A Measure of Systematic Disagreement

Valerio Basile
University of Turin

Abstract

This paper introduces a new metric, called σ, that quantifies the degree of systematicity in inter-annotator disagreement. The metric is inspired by Structural Balance Theory and is designed to approximate the clusterability of annotators in a dataset. When paired with a standard inter-annotator agreement measure such as Krippendorff's α, σ provides a complementary signal designed to capture the extent to which disagreement stems from genuine subjective factors rather than from ambiguity or annotation noise. The metric is applied to over twenty datasets encoding a broad variety of annotations, showing a tendency to produce higher values for tasks conventionally considered subjective.

Keywords: Inter-annotator agreement, Subjective tasks, Perspectivist NLP

1. Introduction

Inter-annotator agreement has been the main lens to quantify, or at least approximate, the quality of a human-annotated dataset (Artstein and Poesio, 2008). While this connection is unchallenged, the recent attention on the phenomenon of Human label variation (Plank, 2022) in the NLP research community has spurred proposals to investigate the aspects that impact observed disagreement and its multiple causes. Basile et al. (2021) argue that annotator disagreement can be traced to a variety of causes, which they cluster in two main sources:

- **Ambiguity**, encompassing all the exogenous factors such as lack of clear annotation guidelines, less-than-ideal annotation interfaces, human distraction, or genuine errors;
- **Subjectivity**, the characteristic of a language annotation task to depend strongly on the individual perception, as well as the personal or cultural background of the annotator.

The study of subjectivity-bound disagreement, already an important aspect mentioned in the seminal work on disagreement by Aroyo and Welty (2015), has led to interesting developments in the NLP research community, including learning with disagreements (Leonardelli et al., 2025; Uma et al., 2022) and the perspectivist turn in NLP (Cabitza et al., 2023).

While disagreement is a measurable signal (Section 2), at the moment we lack a straightforward procedure to determine the contribution of individual factors. The objective of this paper is to introduce a computational tool to measure the degree of systematicity of the disagreement of annotators who expressed judgments on the same data.

2. Related Work

Cohen's κ (Cohen, 1960) and Scott's π (Scott, 1955) are quantitative indexes of the amount of agreement between two annotators. With respect to simpler measures (e.g., percent agreement), κ and π account for the probability of the annotators to agree by chance, just by virtue of imbalanced nature of the label distribution, while differing slightly in the definition of chance agreement. Fleiss' κ extends both Cohen's κ and Scott's π to an arbitrary number of annotators. Krippendorff's α further extends π to cases where the annotation matrix is sparse, i.e., not all annotators annotated every instance, which is a common scenario, e.g., in a crowdsourcing context.

While the aforementioned measures are widespread, we note that they quantify disagreement independently from its origin, or, in other words, disregarding any knowledge about the identity of the annotators. Checco et al. (2017a) identify a set of pitfalls in the use of κ-like metrics, especially in crowdsourcing contexts. Dumitrache et al. (2018) introduce a set of metrics that consider the distribution of the annotated instances and the distribution of the annotators jointly, to account for different annotator behaviors.

Akhtar et al. (2019) introduce the polarization index, a measure of systematic disagreement at the instance level. While the authors note that the average polarization over an annotated dataset can approximate an overall measure of systematic disagreement, the polarization index needs a predetermined partition of the annotator cohort into groups. Recently, Tsirmpas and Pavlopoulos (2026) build on the polarization idea and introduce statistical tools to quantify the level of polarization in relation with determined annotator groups (e.g., by sociodemographic traits). Alacam et al. (2025) propose a method to discriminate the effect of subjectivity vs. uncertainty in hate speech annotation by leveraging the confidence measured through gaze data. To

Proceedings of NLPerspectives @ LREC 2026, pages 56–65
12 May 2026.

cope with the need to know, or somehow box in the identity of the annotators, several works propose methods to learn annotator representations, mainly as a step towards modeling human perspectives in supervised classification contexts. Lo and Basile (2023) apply clustering algorithms to vectors representing the entirety of each annotators activity. Conversely, the approaches of Mostafazadeh Davani et al. (2022) and Mokhberian et al. (2024) learn annotator embedding from the annotated data. These works are highly related to the present paper, where a metric is defined that approximates the clusterability of annotators. More precisely, this paper proposes a metric that validates the assumptions made by Lo and Basile (2023) and others, i.e., that subjective tasks tend to produce annotations with more separable clusters of annotators.

3. Systematicity of Inter-annotator Disagreement

In Social Psychology, like/dislike relationships between humans are modeled through signed undirected graphs, where the nodes represent the individuals and and edge between A and B represent their relationship (if present) as positive (+) or negative (-). The theory of **Structural Balance** (Cartwright and Harary, 1956) models triadic relationships and their possible states. As a classic example, if a person A has a positive relation (e.g. affection) for a person B, and if B is responsible for an entity X (e.g. an event or an artifact, then there will be a tendency for A to like or approve of X. However, if the direct attitude of A towards X (without considering B) is negative, the triangle A-B-X is "imbalanced". At a more abstract level, the theory posits that any three entities in relationship with each other (a *triangle*) in such a graph tend towards a balance achieved by either all edges being + or a situation where one edge is + and the other two are -. The other two possible configurations are instead regarded as imbalanced, as summarized in Figure 1.

Davis (1967) applies the notion of structural balance to graphs, calling a balanced graph an undirected signed graph where all triangles (i.e., cycles of length 3) are balanced, and proving that a balanced graph has a unique clustering. I extend this definition to a degree of balancedness, that is, the rate of triangles in an undirected signed graph that are balanced:

$$\sigma = \frac{(\#\text{balanced triangles})}{(\#\text{triangles})}$$

Next, the outcome of an annotation task is represented as a signed undirected graph, where each node represents an annotator, and the $+-$ sign indicates whether the pair agrees $(+)$ or disagrees $(-)$.

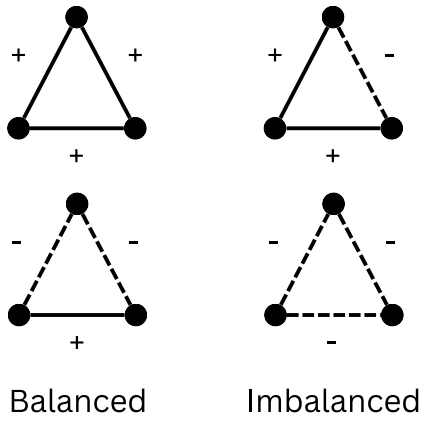

Figure 1: Possible configurations of graph triangles according to the sign of their edges in Structural Balance theory.

The pairwise Krippendorff's α is computed ($\alpha_{\{i,j\}}$ where i and j are two annotators) and compared to overall agreement (α). Formally, the sign s of an edge connecting two annotator-representing nodes i and j is computed as follow:

$$s_{\{i,j\}} = \begin{cases} +, & \text{if } \alpha_{\{i,j\}} \geq \alpha \\ -, & \text{if } \alpha_{\{i,j\}} < \alpha \end{cases}$$

In this paper, I argue that σ, by virtue of approximating the clusterability degree of an annotator graph, will tends to higher values when the annotation is related to more subjective tasks.

4. Experimental validation

σ is tested as a reliable measure of systematic disagreement on a collection of manually annotated datasets. The experiment reported in this section has the goal of verifying the hypothesis that datasets annotated according to subjective phenomena should exhibit higher values of σ by virtue of their disagreement being more systematic. On the contrary, datasets where the agreement is not systematic should exhibit lower σ.

For this experiment, it is crucial to have access to disaggregated datasets, where individual annotations are distributed rather than a single aggregate annotation for each instance. The datasets collected, listed in Section 4.2, have the general form of an $m \times n$ matrix A where m is the number of annotators, n is the number of instances, and $A_{i,j}$ is the label given by annotator i on instance j. A can be fully populated or sparse (typical outcome of crowdsourcing annotation).

4.1. Experimental Setting

All the mainstream IAA measures, and in particular those listed in Section 2 are computed starting from an instance vs. label contingency table. Note that in this process, the identity of the annotators is lost, because the contingency table does not model any relationship between the annotations provided by the same annotator. Therefore, any operation $A_{i,j} := A_{k,j}, A_{k,j} := A_{i,j}$ (swapping two annotations on instance j provided by annotators i and k) results in exactly the same IAA.

This characteristic of IAA measures is exploited to test the hypothesis by generating randomized variations of a dataset that preserve its overall IAA (here computed as α). Firstly, the columns of A are randomly divided into two groups. The rows of the sub-matrix with columns from the first group are shuffled, and so are the rows of the sub-matrix with columns from the second group. The two shuffles are independent from each other. Finally, the entire procedure is repeated ten times, producing the derived annotation matrix A_{rnd}. While it is ensured that $\alpha(A) = \alpha(A_{\text{rnd}})$, the shuffling procedure destroys the systematicity of the annotator agreement. Therefore, if there is a certain degree of systematicity in the original annotation, we should observe $\sigma(A) > \sigma(A_{\text{rnd}})$.

4.2. Data

I collect a number of datasets of varying size and shape, annotated according to different language phenomena, with the only common characteristic of being distributed with disaggregated labels.

I start by retrieving the datasets harmonized in structure and made available by two popular benchmarks for perspectivist classification. From PersEval (Lo et al., 2025), I obtained the following datasets: BREXIT (Akhtar et al., 2020), made of English tweets about Brexit annotated with hate speech (hs), aggressiveness (ag), offensiveness (of), and stereotype (st); MD-Agreement (Leonardelli et al., 2021), English tweets annotated for offensive language; Measuring Hate Speech (Sachdeva et al., 2022), with crowdsourced annotations of hate speech across may targets by a diverse set of annotators; DICES (Aroyo et al., 2024), a collection of human-chatbot conversations annotated for AI safety. The final dataset from PersEval is EPIC (Frenda et al., 2023), containing post-reply pairs annotated for irony, which is replaced by its newer multilingual version MultiPICo (Lo et al., 2024).

I further collected three datasets distributed in the context of the 2025 edition of the Learning with Disagreements challenge (Leonardelli et al., 2025): VariErrNLI (Weber-Genzel et al., 2024) on Natural Language Inference; The Paraphrase Detection dataset released specifically for the challenge, annotated with paraphrastic relation between pairs of questions; Conversational Sarcasm Corpus (CSC) by Jang and Frassinelli (2024), containing short dialogues annotated with perceived sarcasm.

Other three datasets were added to the set selecting from publicly available lists of disaggregated datasets, in particular the Perspectivist Data Manifesto[1] and the Awesome Human Label Variation repository[2]: ConvAbuse (Cercas Curry et al., 2021), a corpus of conversations with AI assistants annotated with abusive language; Tweet Annotation Sensitivity (Kern et al., 2023), made of tweets annotated with hate speech (hs) and offensive language (of); jobQ3MT+ (Liu et al., 2019), a collection of tweets annotated according to the three questions on the interpretation of the job market aspects of the messages (Q1: point of view of job /employment-related information in the target tweet; Q2: employment status of the subject in the tweet; Q3: mention of job/employment transition event in the tweet).

The datasets listed so far have been published with disaggregated labels mostly because they relate to study on annotator disagreement and perspectivist approaches to NLP (Frenda et al., 2025). As a consequence, they mostly cover NLP tasks typically considered subjective, i.e., where the individual perception of the annotator strongly influences their annotation, with the exception of VariErrNLI (natural language inference). Unsurprisingly, analogous corpora annotated for less subjective phenomena are harder to come by. However, I collected three more datasets in this broad category: Frame Disambiguation (Dumitrache et al., 2019) contains crowdsourced annotations for frame disambiguation of sentence-word pairs; Phrase Detectives (Poesio et al., 2019) is a corpus of documents annotated for anaphora with four labels (NR: non-referring; PR: predicative NPs; DN: discourse-new mention; DO: discourse-old mentions). Visual Features (Cheplygina and Pluim, 2018) consists of 100 images from dermoscopic a medical AI challenge, annotated according to four features: asymmetry, border, color, dermoscopic structures.

Table 1 summarizes the datasets along with their size and annotation statistics.

4.3. Results

The result of the metrics computed as defined in Section 4.1 on the dataset listed in Section 4.2 are shown in Table 2. Under my hypothesis, the datasets encoding phenomena whose annotation is more dependent on subjective perception will

[1] https://pdai.info

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

Dataset	Instances	Annotators	Avg. annotations per instance (st. dev.)
BREXIT (Akhtar et al., 2020)	1120	6	6.0(0)
BREXIT-hs	"	"	"
BREXIT-ag	"	"	"
BREXIT-of	"	"	"
BREXIT-st	"	"	"
MD-Agreement (Leonardelli et al., 2021)	10753	819	5.0(0)
MHS (Sachdeva et al., 2022)	39565	7912	3.4(26.96)
DICES (Aroyo et al., 2024)	350	123	123.0(0)
MultiPICo (Lo et al., 2024)	18778	506	5.0(1.53)
VariErrNLI (Weber-Genzel et al., 2024)	480	4	3.9(0.89)
Paraphrase (Leonardelli et al., 2025)	500	4	4.0(0)
CSC (Jang and Frassinelli, 2024)	7036	872	4.5(0.89)
ConvAbuse (Cercas Curry et al., 2021)	2894	8	4.4(7.67)
TAS (Kern et al., 2023)	3013	263	4.1(3.09)
TAS-hs	"	"	"
TAS-of	"	"	"
jobQ3MT+ (Liu et al., 2019)	2000	1185	
jobQ3MT+-Q1	"	"	10.06(0.29)
jobQ3MT+-Q2	"	"	10.06(0.29)
jobQ3MT+-Q3	"	"	10.55(0.96)
Frame (Dumitrache et al., 2019)	433	51	21.03(4.08)
Phrase Detectives (Poesio et al., 2019)			
PD-DN	5997	290	10.90(4.38)
PD-DO	3029	326	14.48(9.42)
PD-NR	155	103	7.41(3.22)
PD-PR	1826	282	7.77(3.98)
Visual Features(Cheplygina and Pluim, 2018)	100	6	
VF-asymmetry	"	"	5.93(0.38)
VF-border	"	"	17.86(0.77)
VF-color	"	"	11.90(0.54)
VF-dermo	"	"	23.83(0.97)

Table 1: Statistics of the datasets used in the experimental validation.

show a more systematic structure (σ). The step of randomizing the annotation matrix has the effect of lowering the systematicity (σ_{rnd}) to an extent proportional to the subjectivity of the task. This is confirmed by ordering the result table by $\sigma_{\text{rnd}} - \sigma$ and grouping the tasks into subjective and "objective"[3]. Most of the former datasets at the top of the table (high subjectivity), while the latter ones cluster at the bottom (low subjectivity). Note that at the bottom of the table the statement $\sigma > \sigma_{\text{rnd}}$ does not hold, signaling that the agreement in the annotation of those dataset is not systematic.

The two outliers are arguably justified. In VF-asymmetry, the task requires annotators to judge the symmetry of certain skin formations in medical imagery. The subjectivity of this task founds confirmation in the scientific literature both in the clinical domain (Kunz et al., 2021) and from a computational modeling perspective (Amirshahi et al., 2017), including highlighting correspondences between the task of judging symmetry and the individual perception of human emotions (Evans et al., 2012). The BREXIT-ag dataset has labels of aggressiveness annotated according to the intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target. Examining the examples in the annotation guidelines (Sanguinetti et al., 2018), this task seems to be correlated, at least partially, with overt lexical features (an example contains the work extermination, for instance), which is arguably a more formal rather than subjective task.

[3]"Objective" is written in quotes in this context following the observation of Cabitza et al. (2023) who prefer the term low intersubjective to highlight the difficulty of considering any annotation task completely objective.

4.4. Visual Analysis

In order to gain better insights into the explanatory potential of the new measure, I visualize

Dataset	α	σ	σ_{rnd}	$\sigma_{\text{rnd}} - \sigma$
Paraphrase	.155	1.000	.350	−.650
BREXIT-hs	.347	1.000	.500	−.500
VF-asymmetry	.344	1.000	.550	−.450
CSC	.121	.494	.120	−.374
BREXIT-of	.364	.800	.470	−.330
ConvAbuse	.578	.714	.386	−.328
jobQ3MT+-Q3	.276	.392	.159	−.233
jobQ3MT+-Q2	.353	.427	.201	−.226
MD-Agreement	.359	.494	.283	−.211
jobQ3MT+-Q1	.247	.339	.154	−.185
PD-NR	.085	.416	.264	−.152
VF-dermo	.072	.600	.450	−.150
BREXIT-st	.294	.600	.480	−.120
TAS-hs	.397	.558	.452	−.106
DICES	.210	.601	.518	−.083
MultiPICo	.264	.496	.439	−.057
Frame	.250	.558	.505	−.053
MHS	.516	.677	.637	−.040
TAS-of	.469	.451	.421	−.030
PD-DO	.040	.384	.377	−.007
VF-border	.112	.500	.510	.010
PD-DN	.076	.376	.423	.047
BREXIT-ag	.299	.500	.580	.080
VF-color	.146	.400	.500	.100
PD-PR	−.048	.532	.645	.113
VariErrNLI	.344	.000	.550	.550

Table 2: Results of the experiment described in Section 4.1, in ascending order of difference between original σ and the same metric computed on randomly shuffled datasets (σ_{rnd}). Datasets of tasks traditionally considered "objective" are highlighted with a darker background.

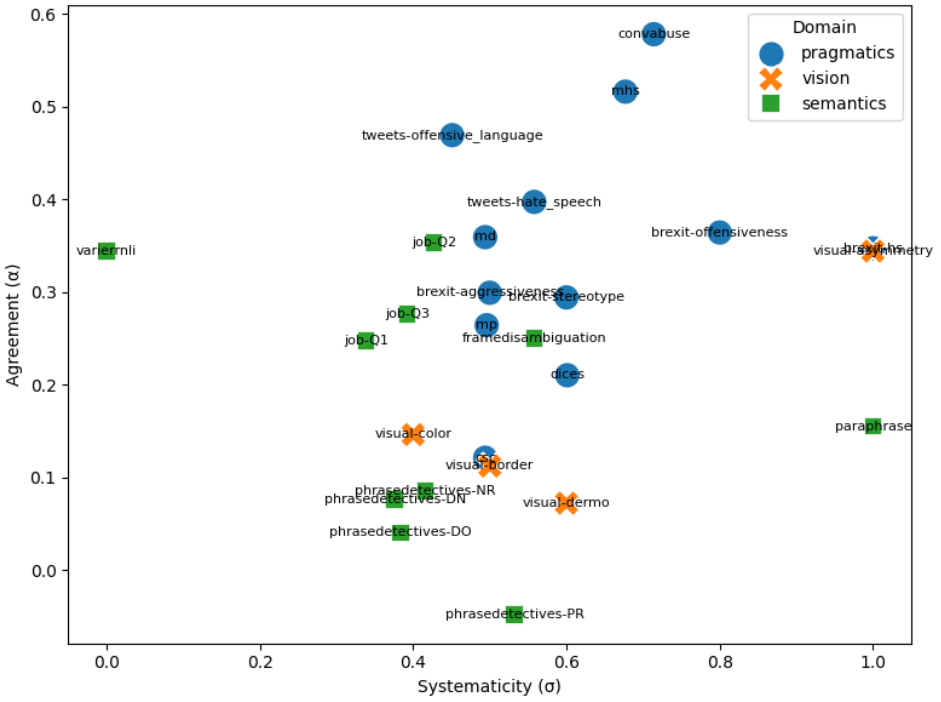

Figure 2: Agreement (α) and systematicity (σ) of the analysed datasets by task domain.

the datasets used in the experiment in a scatterplot (Figure 2). The datasets are coarsely grouped into three domains related to the task they model, namely semantics (VariErrNLI, Paraphrase, jobQ3MT+, Frame, and Phrase Detectives), pragmatics (BREXIT, MD-Agreement, MHS, DICES, ConvAbuse, MultiPICo, CSC, and TAS), and visual (the four VF-* datasets).

Despite a certain amount of variability and the few outliers, the visual analysis shows a clear pattern: the pragmatics dataset generally have a higher agreement (α) than semantics and vision, and the agreement on pragmatics is more systematic (σ).

Besides the visual analysis at the dataset level, Structural Balance Theory provides useful analytical tools for inspecting the inner structure of a dataset annotation. The signed graphs computed as an intermediate steps in the calculation of σ (Section 3) are visualized for some of the datasets involved in the experiment. In the figures, nodes represent annotators, solid lines represent a pairwise agreement above average (i.e., a + edge), and dashed lines represent a pairwise agreement below average (i.e., a - edge).

Figure 3 shows the graph of BREXIT-hs, one of the most subjective datasets according to the analysis. In this graph there are two clear 3-size clusters of annotators, internally connected by + edges and connected with members of the other cluster by - edges. Consequently, the disagreement of the annotators of BREXIT-hs is highly systematic.

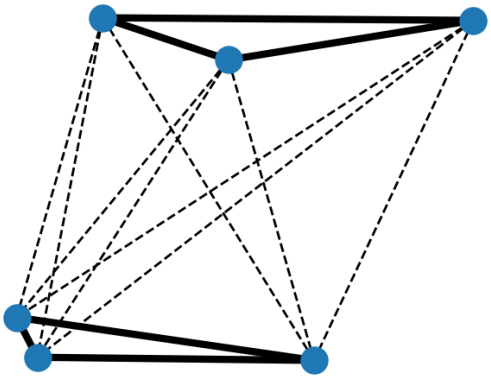

Figure 3: Signed graph of BREXIT-hs.

Figures 4 and 5 show graphs from the annotation of less subjective tasks (VF-color and BREXIT-ag, respectively), which do not exhibit a clustered structure. Interestingly, the distribution of pairwise agreement can vary: while in VF-color 60% of the pairs (9 out of 15) are linked by a +, only 40% of the pairs agree more than the average amount in BREXIT-ag.

Finally, the visualization of the graph of VF-asymmetry reveals a specific structure with a single annotator disagreeing from all the others, who agree among them, possibly contributing to explain the outlier result on this dataset highlighted in Section 4.3.

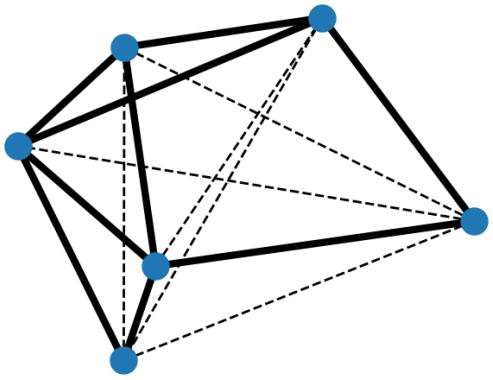

Figure 4: Signed graph of VF-color.

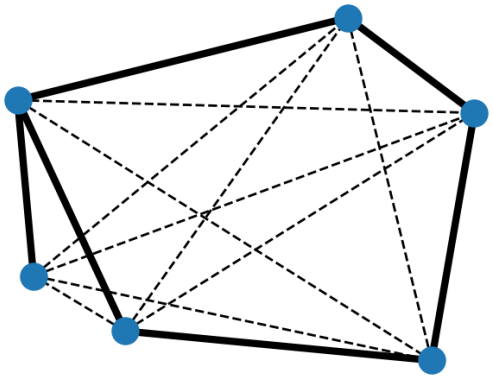

Figure 5: Signed graph of BREXIT-ag.

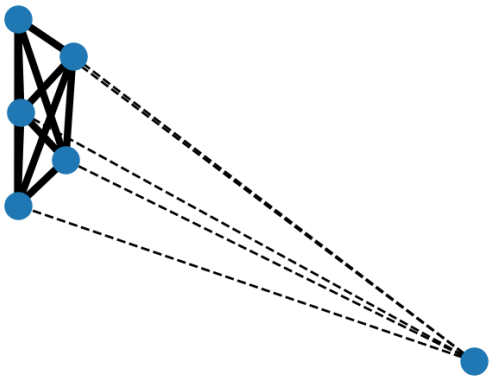

Figure 6: Signed graph of VF-asymmetry.

5. Conclusions

I introduced σ, a quantitative measure of the systematicity of inter-annotator agreement and validated it on a large number of diverse annotated datasets. In particular, the experiment shows that σ captures the systematic patterns of individual annotators and groups, that traditional IAA metrics like Krippenrodffs α are not designed to model.

While in this work σ is validated on a variety of real datasets, its application can be further extended. Intra-annotator agreement could be analyzed under the lens of σ, if data is available with multiple annotations from the same people. Correlations between σ and groups, e.g., by sociodemographics or moral values, are also worth exploring, as well as the impact of persona and perspective-taking prompts in LLM-based annotation.

Besides further tests on more annotated datasets, and extensions to other annotation styles such as rating and ranking, the planned future work also includes the integration of σ into predictive models in order to produce better, more separable representation of the annotators.

6. Limitations

While the experimental section of this paper aims at exploring a wide array of datasets and language phenomena, the resulting figures cannot definitely be grounded in anything other than intuition and the collective consensual experience of a research community. Ironically, there is no objective notion of the subjectivity of a task.

On the practical side, as the number of edges of the graph scales quadratically with the number of nodes, the computational efficiency of σ on a very large dataset with many annotators may dramatically decrease.

Finally, while σ does not need any additional information on the annotators, it relies on Krippendorff's α, which may exhibit an anomalous behavior under particular circumstances (Checco et al., 2017b), a potential limitation already noted by Tsirmpas and Pavlopoulos (2026).

Acknowledgments

I would like to thank the anonymous reviewers who provided valuable insights, some of which made it into the final version. This work also has a debt of gratitude towards Alessandro Mazzei, Daniele Radicioni, and Samuele D'Avenia, for the discussions over the theoretical and experimental aspects which strongly contributed to shape the current version of this paper.

7. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI*IA 2019 - Advances in Artificial Intelligence - XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19-22, 2019, Proceedings*, volume 11946 of *Lecture Notes in Computer Science*, pages 588–603. Springer.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Özge Alacam, Sanne Hoeken, Andreas Säuberli, Hannes Gröner, Diego Frassinelli, Sina Zarrieß, and Barbara Plank. 2025. Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28707–28724, Suzhou, China. Association for Computational Linguistics.

Seyed Ali Amirshahi, Asha Anoosheh, Stella X. Yu, Jakob Suchan, Carl P. L. Schultz, and Mehul Bhatt. 2017. Symmetry in the eye of the beholder. *Journal of Vision*, 17:300–300.

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

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

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn

in ground truthing for predictive computing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Washington DC, USA.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of heider's theory. *Psychological review*, 63 5:277–93.

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.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017a. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada*, pages 11–20. AAAI Press.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017b. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, pages 11–20.

Veronika Cheplygina and Josien P. W. Pluim. 2018. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 105–111, Cham. Springer International Publishing.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

James Allan Davis. 1967. Clustering and structural balance in graphs. *Human Relations*, 20:181 – 187.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement (short paper). In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, volume 2276 of *CEUR Workshop Proceedings*, pages 11–18. CEUR-WS.org.

David W. Evans, Patrick T. Orr, Steven M. Lazar, Daniel Breton, Jennifer Gerard, David H. Ledbetter, Kathleen Janosco, Jessica Dotts, and Holly Batchelder. 2012. Human preferences for symmetry: Subjective experience, cognitive conflict and cortical brain activity. *PLOS ONE*, 7(6):1–9.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: A survey. *Language Resources and Evaluation*, 59(2):1719–1746.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! 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 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Felix Kunz, Matthias Hirth, Tilmann Schweitzer, Christian Linz, Bernhard Goetz, Angelika Stellzig-Eisenhauer, Kathrin Borchert, and Hartmut

Böhm. 2021. Subjective perception of craniofacial growth asymmetries in patients with deformational plagiocephaly. *Clinical oral investigations*, 25(2):525–537.

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.

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

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 68–76.

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

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

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 22345–22370, Suzhou, China. Association for Computational Linguistics.

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

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Barbara Plank. 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.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan*, pages 2798–2895.

William A. Scott. 1955. Reliability of content analysis:the case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Dimitris Tsirmpas and John Pavlopoulos. 2026. Quantifying and attributing polarization to annotator groups.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.