PERSEID - Perspectivist Irony Detection: A CALAMITA Challenge

Valerio Basile[1], Silvia Casola[2], Simona Frenda[3,4] and Soda Marem Lo[1]

[1]University of Turin, Italy
[2]MaiNLP & MCML, LMU Munich, Germany
[3]Interaction Lab, Heriot-Watt University, Edinburgh, Scotland
[4]aequa-tech, Turin, Italy

Abstract

Works in perspectivism and human label variation have emphasized the need to collect and leverage various voices and points of view in the whole Natural Language Processing pipeline.
PERSEID places itself in this line of work. We consider the task of irony detection from short social media conversations in Italian collected from Twitter (X) and Reddit. To do so, we leverage data from MultiPICO, a recent multilingual dataset with disaggregated annotations and annotators' metadata, containing 1000 Post, Reply pairs with five annotations each on average. We aim to evaluate whether prompting LLMs with additional annotators' demographic information (namely gender only, age only, and the combination of the two) results in improved performance compared to a baseline in which only the input text is provided.
The evaluation is zero-shot; and we evaluate the results on the disaggregated annotations using f1.

Keywords

Perspectivism, Irony Detection, Evaluation

1. Challenge: Introduction and Motivation

Recently, researchers have shown a growing interest in human-centered technologies to make Artificial Intelligence (AI) models and products more attentive to the users' sensitivity and needs.

In Natural Language Processing (NLP), works on perspectivism [1] and human label variation [2] have emphasized the intrinsic variability in human annotation and thus the importance of incorporating a diverse set of voices; this aspect affects all phases of the NLP pipeline, including collecting disaggregated datasets [3, 4, 5], analyzing existing disagreement [6], learning from disaggregated data [7, 8], and evaluating considering several voices as valid [9, 1].

During the data collection and annotation phase, works in this area have gone beyond considering disagreement as motivated by noise only and thus as an attribute to be minimized and resolved, e.g., through majority voting. In contrast, research has emphasized the necessity of collecting a variety of voices and considering all such voices as valid. The reason is twofold. On the one hand, researchers have argued that many tasks that are popular in the NLP community (including, for example, hate speech and humor detection) are intrinsically subjective [10], as points of view might differ depending on users' social background, beliefs, and demographics. Using a single aggregated label has thus been increasingly questioned [11, 12, 13], and preserving disaggregated data is preferred. On the other hand, recent work has shown that design choices and biases affect datasets and models and often result in models unexpectedly aligned with a given population segment more than with another [14]; in fact, aggregated data tend to reflect a minority of perspectives, under-representing others [15, 4].

As a result, disaggregated datasets have become more popular, as listed in the Perspectivist Data Manifesto[1] and by Plank [2][2].

Researchers are incresingly reporting annotators' demographics and other metadata when describing the dataset, which was first advised as a good practice to avoid excluding, minimizing, and misrepresenting certain groups of users [16]. Recent work has also explored whether annotators' demographics and background — as described by available metadata — influence their annotation [5, 17, 18, 19, 4] and can help during the modeling of the phenomenon under study [20, 8, 21].

Despite the increasing interest in disaggregated and metadata-rich datasets, few such datasets for irony detection exist. Simpson et al. [22] released a corpus for humor detection in English, used as a benchmark in the first edition of the Learning With Disagreement (LeWiDi) shared task [23]. No annotators' metadata, however, are

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 - 06, 2024, Pisa, Italy
valerio.basile@unito.it (V. Basile); s.casola@lmu.de (S. Casola); s.frenda@hw.ac.uk (S. Frenda); sodamarem.lo@unito.it (S. M. Lo)

[1]https://pdai.info/
[2]www.github.com/mainlp/awesome-human-label-variation

included. Frenda et al. [4] proposed a dataset for irony detection and investigated the influence of the annotators' demographics on their perception [6]. The dataset contains English texts only.

For this challenge at CALAMITA [24], we propose to use the Italian portion of MultiPICo (Multilingual Perspectivist Irony Corpus)[3] [25]. Multipico is a multilingual corpus of short Post-Reply conversational pairs extracted from Twitter and Reddit and annotated as ironic or not ironic by crowdsourcing workers with different demographics and backgrounds. MultiPICo covers 9 languages (Arabic, English, Dutch, French, German, Hindi, Italian, Portuguese, and Spanish) and 25 language varieties[4], ranging from high- to low-resourced ones. Moreover, a rich set of annotators' sociodemographic information (balanced gender, age, nationality, ethnicity, student, and employment status) is provided.

While no perspectivist task leveraging the dataset has been proposed so far, PERSEID is related to the Learning With Disagreement task held at SemEval 2021 [11] and 2023 [13]. In LeWiDi, participant systems were challenged to learn the distribution of labels, tested by cross entropy-based metrics. In contrast, PERSEID aims at stimulating the development of models of human perspectives, in order to explain the label distributions rather than just quantifying them.

2. Challenge: Description

The task of Perspectivist Irony Detection aims to measure models' capability to detect irony in a short verbal exchange for each annotator, conditioned on the knowledge of demographic information about them. To this purpose, we want to look at different model performances if it is informed by one demographic trait or a combination of two. In particular, we focus on the gender and age of the annotator, due to the balanced number of male and female annotators by design 3.2, and due to the fact that age was shown to be one of the most polarized dimensions in [25].

The input to the task does not consist only of a text, but rather of a tuple <PERSPECTIVE, POST, REPLY>.

In this iteration of PERSEID, we considered several variables for the PERSPECTIVE attribute:

- None (Task 0): acting as a baseline, we want to investigate the models' outputs when no information about the annotator is provided.
- Age (Task 1): the PERSPECTIVE is one of four values encoding the age group of the annotator.
- Gender (Task 2): the PERSPECTIVE is the binary self-identified gender of the annotator.
- Age + Gender (Task 3): in this case, both attributes are provided as the PERSPECTIVE.

The POST is a textual post, to which the target REPLY is a reply. The output of the prediction is a binary label indicating whether the REPLY is ironic (or non-ironic) for a human bearing the characteristic of the PERSPECTIVE to the TEXT. The performance of the model is evaluated through a global f1 metric on the disaggregated annotations.

The challenge is zero-shot: no training, fine-tuning, or in-context learning is considered for this version of PERSEID and the whole dataset can be used for inference.

Note that since each annotator can be described by no traits (Task 0), one single trait (Task 1 and Task 2), and two traits (Task 3), we do not aim at optimal performance when considering personalized irony detection; instead, our goal is to understand whether models improve their performance when one or multiple traits is provided and to understand the impact of different configurations.

3. Data description

3.1. Origin of data

The data for the challenge are part of MultiPICo [25], a corpus of 18,778 short conversations collected from Reddit (8,956) and Twitter (9,822) in 9 languages, and a total of 25 varieties.

Data were collected to reproduce the structure of short conversations.

For both Reddit and Twitter, the POST is typically a message initiating a thread and the REPLY a direct reply to that message[5].

Reddit data were retrieved using the Pushshift repository[6] from January 2020 to June 2021. For Italian, data were downloaded from the subreddit /r/Italy.

Pairs having at least one deleted or removed comment were filtered out, and the language of the messages was further validated using the Python library for language identification LangID[7].

Twitter data were collected via Twitter Stream API, using the geolocation service and excluding quotes and retweets. Then, the full conversation was retrieved, and tweets that directly replied to the starting ones were retained.

The data collection resulted in 18,778 instances, together with their metadata, consisting of Post-Reply original IDs, subreddits, and geolocation information.

[3] MultiPICo is available at https://huggingface.co/datasets/Multilingual-Perspectivist-NLU/MultiPICo with a CC-BY 4.0 license.

[4] For example, texts in Austrian, German, and Swiss German are included in the dataset.

[5] For Reddit, second-level replies were collected in a minority of cases; for Twitter, the POST is a reply to a thread-starting message in a minority of cases.

[6] https://redditsearch.io/

[7] https://github.com/saffsd/langid.py

Language	#Annotators	#Annotations	Label rate		#Texts	Sources		Annotation mean
			%not	%iro		#Reddit	#Twitter	
Arabic	68	10,609	68	32	2,181	949	1,232	4.86
Dutch	25	4,991	73	27	1,000	500	500	4.99
English	74	14,171	69	31	2,999	1,499	1,500	4.73
French	50	8,770	70	30	1,760	1,000	760	4.98
German	70	12,510	68	32	2,375	1,042	1,333	5.27
Hindi	24	4,711	65	35	786	286	500	5.99
Italian	**24**	**4,790**	**69**	**31**	**1,000**	**500**	**500**	**4.79**
Portuguese	49	9,754	62	38	1,994	997	997	4.89
Spanish	122	24,036	67	33	4,683	2,183	2,500	5.13
Total	506	94,342	68	32	18,778	8,956	9,822	5.02

Table 1
Number of annotators, annotations, texts per source, and annotation means for each language. For Italian, 1000 pairs were collected, each annotated by 4.79 annotators. Note the label unbalance, with the negative class accounting for 69% of the total annotations.

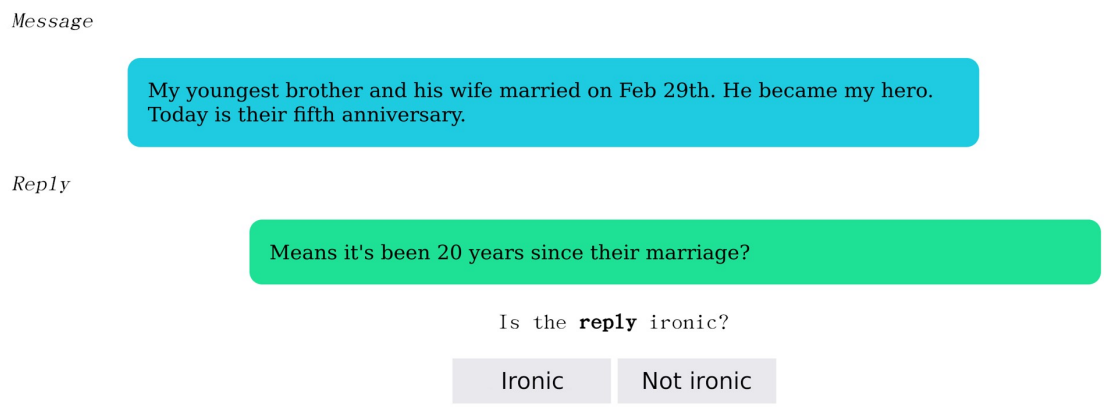

Figure 1: Screenshot of the annotation interface for an English instance of MultiPICo. The Italian interface was similar, with translated question and options.

For Italian, data account for 1000 POST, REPLY pairs, equally sourced from Reddit and Twitter.

3.2. Annotation details

Annotators were asked to read a set of POST and REPLY pairs and answer whether the text of the REPLY was ironic or not, given the context.

The human annotation of the collected data was performed on the crowdsourcing platform Prolific[8], through a custom-built annotation interface designed to collect a diverse and balanced set of annotators. The interface mimicked a message conversation, having the POST as context and asking whether the REPLY was Ironic or Not ironic.

For Italian, 24 native-speaker annotators were hired, who performed 4,790 annotations in total, resulting in a mean of 4,79 annotations per instance (see Table 1).

[8]https://www.prolific.com/

Annotators were selected based on three criteria:

- Their completion rate had to be greater or equal to 99%
- They had to be native speakers of the considered language (i.e., Italian, for the portion of data used in the challenges)
- The set of annotators needed to be balanced across genders.

The quality of the annotation was further assured using attention check questions in the form of *"Please answer X to this question"*. Annotators had 1% probability of receiving these special questions. Annotators who failed to respond correctly to at least 50% of these questions were excluded from the final corpus.

A rich set of metadata is also provided. These include the self-identified Gender (balanced by design), their nationality, their *Age Group* (1 GenX, 15 GenY, 8 GenZ, for Italian), *Ethnicity* (23 white people, 1 mixed person, for

Demographics		Languages								
		English	Spanish	**Italian**	French	Dutch	German	Hindi	Arabic	Portoguese
Age group	Boomer	3	2	–	2	–	5	–	1	–
	GenX	22	17	**1**	7	4	7	3	4	1
	GenY	38	66	**15**	23	10	36	13	36	23
	GenZ	10	37	**8**	17	11	20	8	26	25
Ethnicity	White	47	60	**23**	40	22	66	–	20	37
	Mixed	1	31	**1**	3	2	3	–	13	10
	Asian	18	1	–	1	1	–	22	1	–
	Black	3	2	–	5	–	–	–	2	1
	Other	3	27	–	1	–	1	8	31	1
Student	Yes	13	39	**14**	16	7	14	8	29	30
	No	46	60	**9**	30	16	39	14	25	16
Employment	Full-time	25	41	**9**	24	10	24	10	20	15
	Unemployed	11	24	**7**	5	4	3	1	11	8
	Part-time	11	17	**5**	5	3	10	4	13	6
	Not in paid work	4	4	**1**	5	4	5	–	1	–
	Due to start	–	3	**1**	1	–	2	2	–	2
	Other	1	6	–	6	–	3	1	5	14

Table 2
Sociodemographic information about annotators per language.

Italian), *Student status* (14 yes, 9 no, for Italian), *Employment status* (9 in full-time jobs, 7 unemployed, 5 working part-time, 1 not in paid work and 1 due to start, for Italian), as reported in Table 2.

3.3. Data format

The dataset is in tabular format, one row per annotation. The data contain the text in the form of two fields (POST and REPLY), the binary LABEL, and a series of metadata about the post, reply, and annotator. Here is an example of instance from the Italian section of MultiPICo:

```
'Age': 29.0,
'Country of birth': 'Italy',
'Country of residence': 'Italy',
'Employed': 'Yes',
'Employment status': 'Part-Time',
'Ethnicity simplified': 'White',
'Gender': 'Male',
'Generation': 'GenY',
'GenerationAggregated': 'Young',
'Nationality': 'Italy',
'Student status': 'No',
'annotator_id': 9208155880570654046,
'label': 0,
'language': 'it',
'language_variety': 'it',
'level': 1.0,
'post': 'Ormai il quadro è chiaro: cercare di
   coinvolgere tutti per non farla pagare a
   nessuno. Se non riuscissero a corrompere i
   Pm di Torino andranno in B diretti.',
'post_id': 14071953227682835778,
'reply': '@USER Magari ??',
'reply_id': 2497527360959166890,
'source': 'twitter',
'timestamp': '2022-12-07 15:49:50'
```

3.4. Example of prompts used for zero-shot prediction

The challenge is zero-shot, and the prompt depends on three variables: PERSPECTIVE, POST, and REPLY.

```
Sei {perspective}.
Istruzione: Ti vengono fornite in
input (Input) una coppia di frasi
(Post, Reply) estratte da conversazioni
sui social media. Il tuo compito è
determinare se la Risposta (Reply) è
ironica nel contesto del Post (Post).
Fornisci in output (Output) una singola
etichetta “ironia" o "non ironia".
Input:
Post: {post}
Reply: {reply}
Output:
```

Task 0 No PERSPECTIVE is provided, and the prompt directly starts with the instruction.

Task 1 The PERSPECTIVE variable is a verbalization of the GENERATION, which is expressed as an integer in the dataset. It can be instantiated with the following values[9]:

[9]No workers whose age is > 42, i.e., from the baby boomer generations, participated in the annotation of the Italian portion of the dataset

- "una persona giovane della generazione Z" if GENERATION == GenZ (AGE < 26)
- "una persona giovane della generazione Y" if GENERATION == GenY ($26 \leq$ AGE < 42)
- "una persona adulta della generazione X" if GENERATION == GenX ($42 \leq$ AGE < 58)
- "una persona adulta della generazione baby boomer" if GENERATION == Boomer (AGE > 58)

Task 2 The PERSPECTIVE variable is a verbalization of the GENDER variable, which is expressed as a string in English. It can be instantiated with one of two values:

- "una donna" if Gender == "Female"
- "un uomo" if Gender == "Male"

Task 3 The PERSPECTIVE variable is a verbalization of both the AGE and GENDER variables, e.g., "una giovane donna della generazione Z."

4. Metrics

Inspired by Mokhberian et al. [26], the Perspectivist Irony Detection task is evaluated by means of *global F1*, that is, the F1-score computed across all the individual annotations in the dataset against the predictions of the model.

5. Limitations

Data The sociodemographic information about the annotators is partial, bound to what was available from the crowdsourcing platform, and following a discretization of human personal traits that could be perceived as forced (e.g., representing self-identified gender as a single binary label). Furthermore, as shown by Orlikowski et al. [21], annotators' sociodemographics do not always align with the most relevant grouping of annotators according to the language phenomenon under study.

Annotators of the Italian portion of MultiPICO tend to be young (with no annotators from the baby boomer generation and only one from GenX). This aspect might influence the results.

Similarly to Sachdeva et al. [5], Sap et al. [19], Forbes et al. [27], we noticed the ethnicity of annotators is unbalanced, and all but one annotators are white for the considered data.

In the vast majority (~90%) of cases, the conversation-starting messages and their direct replies were downloaded to capture the full conversational context. In a few cases, the downloaded reply was not direct but rather a second-level reply (a reply to a direct reply); thus, some conversational context might be missing.

Challenge design We describe annotators by no sociodemographic traits (Task 0), one single demographic trait (Task 1 and Task 2), or two demographic traits (Task 3). We evaluate disaggregated annotations at inference time, having the annotators represented only by those traits. Annotators' sociodemographic information does not always align with the most relevant grouping of annotators according to the language phenomenon under study [21, 28], and the limited amount of sociodemographic traits we provide is undoubtedly not enough to describe every single annotator. We are aware of this limitation. In fact, our main aim is to understand whether providing one or more annotator traits makes the model predictions more aligned with annotators having a given characteristic.

6. Ethical issues

This work places itself in an increasing amount of work that calls to consider and include the subjectivity of the annotators in NLP applications, encouraging reflection on the different perspectives encoded in annotated datasets to minimize the amplification of biases. We hope this challenge will be a starting point for investigating and evaluating LLMs in Italian to make them suitable for final users.

The dataset used in the challenge was built by adopting measures to protect the privacy of annotators, and the data handling protocols were designed to safeguard personal information (like anonymization of users' mentions). Although the attention during the collection of data was focused on ironic content spread online, we acknowledge that some of the material contains racist, sexist, stereotypical, violent, or generally disturbing content.

Annotators are balanced through their self-identified gender. However, we are aware that considering gender in a binary form is limited; moreover, a substantial unbalance for some dimensions, like the self-identified ethnicities, is present in the dataset. This pattern suggests the need to interact differently with annotators or social communities if we want a diversity of annotators and perspectives in terms of social background.

7. Data license and copyright issues

Acknowledgments

This work was funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Alexa.

References

[1] 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: Proceedings of the 1st workshop on benchmarking: past, present and future, Association for Computational Linguistics, 2021, pp. 15–21.

[2] B. Plank, 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, 2022, pp. 10671–10682.

[3] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 6860–6868. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25840.

[4] 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: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13844–13857. URL: https://aclanthology.org/2023.acl-long.774. doi:10.18653/v1/2023.acl-long.774.

[5] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 83–94. URL: https://aclanthology.org/2022.nlperspectives-1.11.

[6] S. Frenda, S. M. Lo, S. Casola, B. Scarlini, C. Marco, V. Basile, D. Bernardi, Does anyone see the irony here? Analysis of perspective-aware model predictions in irony detection, in: ECAI 2023 Workshop on Perspectivist Approaches to NLP, 2023.

[7] A. Mostafazadeh Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. URL: https://aclanthology.org/2022.tacl-1.6. doi:10.1162/tacl_a_00449.

[8] S. Casola, S. Lo, V. Basile, S. Frenda, A. Cignarella, V. Patti, C. Bosco, Confidence-based ensembling of perspective-aware models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3496–3507. URL: https://aclanthology.org/2023.emnlp-main.212. doi:10.18653/v1/2023.emnlp-main.212.

[9] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[10] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564. doi:10.1609/aimag.v36i1.2564.

[11] E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, p. 10528–10539.

[12] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, Semeval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 338–347.

[13] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, p. 2304–2318.

[14] S. Santy, J. Liang, R. Le Bras, K. Reinecke, M. Sap, NLPositionality: Characterizing design biases of datasets and models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9080–9102. URL: https://aclanthology.org/2023.acl-long.505. doi:10.18653/v1/2023.acl-long.505.

[15] V. Prabhakaran, A. M. Davani, M. Diaz, On re-

leasing annotator-level labels and information in datasets, in: Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop, 2021, p. 133–138.

[16] E. M. Bender, B. Friedman, Data statements for natural language processing: Toward mitigating system bias and enabling better science, Transactions of the Association for Computational Linguistics 6 (2018) 587–604.

[17] D. Almanea, M. Poesio, ArMIS - the Arabic Misogyny and Sexism Corpus with Annotator Subjective Disagreements, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 2282–2291. URL: https://aclanthology.org/2022.lrec-1.244.

[18] 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).

[19] 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, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5884–5906. URL: https://aclanthology.org/2022.naacl-main.431. doi:10.18653/v1/2022.naacl-main.431.

[20] R. Wan, J. Kim, D. Kang, Everyone's voice matters: Quantifying annotation disagreement using demographic information, in: Proceedings of the 37th AAAI Conference on Anrtificial Intelligence - AAAI Special Track on AI for Social Impact, 2023.

[21] M. Orlikowski, P. Röttger, P. Cimiano, D. Hovy, The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1017–1029. URL: https://aclanthology.org/2023.acl-short.88. doi:10.18653/v1/2023.acl-short.88.

[22] E. Simpson, E.-L. Do Dinh, T. Miller, I. Gurevych, Predicting humorousness and metaphor novelty with Gaussian process preference learning, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5716–5728. URL: https://aclanthology.org/P19-1572. doi:10.18653/v1/P19-1572.

[23] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41. doi:10.18653/v1/2021.semeval-1.41.

[24] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlian, in: Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[25] S. Casola, S. Frenda, S. Lo, E. Sezerer, A. Uva, V. Basile, C. Bosco, A. Pedrani, C. Rubagotti, V. Patti, D. Bernardi, MultiPICo: Multilingual perspectivist irony corpus, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 16008–16021. URL: https://aclanthology.org/2024.acl-long.849.

[26] N. Mokhberian, M. Marmarelis, F. Hopp, V. Basile, F. Morstatter, K. Lerman, Capturing perspectives of crowdsourced annotators in subjective learning tasks, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 7337–7349. URL: https://aclanthology.org/2024.naacl-long.407.

[27] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: Learning to reason about social and moral norms, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 653–670. URL: https://aclanthology.org/2020.emnlp-main.48. doi:10.18653/v1/2020.emnlp-main.48.

[28] S. M. Lo, V. Basile, Hierarchical clustering of label-based annotator representations for mining perspectives, in: G. Abercrombie, V. Basile, D. Bernardi, S. Dudy, S. Frenda, L. Havens, E. Leonardelli, S. Tonelli (Eds.), Proceedings of the 2nd Workshop

on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023, volume 3494 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3494/paper8.pdf.