Class-wise Calibration: A Case Study on COVID-19 Hate Speech

Stephen Obadinma [†], Hongyu Guo [‡], Xiaodan Zhu [†]
[†] ECE & Ingenuity Labs Research Institute, Queen's University
[‡] National Research Council Canada

Abstract

Proper calibration of deep-learning models is critical for many high-stakes problems. In this paper, we show that existing calibration metrics fail to pay attention to miscalibration on individual classes, hence overlooking minority classes and causing significant issues on imbalanced classification problems. Using a COVID-19 hate-speech dataset, we first discover that in imbalanced datasets, miscalibration error on an individual class varies greatly, and error on minority classes can be magnitude times worse than what is suggested by the overall calibration performance. To address this issue, we propose a new metric based on expected miscalibration error, named as Contraharmonic Expected Calibration Error (CECE), which punishes severe miscalibration on individual classes. We further devise a novel variant of temperature scaling for imbalanced data to improve class-wise miscalibration, which re-weights the loss function by the inverse class count to tune the scaling parameter to reduce worst-case minority calibration error. Our case study on a benchmarking COVID-19 hate speech task shows the effectiveness of our calibration metric and our temperature scaling strategy.

Keywords: calibration, imbalanced data, covid-19, deep learning, hate speech

1. Introduction

Calibration aims to gauge and improve a model's ability to provide reliable confidence scores when making predictions, which is critical for many high-stakes problems that have significant impact in social, economic, and health applications, among others. In the past decade, deep neural networks have been shown to achieve impressive prediction performance on a wide variety of problems. To better assess the degree of miscalibration in neural networks [1, 2], there has recently been a surge of interest in improving the calibration metrics. The most widely used metrics include ECE [3, 4] and its variants [2, 5]. Nevertheless, these calibration metrics fail to pay attention to miscalibration on individual classes, hence overlooking minority classes that we really care about and causing significant issues on imbalanced classification problems.

To address this challenge we propose a new metric based on expected miscalibration error, named as Contraharmonic Expected Calibration Error (CECE), which punishes severe miscalibration on individual classes. We further devise a novel variant of temperature scaling for imbalanced data to improve class-wise miscalibration, which re-weights the loss function by the inverse class count to tune the scaling parameter to reduce worst-case minority calibration error. We empirically demonstrate the effectiveness of our metrics and temperature scaling variant using a benchmarking hate speech classification task.

We setup our study on the COVID-19 hate-speech detection task, motivated by the following reasons. First, online hate speech, including that associated with COVID-19 [6], has significant social impact, e.g., amplifying hate crimes happening in our already stressful society during the pandemic [7], making it paramount to find means to address this discrimination and violence. Calibration is critical and challenging here because hate speech is hard to detect as it does not contain discriminative features [8, 9]. Before banning a user or removing a post flagged under hate speech, it is desirable that the classifier can tell if it

[†] {16sco, xiaodan.zhu}@queensu.ca [‡] Hongyu.Guo@nrc-cnrc.gc.ca

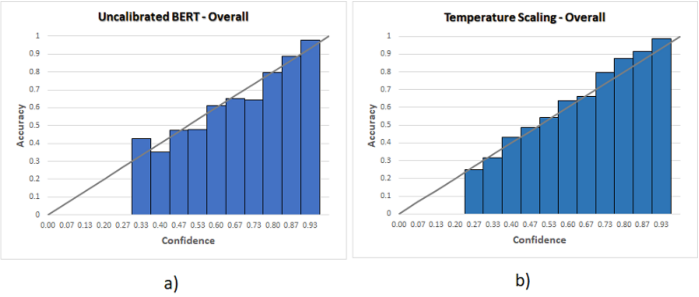

Figure 1. Plotted confidence versus accuracy for the COVID-19 hate speech dataset is shown for both a BERT model with no calibration methods (a), and with temperature scaling applied to the same model (b). The data was sorted into 15 bins based on confidence and the average accuracy in each bin was calculated. The plots are shown for the overall test set. The ideal calibration curve can be seen with the diagonal line, representing perfect calibration as a comparison. (Refer to Section 3.1 for more details of calibration metrics.)

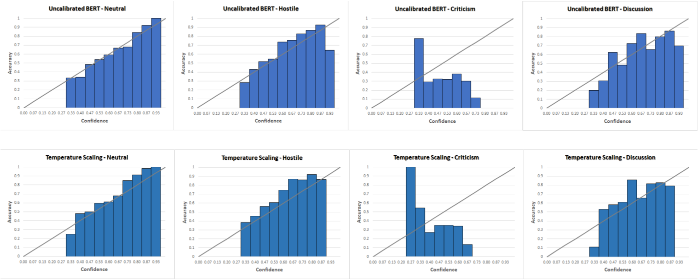

Figure 2. Continuation of Figure 1. Confidence versus accuracy plots on the 4 subsets of the test set from each class consisting of only data from that one class can be seen. The top row is for the uncalibrated model, and the bottom row shows the temperature scaling results. From visual inspection, it is clear that for both the calibrated and uncalibrated baselines, the bar plots for some of the classes appear to deviate significantly more from the ideal line than when all of the classes are considered at the same time in Figure 1, and the deviation is particularly severe for the classes with the fewest number of samples such as the criticism and discussion (the two right columns).

is unsure on a difficult example so that it can forward it to a downstream decision from a human moderator.

Second, hate speech detection is known to suffer from the imbalance problem due to the relative rarity of hate speech when collecting data from websites like Twitter [10]. Class imbalance occurs when there is a skew in the distribution of class data, leading to there being minority classes with far fewer samples than those of the majority. It is imperative that as people spend more time on social media, online spaces become free from abuse [11]. Models trained on imbalanced data tend to prioritize majority class performance at the detriment of the minority classes [12], leading to poorer accuracy on minority classes and as we note, poor calibration as well. Figure 1 and Figure 2 visually demonstrate this issue of there being prominently more calibration error on individual classes (Figure 2) compared to the whole (Figure 1) on detecting online hate speech specifically. However, it is the minority class containing hate that we want to be calibrated the most and that we are most unsure about; making a mistake on them can lead to grave consequences.

In addition, pre-trained language models like BERT [13] and RoBERTa [14] have achieved superior classification accuracy on many language processing tasks [15, 16]. These large pre-trained models, however, are prone to produce overconfident predictions, thus resulting in high miscalibration error [4]; we need better calibration on such models, which have impact on many downstream tasks. Note that while our study is performed on hate speech detetction, the proposed metrics and method can be easily adapted to other imbalanced problems where calibration on minority classes is key.

2. Related Work

Recent research on calibration has been conducted on developing metrics to measure calibration errors, in addition to developing calibration methods. The most widely used metric for gauging the calibration of a model in literature is Expected Calibration Error (ECE) [3, 4]. A great deal of research has been conducted to fix the issues with the standard calibration metrics like ECE [5]. Nixon et al. [2] introduce a new metric named General Calibration Error (GCE) which encompasses a space of various binning-based calibration metrics conditioned on different aspects. ECE is part of the space of functions of GCE, as are additional new metrics that they develop like Adaptive Calibration Error (ACE) and Static Calibration Error (SCE). The latter two metrics do not just consider the probability of the predicted class, but each class probability separately. They also find that calculating the class conditional calibration error yields a much higher result than that of the standard top prediction error, which inspires this paper's examination into individual class calibration as being key to evaluating the true calibration error. For ACE, they use adaptive binning, which is a method of binning that ensures an equal number of data points are present in each bin to eliminate high variance inside some bins with low data. In contrast, ECE bins data points into equally sized bins. However, whilst the probabilities for all classes are considered, the performance specifically on data points of only those classes is not considered in the metrics, meaning there is no means to tell which classes are calibrated worse.

Regarding calibration methods, modern calibration methods generally fall under two methodologies. In the most popular methodology, a recalibration model trained on validation data is applied on the predictions of the classifier after it is already trained to increase the entropy of the predictions while preserving the accuracy. This is preferred most of the time as it does not require longer training times. Deep ensembles are a technique that have also been explored as a recalibration method, but research has shown that model averaging does not necessarily lead to better calibration [17]. The second class of methods modify the training procedure of the classifier so that the model is trained to be less overconfident. Usually either the loss function is modified or some form of label smoothing is utilized as training with soft targets is known to prevent overconfidence [18]. Methods using this approach include Mixup [19] and related methods like Remix [20]. Remix modifies Mixup so that the labels are more biased towards minority classes making the model more lenient on minority samples by making a tighter margin with the majority class, but how it performs in terms of calibration metrics like ECE is not explored.

A modification of temperature scaling [4], called Attended Temperature Scaling [21] is of interest. Standard temperature scaling is not robust to small amounts of validation data nor noisy data. Likewise, it does not perform well when the accuracy is too high or low. Attended temperature scaling can work better in these conditions by generating additional data on an individual class level. The effects of this on individual class calibration is not explored nor do they try to synthesize more minority class samples for imbalanced settings.

Approaches to dealing with multi-class imbalanced data are not very well developed as relationships between classes are not straightforward and simple sampling strategies may not have the intended effect creating a trade-off in performance between classes [12]. Nevertheless, strategies like oversampling the minority classes [22], undersampling majority classes [23], loss re-weighting [24–26], and SMOTE [27] are often deployed to alleviate the issues.

Table 1. Individual class ECEs for each of the 4 classes for the baselines detailed in Section 4.2 on the COVID-19 hate speech dataset. It should be noted that the error for each class are usually significantly higher than what is suggested by the overall number.

Models	Neutral ECE	Criticism ECE	Hostility ECE	Discussion ECE	Overall ECE
Uncalibrated	0.026950	0.388316	0.076012	0.074317	0.027712
TS	0.035463	0.367853	0.097313	0.067996	0.017343
IR	0.054277	0.490876	0.219062	0.117765	0.014545
HB	0.028219	0.475642	0.123105	0.141398	0.031437
VS	0.028219	0.475641	0.121193	0.141398	0.031437
SMOTE	0.017720	0.257550	0.065420	0.137520	0.031960
WTS	0.043935	0.350664	0.116748	0.104876	0.024328

Table 2. Individual class ACEs for each of the 4 classes for the baselines detailed in Section 4.2 on the COVID-19 hate speech dataset. The trend with the per-class errors is still present and is often more severe than seen with ECE as the metric.

Models	Neutral ACE	Criticism ACE	Hostility ACE	Discussion ACE	Overall ACE
Uncalibrated	0.061748	0.347441	0.174086	0.210973	0.022355
TS	0.069531	0.346296	0.187416	0.223486	0.017889
IR	0.086079	0.365669	0.198037	0.240948	0.01512
HB	0.059805	0.362539	0.162407	0.226964	0.017412
VS	0.061930	0.371621	0.166005	0.265200	0.031199
SMOTE	0.060119	0.305535	0.170161	0.220800	0.220800
WTS	0.075495	0.345775	0.195986	0.231392	0.020531

3. Class-Wise Calibration Evaluation

3.1. Calibration and Issues

The goal of calibration is to reduce the error between a classifier's predicted confidences and the classifier's accuracy. That is, a well-calibrated model expects the probability values it outputs for each of the class labels matches the true predictive probability distribution of the given data. For example, if the model has to predict the labels for 100 samples and the model outputs correct predictions on 50% of these samples, then it is expected that the corresponding probability values it outputs on its predicted classes would be 0.5 on average when considering each data point, as the model can only be 50% confident that it can guess the correct answer. This property is of great importance in many high-stakes applications [28], where knowing when the model is likely to be incorrect or inadequate matters for making subsequent downstream decisions.

Formally, let there be a set of input data points $X = \{x_1, x_2, ..., x_N\}$ where N is the size of the evaluation set, and $Y = \{y_1, y_2, ..., y_N\}$ is the corresponding true labels. After the softmax is taken over the logits, the classifier produces a probability distribution $\hat{P}(x_n) = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_K\}$ for each data point x_n where the total sum is 1 and there are K number of classes. The highest probability is taken as the predicted class for that data point $\hat{Y}(x_n)$. A model is considered calibrated if the following is true for each data point x_n in X and for each class k:

$$\mathbb{P}(Y(x_n) = \hat{Y}(x_n) \mid \hat{p}_k(x_n) = p) = p. \tag{3.1}$$

Essentially, this equation is formalizing the notion that if the classifier yields a confidence $\hat{p}_k = p$, then the probability of the actual data distribution for that label should match that probability. In most cases only the confidence for the predicted class is used in this notion. Any deviation between the left equation and right equation is miscalibration. Verifying Equation 3.1 is intractable as the left hand side is continuous function and would require an infinite sample size to correctly estimate. Thus, it is usually approximated by binning-based metrics. Metrics of this type approximate Equation 3.1 by dividing data points by their confidences into a series of discrete bins. They then calculate the expectation of the total error based on the error in each bin. The most popular metric, called Expected Calibration Error [3], bins data points into M evenly sized bins by the confidence of their predicted class and calculates the absolute difference between the average accuracy in each bin and the average confidence score. The errors in each bin are then weighted by the number of data points in them. The equation for ECE is as follows:

$$ECE = \sum_{m=1}^{M} \frac{n_m}{N} |acc(B_m) - conf(B_m)|, \tag{3.2}$$

where B_m are the data points in bin m, and n_m is the number of data points in bin m. ECE is sensitive to the number of bins chosen. The greater the number of bins the less the bias in the estimation is, but the variance increases as there are fewer data points to do the calculation in some bins [2]. The effect of the number of bins on ECE is documented in appendix A. ECE is widely used, but has been known to have issues apart from the sensitivity to the number of bins, including how it only considers the confidence of the predicted class, not of all the classes, which is significant [5]. A bigger problem, however, is that ECE can severely the underestimate calibration error if overconfident and underconfident data points are present in the same bin [2], meaning that the metric when ran on the data as a whole can be unreliable. This motivated looking at the ECE of each class individually to counteract this effect to some extent as overconfident classes could not balance out underconfident classes when they are considered separately.

Hence, we run experiments to examine the calibration error when only looking at data points of a single class. We divide the dataset X, Y into K subsets where each data point x_n in the subset has the true class label $y = k$. We then run the ECE algorithm on each of these subsets by using a consistent number of bins of 15 and analyze the results. In table 1, it can be seen that in the case of four different classes, individual class calibration varies highly. This applies to every method of re-calibration as well, not just the uncalibrated baseline. For most of the methods, particularly the post processing based methods, all of the individual ECEs are higher than that of the test set as a whole. This suggests the presence of the "balancing out" effect we previously mentioned, where overconfident and underconfident data points get cancelled out, yielding a misleadingly low number.

It is important to note that the minority classes, in this case the criticism, hostility, and discussion classes, have a significantly higher ECE than that of the majority neutral class, and in particular, the criticism class has extremely poor calibration. The classifier is not able to predict confidences representative of its ability to classify these samples. Despite this, it has a negligible effect on the overall error due to its scarcity. Some increase in error is expected because binning less data points leads to more variance in the estimations, but the class-wise increases here are drastic. This suggests that using the ECE over the entire data may be misleading, especially for imbalanced classification, where extremely poor performance on minority classes is masked due to exerting smaller influence and being balanced out by better performing classes. If we care about how well calibrated the minority samples are, then we should be examining calibration specifically on them. Furthermore, it can be observed that the class that performs the most poorly is not the class with the fewest number of samples. This signals that in multi-class imbalanced settings, we must consider not just compensating for the low number of samples but the relative difficulty of the class as well. These effects not only apply to ECE but also to a related metric, Adaptive Calibration Error (ACE) [29], as can be seen in table 2. This suggests that this is a problem beyond just a quirk of calculating ECE.

The theoretical causes for this effect likely stem from how the decision boundaries are learned by imbalanced classifiers. Standard training algorithms assume a balanced distribution [30], which creates an inductive bias towards the majority class as there are more data points to update model parameters. Rules are learned for the majority class but ignored for the minority class [31], leading to poor accuracy, especially in the case of a shift in distribution. The classifier essentially struggles to distinguish between noisy samples and minority class samples [31], and so is unable to represent the true distribution of the minority class.

This effect of using the overall value to judge classifier calibration in imbalanced settings is comparable to only using accuracy to evaluate a classifier trained on imbalanced data. The classifier can achieve high accuracy just by guessing the majority class all of the time. The classifier has not learned to classify anything, but still performs well according to the metrics because the majority class overwhelms the minority classes. This problem led to the creation of new metrics like F1 score, which is the harmonic mean of precision and recall, to compensate for this effect [12], since it punishes extremely low values. For example a precision of 1.0 and a recall of 0 would give an F1 score of 0. Similarly here, there is a need to consider alternative metrics that can punish classifiers outputting calibration errors that are exceptionally high on individual classes. This motivated our new calibration metric, which will be discussed in detail next.

3.2. **Class-Wise Calibration Metrics**

To evaluate class-wise calibration, we propose the new metric, Contraharmonic Expected Calibration Error (CECE), based on the contraharmonic mean of the individual class subset ECEs. The contraharmonic mean is a form of mean that is defined as follows for a set of real numbers $C = \{c_1, c_2, ..., c_n\}$:

$$CM(C) := \frac{c_1^2 + c_2^2 + ... \, c_n^2}{c_1 + c_2 + ... \, c_n}. \tag{3.3}$$

The contraharmonic mean follows the rules of a mean. It is always non-negative, and is always at least equal the harmonic mean, and equal to or greater than the arithmetic mean [32]. The squared terms in the numerator mean that the contraharmonic mean will be more biased towards high outliers. For example, for a set of numbers $A = \{0.05, 0.1, 0.6\}$, the arithmetic mean is 0.25. The contraharmonic mean is 0.497, which is much closer to the outlier. As higher calibration error values correspond to worse performance, a mean that punishes extremely high outliers is desirable.

CECE is defined as follows for a set of K number of ECE values, assuming there is an ECE value for each subset of data points with true class labels equal to $k \in K$:

$$CECE = \frac{ECE_1^2 + ECE_2^2 + ... \, ECE_K^2}{ECE_1 + ECE_2 + ... \, ECE_K}. \tag{3.4}$$

As an example, assuming the aforementioned set A represents the ECE for 3 different classes, the CECE would be 0.497, a very high value.

We also introduce other metrics that take the arithmetic and weighted arithmetic mean of these individual ECEs rather than the contraharmonic mean. We call these metrics Macro Subset Expected Calibration Error (MSECE) and Weighted Subset Expected Calibration Error (WSECE) respectively. These metrics give a notion of how well calibrated individual classes are, without punishing extremely high values. MSECE calculates the arithmetic mean of the individual class ECEs, treating each class equally irrespective of the number of samples in each subset:

$$MSECE = \frac{ECE_1 + ECE_2 + ... \, ECE_K}{K}. \tag{3.5}$$

MSECE is useful when there is a need to consider the individual class calibration error with equal weights. For set A the MSECE would be 0.25.

WSECE calculates the weighted arithmetic mean of the individual class ECEs, and each term is weighted by the percentage of samples of that class:

$$WSECE = \frac{n_1}{N}ECE_1 + \frac{n_2}{N}ECE_2 + ... \frac{n_K}{N}ECE_K. \tag{3.6}$$

WSECE is beneficial if it is desired that calibration error on bigger classes is more important.

Lastly, we introduce ECE variance, which calculates the variance between the individual ECE values and the ECE when considering the overall set of data:

$$ECE\,Variance = \frac{\sum_{i=1}^{K}(ECE_i - ECE_{\text{overall}})^2}{K}. \tag{3.7}$$

A high variance can mean there is more of a “cancelling” effect occurring between classes leading to underestimated calibration error. Using set A as an example, if the overall ECE was equal to 0.08, then the ECE variance would be 0.09.

3.3. Novel Temperature Scaling Variant

Most conventional post-calibration methods do not attempt to optimize performance on minority classes in imbalanced settings. Temperature scaling is among the most popular approaches, and it tries to find an ideal temperature, T, on the validation data and training is based on finding T to minimize NLL. However, NLL tends to overfit to majority classes, therefore temperature scaling may ignore minority class calibration to focus on re-calibrating majority class samples more. Motivated by the re-weighting technique [24–26], done during training of classifiers for imbalanced data, the loss function is tuned to give a higher penalty for minority class samples. Usually the loss is weighted by the inverse class frequency or similar [33, 34]. We propose a novel method called Weighted Temperature Scaling (WTS), where we apply temperature scaling as usual but during parameter optimization for T we weigh the cross entropy loss by a factor of $1 - \frac{n_k}{N}$ based on the sample counts of each class in the validation data. By applying this technique while tuning the temperature in temperature scaling, we are willing to introduce more calibration error to the majority class if it means the worst case individual class calibration error is reduced for the minority classes. Reweighting performs poorly when classes are extremely imbalanced as it can cause significant performance downturn in majority classes [35], but nevertheless this method can inspire a new class of methods that prioritize minority classes.

4. Experimental Studies

4.1. Dataset

The dataset we use is the Detecting East Asian Prejudice on Social Media dataset [6]. The primary task for this dataset is to detect hostility and prejudice against East Asian entities like people groups and businesses in the context of the COVID-19 pandemic. The dataset is not explicitly phrased as being about hate speech, but since it is about detecting animosity towards a specific group we treat it as a hate speech detection task as it is largely in the same vein. This dataset contains 4 main classes, a neutral class, a hostility against East Asian related entities class, a criticism of East Asian related entities class, and a meta-discussions of East Asian prejudice class which consists of two sub-classes merged together due to lack of data. The dataset consists of 20,000 tweets with an imbalanced distribution of 67.6% for the neutral class, 7.2% for the criticism class, 19.5% for the hostility class,and 5.7% for the discussion class. The motivation for using this dataset is that it is a multi-class imbalanced dataset in the domain of the high stakes problem of hate speech. Furthermore, some minority classes, like criticism and hostility, are difficult to distinguish from one another. This creates a great emphasis for classifiers not to output overconfident predictions that do not reflect the difficulty in classifying these classes. As such the dataset serves well to demonstrate the issues of imbalanced data calibration.

4.2. Comparison Baselines

The following methods were compared:

- BERT [13]: The uncalibrated baseline model we use is a pre-trained $\text{BERT}_{\text{BASE}}$ uncased with one layer for sequence classification. The rest of the models use the predicted scores from the BERT model excluding SMOTE.
- Temperature Scaling (TS) [4]: A parametric method that finds a temperature T using the validation data that is used to rescale logits produced by the uncalibrated model.
- Isotonic Regression (IR) [36]: A non-parametric approach that transforms the uncalibrated predictions using a piecewise linear function that minimizes the square loss between the predictions and the targets.
- Histogram Binning (HB) [37]: A non-parametric method where uncalibrated predictions are separated into bins and predictions in each bin are given the same confidence score.
- Vector (Platt) Scaling (VS) [4]: A parametric approach that finds two ideal parameters a and b from the validation data that scale the logits of the uncalibrated model. Vector scaling is an extension of Platt scaling [38] into a multi-class setting.
- Synthetic Minority Oversampling Technique (SMOTE) [27]: An oversampling technique that generates additional minority class samples to compensate for data imbalance. A BERT model is trained on this augmented data. No additional calibration methods are used.
- Weighted Temperature Scaling (WTS): TS, but while converging on ideal temperature T, minority classes have a higher weight in determining the value.

4.3. Implementation Details

We use Adam [39] optimizer for all of the experiments and a consistent batch size of 16. The learning rate is initially set to be 2e-5 and a 1cycle learning rate policy [40] is used to adjust it during training for regularization purposes. A 80/10/10 train/validation/test split is used for the COVID hate speech dataset to follow what is originally done in [6]. The number of bins we use as a parameter to each of the calibration metrics is 15.

4.4. Results and Analysis

We compare the baseline and calibration methods with our new metrics, CECE, MSECE, WSECE and ECE variance, and a set of existing metrics for comparison. ECE on the test set is the primary existing metric, and results for ACE, and SCE are included to give extra insight to see if the new metrics we introduce correlate more with the results of the class conditioned metrics. It should be noted that the weighted F1 score of the model was 0.83, with the calibration methods not affecting the F1 of the model to a notable extent. Table 3 reports the results of each baseline on the test set of the COVID-19 hate speech dataset. We can see that the standard calibration methods like temperature scaling, and isotonic regression performed the best in terms of ECE, ACE, and SCE and outperformed the uncalibrated model. Histogram binning had a higher ECE than the uncalibrated model but in terms of ACE and SCE it is better so it can be assumed to be better calibrated than the baseline. Vector scaling performed poorly on both an overall and individual class level, but this method does have a tendency to perform inconsistently.

We can see that the two techniques attempting to address the data imbalance perform better. SMOTE is worse in terms of calibration error when considering the overall data than the baseline. However, it has managed to achieve the best results on an individual level on all the new metrics in terms of reducing worst case calibration error. As SMOTE is not designed primarily as a calibration method, our newly introduced weighted temperature scaling performed the best on CECE and ECE variance out of the other post-calibration methods. It is also better calibrated than the uncalibrated model on the overall metrics as well. This demonstrates that it successfully managed to reduce the worst case individual class calibration error compared to standard temperature scaling. It trades off higher calibration error on the majority classes, but proves that by considering minority classes when tuning the post-calibration methods we can prioritize reducing minority class error.

Table 3. Overall calibration results for the COVID-19 hate speech dataset. Results for all of the models detailed in Section 4.2 in terms of the calibration metrics are given. The results are averaged for 5 random seeds. The two methods addressing the imbalance perform the best in terms of reducing individual class calibration errors compared to the traditional calibration methods. WTS is the most balanced between overall and individual class results.

Metrics	**Uncalibrated**	**TS**	**IR**	**HB**	**VS**	**SMOTE**	**WTS**
ECE	0.0277	0.0173	**0.0145**	0.0314	0.0314	0.0320	0.0243
ACE	0.0224	0.0179	**0.0151**	0.0174	0.0312	0.2208	0.0205
SCE	0.0211	0.0193	**0.0164**	0.0204	0.0308	0.0258	0.0233
Contraharmonic ECE	0.2879	0.2650	0.3467	0.3412	0.3415	**0.1879**	0.2426
Macro Subset ECE	0.1414	0.1422	0.2205	0.1921	0.1916	**0.1196**	0.1541
Weighted Subset ECE	0.0651	0.0731	0.1209	0.0851	0.0847	**0.0493**	0.0835
ECE Variance	0.0336	0.0330	0.0702	0.0545	0.0544	**0.0158**	0.0305

5. Conclusion and Outlook

We contribute a new finding that existing calibration methods fail to pay attention to miscalibration on individual classes, and hence they tend to underestimate the degree of miscalibration on data points belonging to minority classes. To address this issue, we proposed the novel metric CECE, which takes into account the class distribution and punishes exceedingly high error values. We also formulated a new variant of temperature scaling to prioritize minority class calibration. Using a benchmarking COVID-19 hate speech task, we empirically showed the effectiveness of our metric and method.

Although we demonstrate the result using a COVID-19 task, the proposed new metrics and method are general enough to be applied to other imbalanced problems where calibration on minority classes is key. We will investigate such potential in our future work. We believe that our research here will inspire further development of calibration methods that do not just reduce overall calibration error but improve calibration error in important minority class samples in particular.

Acknowledgements

The research of the first and last author is supported by NSERC Discovery Grants, Accelerator Supplement and DND Supplement.

Appendix A. Effect of Number of Bins

The number of bins and the distribution of data within the bins has an impact on binning-based metrics, especially ECE. When binning minority classes, a significantly lower number

Table 4. The effect of varying the number of bins on overall ECE and class-wise ECEs for an uncalibrated BERT model. Increasing the number of bins increases the calibration error. This variance in relation to the number of bins is a problem with binning based metrics.

	5 Bins	**10 Bins**	**20 Bins**	**40 Bins**
ECE	0.027198	0.027198	0.029574	0.032202
Contraharmonic ECE	0.290927	0.295202	0.295264	0.293997
Neutral Class ECE	0.021856	0.027598	0.02782	0.032994
Criticism Class ECE	0.370042	0.400979	0.400979	0.400979
Hostility Class ECE	0.038143	0.073228	0.082316	0.084447
Discussion Class ECE	0.049876	0.061111	0.080737	0.099759

of data points than the data as a whole are binned. This essentially simulates increasing the number of bins as the average number of data points per bin decreases in both cases. Therefore, we conduct an additional analysis to ascertain the effects on calibration error when there are a lower number of data points per bin and the results can be seen in table 4. From it, we observe that increasing the number of bins does increase the error, but that the effects are hardly severe relative to the many magnitude increases in error seen for minority classes compared to the overall error. Therefore, the increase in variance in the estimation due to having bins with a low number of data points is likely not enough to account wholly for why minority classes have higher calibration errors.

Appendix B. Further Analysis of Minority Class Calibration

Here we provide further analysis with a specific example of how the problem of imbalanced data leads to poor calibration of minority classes and why it is crucial we consider class-wise calibration. The following sentence from the dataset demonstrates this imbalance issue, "@markskrikorian this is not the time to blame the chinese communist party for covering up the HASHTAG_EASTASIA+VIRUS or xi jinping for exporting HASHTAG_EASTASIA+VIRUS it's time for #HASHTAG ." The corresponding confidence scores produced by the uncalibrated BERT model are 0.009996, 0.16967, 0.81458, 0.00575 for the neutral, criticism, hostility, and discussion classes respectively. The predicted label by the classifier is therefore hostility, likely due to the presence of negative phrases like "blame the chinese communist party" and "covering up" that correlate to hostility or criticism. In actuality, this sentence is counter hostility and its true label is that of the discussion class. Not only is the classifier wrong in this case, but more importantly, it is highly confident in its wrong decision. The hostility class, which has relatively more samples than the discussion class, has caused the classifier to be biased towards the hostility label by marking the presence of these negative phrases as more important to its decision. The classifier has not managed to learn the properties of the minority class and has effectively ignored the key part of the sentence that is saying the opposite. In fact, the classifier does not even consider it possible to be counter hostility. On the other hand, for a neutral class sample like, "wear the right masks to prevent HASHTAG_EASTASIA+VIRUS" the classifier predicts neutral correctly with a confidence of 0.99. This might be slightly overconfident but it is an easy example that we expect will be classified correctly almost every time. Despite the good calibration on majority classes, it should not overshadow poor performance on minority classes where completely unrepresentative probabilities are generated. Thus we need a closer examination of minority class calibration scores to make it so that classifiers are to achieve any real world value.

References

[1] K. Gupta, A. Rahimi, T. Ajanthan, T. Mensink, C. Sminchisescu, and R. Hartley. "Calibration of Neural Networks using Splines". In: *International Conference on Learning Representations*. 2021.

[2] J. Nixon, M. Dusenberry, G. Jerfel, T. Nguyen, J. Liu, L. Zhang, and D. Tran. *Measuring Calibration in Deep Learning*. 2020. arXiv: 1904.01685 [cs.LG].

[3] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence* 2015 (Apr. 2015), pp. 2901–2907.

[4] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, 2017, 1321–1330.

[5] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, and T. Schön. "Evaluating model calibration in classification". In: *Proceedings of Machine Learning Research*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3459–3467.

[6] B. Vidgen, S. Hale, E. Guest, H. Margetts, D. Broniatowski, Z. Waseem, A. Botelho, M. Hall, and R. Tromble. "Detecting East Asian Prejudice on Social Media". In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Online: Association for Computational Linguistics, Nov. 2020.

[7] A. R. Gover, S. B. Harper, and L. Langton. "Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality". In: *American Journal of Criminal Justice* 45.4 (2020), pp. 647–667. DOI: 10.1007/s12103-020-09545-1.

[8] Z. Zhang and L. Luo. "Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter". In: *Semantic Web* Accepted (Oct. 2018). DOI: 10.3233/SW-180338.

[9] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. "Hate speech detection: Challenges and solutions". In: *PloS one* 14 (Aug. 2019), e0221152. DOI: 10.1371/journal.pone.0221152.

[10] A. Arango, J. Pérez, and B. Poblete. "Hate Speech Detection is Not as Easy as You May Think: A Closer Look at Model Validation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019).

[11] J. Cowls, B. Vidgen, and H. Margetts. *Why content moderators should be key workers*. 2020.

[12] B. Krawczyk. "Learning from imbalanced data: open challenges and future directions". In: *Progress in Artificial Intelligence* 5 (2016), pp. 221–232.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019.

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv* abs/1907.11692 (2019).

[15] M. Mozafari, R. Farahbakhsh, and N. Crespi. "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media". In: Nov. 2019, pp. 928–940. ISBN: 978-3-030-36686-5. DOI: 10.1007/978-3-030-36687-2_77.

[16] P. Alonso, R. Saini, and G. Kovács. "Hate Speech Detection Using Transformer Ensembles on the HASOC Dataset". In: Sept. 2020, pp. 13–21.

[17] R. Rahaman and A. H. Thiery. *Uncertainty Quantification and Deep Ensembles*. 2020. arXiv: 2007.08792 [stat.ML].

[18] R. Müller, S. Kornblith, and G. E. Hinton. "When Does Label Smoothing Help?" In: *NeurIPS*. 2019.

[19] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. "mixup: Beyond Empirical Risk Minimization". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=r1Ddp1-Rb.

[20] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan. *Remix: Rebalanced Mixup.* 2020. arXiv: 2007.03943 [cs.CV].

[21] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné. *Attended Temperature Scaling: A Practical Approach for Calibrating Deep Neural Networks.* 2019. arXiv: 1810.11586 [cs.LG].

[22] C. Padurariu and M. E. Breaban. "Dealing with Data Imbalance in Text Classification". In: *Procedia Computer Science* 159 (2019), pp. 736–745.

[23] A. Fernández, S. Río, N. V. Chawla, and F. Herrera. "An insight into imbalanced Big Data classification: outcomes and challenges". In: *Complex & Intelligent Systems* 3 (2017), pp. 105–120.

[24] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy. "Training deep neural networks on imbalanced data sets". In: *2016 International Joint Conference on Neural Networks (IJCNN).* 2016, pp. 4368–4374.

[25] Y.-A. Chung, H.-T. Lin, and S.-W. Yang. "Cost-Aware Pre-Training for Multiclass Cost-Sensitive Deep Learning". In: *IJCAI.* 2016, pp. 1411–1417.

[26] C. Huang, Y. Li, C. C. Loy, and X. Tang. "Learning Deep Representation for Imbalanced Classification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2016, pp. 5375–5384. DOI: 10.1109/CVPR.2016.580.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[28] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado. "Calibrating Predictive Model Estimates to Support Personalized Medicine". In: *Journal of the American Medical Informatics Association : JAMIA* 19 (), pp. 263–74.

[29] R. Krishnan and O. Tickoo. "Improving model calibration with accuracy versus uncertainty optimization". In: *Advances in Neural Information Processing Systems* (2020).

[30] Yanminsun, A. Wong, and M. S. Kamel. "Classification of imbalanced data: a review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23 (Nov. 2011). DOI: 10.1142/S0218001409007326.

[31] A. Fernández, V. López, M. Galar, M. D. Jesús, and F. Herrera. "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches". In: *Knowl. Based Syst.* 42 (2013), pp. 97–110.

[32] J. Pahikkala. "On contraharmonic mean and Pythagorean triples". In: *Elemente Der Mathematik* 65 (2010), pp. 62–67.

[33] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. "Class-Balanced Loss Based on Effective Number of Samples". In: June 2019, pp. 9260–9269. DOI: 10.1109/CVPR.2019.00949.

[34] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. "Class-Balanced Loss Based on Effective Number of Samples". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2019.

[35] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. "Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss". In: *Advances in Neural Information Processing Systems.* Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[36] B. Zadrozny and C. Elkan. "Transforming classifier scores into accurate multiclass probability estimates". In: *KDD.* 2002, pp. 694–699. URL: https://doi.org/10.1145/775047.775151.

[37] B. Zadrozny and C. Elkan. "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers". In: *ICML.* 2001, pp. 609–616.

[38] J. C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *ADVANCES IN LARGE MARGIN CLASSIFIERS.* MIT Press, 1999, pp. 61–74.

[39] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2015).

[40] L. N. Smith. *A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay.* 2018. arXiv: 1803.09820 [cs.LG].