Embracing Diversity: A Multi-Perspective Approach with Soft Labels

Abstract. Prior studies show that adopting the annotation diversity shaped by different backgrounds and life experiences and incorporating them into the model learning, i.e. multi-perspective approach, contribute to the development of more responsible models. Thus, in this paper we propose a new framework for designing and further evaluating perspective-aware models on stance detection task,in which multiple annotators assign stances based on a controversial topic. We also share a new dataset established through obtaining both human and LLM annotations. Results show that the multi-perspective approach yields better classification performance (higher F1-scores), outperforming the traditional approaches that use a single ground-truth, while displaying lower model confidence scores, probably due to the high level of subjectivity of the stance detection task.

Keywords. Natural Language Processing, Human-Centered AI, Responsible AI, Perspectivism, Crowdsourcing

1. Introduction

Recent advancements in the field of Natural Language Processing (NLP) have underscored the importance of annotator disagreement, redefining it as a meaningful source of information regarding the task, the data and the annotators [1], rather than dismissing it as noise . Research has shown that Large Language Models (LLMs) may exhibit biases that align with dominant Western perspectives [2] exposing inequalities that could negatively impact underrepresented communities, whose voices are often drowned out by majority opinions [3]. As LLMs evolve alongside humans, aligning them with human preferences becomes a crucial aspect of their design process [4,5]. To address this challenge, *Perspectivism* [6], a growing approach in NLP community, leverages disaggregated datasets, where all individual annotators label are included, to capture human disagreements, promoting inclusion of diverse viewpoints. This new schema, rather than relying on aggregated labels- such as those obtained through majority voting- utilizes the diversity of human opinions, allowing models to learn from human disagreement [7,8], thereby avoiding the marginalization of minority voices. In line with the perspectivist paradigm, we propose a new framework to design and further evaluate the multi-perspective approach on stance detection, specifically about controversial and subjective topics. The main goal of this paper is to create perspective-aware by design models that incorporate human disagreement into the model learning phase in a more responsible way.

Contribution Given the context, this study aims to explore the effectiveness of the multi-perspective approach. Specifically, we examine whether this methodology can enhance overall model performance and confidence. Our contributions are as follows:

1. We introduce a new version of the stance detection dataset on controversial topics based on [9], augmented with document summaries and LLM annotations.
2. We employ two distinct methodologies: a baseline approach that utilizes aggregated labels and a multi-perspective approach designed to incorporate minority viewpoints by representing labels in a more nuanced manner i.e. soft labels.
3. We evaluate if the multi-perspective approach leads to improved model performance compared to relying solely on aggregated labels.

2. Related Work

In this section, we set the foundation for our pipeline by combining insights on perspectivism, soft labels, and model uncertainty, while exploring how LLMs can act as annotators to capture diverse perspectives.

Disaggregated datasets In human-labeled datasets, annotations are typically gathered through crowdsourcing, where crowd workers on specific platforms, like MTurk [1] or Prolific [2], are asked to provide their opinion on a given statement. In such contexts, especially when the task is subjective and no single ground truth may exist, crowd workers may disagree for various reasons, such as subjective bias or the ambiguity of the instance [1]. Consequently, recent studies have addressed this issue from a perspectivist standpoint, collecting each annotator's label to account for a range of diverse opinions, leading to the use of disaggregated datasets [10,11]. To promote this approach, the NLP community has recently released a list of publicly available perspectivist datasets[3].

Learning from Soft Labels Soft labels provide a recent alternative to hard labels, i.e. aggregated labels, which are frequently criticized for oversimplifying complex data. While one-hot encoding is used to assign a single, definitive value to each data point in hard labels, soft labels capture a range of possible values. This renders the data more nuanced and better accounts for ambiguities and divergent viewpoints in annotations, reflecting the inherent uncertainty and variability in human judgment. Previous studies modeled human diverse annotations using soft labels [12,13] achieving superior model performances improving also robustness and generalization [14].

Model Uncertainty Different human judgments on subjective tasks introduce uncertainty into LLMs. The black-box nature of LLMs poses challenges in understanding how these models handle disagreements in text classification and generation [15]. [16] identify three sources of model uncertainty: the user input, the model architecture, and the final output. While accuracy is conventionally assessed with respect to the majority class, incorporating model uncertainty through representing multiple perspectives may be more beneficial.

[1] https://www.mturk.com
[2] https://www.prolific.com
[3] https://pdai.info

Leveraging LLMs as annotators Lately, LLMs have demonstrated impressive capabilities in semantic understanding [17,18], as they easily interact with users, both in few-shot and zero-shot scenarios. The current trend involves employing LLMs in various roles, such as acting as annotators to perform a wide range of tasks [19,20]. Although this procedure often requires substantial resources and domain expertise, cutting-edge LLMs like GPT-4 [21], LLama-3 [22] present viable substitutes, albeit with drawbacks of their own [23]. Several works have explored the labeling capabilities of LLMs for subjective tasks, including stance detection, hate speech detection, and narrative analysis [24,25]. While these models are frequently fine-tuned to align with human preferences, current research have investigated whether these models accurately represent human disagreements [26,23]. The cost of using LLMs as annotators is a significant advantage over employing humans, but it is important to recognize that they may introduce biases [27].

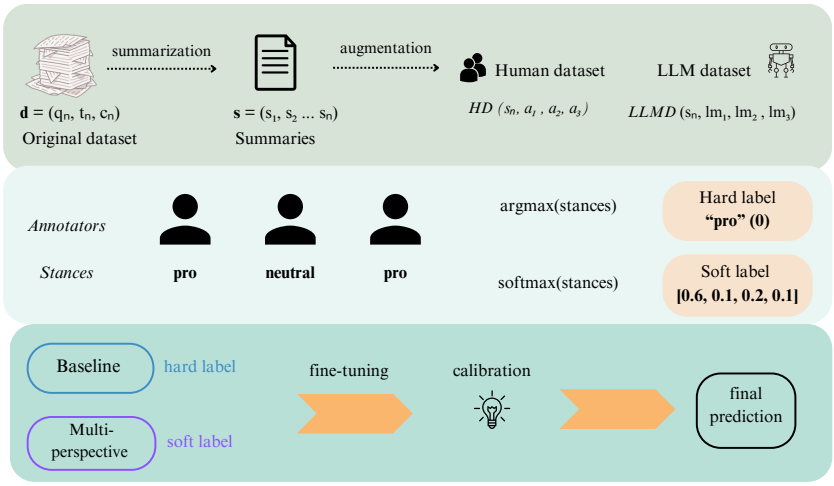

Figure 1. The multi-perspective framework for the stance detection task includes the dataset preparation phase with the summarization and further augmentation steps via obtaining LLM annotations. Then, annotations are transformed into hard and soft labels, model fine-tuning is fulfilled and classifier's final prediction scores are calibrated.

3. Methodology

We propose a multi-stage framework specifically for the stance detection task, as illustrated in Figure 1. The stance detection task which aims to automatically determine the writers' stance (perspective, or viewpoint) towards a target in a given textual content. In the scope of this study, the target is a claim about the corresponding controversial topic and the author may support the claim with a stance of *pro*, disagree with it by taking the stance of *against*, or choose not to have a clear stance which is *neutral*, i.e. neither agree, nor disagree with it. Thus, the stance can take three different values as *pro, against, or*

neutral towards a given claim, and additionally *not-about* which means that the author's statement is not relevant with the given claim (target). To investigate the effect of the perspectivist approach on stance detection task, we use two different methodologies: *Baseline* model with hard labels and *Multi-Perspective* model with soft labels (the overview of these methodologies can be found in Figure 1).

The proposed pipeline[4] consists of the following steps: *Step 1* is the summarization of the documents in the original dataset, *Step 2* is the data augmentation via obtaining LLM annotations to create two distinct datasets as human (HD) and LLM dataset (LLMD) and *Step 3* is the model fine-tuning, along with the calibration (Section 4 for more details).

3.1. Baseline Model

In traditional machine learning settings, label aggregation techniques such as *majority voting* are typically applied to create a single label for each data instance. In the baseline model, we follow the traditional paradigm in which the majority label that is the most frequent label among the multiple annotations provided by the annotators is created and used for each data instance. Majority labels are aggregated using *hard labels* that are in traditional binary classification settings encoded as 0 or 1. In our multi-class scenario, we refer to the majority label as the index of the most common option, represented as the hard label. Each index corresponds to a specific stance label in the following order: *pro* (0), *against* (1), *neutral* (2), and *not-about* (3). An example about data transformation is illustrated in Figure 1.

3.2. Multi-Perspective Model

In the multi-perspective approach, unlike the baseline, a majority label is not generated, instead the multi-perspective model uses disaggregated labels. These disaggregated labels initially represented as discrete values, are converted into continuous values through a softmax function namely *soft labels* [28]. The advantage of using soft labels is that they represent a probability distribution over the possible classes, which can enhance the model performance, particularly in subjective tasks where annotator choices may differ significantly [29]. Since the baseline and multi-perspective approaches handle the dataset design differently, the multi-perspective approach applies the soft loss [28] instead of the standard cross entropy loss. This choice stems from the need to represent the distribution of human labels in a more nuanced way. The soft loss is defined as:

$$-\sum_{i=1}^{n}\sum_{c} p_{\text{hum}}(y_i = c \mid x_i) \log p_\theta(y_i = c \mid x_i)$$

where $p_{\text{hum}}(y \mid x)$ represents the human label distribution (i.e. soft labels) which is obtained by applying the softmax function to the logits produced by the classifier.

[4]This framework is adaptable to different applications, for example summarization can be replaced with paraphrasing if necessary.

4. Experimental Setup

This section outlines the technical details of the conducted experiments. Our code and results are publicly available at `https://anonymous.4open.science/r/perspectivism-0473`. We first describe the overall pipeline as displayed in Figure 1. The original dataset texts (already provided with human hard labels) were first summarized and then further augmented with additional annotations from LLMs. This process resulted in the creation of two distinct datasets: HD, containing the original human-derived annotations, and LLMD, incorporating annotations generated by LLMs. Then, these annotations were converted into hard labels, for the Baseline, and soft labels for the Multi-Perspective model, i.e. with the aim of representing the diverse perspectives in a more fine-grained manner. Subsequently, model fine-tuning was performed, and classifier predictions were calibrated.

4.1. Original Dataset

For this study, we built upon the work of [9]. The dataset comprises the top 10 news search results retrieved from Google and Bing in response to 57 queries on controversial topics. These topics range from education, health, and entertainment to religion and politics, all of which are known to evoke diverse perspectives and opinions. Controversial issues in these domains often touch on deeply held values, beliefs, and societal debates, making the dataset particularly suitable for a study rooted in perspectivism, where differing viewpoints are central to the analysis. Each dataset instance is composed of a query (about a controversial topic), document title and the textual content with respect to a given query of the varying lengths, i.e. 1.200 tokens on average and extend up to 7.000 tokens. Each document (title and content) with respect to the given query has been annotated by three annotators on MTurk (Figure 1). We use this stance detection dataset after the data augmentation steps below to apply our proposed multi-perspectivist approach in the scope of subjective tasks. The reported Fleiss-Kappa[5] score of 0.35 and inter-rater agreement score of 0.49 on the original dataset highlight the subjectivity and ambiguous nature of the stance detection task. Note that each document has not been annotated by the same three annotators due to the design choices which leads to a more enriched dataset with diverse opinions (annotations).

4.2. Dataset Augmentation

Summarization Since the original dataset contains long documents with varying lengths, we decided to first apply summarization due to the maximum input length of transformer-based models, namely BERT [30] and RoBERTa [31]. Both of these models have the maximum input length of 512 and summarization (instead of truncation) can provide more enriched contents especially for the long documents. We applied summarization only on those documents (title and content) longer than 800 tokens (empirically determined) since summarization might lead to information loss on shorter documents.

In our initial summarization experiments, we employed various models, including Pegasus-CNN-DailyMail, BART-large-CNN, and Falcon-7b-Instruct. To assess the qual-

[5]Fleiss Kappa is a statistical measure of agreement which is an extended version of Cohen's Kappa (only for two raters) that takes into account the agreement due to chance as well.

Box 1: Text Summarization

You are a professional summarizer. Create a concise and comprehensive summary of the provided text, as if it were an article. Limit the summary to a maximum of 800 words.

(a) Prompt for text summarization.

Box 2: LLM annotations

You are an expert annotator, chosen for a task of annotating texts on subjective topics. Please annotate the following texts with one of these labels, according to your perspective. Please consider also the related query and title. Labels: 'Pro', 'Neutral', 'Against', 'Not-about'.

(b) Prompt for LLM-based annotations.

Figure 2. GPT4-Turbo prompts for text summarization (a) and for LLM annotations (b).

ity of these summaries, we compared them using the ROUGE score, as shown in Table 4a, Table 4b and Table 4c.

Pegasus-CNN-DailyMail Table 4a shows the summarization performance of Pegasus-CNN-DailyMail. The model exhibits a high precision but low recall and F1 scores across ROUGE metrics, indicating limited content coverage.

BART-large-CNN BART-large-CNN outperforms Pegasus-CNN-DailyMail (Table 4b), with higher recall and slightly lower precision scores. Better recall scores indicate that BART-large-CNN retrieves the content more effectively, while slightly lower precision scores suggests a trade-off, as the summaries may include less relevant details.

Falcon-7B-Instruct Falcon-7B-Instruct is aligned with the precision and recall results of BART-large-CNN, indicating a comparable performance in summarization. These similar results suggest that Falcon-7B-Instruct also effectively balances content coverage and relevance.

GPT-4-turbo Based on the automated evaluation and human evaluation of the generated summaries of the aforementioned models, we decided to use GPT-4-Turbo via OpenAI batch API [6]. For the detailed comparative evaluation results we leveraged different metrics, namely ROUGE, i.e. measuring the overlap of n-grams between the generated summary and the reference, using an automated package [32], BERTScore [33], i.e. semantic similarity, and BLEU score [34], i.e. n-gram similarity. Overall model evaluation results display that the GPT-4 Turbo model has a moderate performance on the ROUGE score, high performance on the BERTScore, whereas it shows a low performance on the BLEU score. Based on these results, we decided to use GPT-4 Turbo for the summarization phase, since BLEU score is typically used for machine translation and not suitable for summarization tasks. The evaluation of the GPT-4 summaries is presented in Table 1.

Since prompting is known to highly affect the model performance on a wide range of tasks[8], we conducted various experiments with prompt engineering for two main pur-

[6]Note that due to the maximum input length constraints of these models, we chunked the input text to be fed into these models except GPT-4 turbo[7] owing to its larger maximum input length.

[8]`https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions`

poses: first to improve summarization using the GPT-4-turbo model, and second, to obtain LLM-based annotations. Two selected prompts after experimenting with different prompts [9] are illustrated, respectively, in Figure 2 . Considering the overall performances of the aforementioned models followed by a human evaluation (due to the repetition of information in previous summaries), we decided to switch to GPT-4. GPT-4-turbo exhibits a modest precision score, yet low recall and F1 scores. However, high BERT-score suggests that the generated summaries are semantically relevant. As expected, the BLUE score is low due to matching n-grams, where word-to-word matching is not typically required.

Metric	Precision	Recall	F1 Score
ROUGE-1	0.4998	0.1774	0.2549
ROUGE-2	0.1770	0.0520	0.0779
ROUGE-L	0.4571	0.1620	0.2329
BERT-Score	**0.8503**	0.8357	0.8429
BLEU Score	0.0171	-	-

Table 1. ROUGE score on GPT-4-turbo's summaries

Augmentation via LLM annotations To establish models that are more responsible and with a higher capability of representing diverse perspectives, we experimented with three different LLMs. We opted for open-source SOTA models, in particular LLama-3-8b [22], Mistral-7b [35] and Olmo-7b [36][10].

Based on the label distribution obtained by the LLMs on the training, validation, and test sets, the *pro* is the most frequently assigned stance label by all three LLMs (37%, 50%, and 69% for Olmo, LLama-3, and Mistral respectively), while Olmo exhibits a significantly higher percentage for the *against* (37% vs. 22% and 15%). The four label distribution charts including the LLM with majority vote can be found in Figure 3 and Figure 4 in Appendix 6.

Furthermore, the percentage of full percentage of agreement, i.e. defined as the proportion of the cases where all annotators concur exactly on the same label, is quite low with 11%. The low agreement score suggests a high level of disagreement among the LLMs, probably due to the difficulty of the annotation task and subjective nature of the dataset, i.e. which discusses debatable controversial topics. Similarly, Cohen's Kappa scores that were calculated for all LLM annotations in a pairwise manner also confirm low level annotator agreement.

Before summarization, each instance d_i consists of $\{q_i, t_i, c_i\}$. From this, we first concatenated the document title t_i and content c_i and summarized it into s_i (Step 1, as detailed in Section 4.2). After obtaining the summary s_i, we designed two distinct datasets: (i) the Human Dataset (HD), which consists of $HD = \{s_i, a_1, a_2, a_3, maj\}$, where a represents crowd-annotated labels (obtained via crowdsourcing) and maj indicates the major-

[9] For LLM annotations collection we designed a prompt in zero-shot settings to minimize any potential bias

[10] We selected the listed LLMs based on their availability and GPU constraints. The models were loaded onto the GPU using half-precision (float16).

ity labels, as described in Section 3.1; and (ii) the LLM-Annotated Dataset (LLMD), defined as $LLMD = \{s_i, lm_1, lm_2, lm_3, maj_{lm}\}$. The structure of LLMD is identical to HD, with the difference being that the annotations lm are obtained from LLMs rather than human annotators, with the majority label maj_{lm} similarly derived from LLM annotations.

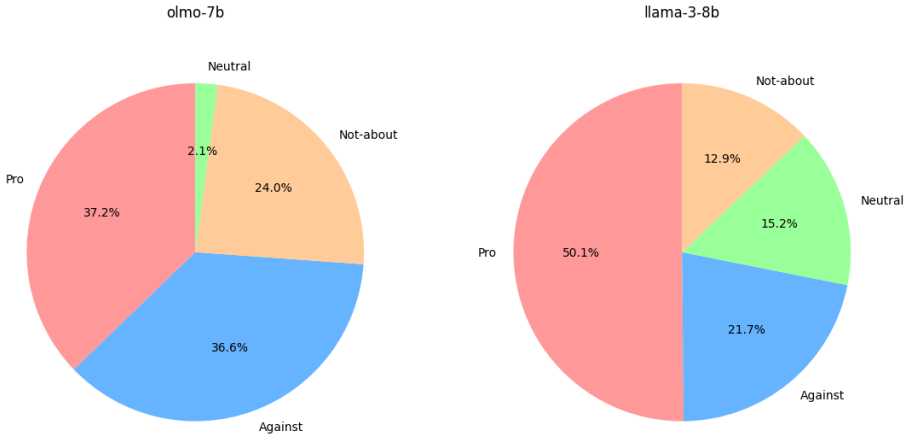

Figure 3. Olmo-7b and Llama 3-8b label distribution (train, val, test)

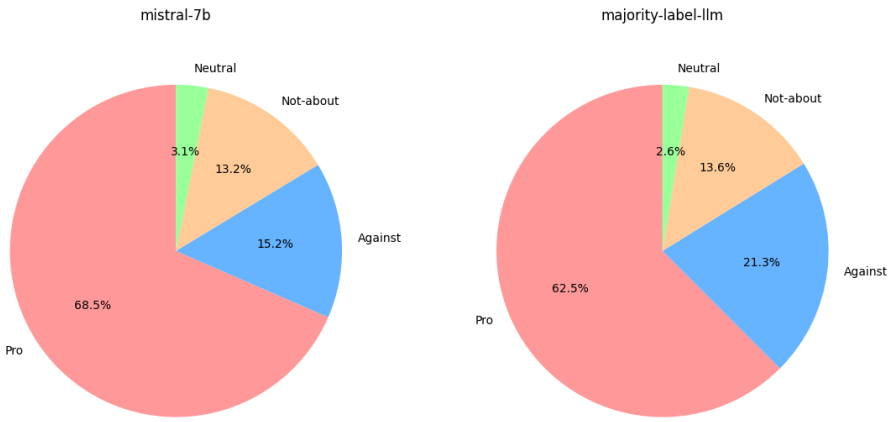

Figure 4. Mistral-7b and Majority label LLM distribution (train, val, test)

Approach	Dataset	Model	Acc.	Prec.	Rec.	F1	Avg. Conf.
Baseline	HD	BERT-large	36.69	39.03	35.93	33.80	40.20
	HD	RoBERTa-large	56.11	61.11	58.04	**57.22**	57.25
	LLMD	BERT-large	60.78	15.50	24.60	19.01	<u>60.59</u>
	LLMD	RoBERTa-large	61.76	15.44	25.0	19.09	60.44
Multi-Perspective	HD	BERT-large	46.76	46.88	47.16	46.75	45.82
	HD	RoBERTa-large	60.43	63.55	62.83	**61.90**	<u>48.76</u>
	LLMD	BERT-large	61.76	15.44	25.0	19.09	30.42
	LLMD	RoBERTa-large	61.76	15.44	25.0	19.09	30.13

Table 2. Comparative evaluation results of different approaches and models.

Dataset Preprocessing The dataset contains 1026 instances in total. Then, we applied the preprocessing steps to clean the dataset. We removed the null documents, then the ones with link-broken label, i.e. that are not accessible, and finally the documents without any majority label, i.e. all annotator disagree on the ground truth. After removing the instances with no majority label, 897 instances left and the dataset was then splitted as 619 for training, 139 instances for validation and 139 instances for testing. This the actual dataset size of HD. Instead, LLMD contains 704 instances in total as 505 instances for training, 97 instances for validation, and 102 for testing. The reason of the dataset size differences of HD and LLMD results from the fact that there were more instances with no majority label based on the LLM annotations. Thus, to make a comparative evaluation of the baseline and multi-perspective approaches, we had to discard more instances for the LLMD.

4.3. Model Learning

We prepared the dataset for fine-tuning by combining the q_i and s_i of each dataset instance as input for both the baseline model (Section 3.1) and the multi-perspective model (Section 3.2). Notably, while the baseline model uses the majority label as the ground truth, the multi-perspective model leverages label probabilities.

4.3.1. Fine-tuning

We fine-tuned the aforementioned LLMs using two different approaches as the baseline, relying on the conventional use of aggregated labels with majority voting, and the multi-perspective model, incorporating diverse perspectives, i.e. without transforming them into a single label, via soft labels to refine the model's learning process. For both of these approaches, we fulfilled model fine-tuning on 2 x 32GB Tesla V100s.

To fine-tune BERT-large and RoBERTa-large with baseline and multi-perspective approaches, we used default hyperparameters. We fine-tuned the both models for 6 epochs, with a learning rate of 1×10^{-15}, *weight decay* of 0.01 and 500 *warmup steps* and a batch size of 8.

Baseline Loss The loss function used for the baseline model is multi-class cross-entropy defined as:

$$\text{Loss} = -\sum_{i=1}^{C} y_i \log(p_i)$$

where C is the number of classes, y is a one-hot encoded vector representing the true class and p is a vector of predicted probabilities. The goal is to minimize the cross-entropy loss during the training process which penalizes the model more when the predicted probabilities deviate from the true distribution of classes.

Baseline The baseline approach shows a lower ECE of 0.03 and 0.05 with HD for the BERT and RoBERTa-large, meaning that the model is well-calibrated, i.e. small deviation from the perfect calibration. On LLMD, the ECE is much higher as 0.35 and 0.37, suggesting that calibration made the uncalibrated models deviate more from the perfect calibration, i.e. model became over- or under-confident on its dataset predictions which signals a ill-calibrated model.

Multi-Perspective BERT-large performed with an ECE of 0.17 on HD, which is significantly higher than the baseline value of 0.03. This suggests that when BERT-large uses the multi-perspective approach, its calibration worsens on HD, as expected according to the findings in section 4.3.2. However, with LLMD, the ECE drops dramatically to 0.05, showing improvement in calibration. In this case, BERT-large seems to make much more reliable predictions according to calibration parameters. Regarding RoBERTa-large: on HD, the ECE increases to 0.30, which is higher than the baseline. This indicates that, like BERT-large, RoBERTa-large performs worse on HD using the multi-perspective approach, becoming less calibrated and thus less reliable according to traditional standards. On LLMD, however, the scores improve to 0.05, showing good calibration for RoBERTa-large as well. In summary, the results indicate a trade-off when applying the multi-perspective approach: it improves model performance on LLMD (with lower ECE values) but leads to worse performance on HD (with higher ECE values). This pattern is consistent across both BERT-large and RoBERTa-large models.

4.3.2. Calibration

After the fine-tuning step, we further applied calibration to adjust the predicted scores from a classifier to better align with the true probabilities which can lead to a fairer comparative evaluation, particularly for the model confidence scores. As a calibration method, we employed temperature scaling [37].

At its core, temperature scaling involves dividing the logits by a small value T and then applying the softmax function to convert the logits (z) into a probability distribution over the possible outputs. The value T is a hyperparameter often tuned on the validation set to minimize specific metrics e.g. negative log-likelihood, and we tuned the T on our validation set for 6 epochs.

5. Results

The model evaluation results are reported using various metrics of accuracy, precision, recall, and F1 score alongside average model confidence scores on the test set in Table 2. Based on the results, multi-perspective models outperform the baseline models in most cases, except the baseline RoBERTa-large model on LLMD, which showed a similar performance with the Multi-Perspective BERT-large and RoBERTa-large models on LLMD. The results confirm that using soft labels improve the model performance.

The best-performing baseline model is RoBERTa-large fine-tuned on HD with the F1-score of 57.22, while the best multi-perspective model is RoBERTa-large fine-tuned on HD with 61.90. Nonetheless, both for the baseline and multi-perspective, HD models show superior performance in comparison to the LLMD models which reflects that humans provide annotations with higher quality in comparison to LLMs. Apart from these, we can observe that baseline models generally exhibit higher model confidence scores (except the BERT-large model on HD) irrespective of the fine-tuning dataset (HD or LLMD). This is probably because the multi-perspective approach introduces higher level of model uncertainty through representing different viewpoints with equal weights. As a result, we argue that confidence score alone may not be the best criterion for evaluating multi-perspective models.

Approach	**HD**	**LLMD**
Baseline		
BERT-large	0.04 (same)	0.35 (same)
RoBERTa-large	0.04 (*U*), 0.06 (*C*)	0.37 (same)
Multi-perspective		
BERT-large	**0.05** (*U*), 0.18 (*C*)	0.20 (*U*) 0.05 (*C*)
RoBERTa-large	**0.12** (*U*), 0.30 (*C*)	0.18 (*U*), 0.05 (*C*)

Table 3. ECE (Expected Calibration Error) values with & without calibration denoted as *U* and *C* respectively, if the values are different.

The secondary focus of this paper is verifying whether model calibration ensures that the estimated class probabilities align closely with the actual outcomes. After applying calibration, we aimed to measure if a given model is well-calibrated, through calculating Expected Calibration Error (ECE) [38] which can be used to quantify how well the predicted output probabilities of the model matches the actual probabilities of the ground truth distribution. Table 3 shows that the uncalibrated baseline models are already well aligned with the perfectly calibrated model (ideal case with ECE of 0 since the lower ECE is, the better), thus calibration did not create a significant effect, while the situation is different for the multi-perspective approach. Calibration on the multi-perspective HD models lead to ill-calibrated models (higher ECE with calibration), while for the LLMD models, calibration helped to decrease the deviation from the perfect calibration (lower ECE with calibration). Based on these results, we applied calibration on all the models except the models with the multi-perspective approach on HD. While the calibration did not affect the overall model performance based on the evaluation metrics (accuracy, precision, recall, and F1 score), it significantly affected the model confidence scores. As Table 3 displays, calibration had a high impact only on the multi-perspective models in which there is a big error difference (ECE) between the uncalibrated and calibrated counterparts. For instance, the multi-perspective BERT-large on HD has the model confidence scores of 0.29 and 0.46, while the LLMD version of the same model has 0.30 and 0.46 with and without calibration respectively. Similar results apply to the multi-perspective RoBERTa-large.

6. Conclusion & Future Work

In this work, we present a pipeline for integrating multi-perspective models into stance detection task on controversial topics, adaptable to various subjective applications. To promote responsibility in NLP systems, we advocate for the implementation of perspectivist models and sharing of disaggregated datasets. We believe these strategies are essential for more inclusive models and for the advancement of this emergent research area in NLP field. We extended previous research by augmenting an existing dataset with summaries using state-of-the-art LLM and open-source LLM-based annotations to capture and preserve diverse viewpoints. We fine-tuned BERT-large and RoBERTa-large models using two methodologies: a baseline approach with hard aggregated labels and a novel approach with multi-perspective soft labels. Results show that multi-perspective models achieve better performances than baseline, with soft labels enhancing hard metrics (i.e. accuracy, precision, recall). However, beyond improving model performance, we also applied calibration on model predictions to properly use them as model confidence scores, then employ these confidence values to compare the baseline and multi-perspective approaches.

This work has potential limitations. Discarding instances without a majority label decreased the dataset size but made our experiments feasible for the comparative evaluation with the baseline model. Nonetheless, we believe that those instances are a valuable source for analyzing the multi-perspective approach which aims to learn from diverse perspectives instead of treating them as noise. One other possible solution is to increase the number of classes and annotators for each dataset instance, which we plan to pursue in future work. In this study, we used LLMs to gather annotations from different perspectives. However, previous research has shown that these models, when used as annotators, do not always align with human label distributions (opinions) accurately [23]. Specifically, state-of-the-art models like GPT-4 tend to be biased towards false positives, often labeling samples as offensive, abusive, or misogynistic. We recommend being mindful of such models' being inherently biased toward particular perspectives. Apart from these, our analysis was constrained by computational resources, affecting batch size and model capacity. In the future, we plan to expand our analysis by applying the current pipeline to a broader range of subjective tasks and datasets. We also aim to increase the number of baselines, following [39], with which we test the effectiveness of our method,in order to make the framework more generalizable. Moreover, we aim to explore deeper confidence-based results using soft labels, which can provide more nuanced insights into model uncertainty and prediction confidence.

References

[1] Sandri M, Leonardelli E, Tonelli S, Ježek E. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; 2023. p. 2428-41.

[2] Davani AM, Díaz M, Baker D, Prabhakaran V. D3CODE: Disentangling Disagreements in Data across Cultures on Offensiveness Detection and Evaluation. arXiv preprint arXiv:240410857. 2024.

[3] Santurkar S, Durmus E, Ladhak F, Lee C, Liang P, Hashimoto T. Whose opinions do language models reflect? In: International Conference on Machine Learning. PMLR; 2023. p. 29971-30004.

[4] Zhi-Xuan T, Carroll M, Franklin M, Ashton H. Beyond preferences in ai alignment. Philosophical Studies. 2024:1-51.

[5] Muscato B, Mala CS, Manerba MM, Gezici G, Giannotti F. An Overview of Recent Approaches to Enable Diversity in Large Language Models through Aligning with Human Perspectives. In: Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024; 2024. p. 49-55.

[6] Basile V, Cabitza F, Campagner A, Fell M. Toward a perspectivist turn in ground truthing for predictive computing. arXiv preprint arXiv:210904270. 2021.

[7] Uma AN, Fornaciari T, Hovy D, Paun S, Plank B, Poesio M. Learning from disagreement: A survey. Journal of Artificial Intelligence Research. 2021;72:1385-470.

[8] Akhtar S, Basile V, Patti V. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. arXiv preprint arXiv:210615896. 2021.

[9] Gezici G, Lipani A, Saygin Y, Yilmaz E. Evaluation metrics for measuring bias in search engine results. Information Retrieval Journal. 2021;24:85-113.

[10] Basile V, et al. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In: CEUR workshop proceedings. vol. 2776. CEUR-WS; 2020. p. 31-40.

[11] Leonardelli E, Uma A, Abercrombie G, Almanea D, Basile V, Fornaciari T, et al. SemEval-2023 task 11: Learning with disagreements (LeWiDi). arXiv preprint arXiv:230414803. 2023.

[12] Peterson JC, Battleday RM, Griffiths TL, Russakovsky O. Human uncertainty makes classification more robust. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 9617-26.

[13] Collins KM, Bhatt U, Weller A. Eliciting and learning with soft labels from every annotator. In: Proceedings of the AAAI conference on human computation and crowdsourcing. vol. 10; 2022. p. 40-52.

[14] Pereyra G, Tucker G, Chorowski J, Kaiser Ł, Hinton G. Regularizing neural networks by penalizing confident output distributions. arXiv preprint arXiv:170106548. 2017.

[15] Baan J, Daheim N, Ilia E, Ulmer D, Li HS, Fernández R, et al. Uncertainty in natural language generation: From theory to applications. arXiv preprint arXiv:230715703. 2023.

[16] Hu M, Zhang Z, Zhao S, Huang M, Wu B. Uncertainty in natural language processing: Sources, quantification, and applications. arXiv preprint arXiv:230604459. 2023.

[17] Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology. 2024;15(3):1-45.

[18] Riccardi N, Desai RH. The two word test: A semantic benchmark for large language models. arXiv preprint arXiv:230604610. 2023.

[19] Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences. 2023;120(30):e2305016120.

[20] Mohta J, Ak K, Xu Y, Shen M. Are large language models good annotators? In: Proceedings on. PMLR; 2023. p. 38-48.

[21] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

[22] Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.

[23] Pavlovic M, Poesio M. The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation. arXiv preprint arXiv:240501299. 2024.

[24] Zhu Y, Zhang P, Haq EU, Hui P, Tyson G. Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:230410145. 2023.

[25] Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? Computational Linguistics. 2024;50(1):237-91.

[26] Lee N, An NM, Thorne J. Can Large Language Models Capture Dissenting Human Voices? arXiv preprint arXiv:230513788. 2023.
[27] Wang X, Kim H, Rahman S, Mitra K, Miao Z. Human-LLM collaborative annotation through effective verification of LLM labels. In: Proceedings of the CHI Conference on Human Factors in Computing Systems; 2024. p. 1-21.
[28] Uma A, Fornaciari T, Hovy D, Paun S, Plank B, Poesio M. A case for soft loss functions. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. vol. 8; 2020. p. 173-7.
[29] Plank B. The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. arXiv preprint arXiv:221102570. 2022.
[30] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
[31] Liu Y. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692. 2019.
[32] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81. Available from: https://aclanthology.org/W04-1013.
[33] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:190409675. 2019.
[34] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics; 2002. p. 311-8.
[35] Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. arXiv preprint arXiv:231006825. 2023.
[36] Groeneveld D, Beltagy I, Walsh P, Bhagia A, Kinney R, Tafjord O, et al. Olmo: Accelerating the science of language models. arXiv preprint arXiv:240200838. 2024.
[37] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International conference on machine learning. PMLR; 2017. p. 1321-30.
[38] Naeini MP, Cooper G, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29; 2015. .
[39] Davani AM, Díaz M, Prabhakaran V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics. 2022;10:92-110.

(a) Pegasus-CNN-DailyMail

Metric	Chunk vs Summary	Doc vs Summary
ROUGE-1		
Recall	0.15	0.15
Precision	0.92	0.92
F1 Score	0.25	0.25
ROUGE-2		
Recall	0.11	0.10
Precision	0.85	0.84
F1 Score	0.18	0.17
ROUGE-L		
Recall	0.15	0.15
Precision	0.92	0.91
F1 Score	0.25	0.25

(b) BART-large-CNN

Metric	Chunk vs Summary	Doc vs Summary
ROUGE-1		
Recall	0.23	0.25
Precision	0.95	0.93
F1 Score	0.36	0.38
ROUGE-2		
Recall	0.18	0.17
Precision	0.86	0.82
F1 Score	0.28	0.28
ROUGE-L		
Recall	0.23	0.25
Precision	0.95	0.92
F1 Score	0.36	0.38

(c) Falcon-7B-Instruct

Metric	Chunk vs Summary	Doc vs Summary
ROUGE-1		
Recall	0.23	0.25
Precision	0.95	0.93
F1 Score	0.36	0.38
ROUGE-2		
Recall	0.18	0.17
Precision	0.86	0.82
F1 Score	0.28	0.28
ROUGE-L		
Recall	0.23	0.25
Precision	0.95	0.92
F1 Score	0.36	0.38

Table 4. Comparison of ROUGE scores between different models.