What is Truth in NLP? Reflecting on Progress, Lessons, and Open Challenges as NLPerspectives turns Five

Gavin Abercrombie
The Interaction Lab
School of Mathematical & Computer Sciences
Heriot-Watt University, Edinburgh, Scotland
g.abercrombie@hw.ac.uk.org

Abstract

This paper reflects on five years of the Workshop on Perspectivist Approaches to NLP (NLPerspectives) and examines how this research community has helped to reconceptualise the notion of ground truth in human-labelled data and its classification. As NLP research has increasingly engaged with social and affective tasks, traditional assumptions about annotation reliability–centred on inter-annotator agreement and single 'gold standard' labels–have proven insufficient for capturing the genuine diversity of human perspectives. I review the developments that have driven the 'Perspectivist Turn,' assess its influence on mainstream NLP practice, and highlight the methodological challenges that arise when modelling disagreement, subjectivity, and annotator variation. In particular, I consider unresolved questions around evaluation paradigms, task formulation, population representation, community norms, and the implications of using pre-trained generative models as classifiers. By synthesising discussions from five years of workshops, keynotes, and related publications, I outline open challenges and propose directions for future work aimed at more rigorous perspectivist NLP. I argue that we should focus on more realistic task formulations and the centering minoritised standpoints, and caution against viewing potentially harmful interpretations as equally legitimate reactions to 'subjective' phenomena.

Keywords: Perspectivism, Variation, Disagreement

1. Introduction

'*What is truth*?' asked Pontius Pilate (John 18:38), but famously, he did not provide or wait for an answer. In recent years a growing section of the Natural language processing (NLP) research community has been asking a similar question. In this paper marking the fifth edition of the Workshop on Perspectivist Approaches to NLP (NLPerspectives),[1] I examine what we mean by *(ground) truth* when it has become (relatively) more common to collect, model, and attempt to represent multiple diverging responses in human label data. Here, I reflect on both progress made, and attempt to take a snapshot of the field, including open questions, challenges, and limitations of current approaches.

2. Background

As a field that grew out of the related discipline of computational linguistics (CL), from the early 2000s, NLP saw a shift away from more well defined linguistic tasks such as part-of-speech tagging and dependency parsing, which had been the focus of research up to and including the turn of the Millenium, to begin considering social and affective tasks like sentiment analysis (e.g. Pang et al., 2002; Wiebe et al., 2005) and emotion recognition (e.g. Mohammad and Turney, 2013).

Throughout this period, methodologies and standards were developed for establishing the reliability of human labels applied to text data. These were believed to provide an indication of the *reliability* of these annotations as markers of the *ground truth* categories to which individual data points belonged.

Chief among these was the measurement of chance-adjusted inter-annotator agreement (IAA). Borrowed from Behavioral Science, this was introduced to CL by Carletta (1996), who pointed out that the quality of previous work had been purely 'judged according to whether or not the reader found the explanation plausible.' In comparison, the application of a rigorous statistical measurement was a common sense way to improve the robustness of data collection methods. An iterative NLP corpus creation methodology was established in which: (1) guidelines were drawn up explaining the phenomena of interest, (2) data was annotated following these instructions, (3) IAA was measured, (4) reasons for disagreement were interrogated, attempts made to iron them out, and guidelines adjusted to avoid such deviance in the future. We then returned to step (1), and repeated until the Cohen's *kappa*, Krippendorf's *alpha*, or another chosen statistic was deemed to be satisfactory, usually by meeting some quite arbitrary threshold (Warrens, 2015). Manuals were written with step-by-step explanations of how to follow this procedure (Artstein, 2017; Pustejovsky and Stubbs, 2012).

One effect of this standardisation was that, as

[1] https://nlperspectives.di.unito.it/

Proceedings of NLPerspectives @ LREC 2026, pages 1–10
12 May 2026.

high IAA scores became synonymous with reliability, and therefore the quality of data collection, it seemed to become almost impossible to publish work that showed even moderate levels of variation in annotator behaviour (as many researchers struggling to achieve high IAA at this time would probably attest). Where disagreement was found, this became reason for devising methods to weed out supposedly unreliable annotators (Hovy et al., 2013) and 'noisy' labels (e.g., aggregation by majority vote), or, as Aroyo and Welty (2013) pointed out, to try to force consensus through over-specified and overly ungeneralisable development.

The 'Perspectivist Turn' There are, in fact, a number of examples of CL and NLP work acknowledging the potential for finding 'signal' in the 'noise' of human label variation going back to the 2010s, and beyond. Aroyo and Welty (2013) proposed to collect the '*Crowd Truth*' distribution of annotator responses believing (dis)agreement levels provided information about the relative 'clarity', 'vagueness', or 'ambiguity' of a labelled item. Further examples are Jurgens (2014), who considered annotator disagreement as a proxy for item difficulty in a word sense labelling task, and Plank et al. (2014), who showed that such disagreements were systematic for part-of-speech labelling. Arguing that the notion of *acceptability* should replace that of 'ground truth', Alm (2011) pointed to work on subjectivity in CL and linguistics going as far back as the 1930s.

However, mainstream NLP methodology continued to prioritise the collection and modelling of single 'gold standard' class labels until the field saw the beginnings of a 'Perspectivist turn' in the early 2020s. Signs of this included talk of the 'end of the gold standard' (Basile, 2020) and the 'need to talk about disagreement' (Basile et al., 2021), the launch of a Perspectivist Data Manifesto urging researchers to follow disaggregated data practices,[2] a growing interest in 'learning with disagreement' (Uma et al., 2021b), including the launch of the Le-Wi-Di shared task (Uma et al., 2021a; Leonardelli et al., 2023, 2025), publication of a prominent survey of relevant resources (Plank, 2022), and an emerging interest in modelling of individual annotators (Cercas Curry et al., 2021; Davani et al., 2022; Vitsakis et al., 2023).

NLPerspectives at Five Which brings us to the launch of the present workshop series. Conceived of in 2021, and with its first edition taking place at LREC in May 2022, NLPerspectives is now celebrating its 5th edition. During this time, it has seen the publication of 68 research papers (up to edition 4, see Figure 1), the presentations of several other non-archival works and research communications, and hosted five keynote talks on topics relevant to and overlapping with perspectivist data practices: Su Lin Blodgett on participatory design (2022 LREC, Marseille), Przemysław Kazienko on personalised NLP (2023 ECAI, Krakow), Barbara Plank on human label variation and model uncertainty (2024 LREC-COLING, Turin), Jose Camacho Collados on cultural factors in multilingual models (2025 EMNLP, Suzhou), and now at LREC 2026 in Palma de Mallorca, Federico Cabitza, author of 'the Perspectivist Turn'.[3]

Each workshop has ended with a panel discussion featuring the organisers, invited speakers, and other researchers from the community reflecting on progress, challenges, the state of the field. In this paper, I attempt to synthesise some of the talking points from these conversations.

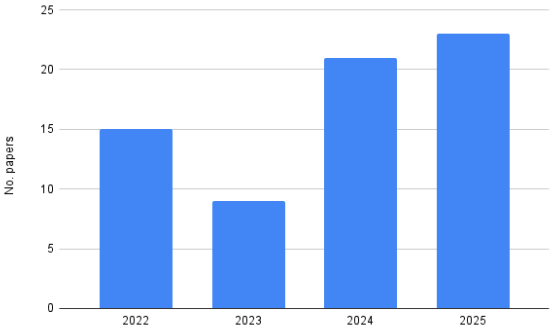

Figure 1: Published papers over time at NLPerspectives. With the exception of 2023, when the workshop was presented outwith the *ACL community (at ECAI), the workshop has seen steady growth in submissions and accepted papers.

Influence on the wider NLP field The idea of collecting, preserving, and modelling multiple labels appears to have become considerably more mainstream than it was five years ago. In a reversal of the situation described previously, anecdotally at least, peer reviewers sometimes now consider a lack of consideration for legitimate annotator disagreement to be a methodological weakness.

There have also been a number of developments that indicate that the field may have been influenced by the workshop and the research of those working in this community. Other workshops and events have sprung up with similar themes, such as Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation (Roth and Schlechtweg, 2025) and the choice of 'subjectivity and disagreement in abusive language data' as special theme for the 7th Workshop on Online Ab-

[2] https://pdai.info/

[3] Information about the workshop is archived at https://nlperspectives.di.unito.it/.

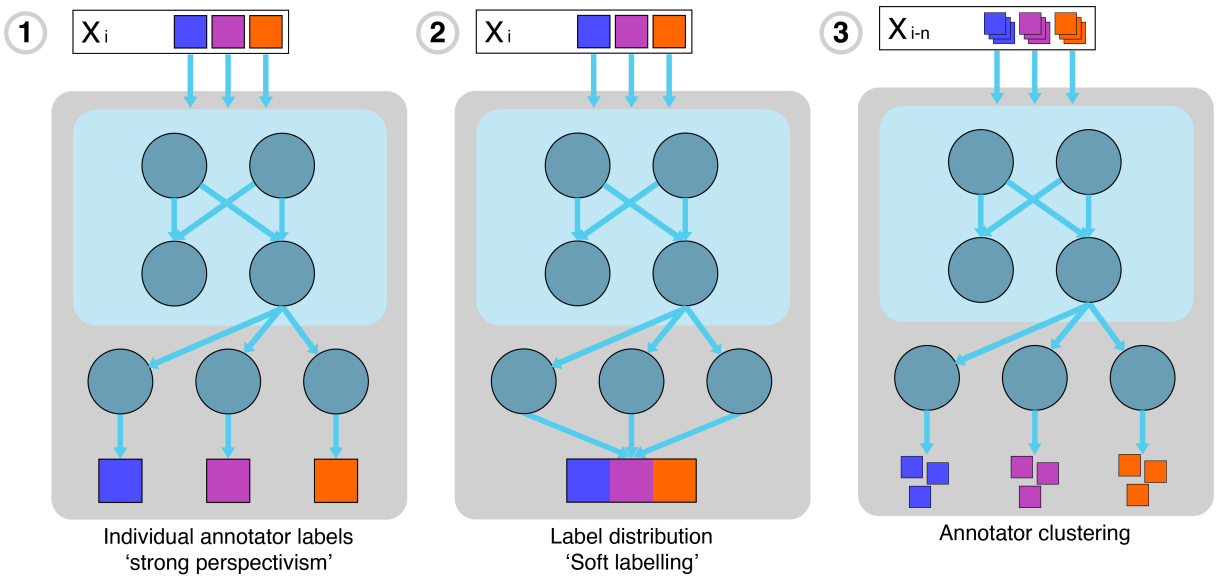

Figure 2: Three perspectivist paradigms. For a given data point X_i, and set of labels (here represented by differently coloured squares), models (1) predict individual annotator labels; 92) output a prediction of the distribution of labels; or (3) focus on grouping annotators according to their labelling behaviour. Figure after Davani et al. (2022).

use & Harms (WOAH) (Chung et al., 2023). Most notably, the ACL 2025 keynote speech titled 'Whose Gold?' was given by founding member of the workshop's organising committee, Verena Rieser, and expounded on perspectivist themes (Che et al., 2025).

3. Open questions and challenges

Although the developments described above indicate a shift in the field towards more mainstream acceptance of the tenets of perspectivist NLP, there remain many challenges. In many ways, the embrace of perspectivism opens more questions than it resolves. While (Frenda et al., 2025) reviewed work conducted under the perspectivist banner, in the rest of this paper, I consider these questions, and make suggestions for approaches.

3.1. Evaluation

From the beginnings of the workshop, perhaps the most frequently posed and least satisfactorily answered question has been 'how (or what) should we evaluate?' Here, three main paradigms have emerged (see Figure 2):

1. The individual annotators responsible for providing the labels are modelled, and a label is output for each that represents a prediction of what they *would* label the item in question (e.g. Cercas Curry et al., 2021; Davani et al., 2022; Lo et al., 2025; Orlikowski et al., 2023). This represents a form of descriptive relativism, in which all individual truths are seen as equally valid.

2. Models are designed to predict the distribution of labels, i.e., in the example illustrated in Figure 2, the soft label is $[0.33, 0.33, 0.33]$. This has been used as the main evaluation metric for the Le-Wi-Di shared task, where it is referred to as 'soft labelling' (Leonardelli et al., 2023). Other examples of this approach include Madeddu et al. (2023), Parappan and Henao (2025), Weerasooriya et al. (2023a), and Weerasooriya et al. (2023b).

3. Some work is concerned primarily with modelling the relationship between the labelling behaviours of different people, often with the aim of finding like-minded groups within the cohort of annotators (Lo and Basile, 2023; Liu et al., 2019; Prabhakaran et al., 2024; Vitsakis et al., 2024) or with finding individuals with particular attitides or beliefs (Chulvi et al., 2023; Jiang et al., 2024).

Each of these formulations has some drawbacks and theoretical weaknesses. For (1), beyond the technical issues of it being difficult to scale up and of problems arising when annotators are represented sparsely in datasets (Leonardelli et al., 2023), it is a little difficult to see many use cases for this approach. One struggles, for instance, to see how this paradigm might fit into any scenario that requires a decision to be taken, such as a content moderation pipeline.

We might make a similar criticism of paradigm (2), which may need some kind of threshold to be established for decision-making applications—in which case we are not far from the majority vote paradigm that perspectivism seeks to avoid. In fact, some work uses the distribution of soft labels primarily as a signal to improve prediction of aggregated labels (e.g., Uma et al., 2020). However, it is easier to see how this output could be informative for downstream tasks, particularly if we have some information about *who* is represented in the distribution (see subsection 3.3).

Paradigm (3) encompasses work that uses demographic information (e.g. Gordon et al., 2022; Prabhakaran et al., 2024), annotator behaviour (e.g. Lo and Basile, 2023; Vitsakis et al., 2024), or underlying beliefs and attitudes of annotators (e.g. Chulvi et al., 2023; Jiang et al., 2024). A weakness here, when compared to similar work in other fields, is that NLP practitioners tend not to consider the populations that they model, limiting both the rigour of the research and the utility of the models (see subsection 3.3).

3.2. Unclear Task Formulation

As discussed in subsection 3.1, there is something of a lack of clarity in the relationship between much of the research published in this area and the real-world tasks that motivate it. For all the faults with traditional 'gold label' modelling that perspectivist NLP has attempted to address, it is far easier to see how systems that output single label predictions might be applied in practice. Perspectivist researchers have so far failed to demonstrate how their approaches—which currently seem more suited to exploratory data analysis—might fit into such decision-making or predictive systems.

One way to do this might be to consider collecting evidence from extrinsic evaluation practices, which have so far been lacking in the field (Reiter, 2025).

3.3. Increasing but Non-rigorous Complexity

With the rise of perspectivism, researchers have been gradually de-simplifying supervised classification tasks, by adding further levels of complexity, realism, and information that needs to be accounted for.

Accepting that some annotation items are subjective or ambiguous, or that readings of them can legitimately differ, researchers became interested in *who* holds *which* perspectives, seeking to harness demographic and other information about them. However, unlike social scientists, we have all but ignored the concept of representing a target population,[4] leaving the research open to accusations of lack of rigour. Extreme examples include contending that three individual annotators can each represent *conservative*, *moderate*, and *liberal* points of view (Almanea and Poesio, 2022).

Once it was accepted that there might be valid variation in annotator responses, we began to look at the causes of these differences, seeking to tease apart fixed opinions, ambiguous and difficult data items, and noisy and erroneous annotation work. For example, in a series of longitudinal experiments, Abercrombie et al. (2023a, 2025) found that annotators are internally inconsistent on repeated annotation items around 75% of the time, which they put down largely to ambiguity in the data. In a field that collects labels primarily from anonymous crowdworkers working in completely uncontrolled environments, there are a number of factors that might make one doubt that these represent any kind of truth, even a subjective one.

As we acknowledge an increasing number of factors that cause label diversity, we should be careful to apply rigour in modelling them.

3.4. The Problem of Noise

A side effect of accepting that label variation can be valid for a wide variety of reasons is that we have lost the tools that we previously believed provided evidence of the reliability of our data and collection practices. In addition to IAA scores becoming variation analysis tools rather quality indicators, it has become very difficult to apply previously common methods such as attention check items. After all, if many reactions are valid, how can we say that annotators should label a particular item a certain way?

In situations where we are *particularly* interested in minoritised perspectives, such as hate speech detection, in which only those with lived experience may be capable of recognising the phenomenon of interest, a valid approach is to seek those people as annotators (Abercrombie et al., 2023b; Fleisig et al., 2024). While one method is to establish a pool of tried and tested annotators (e.g. Jiang et al., 2024), there is a danger of a lack of rigour and that this may be done mainly on 'vibes'. With the recognition that the 'crisis of reproducibility' is firmly embedded in NLP data practices (Belz et al., 2023; Dinkar et al., 2024), and observed specifically in human labelling for supervised classification (Sasidharan Nair et al., 2024), we need to establish new methods for validating the quality of the labels we collect.[5]

[4] Exceptions include Pei and Jurgens (2023), who draw representative population samples, and Eckman et al. (2025), who weight annotations by populations.

[5] See Fleisig et al. (2025) for an exploration of heuristics for this purpose.

3.5. Subjectivity vs Community Norms

Undoubtedly the most researched topic in perspectivist NLP has been that of hate speech and other toxic language. While this is probably due to its inherently contested nature, it has led to often repeated claims that hate speech is a subjective phenomenon (e.g., Akhtar et al., 2021; Almanea and Poesio, 2022; Basile, 2020). This is not only untrue from a philosophical (as well as legal) point of view (Barendt, 2019), but creates the danger of *bothsidesing* the points of view of perpetrators and targets of hate speech. This may be particularly likely in the individual annotator modelling paradigm (subsection 3.1), in which each annotator may be given precisely equal weight (at least in the absence of a well designed task formulation). In fact, hate speech should be defined at the community level, preferably according to the norms of those it impacts (Cercas Curry et al., 2024).

Researchers should be careful to define 'subjectivity' precisely in relation to other phenomena that influence label variation. When working on hate speech phenomena (e.g., racism, sexism), we should consider how to model the community norms of those affected.

3.6. Taking a stand(point)

To do this, we may need to accept that objectivity in research design is neither possible or desirable. As, following *standpoint theory*, only people with relevant lived experience are capable of recognising the phenomena of interest, NLP researchers may need to actively seek to foreground those voices. As one of the theory's principal proponents, Sandra Harding, put it 'a standpoint is not the same as a viewpoint or a perspective, for it requires both science and a political struggle' (Harding, 1998, p.150).

One approach to this is through participatory/co-design with stakeholders. This can take many forms, including focus groups, workshops, and Delphi studies (Wilson et al., 2025), but should aim to involve participants beyond simple consultation and validation of technical goals, and ideally hand over a level of ownership of the research agenda (Caselli et al., 2021; Delgado et al., 2023).

While it may not be easy to fully achieve these aims (due to e.g. conflicting motivations and funding issues (Wilson et al., 2025)), another strand of research investigates how to scale such work up to select specific crowdworkers that share the values of these communities. This is important, as there is growing evidence that demographic information is not a reliable predictor of annotation behaviour (see e.g. Orlikowski et al., 2023, 2025). One approach is to collect information about annotators' underlying attitudes using validated surveys (Jiang et al., 2024).

3.7. Generative Models as Classifiers

As Star and Bowker (1999, p.1) contend, 'to classify is human'. But is it also LLM? The vast majority of work in this area has focused on the type of tasks traditionally solved by discriminative machine learning models. However, since 2022–coincidentally both the year of the first workshop and the release of ChatGPT–we are increasingly undertaking classification tasks with pre-trained generative language models (see e.g. Balestrucci et al., 2025; Plaza-del Arco et al., 2024; Pavlovic and Poesio, 2024), often referred to as 'LLM-as-a-judge'.

While discriminative models were designed to output discrete labels, the latter, if left to their own devices, will emit long screeds of text, expounding on the topic in a vaguely knowledgeable style and marked by formulaic rhetorical devices and bullet points. In many cases this output has been guided through reinforcement learning to appear as helpful and engaging as possible, and models are known to engage in seemingly obsequious behaviours, changing their responses in order to affirm the viewpoints of users (Ranaldi and Pucci, 2025). In short, 'truth' may be a somewhat secondary concern for such models, optimised to satisfy users, and referred to by Hicks et al. (2024) as 'bullshitters' .

As text generation models are now pervasive, perspectivist research needs to take into account these behaviours and consider what it means to use such models for classification tasks, when, in the real-world, users must contend with non-determinism, affirmation bias (Sharma et al., 2023), refusal behaviour (Ouyang et al., 2022), and generation of false information.

At the same time, we should broaden our focus from classification tasks to the human preference ranking data that to a large extent underlies the success of these models. Despite the many sources of disagreement in such data (Dsouza and Kovatchev, 2025), and indications that inclusion of disagreements can lead to better performance (Gooding and Mansoor, 2023), there is currently little evidence that minoritised perspectives are actually maintained in RLHF training data for generative models.

4. Related Work

In addition to the historical work discussed in section 2, two recent surveys take critical looks at perspectivist NLP research. Frenda et al. (2025) provide a systematic review of work published at the workshop and beyond. They focus primarily on thematic analysis of publications up until the writing of the review (2024), and particularly highlight the lack of a clear direction on how to evaluate perspectivist modelling.

Fleisig et al. (2024) take a broader look at the shift to perspectivist methods, highlighting several

issues that overlap with the points we make here. They particularly highlight the need to make normative decisions highlighting annotators with expert knowledge of the phenomenon of interest, including relevant lived experience, and suggest particpatory practices as a means of doing so, as I advocate in subsection 3.5.

Combining the approaches of these two works, in this position paper, I have tried to provide a snapshot of the field on the occasion of the 5th NLPerspectives workshop, and set out a personal view of what is currently lacking in this area.

5. Conclusion

As NLPerspectives marks its fifth edition, we are able to reflect on this research area's growth from a niche and fringe community challenging the orthodoxy of 'ground truth' in NLP to a situation approaching mainstream adoption in the field. At the same time, a number of difficult questions and unsolved challenges have come under discussion at the workshop. In this paper, I have attempted to set out some of these issues, and argued that we need to be more rigorous and deliberate in defining the truths that we seek to model.

Acknowledgements

Thanks to the members of the NLPerspectives Programme Committee for their helpful and insightful comments and suggestions, which I have tried to incorporate in this version of this paper.

Thanks also to the research group at IMS Stuttgart, who invited me to give a talk in March 2026, on which this paper was based.

This work was supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1).

References

Gavin Abercrombie, Tanvi Dinkar, Amanda Cercas Curry, Verena Rieser, and Dirk Hovy. 2025. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 63–74, Suzhou, China. Association for Computational Linguistics.

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023a. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-Abbott, Ioannis Konstas, and Verena Rieser. 2023b. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

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*.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci '13*.

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Pier Felice Balestrucci, Michael Oliverio, Elisa Chierchiello, Eliana Di Palma, Luca Anselma, Valerio Basile, Cristina Bosco, Alessandro Mazzei, and Viviana Patti. 2025. Towards a perspectivist understanding of irony through rhetorical figures. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 27–36, Suzhou, China. Association for Computational Linguistics.

Eric Barendt. 2019. What is the harm of hate speech? *Ethical Theory and Moral Practice*.

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 *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective isms? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. Association for Computational Linguistics.

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors. 2025. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, Paolo Rosso, et al. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *2nd Workshop on Perspectivist Approaches to NLP (NLPerspectives)*, volume 3494. CEUR Workshop Proceedings.

Yi-Ling Chung, Paul Röttger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors. 2023. *Proceedings of the 7th Workshop on Online Abuse & Harms (WOAH)*. Association for Computational Linguistics, Toronto, Canada.

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.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Tanvi Dinkar, Gavin Abercrombie, and Verena Rieser. 2024. ReproHum #0927-03: DExpert evaluation? reproducing human judgements of the fluency of generated text. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 145–152, Torino, Italia. ELRA and ICCL.

Russel Dsouza and Venelin Kovatchev. 2025. Sources of disagreement in data for LLM instruction tuning. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 20–32, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Stephanie Eckman, Bolei Ma, Christoph Kern, Rob Chew, Barbara Plank, and Frauke Kreuter. 2025. Aligning NLP models with target population perspectives using PAIR: Population-aligned instance replication. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 100–110, Suzhou, China. Association for Computational Linguistics.

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.

Eve Fleisig, Matthias Orlikowski, Philipp Cimiano, and Dan Klein. 2025. Balancing quality and variation: Spam filtering distorts data label distributions. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 47–62, Suzhou, China. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, , Raffaella Panizzon, Alessandra Teresa Cignarella, and Davide Marco, Cristina Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

Sian Gooding and Hassan Mansoor. 2023. The impact of preference agreement in reinforcement learning from human feedback: A case study in summarization.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Sandra Harding. 1998. *Is science multicultural?: Postcolonialisms, feminisms, and epistemologies*. Indiana University Press.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3006–3012, Reykjavik, Iceland. European Language Resources Association (ELRA).

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, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019. Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1111–1120, New York, NY, USA. Association for Computing Machinery.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. PERSEVAL: A framework for perspectivist classification evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22334–22359, Suzhou, China. Association for Computational Linguistics.

Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti, and Valerio Basile. 2023. DisaggregHate it corpus: A disaggregated Italian dataset of hate speech. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 243–250, Venice, Italy. CEUR Workshop Proceedings.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20322–20338, Suzhou, China. Association for Computational Linguistics.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds. exploring model label variation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 19–30, Torino, Italia. ELRA and ICCL.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. 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 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O' Reilly.

Leonardo Ranaldi and Giulia Pucci. 2025. When large language models contradict humans? large language models' sycophantic behaviour.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

Michael Roth and Dominik Schlechtweg, editors. 2025. *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*. International Committee on Computational Linguistics, Abu Dhabi, UAE.

Sachin Sasidharan Nair, Tanvi Dinkar, and Gavin Abercrombie. 2024. Exploring reproducibility of human-labelled data for code-mixed sentiment analysis. In *Proceedings of the Fourth Workshop*

on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024, pages 114–124, Torino, Italia. ELRA and ICCL.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. In *Proceedings of the International Conference on Learning Representations*.

Susan Leigh Star and Geoffrey Bowker. 1999. Sorting things out. *Classification an its consequences The MIT Press, Cambridge, Massachusetts, London, England*.

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 Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

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.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669, Toronto, Canada. Association for Computational Linguistics.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Matthijs J Warrens. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, 5.

Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher Homan. 2023a. Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–966, Toronto, Canada. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023b. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with DisCo. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, Toronto, Canada. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*.

Marianne Wilson, David M. Howcroft, Ioannis Konstas, Dimitra Gkatzia, and Gavin Abercrombie. 2025. Participatory design for positive impact: Behind the scenes of three NLP projects. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 252–263, Vienna, Austria. Association for Computational Linguistics.