Calibration as a Proxy for Fairness and Efficiency in a Perspectivist Ensemble Approach to Irony Detection

Samuel B. Jesus
Federal University of Minas Gerais
samuelbrisio@dcc.ufmg.br

Guilherme Dal Bianco
Universidade Federal da Fronteira Sul
guilherme.dalbianco@uffs.edu.br

Wanderlei Junior
Federal University of Minas Gerais
wanderlei-junior@ufmg.br

Valerio Basile
University of Turin
valerio.basile@unito.it

Marcos André Gonçalves
Federal University of Minas Gerais
mgoncalv@dcc.ufmg.br

Abstract

Identifying subjective phenomena, such as irony in language, poses unique challenges, as these tasks involve subjective interpretation shaped by both cultural and individual perspectives. Unlike conventional models that rely on aggregated annotations, perspectivist approaches aim to capture the diversity of viewpoints by leveraging the knowledge of specific annotator groups, promoting fairness and representativeness. However, such models often incur substantial computational costs, particularly when fine-tuning large-scale pre-trained language models. We also observe that the fine-tuning process can negatively impact fairness, producing certain perspective models that are underrepresented and have limited influence on the outcome. To address these, we explore two complementary strategies: (i) the adoption of traditional machine learning algorithms—such as Support Vector Machines, Random Forests, and XGBoost—as lightweight alternatives; and (ii) the application of calibration techniques to reduce imbalances in inference generation across perspectives. Our results demonstrate up to 12× faster processing with no statistically significant drop in accuracy. Notably, calibration significantly enhances fairness, reducing inter-group bias and leading to more balanced predictions across diverse social perspectives.

1 Introduction

In subjective tasks, such as hate speech or irony detection, (text) classification depends on cultural knowledge and the individual impact of the speech on each individual (Basile et al., 2021). An inherent characteristic of this type of problem is label disagreement (e.g., hate vs. non-hate or ironic vs. non-ironic) (Aroyo and Welty, 2015). Therefore, modeling the individuality of perception, reflected in the labels, can provide valuable information for the task of automatic hate speech detection (classification).

Traditional classification methods aggregate multiple annotations through strategies such as choosing the majority class and discarding minority or less representative views (Fleisig et al., 2023). The proposal of **perspectivism** (Cabitza et al., 2023) is to preserve multiple annotations to capture different views, promoting fairness between the models (Frenda et al., 2024a). By training independent models per cultural group, each reflecting specific interpretations, the cultural diversity in the data is considered. A desirable consequence of the perspectivist approach is the mitigation of biases against historically marginalized groups, such as LGBTQ+, black, and religious minorities, among others (Akhtar et al., 2021). In particular, in Casola et al. (2023), a perspectivist method is proposed, combining (or ensembling) models fine-tuned by each perspective, whose results indicate promising combinations. Despite the good effectiveness of the results, fine-tuning multiple language models imposes high computational demands.

In this context, this work has two central objectives. The first objective is to enhance the **efficiency** of the perspectivist approach proposed by Casola et al. (2023) — hereinafter referred to as Confidence-based EnseMble (CEM) — through integration with traditional machine learning models, aiming to maintain effectiveness while reducing computational cost. The second is to improve the **fairness** between perspectivist models through calibration techniques. In the base method, it was observed that some perspectives showed low representativeness (low confidence in predictions), which limits or makes their contribution to the final label unfeasible, compromising the fair principle of perspectivism. We hypothesize that this effect results from miscalibration. In a properly calibrated classification model, the posteriori probability estimated by the classifier should present a higher correspondence with the empirical frequency of hits. Thus, a **calibration** step was incorporated to

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 133–141

increase the reliability of the methods.

The experimental results demonstrate that *combining the CEM with* traditional models reduces execution time by up to 12 times without a statistical loss in effectiveness. We also demonstrate, by means of a sustainability metric that integrates effectiveness and carbon footprint, that a reduction of over 34% is achievable using traditional models (e.g., logistic regression) as the classifier.

Finally, calibration promotes a greater balance in the contribution of the different perspectives in the final result, generating *fairer* models from a perspectivism viewpoint. Indeed, this approach significantly improves the alignment between each perspective's contribution, shifting the contribution distribution closer to its actual perspective´s representations in the training data and reducing unfair imbalances introduced by miscalibrated probabilities. Compared to the original method, the calibrated approach achieved a relative improvement of approximately 39% in fairness, yielding more balanced outcomes across diverse social and linguistic groups while preserving competitive performance. These findings highlight the value of calibration as a key mechanism for ensuring equity in perspectivist modeling.

The rest of the paper is organized as follows. Section 2 covers related work. Section 3 details the proposed approach. Section 4 presents the experimental protocol and discusses the experimental results. Section 5 concludes the paper.

2 Related Work

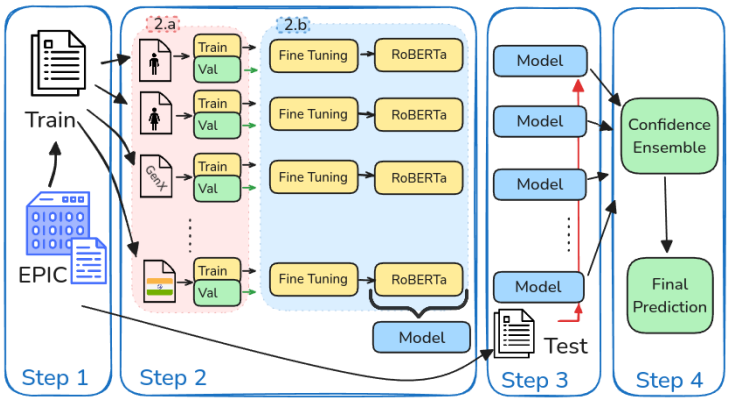
In subjective NLP tasks — such as detecting hate speech, irony, sentiment, and abusive language — obtaining multiple rater annotations is often necessary due to the inherent ambiguity and variability of human judgment (Frenda et al., 2024b). In traditional approaches, disagreements between annotators are frequently treated as noise (Fleisig et al., 2023), with the final label determined by a majority vote scheme that disregards the perspectives of potentially affected minority groups (Akhtar et al., 2021). In contrast, the perspectivist approach advocates valuing this diversity by explicitly modeling individual variations rooted in demographic and cultural characteristics (Basile et al., 2021). This paradigm has gained prominence amid growing demands for fair, inclusive, and bias-aware NLP models (Basile et al., 2021; Fleisig et al., 2023; Akhtar et al., 2021).

Several recent studies have operationalized this concept in practice. In Casola et al. (2023), for example, the authors divided the training data into distinct subsets aligned with specific social or demographic groups (e.g., male and female annotators), fine-tuning a dedicated language model for each group to capture their characteristic patterns. The individual outputs were then combined through a confidence-based ensemble method, yielding a final prediction. Similarly, Fleisig et al. (2023) proposed an approach that explicitly incorporates the target group of an ironic statement by leveraging a dual-module architecture: GPT-2 to identify the group at which the statement is aimed, and RoBERTa to estimate the annotators' scores, with both models adjusted for the specific classification task. Meanwhile, Ngo et al. (2022) introduced a technique that captures individual annotators' patterns by concatenating texts associated with the same annotator and including this information alongside the input for the language model, thereby embedding the annotators' belief profiles within the prediction process.

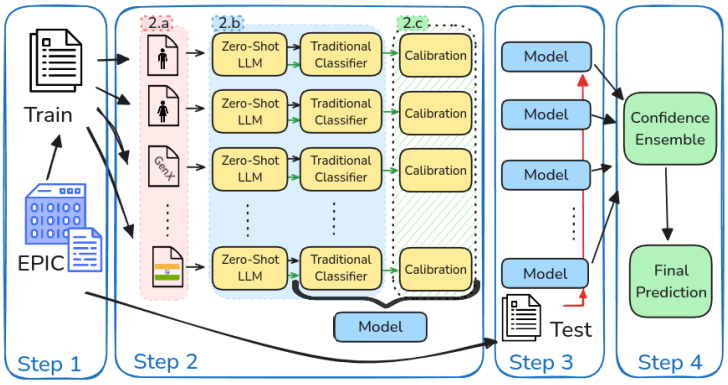
In machine learning, model bias can lead to unfairness and discrimination against specific groups (Ferrara, 2024). Calibration approaches ensure that the balances of positive predictions align with the proportions of positive examples in the training set (Huang et al., 2024). See Kheya et al. (2024) for a survey on methods to reduce the bias. Platt Scaling, for example, is a widely used calibration technique that adjusts a model's output scores into well-calibrated probabilities using a logistic regression model, thereby promoting fairer outcomes (Guo et al., 2017). In recent work, **?** integrates ensemble-based uncertainty estimation with calibration constraints using a multi-objective loss function to address fairness and calibration jointly.

Taken together, these works underscore the growing focus in NLP on recognizing, preserving, and leveraging the richness of diverse human perspectives, yielding advances in both the fairness and reliability of models applied to highly subjective and context-dependent tasks.

That said, to the best of our knowledge, no prior study has examined the impact of calibration in perspectivism or its influence on the accurate and fair representation of social dimensions in the resulting models.



(a)



(b)

Figure 1: Original CEM method (Casola et al., 2023) (a) and CEM method with the proposed changes (b) .

3 Proposed Approach

In this section, we describe how traditional machine learning techniques can be effectively combined with a perspectivist approach to improve computational efficiency. In addition, we present the incorporation of a calibration step designed to promote greater fairness across perspectives, ensuring that their contributions to the final prediction align more closely with their distribution in the training data.

Figure 1a illustrates the perspectivist approach (CEM) introduced in Casola et al. (2023), which is comprised of four sequential steps:

1. The training data are divided into distinct perspectivist subsets based on annotator metadata (*Step 1*), such as gender or nationality.

2. Dense representations are generated for each subset by fine-tuning a pre-trained language model, with this step representing the primary computational cost of the approach (*Step 2*).

3. All resulting models are applied to the same test set, producing independent inferences for each perspectivist subgroup (*Step 3*).

4. The final prediction is computed through an aggregation method, such as: (i) Maximum Confidence (MC), selecting the label with the highest individual confidence score; (ii) Sum of Confidences (SC), summing cross-group scores and selecting the highest total; or (iii) Majority Vote, adopting the label most frequently assigned across perspectives (*Step 4*). Note that the confidence score is computed using the difference between the output probability of the model.

Figure 1b illustrates the proposed adaptations to the baseline approach, introducing a modified Step 2.b and an additional Step 2.c. In Step 2.b, traditional classification algorithms (such as SVM, logistic regression, and XGBoost) are employed in place of the fine-tuned language models used in the original method. Since these algorithms require fixed-length numerical inputs, a pre-trained language model (in this case, RoBERTa) is used as an encoder exclusively to extract features (i.e., using the average from the last four layers), leveraging its ability to encode complex syntactic and semantic patterns into dense vector representations.

The generated embeddings are then used as inputs for training traditional classifiers, allowing the approach to leverage a rich data representation while significantly reducing computational overhead. All subsequent steps remain identical to those defined in the baseline approach, preserving the overall structure of the pipeline while making it more computationally efficient and broadly applicable[1].

The calibration procedure employed in this study is based on *Platt Scaling*, a widely adopted post-processing technique that leverages logistic regression to recalibrate the output scores or probabilities generated by base classifiers (Guo et al., 2017). This method addresses the common issue of miscalibrated probability estimates in machine learning models, where raw output scores (often referred to as logits) do not correspond well to true class membership likelihoods. We also explore the Isotonic calibration approach, which is a non-parametric method that avoids making assumptions about the form of the relationship between the model's scores and the true probabilities (Leathart et al., 2017). However, as noted by Ojeda et al. (2023), a significant drawback of Isotonic calibration is its propensity to overfit, which necessitates a larger calibration set to mitigate this risk.

Concretely, *Platt Scaling* involves fitting a parametric sigmoid function to the scores produced by the classifier on a validation set, distinct from the training data used to build the model. The function is defined as:

$$P(y = 1 \mid s) = \frac{1}{1 + e^{(A \cdot s + B)}}, \qquad (1)$$

where s represents the uncalibrated score or logit output of the classifier, and the parameters A and B are learned by optimizing the logistic regression on the validation set to minimize the difference between predicted probabilities and observed outcomes. The intuition behind this formulation is to transform the classifier's raw output into calibrated probabilities that better reflect the true empirical likelihood of positive class membership. By fitting the sigmoid function, *Platt Scaling* effectively corrects for systematic overconfidence or underconfidence in the model's predictions.

In the context of perspectivist models, where multiple classifiers trained on distinct annotator subgroups contribute to final decisions, such calibration is particularly critical. It ensures that each perspective's predicted probabilities are harmonized, facilitating fairer aggregation and reducing potential biases that arise from disproportionate confidence levels across perspectives.

Calibration introduces an additional step (Step 2.c) that adjusts the probabilities generated by each prediction model to ensure they are on comparable scales. This adjustment prevents any single uncalibrated perspective from disproportionately dominating the label assignment process. In Figure 1b, calibration uses the probabilities derived from inference on the validation set (indicated by the green arrow). Importantly, Step 2.c is orthogonal to the model type and can be applied regardless of whether the underlying classifier is a language model or a traditional machine learning algorithm. This flexibility allows calibration to enhance the reliability and fairness of the ensemble predictions without altering the base classifiers.

4 Experiments

We report the experimental results corresponding to the two primary research objectives: (1) to quantify the computational efficiency gains achieved by **combining zero-shot Roberta for tokenization with traditional machine learning classifiers instead of the RoBERTA finetuning process** and (2) to evaluate the effects of calibration on enhancing fairness within the perspectivist framework. Experiments were conducted on a computing environment comprising an AMD 2990WX processor (64 threads, 3 GHz), a GeForce RTX 2080 GPU (8 GB), and 128 GB of RAM. The source code supporting this work will be made publicly available in the repository at https://...[to be released upon acceptance]. We begin by describing the experimental protocol, including details of the dataset, the evaluation metrics utilized, and our novel metric designed to quantify fairness in perspectivist classification scenarios.

4.1 Dataset

For the experimental evaluation, the *English Perspectivist Irony Corpus* (EPIC) (Frenda et al., 2023) was used. EPIC contains 3,000 records of short messages from *Reddit* and *Twitter*, labeled as ironic or not. Each message is represented by the combination of the post and the reply. The messages were annotated, on average, by five individuals, allowing the capture of variations associated with the

[1]Test data undergoes the same encoding process using the zero-shot language model, ensuring representation consistency and that there is no influence of the training data on the generation of the test representation and vice-versa.

generation, gender, and geographic location of the annotators. It was built to analyze how different cultural and demographic perspectives affect the perception of irony in short online conversations. We chose the EPIC dataset because it allows for a direct comparison with CEM and is one of the few disaggregated datasets that includes annotator metadata.

4.2 Experimental Protocol and Evaluation Metrics

Protocol and Statistical Analysis Each experiment was repeated ten times using different seeds. Each seed produces a random split into training (60%), validation (20%) and test (20%) sets. For replication purposes, seeds were used from the range of 10 to 20. The results include a 95% confidence interval and statistical analysis using the Wilcoxon test with a Bonferroni correction for multiple comparisons. To ensure a consistent training set size across all scenarios, we generated a validation dataset in every case, even when it was not strictly necessary (i.e., discarded).

Effectiveness Effectiveness is measured by the *macro F1-score* (Sokolova and Lapalme, 2009), corresponding to the simple average of the F1-scores per class, giving equal weight to all. We chose macro-F1 as the data is very skewed, with approximately 70% of the instances belonging to the non-ironic class. The F1-score is computed on the aggregated test set – average of the results on the 10 test sets.

Efficiency Efficiency is evaluated based on the total time (in seconds) equivalent to the sum of the times of the tokenization, training, prediction, and calibration processes (when applicable), comparing approaches with and without perspectivism.

Sustainability To measure the tradeoff between effectiveness and the eco-sustainability of the approaches, we use the *Carburacy metric* (Moro et al., 2023). Such a metric combines the effectiveness score and CO2 emissions into one score, considering the eco-sustainability of each approach. The definition of the *Carburacy metric* is described below:

$$\Upsilon = \frac{e^{\log_\alpha \mathcal{R}}}{1 + \mathcal{C} \cdot \beta} \tag{2}$$

where the effectiveness (R), represented by the F1-score, is combined with the normalized carbon cost (C). The trade-off between R and C is governed by the hyperparameters α and β, which weigh the F1-score and the carbon penalty, respectively. We define α and β as 10 and 1, as suggested in the original work. We measure the carbon cost (C) using the *eco2AI library* (Budennyy et al., 2022), which estimates emissions based on CPU and GPU energy consumption.

Fairness From a Perspectivist Approach To evaluate fairness in the presence or absence of calibration, we introduce a metric designed to assess the relative contribution of each perspective within the ensemble to the final label assignment, grounded in the distribution of samples across perspectives in the training set. The core intuition behind this metric is to compare the expected influence of each perspective—based on its prevalence in the training data—with its actual impact on the ensemble's output, as derived from classifiers trained independently for each perspective. In a fair system, perspectives that are underrepresented in the training data should naturally exert less influence on the ensemble decision. In contrast, more frequently represented perspectives should have a proportionally greater impact. Thus, fairness is characterized by the alignment between a perspective's frequency in the training set and its corresponding contribution to the final prediction. Calibration plays a key role in achieving this proportionality, mitigating distortions that may arise from imbalanced, overfitted, or overtuned individual classifiers within the ensemble.

Concretely, the first step in applying the proposed metric involves computing the ideal contribution of each perspective to the ensemble's predictions, based on their representation in the training set. When the dataset is structured along multiple dimensions—for example, *Gender*, *Generation*, and *Nationality*—each dimension is expected to contribute equally to the overall decision. In a scenario with three such dimensions, each would ideally account for approximately 33% of the ensemble's output. Within each dimension, the expected contribution of individual perspectives (e.g., *male* and *female* under the *Gender* dimension) is determined by their relative frequency in the training data. For instance, if the training set consists of 40% *male* and 60% *female* samples, then a fair ensemble should reflect this distribution in its predictions for that dimension. Table 6 presents the computed "Ideal" contributions for each perspective,

Table 1: F1-score (± CI) without calibration for each aggregation strategy and model. '*' and '↓' indicate a statistical tie or loss compared to RoBERTa.

Without Calibration	RoBERTa	LR	XGB	SVM
Maximum Confidence (MC)	67.4 ± 1.5	$64.3 \pm 1.0 \downarrow$	$54.0 \pm 0.8 \downarrow$	$48.7 \pm 1.2 \downarrow$
Sum of Confidences (SC)	66.6 ± 1.7	64.7 ± 1.2 *	$54.0 \pm 1.0 \downarrow$	$48.6 \pm 0.8 \downarrow$
Majority Vote	65.0 ± 2.0	64.2 ± 1.3 *	$53.7 \pm 0.6 \downarrow$	$48.7 \pm 1.2 \downarrow$
Without Perspectives	64.5 ± 2.5	63.5 ± 1.2 *	$55.5 \pm 1.3 \downarrow$	$49.1 \pm 1.3 \downarrow$

Table 2: F1-score (± CI) with Platt calibration for each aggregation strategy and model. '*' and '↓' indicate a statistical tie or loss compared to RoBERTa.

With Platt Calibration	RoBERTa	LR	XGB	SVM
Maximum Confidence (CM)	67.0 ± 1.8	64.7 ± 1.1 *	$60.9 \pm 1.3 \downarrow$	63.1 ± 0.5 *
Sum of Confidences (SC)	67.2 ± 1.7	$\mathbf{65.2 \pm 1.0}$ *	$62.9 \pm 1.2 \downarrow$	64.3 ± 0.9 *
Majority Vote	67.0 ± 1.5	64.8 ± 1.5 *	$62.2 \pm 1.3 \downarrow$	64.4 ± 0.7 *
Without Perspectives	65.1 ± 1.7	62.1 ± 1.6 *	$60.4 \pm 1.6 \downarrow$	$60.0 \pm 1.4 \downarrow$

capturing the expected number of predictions proportionally aligned with both dimension-level balance and intra-dimension frequency distributions.

$$\sqrt{\sum_{i=1}^{n} (Ideal - X_i)^2} \tag{3}$$

Equation 3 formalizes this assessment by quantifying the squared deviations between the actual contribution of each perspective (X) and its ideal distribution ($Ideal$) for both calibrated and uncalibrated models. The use of squared differences serves to emphasize larger discrepancies, ensuring that significant imbalances have a proportionately greater influence on the resulting metric. In this formulation, a value of 0 denotes the ideal case, where every perspective's contribution to the final prediction is fully aligned with its expected distribution, indicating a balanced and fair decision-making process across all perspectives.

4.3 Experimental Results

4.3.1 Effectiveness

Table 1 summarizes the effectiveness comparison between the CEM approach—employing RoBERTa **with fine-tuning and zero-shot RoBERTa for tokenization combined with traditional machine learning classifiers**, including Logistic Regression (LR), XGBoost, and Support Vector Machines (SVM). For simplicity, when we refer to traditional machine learning classifiers, we use zero-shot RoBERTa to generate embeddings as features for the classifier. Both RoBERTa and LR achieved the highest *Macro F1-score* values, with no statistically significant difference between them, except under the Maximum Confidence aggregation method, where RoBERTa demonstrated a modest but statistically significant advantage of 2.8 percentage points. The effectiveness of LR can be attributed to its ability to model linear relationships between features (Hassan et al., 2022). Conversely, XGBoost and SVM exhibited inferior performance, with reductions exceeding 9% relative to the top-performing models. The superior performance of LR is likely attributable to its efficacy in modeling linear relationships.

Table 2 present the effectiveness results following the application of *Platt Scaling* The *Platt Scaling* demonstrates improvements across all evaluated models except for LR, which inherently produces calibrated outputs (Cunha et al., 2025). Nevertheless, calibration contributed to a reduction in variance for LR in certain scenarios, notably under the Sum of Confidences aggregation method. Traditional classifiers, specifically *XGBoost* and *SVM*, exhibited the most substantial gains, with increases in *F1-score* reaching up to 10 and 16 percentage points, respectively. RoBERTa showed a marginal improvement with the Majority Vote method, although this increase did not achieve statistical significance. Notably, post-calibration, LR achieved a statistical tie with RoBERTa across all aggregation strategies. Additionally, calibration consistently enhanced the performance of the Sum of Confidences (SC) aggregation method, which yielded the highest results among all classifiers.

4.3.2 Efficiency

Table 3 details the computational time required by the evaluated methods, comparing both the perspectivist and non-perspectivist approaches employing RoBERTa or a **zero-shot RoBERTa with traditional machine learning classifiers**. The results indicate that traditional models achieve

Table 3: Execution time (in seconds) with 95% confidence intervals without calibration.

Time	RoBERTa	LR	XGB	SVM
Without-Perspectives	239.8 ± 13.7	16.5 ± 0.0	18.3 ± 0.1	22.8 ± 0.1
With-Perspectives	1904.7 ± 70.3	136.2 ± 0.5	154.1 ± 0.3	164.2 ± 0.5

Table 4: Execution time (in seconds) with 95% confidence intervals with calibration.

Time - Calibration	RoBERTa	LR	XGB	SVM
Without-Perpectives	245.2 ± 13.8	16.6 ± 0.1	18.9 ± 0.1	20.3 ± 0.1
With Perspectives	1951.4 ± 70.3	137.3 ± 0.4	161.0 ± 0.5	155.9 ± 0.8

processing speeds up to twelve times faster than RoBERTa. A comparative analysis between Tables 3 and 4 reveals that incorporating calibration incurs a negligible increase in execution time, with only a 47-second (approximately 2%) overhead for RoBERTa and a 7-second (approximately 4%) increase for XGBoost. Interestingly, SVM exhibits a reduction of 8.3 seconds (approximately 5%) in runtime, which may be attributed to a decreased training set size due to the allocation of a data portion for calibration via *Platt Scaling*.

In summary, the experimental findings demonstrate the feasibility of significantly reducing computational time without compromising predictive performance. Concurrently, the integration of calibration facilitates the generation of more equitable inferences that appropriately represent minority perspectives, thereby advancing the core objective of perspectivist methodologies.

4.4 Sustainability

Table 5 summarizes the sustainability score obtainable using the Carburacy metric (which combines the F1-score and carbon footprint). Due to space constraints, we report here only the two best-calibrated models (LR and RoBERTa) based on the highest F1 score (Section 4.3.1). The table shows a significant advantage of using LR over RoBERTa in all aggregation strategies with an average difference of 35%. This is explainable by the fact that LR uses only a fraction of the time needed to fine-tune the LLM, resulting in a much lower carbon footprint with a statistically tied F1 value. In summary, considering the combined benefits (model effectiveness and environmental impact), the cheaper and more effective LR model offers a considerable gain over the original approach.

4.4.1 Fairness

Table 6 presents the distribution of final label assignments across the different perspectives, based on the *Max Confidence* decision rule, where the perspective yielding the highest estimated probability

Table 5: Carburacy for LR and RoBERTa using the MC, SC, and Majority Vote.

Method	LR	RoBERTa
MC	0.775	0.422
SC	0.778	0.423
Majority Vote	0.776	0.422

determines the prediction. Columns "Non-Calibrated" and "Calibrated" correspond, respectively, to the original and calibrated approaches, both implemented using the RoBERTa classifier. The results reveal that certain perspectives (specifically, *Boomer*, *Female*, and *GenY*) have no measurable influence on the final prediction, indicating that they are effectively ignored by the classifier and only introduce unnecessary computational cost. In contrast, the decision process is dominated almost exclusively by two perspectives (*Ireland* and *GenX*), suggesting an implicit bias toward these dimensions. Understanding the reasons for this disproportionate utilization of specific perspectives constitutes an intriguing open question, which we leave as a direction for future investigation.

Following the application of the proposed metric, the original (non-calibrated) approach yielded an overall fairness score of 53, whereas the calibrated approach achieved a score of 33. Recall that, for this metric, the lower, the fairer. This result reflects a relative improvement of approximately 39% in fairness when calibration is applied. For example, while the ideal contribution for the *India* perspective is 4.5%, in the non-calibrated model, this perspective contribution is null, compared to 2.3% in the calibrated model. On the other hand, the influence of the *Ireland* and *Gen X* perspectives in the final decision decreased significantly, becoming closer to the ideal values, according to the proposed reasoning. These findings indicate that calibration significantly improves alignment between observed and ideal contributions, resulting in a more balanced and equitable prediction process across perspectives.

In summary, the results demonstrate that calibration promotes a more balanced and representative contribution from all perspectives to the final prediction, yielding a fairer and more inclusive outcome. By aligning the influence of each perspective with its actual representation in the training data, the calibrated approach mitigates disproportionate dominance by specific groups and reduces the risk of systemic bias. This improvement not only strengthens the reliability and interpretability of the model's decisions but also advances its suitability for applications where equitable treatment across diverse groups is a critical requirement.

Table 6: Training set size for each perspective, its ideal and actual contributions to the final predictions for both, non-calibrated and calibrated approaches.

Perspective	|Training| Size|	Ideal	Non-Calib.	Calib.
Australia	1363	5.2	3.6	17.9
India	1180	4.5	0	2.3
Ireland	1288	4.9	48.9	26.9
Male	2026	7.8	3.2	8.1
United kingdom	1369	5.3	6.7	1.3
United State	1368	5.2	9.9	14.6
GenX	1755	10.9	25.8	17.9
GenY	1971	12.2	0	4.5
GenZ	1151	7.1	1.8	0.5
Boomer	447	2.8	0	0.4
Female	1971	16.3	0	5.6
Male	2026	16.7	3.2	8.1

5 Conclusions

We propose integrating traditional classification methods as a way to simultaneously foster greater fairness and improved computational efficiency within recent perspectivist approaches. Our results demonstrate that, due to miscalibrated probabilities, the method introduced by Casola et al. (2023) tends to produce biased outcomes, under-representing certain perspectives and, as a result, falling short of its central objective of promoting inclusivity and equity. To mitigate this limitation, we incorporated a calibration step as an orthogonal layer, allowing the model to more accurately align its final prediction distribution with the actual representation of each group in the training data. This adjustment not only improves balance across perspectives, yielding a fairer and more representative outcome, but also achieves competitive levels of effectiveness when compared with state-of-the-art approaches. In this way, the proposed method advances the state of the art by reconciling the often competing demands of efficiency, performance, and fairness.

Looking ahead, we intend to investigate the application of our framework to other perspectivist datasets, exploring a broader range of social, linguistic, and cultural contexts. Also, more recent language models, including Llama and its variants, as well as modern BERT-based architectures. We want to dig deeper into the reasons why certain perspectives seem to dominate the ensemble's decision, not reflecting their ideal contributions. We will also evaluate other supervised stacking techniques (Gioacchini et al., 2024) as a means to further optimize effectiveness while reducing computational overhead and improving fairness even further, thereby supporting the design of more equitable and resource-efficient NLP systems.

Limitations

While the training data and the learning process consider different perspectives through disaggregated labels, the evaluation is conducted on an aggregated test set. This limitation may have some impact on the experimental results; however, we have chosen to follow the original work's methodology of Casola et al. (2023). One possible future direction to avoid using the aggregate test set is to evaluate the individual predicted labels for each instance (Mostafazadeh Davani et al., 2022). For instance, the predictions produced by the model trained with "GenX" will be matched with annotators who belong to the same perspective.

Acknowledgment

This work was supported by CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde, and the National Institute of Science and Technology in Artificial Intelligence Responsible for Computational Linguistics, Information Processing, and Dissemination (INCT-TILD-IAR; 408490/2024-1).

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896.

L.M. Aroyo and C.A. Welty. 2015. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and*

Future, pages 15–21, Online. Association for Computational Linguistics.

S. A. Budennyy, V. D. Lazarev, N. N. Zakharenko, A. N. Korovin, O. A. Plosskaya, D. V. Dimitrov, V. S. Akhripkin, I. V. Pavlov, I. V. Oseledets, I. S. Barsola, I. V. Egorov, and 1 others. 2022. Eco2ai: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*, 106:S118–S128.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA*, pages 6860–6868. AAAI Press.

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Washington Cunha, Alejandro Moreo Fernández, Andrea Esuli, Fabrizio Sebastiani, Leonardo Rocha, and Marcos André Gonçalves. 2025. A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2).

Emilio Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1).

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024a. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024b. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Luca Gioacchini, Welton Santos, Barbara Lopes, Idilio Drago, Marco Mellia, Jussara M. Almeida, and Marcos André Gonçalves. 2024. Explainable stacking models based on complementary traffic embeddings. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 261–272.

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

Sayar Ul Hassan, Jameel Ahamed, and Khaleel Ahmad. 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, 3:238–248.

Yu Huang, Jingchuan Guo, Wei-Han Chen, Hsin-Yueh Lin, Huilin Tang, Fei Wang, Hua Xu, and Jiang Bian. 2024. A scoping review of fair machine learning techniques when using real-world data. *Journal of Biomedical Informatics*, 151:104622.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. The pursuit of fairness in artificial intelligence models: A survey. *arXiv preprint arXiv:2403.17333*.

Tim Leathart, Eibe Frank, Geoffrey Holmes, and Bernhard Pfahringer. 2017. Probability calibration trees. In *Asian conference on machine learning*, pages 145–160. PMLR.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14417–14425.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. 2022. StudEmo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association.

Francisco M Ojeda, Max L Jansen, Alexandre Thiéry, Stefan Blankenberg, Christian Weimar, Matthias Schmid, and Andreas Ziegler. 2023. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, 42(29):5451–5478.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.