Hierarchical Clustering of Label-based Annotator Representations for Mining Perspectives

Soda Marem Lo[1,*,†], Valerio Basile[1,†]

[1]Computer Science Department, University of Turin, Turin, Italy

Abstract

Modeling annotator perspectives has emerged as a technique to model subjective linguistic phenomena more accurately. Authors in the NLP community approached this issue by creating perspective-aware and personalized models, where demographic data or previous annotations are needed. In this paper, we explore two methodologies to represent annotators solely on the basis of the labels they assigned: label agreement and Kernel PCA. For both these techniques, we computed respectively 5 and 4 clusters, trained perspective-aware models on each of them, and finally implemented majority vote ensembles. The results show that clusters obtained by the first mining technique are more balanced and homogeneous in terms of annotators' demographic traits, while those obtained by KPCA tend to correlate more with their nationalities. Despite these differences, both ensemble models outperform the baseline, confirming that leveraging annotation using clustering techniques is advantageous for the classification of a subjective phenomenon such as irony. We sustain that this approach can be beneficial for taking into account annotators' perspectives when demographic data are not known, together with the possibility that their annotations might be influenced by factors other than given demographics.

Keywords

Perspectivism, clustering, irony detection

1. Introduction

Subjective tasks in Natural Language Processing face the issue of correctly modeling the perception of the humans involved in the process, e.g., producing language resources used to train and evaluate models. In recent years, several authors have started considering the importance of disagreement, criticizing the idea of a single valid truth [1], and examining its potential impact on several aspects of NLP [2]. Such observation is fundamental especially considering highly subjective tasks where annotators' opinions may differ in relation to their cultural and social background, or their personal experiences [3].

To this aim, the *perspectivist* approach[1] works towards modeling raters' perspectives, keeping all human labels during the training phase of the classifier [4].

Authors moving along this paradigm shift have often pointed out the necessity to publish disaggregated [5], and well-documented data, with as much meta-data as possible [6]. This information has been used in [7] to build perspective-aware models, based on demographic traits such as gender, nationality and generation, which resulted to be more confident in detecting irony in respect to the non-perspectivist ones.

On the other hand, it is important to notice that annotators' opinions are not necessarily linked to these traits only, especially when considering phenomena where both demographic-depending aspects such as cultural background and culturally-shared linguistic expressions can be key elements to their definition and individuation. This is what happens with irony, influenced by elements as the origin of the speaker [8, 9] and linguistic patterns sometimes shared across languages [10].

Moving from the idea that human labels hide values and possible interpretations of a linguistic phenomenon [6], we want to explore whether annotators choices overlap with their demographics, or might be linked to other traits that influence a similarity of opinions despite the different backgrounds. Specifically, we mined annotators' perspectives to see how they group together on the base of their annotations only. We propose two methods to vectorize annotators leveraging the set of their labels. Then, we trained cluster-based models and built a majority voting ensemble to validate our representation techniques in a in-dataset and cross-dataset setting.

The main contributions of this papers are the following:

- Two techniques to model annotators as vector representations and automatically cluster them;
- A quantitative and qualitative analysis of the automatically predicted clusters of annotators, both in terms of quality of the clusters and mapping between clusters and divisions of annotators based on demographic traits;
- Experimental evidence that leveraging automatically grouping of the annotators in a disaggre-

2nd Workshop on Perspectivist Approaches to NLP

*Corresponding author

†These authors contributed equally.

sodamarem.lo@unito.it (S. M. Lo); valerio.basile@unito.it (V. Basile)

0000-0002-5810-0093 (S. M. Lo); 0000-0001-8110-6832 (V. Basile)

[1]Data Perspectivism Manifesto

gated dataset is beneficial for the predictive power of an ensemble of classifiers for irony detection.

The experiments are conducted on EPIC (English Perspectivist Irony Corpus) [7], a disaggregated corpus for irony detection, described in Section 3. The methods are introduced in Section 4 where the results of the clustering are analysed, and applied to irony detection in Section 5.

2. Related works

The correlation between annotators' choices, their demographic traits, beliefs and social backgrounds has become subject of attention in tasks such as offensive language [11, 12], hate speech [13, 3] and toxicity detection [14]. These works have demonstrated how the identity of the annotators, their social groups and their beliefs can play a role in the annotation phase.

Taking into account raters' backgrounds can be of fundamental importance to avoid building machines biased toward the opinions of a majority [4, 15], especially when working on phenomena that cannot be objectively defined.

The perspectivist approach aims at leveraging disagreement to model annotators' points of view and culturally-driven perspectives [5]. In [16] the authors grouped annotators by measuring polarization of their judgments on hate speech content, then created a gold standard of each group to obtain perspective-aware models, eventually including the learned perspectives in an ensemble classifier. Inspired by this work, authors in [7] implemented perspective-aware models based on annotators' demographic characteristics, and proposed to evaluate them on the confidence [17] of their predictions. The perspective-aware models resulted to be more confident than non-perspectivist ones.

Techniques for modeling annotators' perspectives have also been developed using personalization methods, recently applied to NLP with the aim of processing diversity among annotators [18] in several subjective tasks, such as offensive content, sense of humor and emotion detection [19, 20, 21], but also in the classification of interpersonal conflict types [22]. This approach tends to consider not always demographic data, but also personal beliefs and opinions obtained by historical posts of the same user [22, 23]. For example, in [21], the authors developed a measure of the human bias to model individual human perspectives, i.e. how a user's perception differ from others, to obtain a representation of the subjectivity of each annotator. Authors in [12] propose both a mesoscopic (group-based) and microscopic (user-based) approach to predict annotators' beliefs, considering their metadata, the annotator identifier (id), and previous annotations, demonstrating improved performance of classifiers as users' information increased. Moreover, they grouped annotators based on their agreement level, to extract social groups and analyze the impact of group profile on the task of offensive content recognition. Interestingly, when testing the agreement measure on demographic groups, no significant correlation was found, showing that there might be other factors conditioning users' perceptions of aggressiveness.

Agreement was already used to mine annotators' perspective in [24], where the authors measured label and features agreement, in order to cluster together those who shared a perspective for similar reasons. Influenced by this work, this paper wants to explore how annotators are clustered based on their annotations about ironic content. Thus, we compared two methodologies to mine raters' opinions, observing whether these choices coincide with their demographic data; finally we implemented cluster-based models inspired by [16] and [7].

3. Corpus description

In this section we present the two corpora used for our in-dataset and cross-dataset experiment, respectively EPIC (English Perspectivist Irony Corpus), released by [7]; and the corpus used for SemEval-2018 Task 3 "Irony Detection in English Tweets" [25].

3.1. EPIC

For the in-dataset setting we trained and tested our models on the English Perspectivist Irony Corpus [7, EPIC], a disaggregated corpus consisting of $3,000$ *Post, Reply* pairs from Reddit $(1,500)$ and Twitter $(1,500)$ collected across five English-speaking countries: Australia, India, Ireland, United Kingdom and United States. Regarding Twitter, authors used the API geolocation service to identify the five English varietes. With respect to Reddit, they collected data from the following subreddits, assuming the origin of the texts: r/AskReddit (United States), r/CasualUK (United Kingdom), r/britishproblems (United Kingdom), r/australia (Australia), r/ireland (Ireland), r/india (India). The 74 annotators were balanced across both gender and nationality, with a total of ~ 15 raters for each of the aforementioned nationalities, who labelled around 200 instances each. Thus, the corpus consists of $14,172$ annotations, with a median of annotations per instance of 5.

The authors collected demographic information about the annotators (gender, age group, nationality, ethnicity, student status and employment status), and used data related to gender (female, male), age (boomer, Generation X, Generation Y, Generation Z) and nationality (Australian, British, Indian, Irish and US-American) to build 11 demographic-based models, each trained only on the labels provided by one group, and tested

on both a demographic-independent aggregated test set and perspective-based test sets. The former, to which we will refer as GOLD TEST SET, was obtained applying a majority voting strategy on the entire corpus. The authors discarded those instances for which a majority was not available resulting in an aggregated set of $2,767$ instances. This set was split into training (80%, 440 ironic, 1331 not ironic) and test set (20%, 110 ironic, and 443 not ironic), thus obtaining the GOLD TEST SET of 553 instances (246 from Reddit and 307 from Twitter).

We replicated this methodology to train and test the non-perspectivist (NP) model on this split, as in [7].

3.2. SemEval-2018 Task 3

To verify the robustness of our cluster-based models, we tested their performances in a cross-dataset setting on the corpus used for the SemEval-2018 shared task on iroy detection [25].

It consists of 4,792 tweets, collected between December 2014 and January 2015, and annotated by three students in linguistics, who spoke English as a second language (other demographic data were not collected). For the shared task the corpus was randomly split into training (1445 ironic, 1417 not ironic) validation (456 ironic, 499 not ironic) and test set, (784 instances, 311 ironic, and 473 not ironic).

For the experiment in the cross-dataset setting we tested our models, previously trained on EPIC, on SemEval-2018 test set.

4. Mining perspectives

This section introduces the methodology used to automatically compute clusters of annotators. The core of our approach is to vectorize each annotator based on the labels assigned for each of the $3,000$ instances i. Given n raters annotating k instances, we obtained a matrix $V^{n \times k}$, which will be called *label matrix*.

Considering that each (*Post*, *Reply*) pair has an average of 4.72 and a median of 5 annotations, annotators can have three possible opinions: 0 (not ironic), 1 (ironic), or a missing value. Thus, for each annotator, we obtain a vector with the dimensionality of the number of instances *i*, where the combination of the assigned label represents rater's perspective. Since annotators have annotated around 200 instances each, there are at least $2,800$ missing values per annotator. For this reason we have chosen to adopt two methods to represent the annotators as vectors.

First representation technique: label agreement (α) We computed a pairwise similarity matrix using Krippendorff's alpha (α) [26] as a metric to handle missing values in the annotation when estimating how much each couple of annotators agrees between each other.

Second representation technique: dimensionality reduction (KPCA) We opted for reducing the dimensionality of the label matrix adopting a nonlinear form of Principal Component Analysis (Kernel PCA) [27], then computing the pairwise distance matrix among annotators.

The two methodologies will be explained and discussed in the following paragraphs.

4.1. Label agreement

Following [24], we measured label agreement in terms of Krippendorff's α, since it has been developed both to take into account that some agreement can arise by chance (as the more common Cohen's Kappa agreement score), and to measure agreement among raters with incomplete annotations, in contrast with Kappa measures (Cohen's and Fleiss') that rely on a complete annotation matrix.

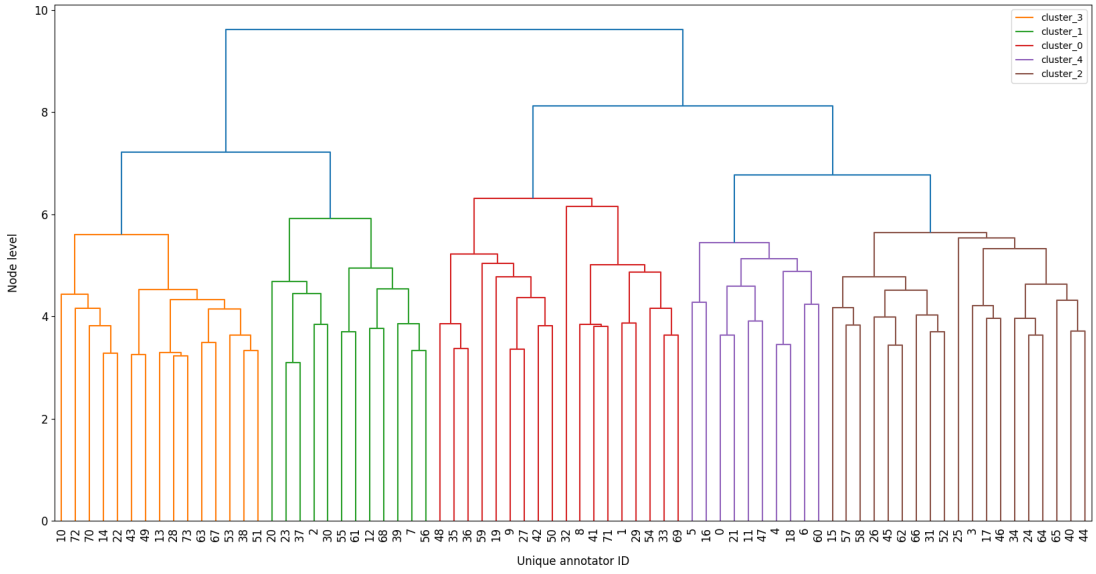
Considering n annotators labeling k instances, we firstly obtained the the label matrix $V^{n \times k}$. We used the α to compute the pairwise agreement between annotators *i* and *j*, resulting in the similarity matrix $A \in \mathbb{R}^{n \times n}$, computed as $A_{i,j} = \alpha(V_{i:}, V_{j:})$. Finally, we obtained a distance matrix $D = 1 - A$, used as input for the unsupervised clustering algorithms.

Given the high sparsity of the matrix, and the annotation distribution already discussed in Section 3, we have encountered 82 cases in which annotators did not have any common annotation. Since missing values are not acceptable in agglomerative clustering, we decided to assign $\alpha = 0$. As a consequence, we assumed no correlation between the two in the clustering phase, totally relying on the similarities that these annotators might have with other raters. While this is a strong assumption, made for practical reasons, the incidence of such pairs of annotators is very low, i.e., about 1% of all the pairs.

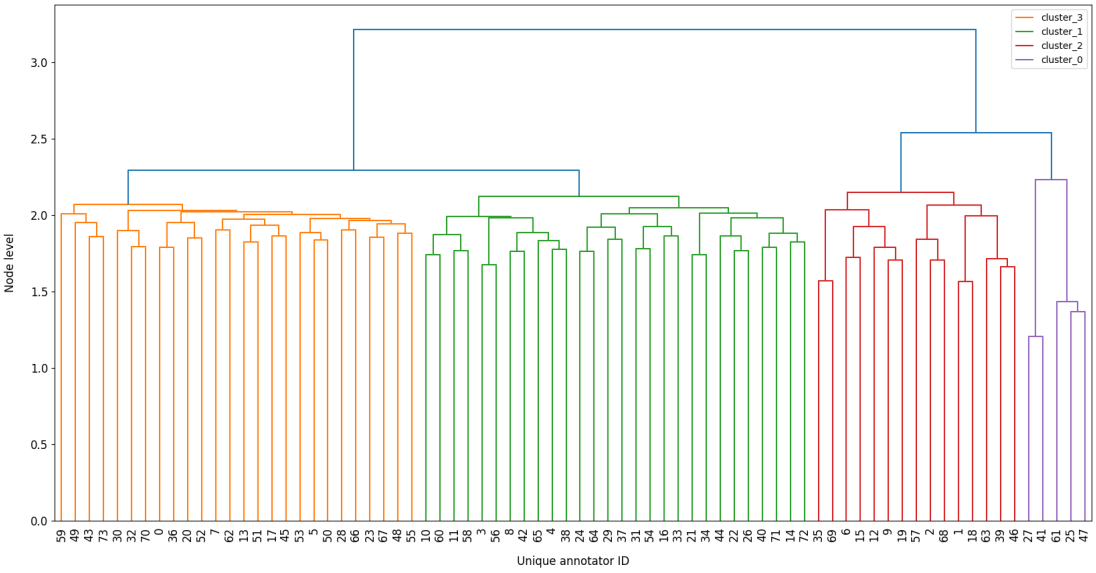
Moreover, in computing α we have encountered a major limitation of the metric itself, already pointed out by Checco et al. [28] as a "paradox" that makes systematic agreement less reliable than random guessing. In fact, in 158 cases, although there was perfect agreement between pairs of annotators, the number of samples was not enough for the α to be well-defined. In these cases, we relaxed this constraint by setting $\alpha = 1$ for the sake of the further clustering steps.

4.2. Nonlinear PCA

As a second method to vectorize annotators' perspective, we have performed a dimensionality reduction of the label matrix $V^{n \times k}$. Since it was a sparse matrix with a highly number of missing values, we have firstly applied



(a) Label agreement (α)



(b) Dimensionality reduction (KPCA)

Figure 1: Dendrograms obtained via the two representation thechniques: Label agreement (a), and Dimensionality reduction (b).

a one-hot encoding considering the three possible categories: ironic (encoded as 01), not ironic (encoded as 10) and missing value (encoded as 00). We obtained a new matrix with twice as many columns as the original label matrix, which has been reduced via Kernel Principal Component Analysis, using the Scikit-learn decomposition package.

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of data by applying an orthogonal linear transformation into a low dimensional subspace, keeping the largest variance as possible in order to avoid loosing relevant information. As an extension of it, Kernel PCA makes possible to apply a nonlinear mapping of the data into a high-dimensional

feature space [27] using kernel methods.

We have firstly tried to apply regular Principal Component Analysis selecting 59 components to keep the 85.7% of the variance. When computing the pairwise distance of the reduced matrix with either euclidean, cosine or manhattan metrics, we obtained a poorly informative dendrogram, suggesting that our data might not be linearly separable.

For this reason we opted for a nonlinear PCA; we computed a dendrogram for multiple kernels, and eventually we chose the cosine similarity as the kernel that resulted in the most balanced clustering. For the number of components, we calculated the ratio between the sum of the eigenvalues λ_i of k components, and the sum of the eigenvalues λ_j of all non-zero components n:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_j}$$

We tried with multiple fixed dimensionalities k, and stopped at 60 components to explain the 85.5% of the variance. Then, we obtained a distance matrix computing the pairwise distance of our reduced matrix, calculated via the euclidean metric.

4.3. Hierarchical clustering

After obtaining a distance matrix of the annotators for each of the two representation techniques described in previous sections, we used the library Scikit learn to perform hard clustering on both data. Specifically, we computed a clustering to have a graphical representation of how the annotators join together, and how clusters themselves are connected to each other by analyzing the resulting nodes.

In both cases, we opted for Ward's linkage criterion, calculating the linkage with the euclidean distance metric, as the method requires, and computing the full tree. It resulted in the clusters illustrated by the dendrograms in Figure 1. DBSCAN and Affinity Propagation were also tried as clustering algorithms, however they did not converge to usable clusters on our dataset.

Choosing the number of clusters Once the two clusterings are computed, we applied the Calinski Harabaz [29] and Davies Bouldin Indexes [30] to respectively measure their density and their similarity. We used these intrinsic evaluation metrics to assess the best number of clusters between 2 and 5, adding a further analysis with 11 clusters as the sum of the number of demographic traits considered for the perspective-aware models in [7]. Since these two metrics do not need any ground truth labels, we were able to perform an intrinsic clustering validation comparing the scores among clusters of the same representation technique, and considering the combination of the two measures together with the computed dendrograms. The results show that to a lower number of clusters corresponds an increase in density and separation (higher Calinski Harabaz Index), together with an increasing generalization, thus having clusters more similar among each other (higher Davies Bouldin Index). We tried to balance these two effects, by minimizing the ratio between the two metrics, and assigned a number of 5 clusters to the clustering obtained with α, and a number of 4 for KPCA.

Cluster	Node level	N. annotators	Label rate iro	Label rate not
0	6.307	18	17.1%	82.9%
1	5.919	12	43.6%	56.4%
2	6.639	19	31.9%	68.1%
3	5.600	15	44.7%	55.3%
4	5.444	10	19.8%	80.2%

Table 1
Quantitative information about clusters obtained via Krippendorff's alpha (first representation technique).

Cluster	Node level	N. annotators	Label rate iro	Label rate not
0	2.232	5	24,1%	75.9%
1	2.122	28	26,2%	73.8%
2	2.149	15	29,4%	70.6%
3	2.070	26	40,3%	59.7%

Table 2
Quantitative information about clusters obtained via Kernel PCA (second representation technique).

4.4. Quantitative analysis

Comparing the two figures, it is possible to notice that in the second representation technique, in cluster 1 and cluster 3 (Figure 1 (b)) the first nodes formed when the two most similar items joined together are almost at the same level of the cluster formation. Moreover, as illustrated in Table 2, the four clusters join nearly at the same level, showing a lower distance between them. This is reflected by a systematically lower Silhouette score for the clusters obtained applying the Kernel PCA, in respect to the first representation technique Figure 1 (a), where the distance between the clusters is well defined and reflected by the different height of all the nodes, including the ones where the clusters are formed (Table 1).

Looking at the positive label rate, it is higher in cluster 1, 2 and 3 from the α representation technique (Table 1) and cluster 3 from the KPCA representation technique (Table 2), indicating a major sensitivity of these annotators to irony.

Representation technique	Demographics	ARI	AMI
α	Gender	0.030	0.032
	Nationality	-0.007	-0.007
	Generation	-0.002	-0.009
KPCA	Gender	-0.001	0.007
	Nationality	0.104	0.195
	Generation	-0.004	0.052

Table 3
Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) in respect to annotators' demographics for both first and second representation techniques.

4.5. Qualitative analysis

To see whether there was a correlation between the obtained clusters and demographics, we firstly leveraged the Rand index (ARI) [31] and the Mutual information (AMI) [32] both adjusted by chance. The former estimates the similarity between two clusterings, while the latter is a measure of similarity between two labels. Both metrics are typically used to validate the output of a clustering algorithm. However, in this work they were used to infer a mapping between our cluster and each of the annotators' demographics (gender, generation and nationality), treated as the ground truth. The results in Table 3 show a negative correlation for at least one of the two measures in most of the cases, with the exception of gender for the α representation technique, and nationality for the KPCA-based one. Especially in the latter, both the ARI and AMI tend to be higher than other scores, which instead are always very close to zero. This result is in line with recent observations that using demographic information about the annotators does not necessarily guarantee a better performance in terms of perspective modeling [33].

Consequently, we further explored the correlation with demographic data: we looked at the composition of the clusters with respect to gender, nationality and generation,[2] as illustrated in Table 4 and Table 5.

From the clusters obtained via Krippendorff's alpha (α), we did not find any systematic mapping between demographic traits and the clusters. In particular, in Table 4 there are cases in which a cluster represents one social group more than another, as focusing on gender, is possible to notice that cluster 3 has a small percentage of female annotators. Considering nationality, this same cluster has the 40% of the British annotator, totally absent in cluster 1, and most of the indian and irish annotators are represented in the first three clusters. An unbalanced representation can be individuated also when looking at generations, especially in respect to the boomer an GenZ annotators: the former are totally absent in cluster 0 and 1, and the latter are concentrated especially in cluster 0 and cluster 2 in respect to the remaining two. Nevertheless, no partition of demographic group can be highlighted, since none of the considered social groups merges homogeneously into specific clusters.

Dem. data	Cl 0	Cl 1	Cl 2	Cl 3	Cl 4	tot.
Female	34.3%	11.4%	31.4%	8.6%	14.3%	100%
Male	15.4%	20.5%	20.5%	30.8%	12.8%	100%
Australia	26.7%	13.3%	20%	26.7%	13.3%	100%
India	33.3%	33.3%	20%	6.7%	6.7%	100%
Ireland	26.7%	20%	33.3%	6.7%	13.3%	100%
UK	20%	0%	26.7%	40%	13.3%	100%
US	14.3%	14.3%	28.6%	21.4%	21.4%	100%
Boomer	0%	0%	33.3%	33.3%	33.3%	100%
GenX	36.4%	18.2%	13.6%	13.6%	18.2%	100%
GenY	18.4%	15.8%	31.6%	21.1%	13.2%	100%
GenZ	30%	10%	30%	30%	0%	100%

Table 4
Distribution of annotators in each cluster considering each demographic trait (first representation technique).

We obtained different results with the KPCA-based representation technique, especially looking at the nationality and generation traits (Table 5). Coherently with the correlation showed in Table 3, annotators from all nationalities are divided almost perfectly between two clusters (cluster 1 and 3 for Australian and British, cluster 0 and 2 for Indian, cluster 2 and 3 for Irish raters), with the exception of annotators from the US, almost completely clustered in 1. A similar pattern can be found looking at generations: Boomers are entirely represented in cluster 1, which also hosts 60% of GenZ annotators. Note however that these two cohorts of annotators are the less numerous.

Dem. data	Cl 0	Cl 1	Cl 2	Cl 3	tot.
Female	11.4%	34.3%	14.3%	40%	100%
Male	2.6%	41%	25.6%	30.8%	100%
Australia	0%	46.7%	0%	53.3%	100%
India	33.3%	13.3%	40%	13.3%	100%
Ireland	0%	13.3%	53.3%	33.3%	100%
UK	0%	46.7%	6.7%	46.7%	100%
US	0%	71.4%	0%	28.6%	100%
Boomer	0%	100%	0%	0%	100%
GenX	0%	36.4%	22.7%	40.9%	100%
GenY	13.2%	28.9%	15.8%	42.1%	100%
GenZ	0%	60%	30%	10%	100%

Table 5
Distribution of annotators in each cluster considering each demographic trait (second representation technique).

These results show that the two methods subject of our experimentation lead to conceptually different results. In the first representation technique, we interpreted the

[2]For this analysis, we excluded a single annotator for whom age was not disclosed, clustered in cluster 1 (α), and cluster 2 (KPCA).

Representation technique	Cluster	#Instances
α	0	1,570
	1	1,216
	2	1,693
	3	1,431
	4	1,214
KPCA	0	534
	1	1,901
	2	1,269
	3	1,768

Table 6
Datasets extracted from the clusters for each representation technique.

similarity between annotator pairs in terms of inter annotator agreement, while in the second we worked directly on the vectors of the annotators, positioning them on a feature space and calculating their label-based distances. This second approach, in particular, seems better at capturing the impact of nationality and generation in defining what a rater considers ironic.

5. Modelling mined perspectives

In this section we present experiments carried out to validate our methodology. In particular, we created perspective-aware models [16] based on the automatically extracted clusters of annotators, ensembled them, and explored the difference between non-perspectivist and cluster-based ensemble models both in-dataset and cross-dataset.

As regarding the experimental setup, we fine-tuned the uncased version of BERT [3] [34] for sequence classification. The input consisted in the *Post*, *Reply* pairs. We set a batch size of 16 and a learning rate of $5 \cdot 10^{-5}$ and, to prevent overfitting, we customized the model to implement the Focal Loss [35]. Finally, we set early stopping with a patience of 2 epochs on the validation loss (using 20% of the training data as validation set).

As a baseline (called NP for non-perspectivist), we aggregated the annotations via majority voting and discarded those where a majority was not found, adopting the methodology explained in Section 3. Thus, we trained the model on the aggregated set of 1,771 instances, and tested it on the GOLD TEST SET. For the models based on the two clustering techniques, we implemented an ensemble strategy, inspired by [16]: for each cluster we created a gold standard to train a perspective-aware model, and applied majority voting on their predictions, obtaining an ensemble classifier per technique. We tested the models on the GOLD TEST SET and compared the results with the baseline.

To train the cluster-based models, we firstly excluded the GOLD TEST SET, and grouped the remaining label-texts pairs according to each of the obtained clusters, extracting 5 and 4 datasets respectively for the first and second representation technique. Eventually, we applied a majority voting strategy and excluded those instances where a majority was not present. Table 6 illustrates the number of instances per dataset.

After training, we tested the models both in a in-dataset (on EPIC's GOLD TEST SET) and cross-dataset setting, specifically on SemEval 2018 Task 3 test set [25], previously described in Section 3.2. Finally, we implemented a majority voting ensemble (*M-ENS*), that returns a final label by applying majority vote over the predictions of each cluster-based classifier. Table 7 shows the average precision, recall and F1-score over 10 runs. We found low variation in the scores, as illustrated by the standard deviation in parenthesis.

Looking at Table 7, we can notice that the two majority ensembles obtained from the explored representation techniques always outperform the baseline, both in-dataset and cross-dataset. In the first setting, the macro-averaged F1 score of the *M-ENS* α gives the best results, while *M-ENS* KPCA presents the best performance cross-dataset. Results demonstrate that modelling annotators' opinions is necessary when working on highly subjective phenomena as irony, as strongly confirmed by the performance of cluster-based ensembles in a cross-dataset setting. More importantly, these experiments prove that training perspective-aware models based on annotators' mined opinion can be an effective instrument to capture a diversity of points of view.

Notably, the increase in macro-F1 score is a reflection of a better prediction of the positive class. Considering that the classes were highly unbalanced (see Section 3.1) the accuracy measure is higher for the baseline model, which is less sensitive to the presence of irony and therefore over-predicts the negative class.

Despite the clusters obtained in the two representation techniques being very different in terms of methodology (Section 4.1, Section 4.2) and composition (Section 4.3), the models exhibit comparable performance. In-dataset, the ensemble based on α clusters gives slightly better scores than KPCA; but this trend is inverted in the second setting.

These results confirm the idea that by mining annotator perspectives we can let the annotators' opinions emerge regardless of their demographics, observing how social background can influence the individual's definition of what is ironic, shared among characteristics that might go beyond common demographic traits.

[3] https://huggingface.co/bert-base-uncased

setting	model	negative class			positive class			macro-average			
		prec.	rec.	F1	prec.	rec.	F1	prec.	rec.	F1 (std)	Acc.
in-dataset	NP	.862	.764	.801	.335	.481	.370	.598	.622	.585 (.055)	.708
	M-ENS α	.880	.727	.795	.357	.598	.443	.618	.663	**.619** (.018)	.702
	M-ENS KPCA	.865	.709	.775	.331	.550	.404	.598	.630	.589 (.036)	.678
cross-dataset	NP	.597	.658	.603	.308	.327	.292	.452	.492	.448 (.053)	.527
	M-ENS α	.590	.515	.536	.355	.453	.388	.472	.484	.462 (.047)	.490
	M-ENS KPCA	.636	.468	.501	.394	.556	.438	.515	.512	**.470** (.038)	.503

Table 7
In-dataset and cross-dataset performance of cluster-based models, trained on EPIC, and tested respectively on EPIC and SemEval2018 Task 3 test set. In parenthesis the standard deviation of F1 score over ten runs, in-dataset and cross-dataset.

6. Conclusion

In this paper, we implemented and tested two techniques to mine annotator perspectives, moving from the idea that the set of their annotations can be used as a representation of their opinion on the topic they are annotating, in our case ironic content in social media platforms. We chose to perform this analysis on irony since it is a highly subjective phenomenon where not only demographic, but also linguistic and social aspects can influence annotators' interpretation and judgement. For this reason, we used the recently published English Perspectivist Irony Corpus (EPIC).

For mining annotators' perspectives we proposed two methodologies. The former, inspired by [24], was to interpret similarity of opinions in terms of inter-annotator agreement, adapting Krippendorff's alpha and overcoming its structural limitations. The latter consisted in a dimensionality reduction of annotator vectors, using Kernel Principal Component Analysis, thus applying a non-linear mapping of our data. Then, we applied a hierarchical clustering algorithm to analyse how annotators group together. Looking at the composition of clusters in respect to annotators' demographic data, results demonstrate how different the two mining techniques are. In fact, Kernel PCA highlights the correlation between annotators' nationality and irony perception, while the first method returns more heterogeneous and better balanced clusters.

In the experimental phase, we trained perspective-aware models for each cluster obtained via the two representation techniques, and implemented an ensemble strategy to select the predicted labels, based on majority voting. Both in-dataset and cross-dataset performance showed that the ensemble models always outperform the baseline, demonstrating the robustness of our method also when tested on a different corpus.

Considering these promising results, we believe that this approach can be of fundamental use for future research in the perspectivist field. Firstly, it makes possible to mine annotators' opinions when demographic information are not known. Secondly, it can help to avoid built-in biases in creating perspective-aware classifiers, testing whether annotators' choices might be driven by factors uncorrelated to given demographics, but rather linked to other elements of their social and individual background.

Although we tackled the Krippendorff's alpha paradox described in Section 4.1, there are other abnormalities of the measure itself extensively described in [28], which might had a negative impact on the clusters obtained via the first representation technique.

Moreover, in this work we group annotators using a hard clustering algorithm. However, as reality is more nuanced and many dimensions interact in describing human variability, a soft clustering approach could lead to more accurate representations, although its application is computationally more complex in this context.

For the future, we plan to perform the same experiments on multiple pre-trained language models, to further test the consistency of our results, and test other representation techniques such as autoencoders. Our analysis of the composition of the annotator clusters indicates some degree of intersectionality of demographic traits with respect to the annotation of irony, which we consider a research direction to pursue further. Another aspect worth investigating is the relative position of individual annotators among their assigned clusters, checking whether it correlates with factors like annotation quality. Finally, while our results are very encouraging, it must be noted that the experimental task still involved an aggregated test benchmark. We expect that our method will produce more impactful results when measured on a perspectivist, disaggregated benchmark, which we aim to develop in the next steps of our research.

References

[1] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24.

[2] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, et al., We need to consider disagreement in evaluation, in: Proceed-

ings of the 1st workshop on benchmarking: past, present and future, Association for Computational Linguistics, 2021, pp. 15–21.
[3] S. Akhtar, V. Basile, V. Patti, Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, arXiv preprint arXiv:2106.15896 (2021).
[4] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Washington DC, USA, 2023.
[5] V. Basile, 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, volume 2776, CEUR-WS, 2020, pp. 31–40.
[6] B. Plank, The'problem'of human label variation: On ground truth in data, modeling and evaluation, arXiv preprint arXiv:2211.02570 (2022).
[7] S. Frenda, A. Pedrani, V. Basile, S. M. Lo, A. T. Cignarella, R. Panizzon, C. Marco, B. Scarlini, V. Patti, C. Bosco, D. Bernardi, Epic: Multi-perspective annotation of a corpus of irony, in: ACL 2023, 2023. URL: https://www.amazon.science/publications/epic-multi-perspective-annotation-of-a-corpus-of-irony.
[8] A. Joshi, P. Bhattacharyya, M. J. Carman, Investigations in computational sarcasm, Springer, 2018.
[9] R. Ortega-Bueno, F. Rangel, D. Hernández Farıas, P. Rosso, M. Montes-y Gómez, J. E. Medina Pagola, Overview of the task on irony detection in spanish variants, in: Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org, volume 2421, 2019, pp. 229–256.
[10] J. Karoui, F. Benamara, V. Moriceau, V. Patti, C. Bosco, N. Aussenac-Gilles, Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 262–272.
[11] E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, arXiv preprint arXiv:2109.13563 (2021).
[12] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, P. Kazienko, Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach, Information Processing & Management 58 (2021) 102643.
[13] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.
[14] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, N. A. Smith, Annotators with attitudes: How annotator beliefs and identities bias toxic language detection, arXiv preprint arXiv:2111.07997 (2021).
[15] V. Prabhakaran, A. M. Davani, M. Diaz, On releasing annotator-level labels and information in datasets, arXiv preprint arXiv:2110.05699 (2021).
[16] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 8, 2020, pp. 151–154.
[17] A. A. Taha, L. Hennig, P. Knoth, Confidence estimation of classification based on the distribution of the neural network output layer, arXiv preprint arXiv:2210.07745 (2022).
[18] L. Flek, Returning the n to nlp: Towards contextually personalized classification models, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 7828–7838.
[19] J. Bielaniewicz, K. Kanclerz, P. Miłkowski, M. Gruza, K. Karanowski, P. Kazienko, J. Kocoń, Deep-sheep: Sense of humor extraction from embeddings in the personalized context, in: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2022, pp. 967–974.
[20] J. Kocoń, M. Gruza, J. Bielaniewicz, D. Grimling, K. Kanclerz, P. Miłkowski, P. Kazienko, Learning personal human biases and representations for subjective tasks in natural language processing, in: 2021 IEEE International Conference on Data Mining (ICDM), IEEE, 2021, pp. 1168–1173.
[21] P. Kazienko, J. Bielaniewicz, M. Gruza, K. Kanclerz, K. Karanowski, P. Miłkowski, J. Kocoń, Human-centred neural reasoning for subjective content processing: Hate speech, emotions, and humor, Information Fusion (2023).
[22] J. Plepi, B. Neuendorf, L. Flek, C. Welch, Unifying data perspectivism and personalization: An application to social norms, arXiv preprint arXiv:2210.14531 (2022).
[23] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, P. Kazienko, What if ground truth is subjective? personalized deep neural hate speech detection, in: Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022, 2022, pp. 37–45.
[24] M. Fell, S. Akhtar, V. Basile, Mining annotator perspectives from hate speech corpora., in: NL4AI@ AI* IA, 2021.
[25] C. Van Hee, E. Lefever, V. Hoste, Semeval-2018 task 3: Irony detection in english tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, 2018, pp. 39–50.

[26] K. Krippendorff, Computing krippendorff's alpha-reliability (2011).
[27] B. Schölkopf, A. Smola, K.-R. Müller, Kernel principal component analysis, in: Artificial Neural Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings, Springer, 2005, pp. 583–588.
[28] A. Checco, K. Roitero, E. Maddalena, S. Mizzaro, G. Demartini, Let's agree to disagree: Fixing agreement measures for crowdsourcing, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 5, 2017, pp. 11–20.
[29] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 3 (1974) 1–27.
[30] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence (1979) 224–227.
[31] L. Hubert, P. Arabie, Comparing partitions, Journal of classification 2 (1985) 193–218.
[32] N. X. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: is a correction for chance necessary?, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 1073–1080.
[33] M. Orlikowski, P. Röttger, P. Cimiano, D. H. B. University, U. of Oxford, C. S. Department, B. University, Milan, Italy., The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics, ArXiv abs/2306.11559 (2023).
[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.