to measure the distributional shift and (2) how to achieve new T based on the measured quantity. In this section, we introduce two kinds of metrics to estimate the distributional shift and propose a simple yet effective transfer method of T.

Given a pretrained model $f(\theta)$, we let the subscripts c and n denote the **clean** and **noisy** statuses. Given a clean x_c

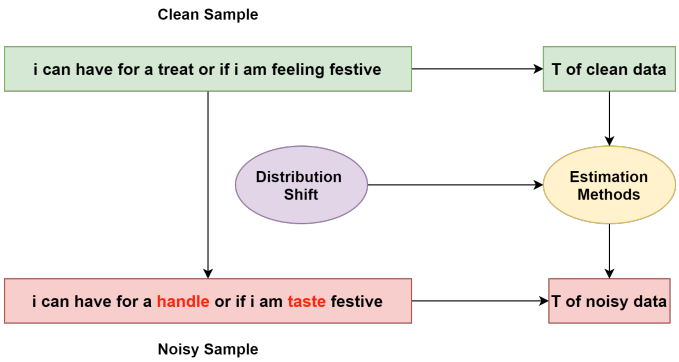


Figure 2: Overview of transferable TS with two key issues: (1) How can the distributional shift be measured? (2) How can T be transferred to TS to boost calibration under noisy settings?

and noisy sample x_n, we can obtain the output logits $z_c = \sigma_s[f(x_c)]$ and $z_n = \sigma_s[f(x_n)]$, respectively.

How is Distribution Shift Measured?

We measure the distribution shift from two perspectives: (1) output distribution distance, as shown in Fig 1, in which the MMD between the outputs of the clean and noisy datasets in the Y-axis increases along with the noise proportion shown in the X-axis; and (2) output uncertainty, inspired by the conclusion that a classifier exhibits more uncertainty when it processes more difficult samples(Hendrycks and Gimpel 2016), such as noisy samples.

Maximum Mean Discrepancy Maximum mean discrepancy (MMD) is often used in measuring the distance between two distributions and is defined as follows:

$$MMD(P, Q) = \|\mu_P - \mu_Q\|_H . \tag{6}$$

where P and Q denote two probability measures. μ_P and μ_Q represent the kernel mean embedding of P and Q, where $\mu_P = \int_X k(\cdot, x)dP(x)$ with Gaussian kernel k (Tolstikhin, Sriperumbudur, and Schölkopf 2016).

We measure the distance between the output distribution of a clean x_c and noisy sample x_n as follows:

$$ds(x_c, x_n) = MMD(z_c, z_n). \tag{7}$$

where ds denotes the distributional shift intensity, and larger MMD score indicate a larger shift in distribution.

Entropy Entropy is a frequently used metric for measuring model output uncertainty, and we calculate the output distributional shift of a clean x_c and noisy sample x_n based on entropy as follows:

$$ds(x_c, x_n) = \frac{ent(z_n) - ent(z_c)}{ent(z_c)}. \tag{8}$$

where $ent(z) = \sum -z \log(z)$, and a larger score denotes a larger distributional shift.

Distributional Shift We use the averaged score of all the samples in the batch or dataset to measure the distributional shift as follows:

$$ds(X_c, X_n) = \frac{1}{N} \sum_{i=1}^{N} ds(x_{c,i}, x_{n,i}). \tag{9}$$

where X_c and X_n denote the samples of clean and noisy batches or datasets with size N, respectively. Subscript i denotes the i-th sample.

Transfer T Based on Measured Distributional Shift

Given the measured distribution shift of the noisy dataset $ds(X_c, X_n)$, we propose a simple transfer method as follows:

$$T_n = [1 + ds(X_c, X_n)] * T_c. \tag{10}$$

We can view the distributional shift distance as a magnification factor of T_c; if the distance increases, we readjust T_c in the increasing direction. The new calibrated confidence of noisy sample x_n can be obtained as follows:

$$\hat{q}_n = \max_k \sigma_s \left[f(x_n)/T_n\right]^{(k)} . \tag{11}$$

The time complexity of our method is $\mathcal{O}(N)$, where N denotes the dataset size. We denote the proposed method based on two metrics as **transfer-entropy** and **transfer-mmd**.

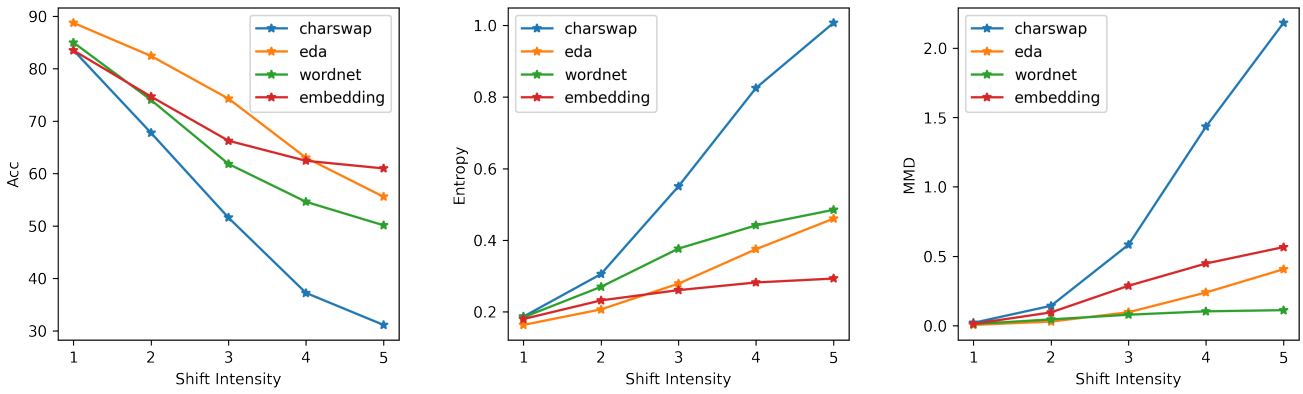
Experiments

Dataset	Train	Validate	Test	Noise	#class
QNLI	99,278	5,465	5,465	5,465	2
AG-News	112,400	7,600	7,600	7,600	4
Emotion	12,000	4,000	4,000	4,000	6

Table 2: Dataset split in QNLI, AG-News and Emotion. Column Noise means the noisy dataset size for each noise level.

Dataset	Selected Example
Clean	**im feeling quite sad and sorry for myself but ill snap out of it soon**
+ CharSwap	**im feeIling quite sMda and sorry for myself but iCl snxap out of it sgon**
+ EDA	**im feeling quite pitiful and gloomy for myself but extinct ill snap out of rather it soon**
+ WordNet	**im look quite pitiful and no-count for myself but badly breeze out of it shortly**
+ Embedding	**im feeling rather pitiable and apologise for myself but ill snapshot out of it early**

Table 3: Selected samples from Emotion dataset with four typical noise types in TextAttack.



(a) Accuracy vs Shift Intensity (b) Entropy vs Shift Intensity (c) MMD vs Shift Intensity

Figure 3: The relationship between accuracy (a), entropy (b), MMD (c) and shift intensity of noisy datasets for the Emotion task with the BERT model.

In this section, we conduct empirical experiments to answer the following questions:

(1) Are entropy and MMD effective in measuring the distribution shift raised by noise?

(2) Can the proposed methods improve the calibration performance under noisy settings?

Experimental Setting

Tasks and Datasets We choose three text classification tasks: QNLI[2] (Wang et al. 2018a), AG-News[3] (Zhang, Zhao, and LeCun 2015), and Emotion[4] (Saravia et al. 2018). The dataset splits are shown in Table 2. We use the original validation set with true labels as our test set, the new version of validation set is sampled from the original training set in the benchmark, which has the same size as the test set.

QNLI QNLI (Wang et al. 2018b) is a natural language inference dataset from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016). The task is to judge whether the context sentence includes the answer to the question.

AG-News AG-News (Zhang, Zhao, and LeCun 2015) is a subdataset of AG's corpus of news articles from the four largest classes ("World", "Sports", "Business", "Sci/Tech").

Emotion Emotion (Saravia et al. 2018) is a dataset of English Twitter messages with six types of emotions: anger, fear, joy, love, sadness, and surprise.

Noisy Dataset We use four typical text augmentation methods in TextAttack, CharSwap, EDA, WordNet, and Embedding, to generate noisy datasets based on the clean test sets, which consist of five levels of data shift intensity for each noise type. The selected noisy samples from Emotion dataset are shown in Table 3.

Pretrained Models In this paper, we evaluate the calibration performance of two pretrained models, BERT (Devlin et al. 2018) and RoBERTa (Liu et al. 2019).

Baselines

Base Base, which is abbreviated as **base**, denotes the models without any calibration.

TS TS (Guo et al. 2017), which is abbreviated as **ts**, was introduced in the Section Preliminary.

ETS Ensemble TS (Zhang, Kailkhura, and Han 2020), which is abbreviated as **ets**, calibrates the model via a weighted sum of three T: 1, infinite and learnable T of TS. We use the official ETS implementation code [5].

[2]https://huggingface.co/datasets/glue

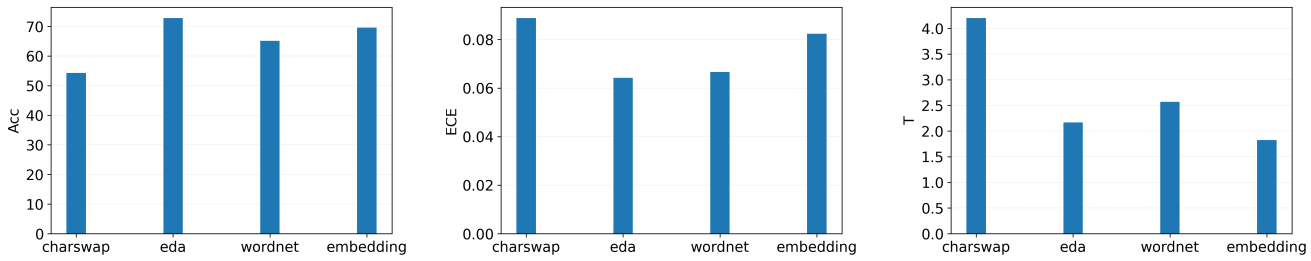
[3]https://huggingface.co/datasets/ag_news

[4]https://huggingface.co/datasets/emotion

[5]https://github.com/zhang64-llnl/Mix-n-Match-Calibration

Task	Model	base	mcd	ts		ets		transfer-entropy (ours)			transfer-mmd (ours)		
		ECE (↓)	ECE (↓)	ECE (↓)	T	ECE (↓)	T	ECE (↓)	T	ΔECE(v.s. ets)	ECE (↓)	T	ΔECE(v.s. ets)
QNLI	**BERT**	0.133	0.0972	0.079	1.52	0.0642	1.68	0.0401	3.51	37.51%	0.0266	2.3	58.63%
	RoBERTa	0.1345	0.1038	0.0891	1.47	0.0794	1.58	0.035	3.2	55.94%	0.0342	2.24	56.88%
AG-News	**BERT**	0.0828	0.0815	0.0493	1.71	0.0455	1.84	0.0289	2.53	36.51%	0.039	1.82	14.22%
	RoBERTa	0.0803	0.0757	0.0512	1.46	0.0454	1.56	0.0315	2.14	30.71%	0.0375	1.58	17.5%
Emotion	**BERT**	0.2167	0.1939	0.2118	1.03	0.1643	1.32	0.0755	2.69	54.05%	0.1545	1.39	5.93%
	RoBERTa	0.2437	0.231	0.2239	1.17	0.1767	1.52	0.0675	2.98	61.82%	0.163	1.62	7.73%
Average		0.1485	0.1305	0.1174	1.39	0.0959	1.58	**0.0464**	2.84	**46.09%**	0.0758	1.83	26.82%

Table 4: Overall performance of all the baselines and our proposed methods. The results are averaged across the noise types and shift intensities with five single runs. The average row denotes the averaged results of the corresponding column. The best results are highlighted in bold and underlined. For ECE, the smaller, the better.



(a) BERT Emotion Acc (b) BERT Emotion ECE (c) BERT Emotion T

Figure 4: Accuracy (a), ECE score (b) and T (c) on different noise types of the Emotion task with the BERT model based on the transfer entropy. The bars denote CharSwap, EDA, WordNet and Embedding.

MCD MC dropout (Gal and Ghahramani 2016), which is abbreviated as **mcd**, implements multiple model predictions to boost the calibration performance by maintaining an activated dropout layer in the inference phase. The dropout rate is set to 0.2, and the sample size of MCD is 5.

Metric We use ECE to evaluate the calibration performance, and the number of bins in ECE is set to 10.

Experimental Results and Analysis

Measure Distributional Shift In this experiment, we focus on validating the effectiveness of distribution shift metrics: entropy and MMD. We plot the relationship between accuracy (a), entropy (b), MMD (c), and shift intensity of noisy datasets for the Emotion task with the BERT model in Fig 3. For all noise types, the accuracy decreases with the shift intensity, while the entropy and MMD increase.

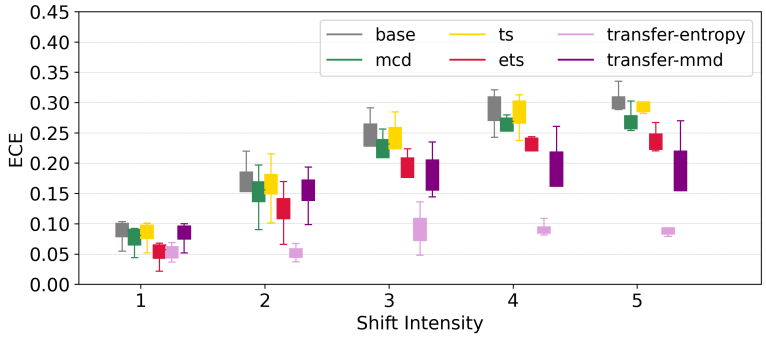
The varied intensity of CharSwap is far larger than other types, which denotes that the distributional shift raised by CharSwap is more obvious. As observed in subfigures (b) and (c), we discover that entropy is more sensitive to the distribution shift than MMD for CharSwap (blue line), EDA (orange line), and WordNet (green line) noises, while MMD performs more sensitively for Embedding noise (red line). **In summary, entropy and MMD can effectively measure the distribution shift of noisy datasets. Both metrics are sensitive to different noise types.**

Overall Performance We report the overall performance of all methods in Table 4. The ECE scores in the table are averaged across the noise types and shift intensities with five single runs. As observed in the average row, our proposed transfer-entropy and transfer-mmd perform better than all the baselines. Compared with the best baseline ETS, our proposed method can decrease the ECE scores by 46.09% and 26.82%. Transfer-entropy can achieve the lowest ECE scores in the AG-News and Emotion tasks, while transfer-mmd can achieve the best results in the QNLI task.

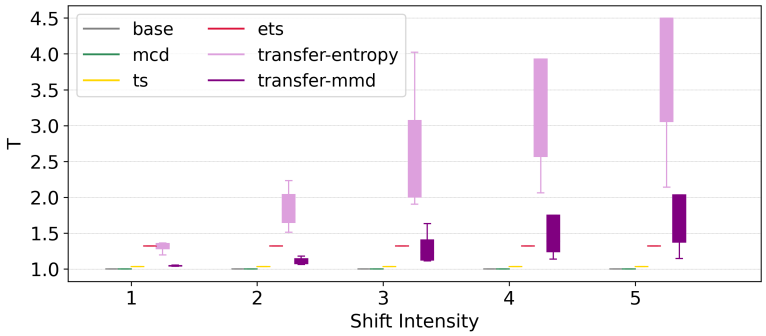
Observed from the perspective of T, we find that the averaged T values of TS and ETS are smaller than those of transfer entropy and transfer-mmd. A larger T indicates that the output logits are more uniform, namely, more uncertain about the predictions. The larger T in our methods are obtained based on the measured distribution shift of noisy datasets. **From the overall perspective, our proposed methods can improve the calibration performance of post-hoc methods under noisy settings.**

Performance on Different Noise Types We plot the accuracy, ECE score, and T on different noise types of the Emotion task with the BERT model based on the transfer-entropy in Fig 4. All the results are averaged across the shift intensity with five single runs.

In subfigure (a) of Fig 4, for the model accuracy, the CharSwap dataset is the lowest, while the EDA is the highest, which indicates that the datasets perturbed by CharSwap are more threatening to the transformer models. This is consis-



(a) BERT Emotion ECE



(b) BERT Emotion T

Figure 5: Performance on different shift intensities of the Emotion task with the BERT model. For each method, we show the averaged results on the dataset and summarize the results on different shift intensities with a box plot. Each box shows the quartiles across all noise types, while the error bars denote the minimum and maximum values across different noise types.

tent with the conclusion in Fig 3.

In subfigure (b) of Fig 4, the ECE scores of EDA and WordNet are better than those of CharSwap and Embedding. The T from the CharSwap datasets is far larger than that of other noise types, indicating that entropy is more sensitive to the CharSwap datasets. T of the Embedding datasets is the lowest, which is consistent with Fig 3. **From the noise type perspective, CharSwap datasets are more threatening and harder to calibrate compared with other noise types.**

Performance on Different Shift Intensities We report the performance on different shift intensities of the Emotion task with the BERT model in Fig 5. We show the averaged results on the dataset for each method and summarize the results on different shift intensities with a box plot. Each box shows the quartiles across all noise types, while the error bars denote the minimum and maximum values across different noise types.

In subfigure (a) of Fig 5, we find that our proposed transfer-entropy achieves similar performance as ETS in a small shift intensity (level 1) but performs better than ETS to a great extent in larger shift intensities (level 2-5). The transfer-mmd can perform competitively compared with ETS when the shift intensity increases. Our proposed methods can improve calibration when the shift intensity is larger.

In subfigure (b) of Fig. 5, the obtained T increases with the shift intensity for transfer-entropy and transfer-mmd. The incremental extent of transfer-entropy is far larger than transfer-mmd in each shift intensity, which denotes that entropy is more sensitive to the shift intensity than MMD. **From the shift intensity perspective, our proposed transfer-entropy can surpass all the baselines, and transfer-mmd can perform competitively when the distribution shift is large.**

Conclusion and Future Work

In this paper, we aimed to improve the calibration performance of post-hoc methods under noisy settings in text classification tasks. We proposed a simple yet effective transfer method based on two distribution shift metrics: predictive entropy and MMD. To evaluate the effectiveness of the proposed methods, we constructed a noisy text classification benchmark with four perturbed types: CharSwap, EDA, WordNet, and Embedding. Empirical experiments conducted on the generated datasets demonstrate that (1) entropy and MMD can effectively measure the distribution shift between the clean and noisy datasets and (2) the transferred T based on the measured distribution shift can significantly improve the calibration performance of TS compared with the competitive baselines. In the future, we would like to study calibration under few-shot settings, such as prompt learning.

References

Bayer, M.; Kaufhold, M.-A.; and Reuter, C. 2022. A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, 55(7): 1–39.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1613–1622. JMLR.org.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Cao, L.; Zhang, H.; and Feng, L. 2020. Building and Using Personal Knowledge Graph to Improve Suicidal Ideation Detection on Social Media. *IEEE Transactions on Multimedia*.

Desai, S.; and Durrett, G. 2020. Calibration of Pre-trained Transformers. In *Empirical Methods in Natural Language Processing*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 1050–1059.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 1321–1330.

Hendrycks, D.; and Gimpel, K. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*.

Huq, A.; and Pervin, M. T. 2020. Adversarial Attacks and Defense on Texts: A Survey. arXiv:2005.14108.

Karandikar, A.; Cain, N.; Tran, D.; Lakshminarayanan, B.; Shlens, J.; Mozer, M. C.; and Roelofs, B. 2021. Soft Calibration Objectives for Neural Networks. *Neural Information Processing Systems*.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. *Neural Information Processing Systems*.

Kumar, A.; Sarawagi, S.; and Jain, U. 2018. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. In *International Conference on Machine Learning*, 2805–2814.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 6402–6413.

Li, P.; Zhong, P.; Mao, K.; Wang, D.; Yang, X.; Liu, Y.; Yin, J.; and See, S. 2021. ACT: an Attentive Convolutional Transformer for Efficient Text Classification. In *AAAI Conference on Artificial Intelligence*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, volume 2015, 2901–2907.

Oluwatobi, O.; and Mueller, E. 2020. DLGNet: A Transformer-based Model for Dialogue Response Generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 54–62. Online: Association for Computational Linguistics.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 13969–13980.

Pearce, K.; Zhan, T.; Komanduri, A.; and Zhan, J. 2021. A Comparative Study of Transformer-Based Language Models on Extractive Question Answering. arXiv:2110.03142.

Pereyra, G.; Tucker, G.; Chorowski, J.; Łukasz Kaiser; and Hinton, G. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *ICLR (Workshop)*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *empirical methods in natural language processing*.

Saravia, E.; Liu, H.-C. T.; Huang, Y.-H.; Wu, J.; and Chen, Y.-S. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3687–3697. Brussels, Belgium: Association for Computational Linguistics.

Tolstikhin, I.; Sriperumbudur, B. K.; and Schölkopf, B. 2016. Minimax estimation of maximum mean discrepancy with radial kernels. In *Neural Information Processing Systems*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *neural information processing systems*.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018a. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, W.; Wang, R.; Wang, L.; Wang, Z.; and Ye, A. 2023. Towards a Robust Deep Neural Network Against Adversarial Texts: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 35: 3159–3179.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Empirical Methods in Natural Language Processing*.

Welling, M.; and Teh, Y. W. 2011. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 681–688.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Zhang, J.; Kailkhura, B.; and Han, T. Y.-J. 2020. Mix-n-Match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Zhang, X.; Zhao, J. J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.