PERSEVAL: A Framework for Perspectivist Classification Evaluation

Anonymous ACL submission

Abstract

Natural Language Processing systems must be able to understand and adapt to diverse perspectives, distinguishing between varying viewpoints and subjective interpretations — an approach often referred to as perspectivism. While previous research has highlighted the need for moving beyond a single gold standard in evaluation, current practices remain fragmented and do not fully capture the complexity of perspectivist classification. To address this gap, we introduce PERSEVAL, the first unified framework for evaluating perspectivist models in NLP. A key innovation of this framework is its treatment of annotators and users as disjoint entities. This mirrors real-world scenarios where the individuals providing annotations to train models are distinct from the end-users whose perspectives the system must learn and accommodate. We instantiate PERSEVAL by experimenting with several encoder-based and decoder-based approaches. The results consistently show improvements when the models are informed with knowledge about the users.

1 Introduction

Recently, part of the Natural Language Processing (NLP) community has seen what Cabitza et al. (2023) called a *perspectivist turn*. Researchers have increasingly questioned data harmonization techniques such as majority vote, and even the aggregation of human labels itself, in favor of taking into account multiple perspectives as legitimate ground truths instead (Basile, 2020). This shift marks a paradigm change across the entire Machine Learning pipeline (Plank, 2022a): from collecting disaggregated, well-documented corpora (Bender and Friedman, 2018), to leveraging annotator disagreement in model training to better account for user diversity (Prabhakaran et al., 2021), and, importantly, to adopting evaluation strategies capable of embracing this disagreement (Uma et al., 2021b).

The open challenges in perspectivist approaches are still many (Fleisig et al., 2024), but research is proliferating, as demonstrated, for example, by the dedicated task Learning With Disagreements (LeWiDi) on its second edition,[1] and the NLPerspectives Workshop, on its third edition.[2]

In a landscape where many techniques are being developed to model human label variation, ensuring comparability across different approaches is of paramount importance. However, practices in previous work vary widely, with different paradigms. Inspired by early work on understanding and predicting annotator disagreement, one popular approach to perspectivist evaluation considers all annotators as known at training time (Mostafazadeh Davani et al., 2022). Although this is an sensible research scenario, it is far from real-world applications. In practical settings, models are typically trained on a fixed set of annotators and then made available to new users, with adaptation occurring through limited user interactions or feedback. Other works might be more flexible, and partially account for unseen annotators (Deng et al., 2023; Kazienko et al., 2023). Finally, some previous work has also explored perspective-based evaluation, targeting the majority vote of a subgroup of annotators sharing explicit or implicit characteristics (Frenda et al., 2023).

With the overarching goal of rationalizing and streamlining perspectivist evaluation, this paper presents PERSEVAL (Perspectivist Evaluation), a framework for perspectivist classification evaluation. To better mirror real-world scenarios, we consider annotators, who provide the bulk of the annotation for training models, as disjoint from system users, for which performance is tested and are assumed to be known at test time only. Relaxing our working hypothesis, we also define two

[1] https://le-wi-di.github.io/
[2] https://nlperspectives.di.unito.it/

scenarios for which minimal test users' annotations are available, for example from user feedback or human-in-the-loop approaches. The first, inspired by Kocoń et al. (2021b), accounts for cases in which only little information about test users' preferences is available during training; the model can thus use this information to learn a user-specific bias. The second scenario assumes a system has been already trained and deployed, and allows using test user information for adaptation. All the variants of PERSEVAL are explained in Section 3.

Moreover, we consider two different scenarios depending on the availability of explicitly defined user characteristics: users can either be known by their identifier only, or they can be represented as a set of metadata, for example describing their socio-demographic information or declared preferences.

Evaluation within PERSEVAL occurs at the annotation level, and incorporates both global and fine-grained metrics — evaluating at the user, text, and trait levels (Section 5). This enables a comprehensive comparison and analysis across different perspective models.

We showcase our evaluation framework on encoder- and decoder-based models, considering 5 disaggregated datasets focused on phenomena ranging from irony and offensive speech detection to AI safety, and with a diverse design concerning the number of annotators, sparsity of the dataset and provided demographics.

In summary, the contributions of this work are the following:

- We present PERSEVAL, an evaluation framework for perspective systems. We rationalize the user representation, the user splitting, and the evaluation functions.
- We collect and harmonize five disaggregated datasets with diverse domain, classification tasks, and user representations.
- We test two baseline models, and compare their performance in the proposed settings.
- We developed and share a user-friendly library providing functionalities facilitating the comparison and analysis of different approaches, making PERSEVAL an intuitive tool to rationalize the development and evaluation of perspectivist classification models.

2 Related works

Researchers have framed perspectives as tied to cultural background (Akhtar et al., 2021), demographic information (Frenda et al., 2023; Casola et al., 2024), a combination of attitudes and behavior (Chulvi et al., 2023), a set of psychological characteristics (Mieleszczenko-Kowszewicz et al., 2023) or beliefs (Kazienko et al., 2023), moving in a continuum from a mesoscopic (group-based) to a microscopic (individual) perspectives (Kocoń et al., 2021a). These diverse approaches to perspectivism in NLP are reflected in both data collection and modeling.

Disaggregated datasets are growing in number, as detailed in the repository of Plank (2022b)[3] and in the Perspectivist Data Manifesto.[4]

Moving from the differences in dataset design, number of annotators, available metadata, and corpus size, researchers have developed different ways to tackle the problem of modeling annotator perspectives. Works situated near the middle of an ideal continuum between data- and human-centric approaches focus on modeling groups of annotators (Frenda et al., 2023; Casola et al., 2023; Lo and Basile, 2023). Moving along this continuum, Mostafazadeh Davani et al. (2022) propose a multi-task-based approach where the goal is to predict each annotator's label. Personalization techniques have also been inspired by recommender systems methods (Kazienko et al., 2023; Heinisch et al., 2023; Mokhberian et al., 2024).

Despite an increasing number of modeling techniques, perspectivist evaluation remains an open problem (Basile et al., 2021). Previously cited works have adopted various approaches, such as evaluating single annotators (Mostafazadeh Davani et al., 2022; Mokhberian et al., 2024). The only structured framework of evaluation comes from LeWiDi task (Uma et al., 2021a; Leonardelli et al., 2023). Nevertheless, their proposal to evaluate the impact of disagreement via cross-entropy does not directly address the challenge of evaluating the models' ability to capture human perspectives.

To the best of our knowledge, the only previous benchmark in this field of research is The Inherent Disagreement 8 dataset (TID-8) by Deng et al. (2023). It is a collection of 8 language-understanding disaggregated datasets with a vary-

[3] https://github.com/mainlp/awesome-human-label-variation

[4] https://pdai.info/

ing number of annotators. The authors tested on modeling annotators' perspective through annotator and annotation embeddings. They presented the results in terms of Exact Match Accuracy score Macro F1 by both testing on the same annotators of the training set (annotation split), and on new annotators (annotator split). In respect to this work, we want to cover a larger diversity of approaches, unresolved issues, and possible lines of research. PERSEVAL is the result of efforts in this direction, a framework that manages different real-world scenarios in terms of data splitting, annotators' metadata, and thus perspective framing.

3 PERSEVAL: the Framework

We propose a conceptual framework for the development and evaluation of perspectivist text classification models, which can be characterized along several key dimensions. These dimensions — data split, user representation, and adaptation strategies — are discussed in the following sections.

3.1 Data Split

Previous research on disaggregated datasets has taken different approaches to data splitting. Most perspectivist evaluation practices rely on a fixed set of annotators, who typically annotate every text in the corpus; a standard text-based split is then adopted. While this approach is useful from a theoretical standpoint, it does not reflect real-world scenarios. In practice, a system is trained on annotations provided by one group of individuals (the *annotators*), while its inference is run on a set of instances encoding the perspectives of a distinct set of individuals (the *users*). This distinction is especially important in personalization tasks, where evaluating a model's performance on a hold-out group of users is crucial to ensure that the system can effectively generalize to new users' perspectives and preferences.

Starting from these considerations, we conceptualize the data split in PERSEVAL under the assumption that annotators, who provide the training annotations, are disjoint from the test users.

When explicit knowledge about the users is available — for example, in the form of socio-demographic information or preferences —, a model can attempt to learn biases toward such characteristics. When no such information is available, however, inferring preferences for completely unknown users is unfeasible. As a consequence, we define two adaptation scenarios:

- **Adaptation at training time**: This variant mirrors a situation in which a minimal annotation from users can be obtained before training the system. In this scenario, a few annotations from test users are included in the training split.
- **Adaptation at inference time**: This variant mirrors a situation in which an already trained system should be adapted to new users. In this case, test users instances should not be used at training time, but rather as a way to adapt an existing model.

In both cases, we assume that a minimal amount of information about users in the test set (provided in terms of their annotation) is available. For example, we can assume that annotations have been collected through the user interaction with the system or through other human-in-the-loop approaches.

3.2 User representation

The degree of information available about users varies across datasets. Users can be represented by different levels of metadata, from a simple identifier to a rich set of attributes.

When metadata are available, we represent the user through a set of traits, which may include socio-demographic information or explicit preferences. This representation enables models to learn user-specific perspectives based on these traits. We refer to this as the *named* representation. Given our proposed data split, which divides training from test users, the challenge is to learn perspectives based on a set of annotators and generalize the model of the perspectives to unseen users based on their traits only. In named perspectives text classification tasks, much of the information that models can use to learn a representation of human perspectives is encoded in the users' explicit traits.

In cases where user metadata are not available, the user is represented only by an identifier. While this restricts the model's ability to personalize predictions, it is a common scenario in many real-world applications. We call this representation *unnamed*. In the unnamed perspectives task, the model must classify perspectives without any knowledge of the user's traits. This variant necessitates adaptation techniques to infer user perspectives from the available annotations: the strict hypothesis for which test users are completely disjoint

Task	Adaptation	Adapt. Phase
Named	No adaptation	Never
	Adaptation-T	At training time
	Adaptation-I	At inference time
Unnamed	Adaptation-T	At training time
	Adaptation-I	At inference time

Table 1: Task variants proposed in PERSEVAL.

from training annotators must be relaxed (Section 3.1). As a consequence, the named classification task can be performed with and without adaptation, while the unnamed task requires some form of adaptation. Table 1 summarized the available variants.

3.3 Extended training set

Instances in PERSEVAL are <text, users> pairs. This is fundamentally different from traditional evaluation methodologies in NLP where instances are typically just textual. A corollary of this observation is that in PERSEVAL it would be perfectly acceptable to have a training instance and a test instance sharing the same text, because associated with labels from different perspectives. However, this behavior is not always desirable, as the knowledge learned by a model from a training text may affect the inference on an instance with the same text in unpredictable ways.

Moreover, practical considerations arise. When collecting a perspectivist datasets, two approaches are common: in some cases, researchers hire a small set of expert annotators, who typically annotate each instance in the dataset. This results in a dense annotation matrix. Alternatively, many diverse annotators can be hired, for example through crowd sourcing platforms. The resulting datasets are typically very sparse.

In scenarios where the annotators' and test users' annotations overlap, there is a higher chance of the same text appearing in both the training and test sets, potentially with different labels. To ensure fair evaluation and avoid data leakage, we follow standard practice and exclude any text instances in the test split that have been annotated by the training annotators.

However, we also explore a variant where texts that appear in both training and test sets, but annotated by different users, are allowed in the training data (we call this variant *extended*). This variant tests the model's ability to learn from systematic disagreements among users who annotate the same text differently, capturing the diversity of perspectives inherent in the data.

All the task variants previously described can make use of the extended training set variant.

The difference between the task variants described so far manifests in different training splits (or in some cases additional sets when using adaptation at inference time). However, the test set remains consistent across all task variants to ensure a fair comparison of model performance.

4 Datasets

PERSEVAL incorporates a diverse range of datasets, encompassing both dense and crowdsourced, sparse datasets. Dense datasets, where each instance is annotated by all annotators, are represented by the BREXIT (Akhtar et al., 2021) and DICES (Aroyo et al., 2024) datasets. In contrast, the sparse datasets — where annotations are provided by a larger pool of annotators but each annotator annotates only a subset of instances — are represented by the EPIC (Frenda et al., 2023), MHS (Sachdeva et al., 2022), and MD-Agreement (Leonardelli et al., 2021) datasets, all collected by crowdsourcing annotations. Moreover, the incorporated datsets vary in task, domain, and available annotator information.

4.1 BREXIT

This is a dataset for abusive language detection, consisting of 1,120 English tweets. The dataset is annotated by 6 annotators from 2 groups: 3 Muslim immigrants in the UK (target group), and 3 researchers as a control group. Each annotator annotated the entire corpus giving a binary label at multiple levels, specifically hate speech, aggressiveness, offensiveness, and stereotype. The only available user trait is the group each annotator belongs to i.e., either target or control. In PERSEVAL the positive class is *hate speech*, which is highly umbalanced toward the negative class.

4.2 EPIC

The English Perspectivist Irony Corpus consists of 3,000 texts collected from Twitter and Reddit in 5 English-speaking countries and annotated by 74 crowd workers. Each annotator labeled around 200 texts, for a total of 14,172 annotations. The authors also released annotators' demographic information (Appendix B), balanced across gender and nationality. In PERSEVAL the target class is *irony*.

4.3 MHS

The Measuring Hate Speech corpus contains 39,565 English comments extracted from YouTube, Twitter, and Reddit. It has been annotated by 7,912 people, resulting in 135,556 annotations with both a specific label and multiple hate-informative labels to capture the degree of hatefulness in a continuum. The annotators shared their demographics, reported in Appendix B. In PERSEVAL the positive class is *hate speech*.

4.4 MD-Agreement

The Multidomain Agreement dataset, recently used in the LeWiDi task (Leonardelli et al., 2023), comprises 10,753 English tweets from three domains associated with the hashtags #BlackLivesMatter, #Election2020 and #Covid-19. Each text has been annotated 5 times by 819 annotators, for a total of 53,765 annotations. This is the only dataset that does not provide any demographic traits. In PERSEVAL the positive class is *offensiveness*.

4.5 DICES

The DICES (Diversity in Conversational AI Evaluation for Safety) dataset focuses on conversational AI safety. It is a multi-turn conversation corpus generated by humans interacting with a generative AI-chatbot, provoking it to respond with an undesirable or unsafe answer. For PERSEVAL we opted for DICES-350, designed to study in-depth cross-demographic differences within the US. Specifically, it consists of 350 multi-turn conversations (within a maximum of 5 turns), fully annotated by 123 people, having a total of 43,050 annotations. This is the only dataset with a non-binary label (with values harmful, not harmful and unsure). The author released annotators' traits, reported in Appendix B.

5 Evaluation metrics

Our evaluation setting is inspired by previous work in personalization. Given a specific true annotation for an annotator and a text instance, we compute standard classification metrics, i.e., precision, recall, and F1-score (referred to as global metrics in the following).

Moreover, the annotator-based characteristic of the disaggregated labels gives us the chance to gain further insights into the models' capability to learn from diverse human perspectives. Inspired by Mokhberian et al. (2024), we also also report *user-level* metrics. These metrics are computed individually for each test user and then averaged; they provide a fairer evaluation regardless of the extent of the contribution in terms of annotations of each annotator to the dataset.

We also report *text-level* metrics, computed individually for each text in the test set and averaged. The analysis of these metrics help understanding whether some texts are easier to classify for a given model and whether having instances with the same textual content (but different users, and thus, different annotations, in the extended version of the dataset), helps the model in the classification.

Finally, for the named task, we also report *trait-level* metrics. These metrics, computed for each trait and then averaged for each dimension, are meant to describe if the preference of all groups of people is fairly learned by the model or if the model underperforms when considering users with certain characteristics.

6 Baseline Models and Evaluation

We benchmarked a series of approaches for perspectivist classification, using encoder- and decoder-based models, covering all task variants proposed in Section 3.

We focus on an intra-model comparison of the results, in order to highlight which settings are most effective for the encoder-based and decoder based appoaches separately. This is also motivated by the different settings supported by each approach. In particular, when working with the encoder-based model, we did not include inference-time adaptation since this architecture does not support it. On the other hand, performing zero- and few-shot learnig by prompting the LLM, we did not cover the Adaptation-T variant either for the Named or the Unnamed Task.

Section 6.1 presents the experimental results with the encoder-based model, while in Section 6.2 we discuss the results with the decoder-based model. In all cases, we report the metrics related to the prediction of the positive class, with the exception of DICES, the only multi-class dataset for which we present the macro-averaged metrics.

Dataset		Adapt	Extended	Global			User f1	Text f1
				Precision	Recall	f1		
EPIC	baseline	-	No	.513	.605	**.555**	**.538**	**.376**
	Named	None	No	.517	.570	.542	.527	.364
		Train	No	.510	.597	.550	.534	.371
	Unnamed	Train	No	.524	.545	.534	.518	.352
BREXIT	baseline	-	No	.465	.727	**.567**	.519	.403
	Named	None	No	.429	.798	.558	**.524**	**.416**
		Train	No	.418	.778	.544	.512	.405
	Unnamed	Train	No	.386	.889	.514	.484	.378
MHS	baseline	-	No	.649	.739	.691	.643	.521
	Named	None	No	.647	.748	**.694**	**.727**	**.526**
		Train	No	.682	.692	.685	.639	.511
	Unnamed	Train	No	.687	.686	.681	.634	.504
MD	baseline	-	No	.581	.778	**.665**	.591	**.500**
	Unnamed	Train	No	.581	.778	**.665**	**.597**	.499
EPIC	baseline	-	Yes	.519	.654	.579	.559	.405
	Named	None	Yes	.539	.617	.575	.560	**.567**
		Train	Yes	.562	.594	.578	.564	.398
	Unnamed	Train	Yes	.533	.650	**.586**	**.571**	.405
BREXIT	baseline	-	Yes	.478	.778	**.592**	.543	.427
	Named	None	Yes	.449	.848	.587	**.557**	**.455**
		Train	Yes	.388	.889	.540	.509	.424
	Unnamed	Train	Yes	.398	.909	.554	.524	.436
MHS	baseline	-	Yes	.667	.731	**.698**	**.652**	.528
	Named	None	Yes	.674	.717	**.698**	.649	.528
		Train	Yes	.654	.746	**.698**	.650	**.532**
	Unnamed	Train	Yes	.663	.732	.696	.647	.527
MD	baseline	-	Yes	.602	.748	.667	.603	.495
	Unnamed	Train	Yes	.591	.802	**.681**	**.620**	**.518**

Table 2: Encoder model's global precision, recall and f1 score, and user- and text- level f1 scores for the positive class for binary datasets.

6.1 Encoder-based Model

We fine-tuned RoBERTa (Zhuang et al., 2021)[5], customized implementing Focal Loss (Lin et al., 2017) to prevent overfitting in case of unbalanced datasets. All splitting and training parameters are reported in Appendix A. Inspired by the personalized User-ID model from Ferdinan and Kocoń (2023), we added identifiers and traits of the annotators to the text embedding as a special token. The input thus concatenates the annotator id, a special token for each of the annotator's traits, and the input text to classify. The special tokens explicitly encode the annotator's identity and characteristics and are used by the model to learn annotator- and trait-specific features in the classification. The model is then trained with a classification head to predict the binary label. We also computed a baseline without any additional special token.

Looking at the results on the datasets with binary labels (Table 2), when we force the constraint on text being annotated by training users or test users but not both (*non-extended training set*), the baseline tends to have higher scores in terms of global F1. With the extended training set, the trend is the opposite. This indicates that the model learns the relation between latent features of the text and the users labeling them, to some extent. The user and text-based F1 scores highlight the benefit of including demographic traits in training time, especially in the setting without adaptation set (*Adaptation-None*). When demographics are not available, such as, e.g., MD-Agreement, providing the user-id still results in being beneficial.

Finally, we present DICES separately as the only multi-class dataset, and report the macro-averaged metrics (Table 3). This dataset confirms the positive impact of providing demographics with the encoder-based model, with improved results in all settings in terms of global, user-level and text-level F1 score. Moreover, performing adaptation helps the performance across all the metrics.

6.2 Generative Models

For the decoder-based model, we focus on open-source models and benchmark the performance of Llama-3.1 8B[6], instruction tuned.

[5]https://huggingface.co/FacebookAI/roberta-base

[6]https://huggingface.co/meta-llama/Llama-3.1-8B

	Adapt	Extended	Global			User F1	Text F1
			Precision	Recall	F1		
baseline	-	No	.417	.480	.340	.311	.245
Named	None	No	.423	.458	.400	.391	.361
	Train	No	.438	.506	.420	.407	.373
Unnamed	Train	No	.443	.511	**.434**	**.424**	**.389**
baseline	-	Yes	.451	.513	.440	.448	.335
Named	None	Yes	.480	.538	.453	.439	.378
	Train	Yes	.479	.547	**.457**	.445	**.389**
Unnamed	Train	Yes	.481	.540	.456	**.446**	.388

Table 3: Macro-averaged global precision recall and F1, user- and text-level F1 scores for both the Encoder model with the DICES dataset.

We consider several settings:

- **Base-zero**: We prompt the models to classify the test set examples, without providing any additional information.

- **Perspective**: Inspired by work on role-based sociodemographic prompting (Cheng et al., 2023; Beck et al., 2023), we ask the models to impersonate each user's trait. To do so, we prepend the given trait to the prompt (for example *You are a person from Generation X.*). We use this variant to test models without adaptation with a named user representation. We prompt the model for each available user trait, and computing the final labels by by a majority-vote over the predicted labels.

- **In-Prompt Augmentation (IPA)**: We reproduced Salemi et al. (2024)'s approach, using the In-Prompt Augmentation (*IPA*) strategy. This strategy consists of prompting the model with user-specific input selected via retrieval augmentation, a framework which extracts pertinent texts, relevant to the classification of the unseen test case. Using the authors' terminology, given a sample (x_i, y_i) and a user u, a query generation function ϕ_q transforms the input x_i into a query q for retrieving the user profile P_u (i.e. the user's historical data) from the Adaptation set. To do so, we used the FContriever model (Izacard and Grave, 2021), a pre-trained dense retrieval model $\mathcal{R}(q, P_u, k)$ that retrieves the k most pertinent entries. Finally, the prompt construction function ϕ_p assembles the personalized prompt. Specifically, we selected a k of 5 entries. We used this approach both giving information about the user's trait value (named user representation, with adaptation at inference time) and without providing demographic information.

Since some outputs could not be properly parsed — such as when the model refused to provide an answer, particularly for datasets related to hate speech — we assigned a third label to these cases. In these experiments we only use the test sets, since we performed zero and few-shot prompting. Therefore, considerations about the extended set (Section 3.3) do not apply.

Table 4 presents the results.

Across all datasets, all the approaches outperform the baseline. Looking at the scores on named tasks, it is possible to notice that adding the demographic information about the annotators consistently helps the model in the prediction. As expected, the unnamed task is harder, however the user-based selection of few-shot examples of IPA significantly outperform the baseline. Indeed, IPA is the most effective strategy for perspectives classification with generative models, with the exception of BREXIT. We postulate that this is due to the high polarization of the annotations in this dataset, and the narrow characterization of the annotators.

The same pattern can be seen when looking at DICES dataset (Table 5), where IPA shows to be the best approach. The positive influence of adding sociodemographic information is also confirmed.

7 The PERSEVAL Python Library

PERSEVAL is implemented as a Python library to facilitate access to the data, the different splits related to task variants, and the evaluation metrics. The main interaction starts by instantiating a dataset from the `data` submodule. The user can then request the training, test, and optionally adaptation data splits with the `get_splits()` method, indicating whether the adaptation data (`user_adaptation`) is absent (`False`), available at training time (`train`) or at inference time (`test`). Additionally, the user chooses whether to extend the training split including texts also

Dataset	Approach		Adapt	Global			User F1	Text F1
				Precision	Recall	F1		
EPIC	base-zero	baseline	-	.473	.600	.529	.511	.363
	perspective	Named	None	.474	.498	.484	.467	.322
	IPA	Named	Test	.402	.857	**.547**	.528	**.387**
		Unnamed	Test	.391	.902	.546	**.530**	.386
BREXIT	base-zero	baseline	-	.466	.545	.502	.476	.340
	perspective	Named	None	.472	.596	**.527**	**.502**	**.371**
	IPA	Named	Test	.227	.909	.364	.362	.238
		Unnamed	Test	.201	.929	.364	.362	.238
MHS	base-zero	baseline	-	.489	.755	.593	.545	.425
	perspective	Named	None	.456	.591	.514	.453	.353
	IPA	Named	Test	.542	.773	**.637**	.537	**.467**
		Unnamed	Test	.506	.820	.626	**.570**	.456
MD	base-zero	baseline	-	.567	.545	.556	.515	.381
	IPA	Unnamed	Test	.469	.884	**.613**	**.535**	**.451**

Table 4: Decoder-based approach global precision, recall and F1 score, and user- and text-level F1 scores for the positive class.

Dataset	Approach		Adapt	Global			User F1	Text F1
				Precision	Recall	F1		
DICES	base-zero	baseline	-	.309	.303	.290	.282	.310
	perspective	Named	None	.320	.310	.289	.281	.276
	IPA	Named	Test	.381	.364	**.368**	**.357**	**.428**
		Unnamed	Test	.287	.264	.266	.319	.418

Table 5: Macro-averaged global precision recall and F1, user- and text-level F1 scores for both the Encoder model and the Large Language Models with the DICES dataset.

in test instances (`extended=True`) or to exclude them (`extended=False`). The dataset object contains a series of metadata about the dataset, such as its name, label names, and a dictionary of the annotator traits. Moreover, it contains the three splits, instantiated as objects of the same `PerspectivistSplit` class. These objects, called `training_set`, `test_set`, and `adaptation_set`, contain the list of users, texts, and the annotations, for the respective split. The `User` objects contain a unique identifier and a dictionary of traits. The `Text` objects contain a dictionary with the textual content of an instance, depending on the structure of the dataset. The `annotation` property is a dictionary where the keys are a pair (*User id*, *Text id*), and the value is a dictionary containing a value for each annotated label.

Besides providing access to the datasets and appropriate splits of the data for each task variant, the PERSEVAL library facilitates the automatic evaluation of models. The library implements the class `Evaluator`, which can be instantiated by passing the path of a file containing the predictions, a test set, and a target label name. The `Evaluator` object implements the functions to calculate the evaluation metrics described in Section 5. The output of the global, annotator-, text-, and trait-level metrics can be visualized in their aggregated forms and can be accessed (also at the level of each individual annotator, text, and trait) programmatically for a deeper analysis.

8 Conclusion

We introduced PERSEVAL, the first unified framework for the evaluation of perspectivist text classification. We assume train annotators and test users are different, and design a named perspectivist classification task where users are represented by their explicit traits and an unnamed task where only their identifier is available.

We included five datasets, two dense and three crowd-sourced, with different user traits, and implemented two baseline models. We thoroughly tested several variants of perspectivist classification presenting a robust benchmark for complex real-world applications.

PERSEVAL is also implemented in an intuitive and easy-to-use Python library, facilitating the access to the data and automatic evaluation.[7]

Limitations

In this paper, we primarily focus on presenting a comprehensive framework for evaluating perspec-

[7]The code will be released with a free software license upon acceptance.

tivist models. Our goal was not to test an extensive range of models; instead, we conducted experiments on just two baseline models. We believe that the framework and library introduced here will serve as a valuable resource for future research in evaluating real-world systems within similar contexts. While we considered multiple datasets, all are in English and most feature binary labels. In future work, we plan to expand this work by incorporating disaggregated datasets in various languages.

Ethical statement

The work presented in this paper is in the context of a broader initiative to consider the subjectivity of the annotators in NLP applications, encouraging reflection on the different perspectives encoded in annotated datasets to minimize the amplification of biases. The proposed benchmark can be used as a basis for evaluating a wide range of NLP models, including LLMs, according to their capability of representing the variability of human perspectives.

The language resources included in the benchmark were built adopting measures to protect the privacy of annotators and data handling protocols designed to safeguard personal information. Some of the material could contain racist, sexist, stereotypical, violent, or generally disturbing content.

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. *arXiv preprint arXiv:2106.15896*.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2024. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36.

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, page 441–453, Berlin, Heidelberg. Springer-Verlag.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective NLP tasks. *CoRR*, abs/2309.07034.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Silvia Casola, Soda Lo, Valerio Basile, Simona Frenda, Alessandra 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.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, Paolo Rosso, et al. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *CEUR WORKSHOP PROCEEDINGS*, volume 3494.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Teddy Ferdinan and Jan Kocoń. 2023. Personalized models resistant to malicious attacks for human-centered trusted ai. *Emotion*, 40000:50000.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

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.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Przemysław Kazienko, Julita Bielaniewicz, Marcin Gruza, Kamil Kanclerz, Konrad Karanowski, Piotr Miłkowski, and Jan Kocoń. 2023. Human-centered neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion*, 94:43–65.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021a. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021b. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP*.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanislaw Wozniak, Ewa Dzieciol, Przemyslaw Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP*.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

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.

Barbara Plank. 2022a. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Barbara Plank. 2022b. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR)*

Workshop, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Training parameters

Table 6 presents the training parameters for the Encoder model.

B User traits

Table 7 shows the traits in the BREXIT, DICES, EPIC and MHS datasets.

Parameter	Value
eval_strategy	epoch
greater_is_better	False
learning_rate	$5e^{-6}$
load_best_model_at_end	True
metric_for_best_model	eval_loss
num_train_epochs	5
per_device_eval_batch_size	32
per_device_train_batch_size	16

Table 6: Model parameters for "RoBERTa base".

Dataset	Traits	Values
BREXIT	Group	Target, Control
DICES	Gender	Male, Female
	Age	GenX+, GenY, GenZ
	Education	College degree or higher, High school or below
	Ethnicity	Asian, Black, Latinx, White
EPIC	Gender	Male, Female
	Age	19-64 y/o, grouped in Boomer, GenX, GenY and GenZ
	Nationality	Australia, India, Ireland, United Kingdom, United States
MHS	Gender	Male, Female
	Age	18-81 y/o, grouped in Boomer, GenX, GenY, GenZ
	Education	College degree or higher, High school or below
	Income	less than 50k annual income, more than 50k annual income

Table 7: The sets of user traits included in PersEval for the BREXIT, DICES, EPIC and MHS datasets.