

LREC-COLING 2024

**3rd Workshop on Perspectivist Approaches to NLP
(NLPerspectives)
@LREC-COLING 2024**

Workshop Proceedings

Editors

Gavin Abercrombie, Valerio Basile, Davide Bernardi,
Shiran Dudy, Simona Frenda, Lucy Havens,
Sara Tonelli

21 May, 2024
Torino, Italia

Proceedings of NLPerspectives: The 3rd Workshop on Perspectivist Approaches to NLP @LREC-COLING 2024

Copyright ELRA Language Resources Association (ELRA), 2024
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-23-4
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association and the International Committee on Computational Linguistics

Message from the organizers

This volume documents the Proceedings of the 3rd Workshop on Perspectivist Approaches to Disagreement in NLP, held on May 21st as part of the LREC-COLING 2024 conference (the joint international conference on Computational Linguistics, Language Resources and Evaluation) in Turin, Italy.

Until recently, the dominant paradigm in natural language processing (and other areas of artificial intelligence) has been to resolve observed label disagreement into a single “ground truth” or “gold standard” via aggregation, adjudication, or statistical means. However, in recent years, the field has increasingly focused on subjective tasks, such as abuse detection or quality estimation, in which multiple points of view may be equally valid, and a unique ‘ground truth’ label may not exist. At the same time, as concerns have been raised about bias and fairness in AI, it has become increasingly apparent that an approach which assumes a single “ground truth” can erase minority voices. Perspectivism in NLP pursues the spirit of recent initiatives such as Data Statements, extending their scope to the full NLP pipeline, including the aspects related to modelling, evaluation and explanation.

In line with the first and second editions, the Workshop on Perspectivist Approaches to NLP explores current and ongoing work on the collection and labelling of non-aggregated datasets, and approaches to modelling and including these perspectives, as well as evaluation and applications of multi-perspective Machine Learning models.

The first edition was held at the Language Resources and Evaluation Conference (LREC) in Marseille in 2022, and the second was held at the 26th European Conference on Artificial Intelligence (ECAI) in Kraków in 2023.

In this third edition, the workshop received 28 submissions, including 25 research papers (three of which non-archival) and three research communications. Of these, 22 contributions were accepted. The proceedings are composed of the 16 accepted archival research papers.

Finally, we want to thank the members of the committee for their commitment to the review process and the authors of these contributions for their valuable investigations and for making this community more vibrant.

Organizing Committee

Gavin Abercrombie – Heriot-Watt University
Valerio Basile – University of Turin
Davide Bernardi – Amazon Alexa
Shiran Dudy – Northeastern University
Simona Frenda – University of Turin
Lucy Havens – University of Edinburgh
Sara Tonelli – Fondazione Bruno Kessler

Program Committee

Riza Batista-Navarro
Federico Cabitza
Agostina Calabrese
Silvia Casola
Amanda Cercas Curry
Teddy Ferdinan
Annette Hautli-Janisz
Cassandra L. Jacobs
Anna Koufakou
Sofie Labat
Marta Marchiori Manerba
Michele Mastromattei
Massimo Poesio
Julia Romberg
Pratik Sachdeva
Manuela Sanguinetti
Erhan Sezerer
Zeeraq Talat
Tiago Timponi Torrent
Nikolas Vitsakis
Tharindu Cyril Weerasooriya
Fabio Massimo Zanzotto

Table of Contents

| | |
|---|-----|
| <i>Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models</i> Alicia Parrish, Susan Hao, Sarah Laszlo and Lora Aroyo | 1 |
| <i>Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation</i> Flor Miriam Plaza-del-Arco, Debora Nozza and Dirk Hovy | 19 |
| <i>Revisiting Annotation of Online Gender-Based Violence</i> Gavin Abercrombie, Nikolas Vitsakis, Aiqi Jiang and Ioannis Konstas | 31 |
| <i>A Perspectivist Corpus of Numbers in Social Judgements</i> Marlon May, Lucie Flek and Charles Welch | 42 |
| <i>An Overview of Recent Approaches to Enable Diversity in Large Language Models through Aligning with Human Perspectives</i> Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici and Fosca Giannotti | 49 |
| <i>Disagreement in Argumentation Annotation</i> Anna Lindahl | 56 |
| <i>Moral Disagreement over Serious Matters: Discovering the Knowledge Hidden in the Perspectives</i> Anny D. Alvarez Nogales and Oscar Araque | 67 |
| <i>Perspectives on Hate: General vs. Domain-Specific Models</i> Giulia Rizzi, Michele Fontana and Elisabetta Fersini | 78 |
| <i>Soft metrics for evaluation with disagreements: an assessment</i> Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso and Elisabetta Fersini | 84 |
| <i>Designing NLP Systems That Adapt to Diverse Worldviews</i> Claudiu Creanga and Liviu P. Dinu | 95 |
| <i>The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation</i> Maja Pavlovic and Massimo Poesio | 100 |
| <i>What Does Perspectivism Mean? An Ethical and Methodological Countercriticism</i> Mathieu Valette | 111 |
| <i>OrigamLM: A Dataset of Ambiguous Sentence Interpretations for Social Grounding and Implicit Language Understanding</i> Liesbeth Allein and Marie-Francine Moens | 116 |
| <i>Linguistic Fingerprint in Transformer Models: How Language Variation Influences Parameter Selection in Irony Detection</i> Michele Mastromattei and Fabio Massimo Zanzotto | 123 |

Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI
Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor and Ding Wang 131

A Dataset for Multi-Scale Film Rating Inference from Reviews
Frankie Robertson and Stefano Leone 142

Is a picture of a bird a bird? A mixed-methods approach to understanding diverse human perspectives and ambiguity in machine vision models

Alicia Parrish, Susan Hao, Sarah Laszlo, Lora Aroyo

Google Research

alicia.v.parrish@gmail.com

Abstract

Human experiences are complex and subjective. This subjectivity is reflected in the way people label images for machine vision models. While annotation tasks are often assumed to deliver objective results, this assumption does not allow for the subjectivity of human experience. This paper examines the implications of subjective human judgments in the behavioral task of labeling images used to train machine vision models. We identify three primary sources of ambiguity: (1) depictions of labels in the images can be simply ambiguous, (2) raters' backgrounds and experiences can influence their judgments and (3) the way the labeling task is defined can also influence raters' judgments. By taking steps to address these sources of ambiguity, we can create more robust and reliable machine vision models.

Keywords: Disagreements, Ambiguity, Machine vision

1. Introduction

Computer vision models rely on human annotations, and the default assumption when creating training and evaluation datasets is often that there is a single correct answer about what concepts or objects are present in an image. Though there is growing acceptance that human disagreements are common with respect to inherently ambiguous data [Kairam and Heer \(2016\)](#), the role of human disagreements as a general property of *any* annotation task is much less accepted. In image annotation, even the annotation of concrete concepts (e.g., *bird*) in clear, high quality, unobscured imagery can lead to disagreement between raters that we should seek to understand. The interplay of annotator, concept, and image characteristics in labeling tasks should inform how we analyze human ratings, leverage disagreement insights to train and evaluate models, and translate findings into best practices.

To understand individual human behavior in image annotation, we focus on large label space models for computer vision. *Large label space models* are machine vision models that predict the probabilities of many entities in an image, in contrast to *binary classification models* that predict the presence or the absence of a single entity and *segmentation models* that identify pixels corresponding to an entity. Most image models require labeled training data to learn to classify accurately (e.g., [Ji et al. \(2019\)](#)). This requirement typically consists of a training set of images labeled with their contents, usually by human annotators. For example, to learn to classify birds in images, a large label space model would need to see many (usu-

ally at least tens of thousands) of images of birds, depicted in a range of different environments and positions with the inclusion of rare species. Human annotators are employed to label each image, providing the "ground truth" needed to train the model.

We know, however, that *human raters disagree*. Bird experts may disagree on which species of bird an image belongs to. Non-experts may be unsure about taxonomic classifications of certain bird species. People can disagree whether the concept of "bird" applies in a given case (e.g., *pictures* of birds). Some reasons, like poor image quality, can indicate problems with a specific image. However, many cases of human disagreements are due to ambiguity in the label or the labeling task. Label ambiguity can arise from many factors, including similar-looking labels (*birds* and *bats* look similar), regional naming differences (*robin* in the US, vs. *redbreast* in the UK), and different understandings of the task. Label ambiguity is a challenge for machine vision models because it can lead to inaccurate predictions. For example, if a machine vision model is trained on a dataset of *bird* images with ambiguous labels, it may not be able to accurately identify birds in new images (see [\(Karimi et al., 2020\)](#) for an analysis of the impact of label noise on medical image analysis models).

In order to better understand the human factors that influence label ambiguity on large label space model performance, we developed an open data challenge to crowdsource *adversarial image-labels pairs* for machine vision models. In this online challenge, participants competed to identify edge case images that state-of-the-art machine vision models might incorrectly classify. The goal was to understand systematic failures of these models, with

an eye towards augmenting the data used to train these models to better cover such failure cases. Our challenge included 2 tasks. In Task 1, annotators were asked whether a given label applied to each image-label pair. In Task 2, image-label pairs collected during the challenge were tested against multiple state-of-the-art classification models and surfaced (i) pairs with clear human-machine disagreements and (ii) pairs where multiple *human* annotators couldn't reach clear agreement. This challenge required a data analysis strategy designed to identify patterns in the misclassified images to better understand the human factors that contribute to label ambiguity and to develop new mitigation methods to improve machine classification accuracy. The adversarial data from this challenge and the resulting analysis have the potential to make a significant contribution to the development of more robust and reliable large label space models. In this paper, we present the results of the public adversarial data challenge, analyze the ambiguities in the resulting data, and organize them into a theoretical framework to provide recommendations for human annotation and data collection policies that best address the types of ambiguities we observed.

2. CATS4ML challenge

The CATS4ML (Crowdsourcing Adverse Test Sets for Machine Learning) challenge ran online for four months, under the CrowdCamp umbrella of the HCOMP 2021. The challenge used the Open Image Dataset¹ (OID V4; Krasin, 2017) as source material. It contains ~9M images annotated with 20k possible image-level labels, object bounding boxes and segmentation masks. Importantly, the labels, bounding boxes, and segmentation masks are provided by a machine, with only a small portion verified by humans. The challenge was designed on the premise that, likely, the machine labeler makes mistakes, these mistakes are systematic, and studying systematic machine failures can improve machine labelers in the future. In this challenge, we aimed to identify adversarial image-label pairs in OID V4 that would yield human-model disagreement.

Challenge participants examined the machine-labeled subset of OID V4 images, focusing on a selected set of 23 entities - *Bird, Canoe, Lipstick, Chopsticks, Muffin, Pizza, Croissant, Child, Smile, Selfie, American football, Athlete, Physician, Nurse, Teacher, Chef, Firefighter, Coach, Construction Worker, Bus driver, Funeral, Thanksgiving, or Graduation* - and submit image-label pairs where they thought the image classification machine algorithm was wrong. Limiting the label set to 23 was

¹https://storage.googleapis.com/openimages/web/factsfigures_v4.html

necessary to make the scope of the competition tractable—human participants were unlikely to be able to examine all 20k labels in the OID. These 23 labels were selected to represent a neutral (non-controversial, non-sensitive) set of topics across different types: 8 objects, 3 events, 9 roles and professions, and 3 abstract concepts. Another criteria for selection was to have a good representation of different levels of ambiguity of the label, e.g. “child” is a broad concept and could be interpreted in different ways; “athlete” could mean different things for different cultures; “physician” and “nurse” could have ambiguous visual representations.

Ten individuals submitted image-label pairs to the challenge, submitting more than 14,000 image-label pairs. Of these, 13,683 image-label pairs were “valid” (i.e., the pairs were drawn from OID V4 and used one of the 23 challenge labels). After removing duplicate submissions, 10,668 unique pairs remained. Participants could choose for which labels of the 23 to submit and how many images. The 10,668 unique image-label pairs were further validated by engaging two globally-diverse crowds of human annotators in three different locales and two in-house experts (described in the Methods section). The image-label pairs were also submitted to six machine vision models to examine how human judgements aligned with state of the art model judgements and to identify cases of human-model disagreements. The challenge data is already available publicly on github;² additional human annotations collected for this study will be made available via the same resource.

3. Methods

Here, we provide a detailed description of the materials (datasets and models) used, annotation task procedures, task annotators, and data analysis and score computation decisions that we made in arriving at the results, all summarized in Figure 1.

3.1. Materials

Dataset description. As described above, the CATS4ML dataset was composed of 10,668 unique submissions made by challenge participants. Appendix Figure 5 shows the distribution of images across all 23 target labels - most images were submitted for the label ‘bird’ (26% of the data) with an exponential long-tail distribution across all other labels (e.g., seven labels with between 500-1100 images per label, nine with between 140-350 images per label and six labels with 100 or fewer images per label).

²<https://github.com/google-research-datasets/cats4ml-dataset>

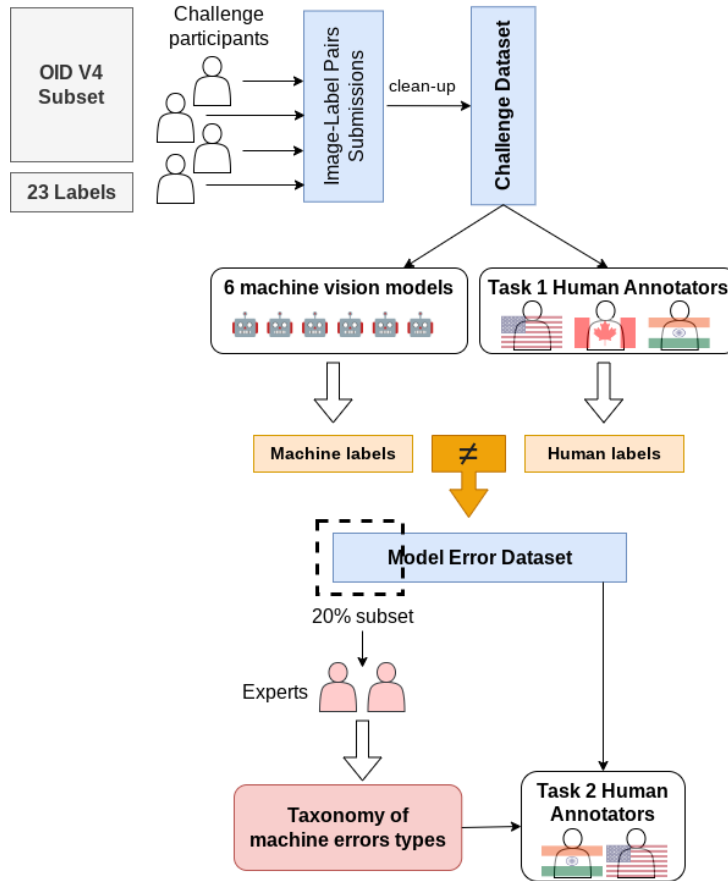


Figure 1: **Adversarial data collection from the CATS4ML challenge and the follow-up annotation tasks.** First, challenge participants used a subset of the OID V4 dataset to discover image-label pairs and submit them to the challenge. After cleaning the data to remove duplicates and invalid submissions, we labeled the data with state-of-the-art machine vision models and human annotators from three different locales. From their labels, we constructed a machine error dataset that consisted only of the image-label pairs with human-model disagreements. Two members of the research team qualitatively analyzed 20% of this dataset to create a *taxonomy of reasons* for the machine errors, which was then used by human annotators from two different locales to annotate the entire machine error dataset.

Vision models. To provide machine labels of each CATS4ML dataset image, we used an ensemble of six machine vision models, each of which were state-of-the-art when they were released. These models are all non-public variants of the InceptionV2-based image classifier (Ioffe and Szegedy, 2015) developed in the period of 2015-2022 (including models used in OID-V4 and OID-V3 (Krasin, 2017), publicly available through Open Images Dataset).

Model error dataset. Based on human annotation Task 1 and qualitative validation by experts, we constructed a subset of 8,326 image-label pairs to have labeled by humans in Task 2. Image-label pairs included in Task 2 met at least one of the following criteria: (i) at least one vision model disagreed with the human majority vote from Task 1, or (ii) there was significant disagreement among the human annotators in Task 1. This smaller dataset

allows for a targeted qualitative analysis of the reasons for human-model disagreements.

3.2. Annotation task procedure

Human annotation task 1—Label verification.

In Task 1, annotators indicated whether a given label applied to an image for each image-label pair in the CATS4ML dataset. No specific training was provided to annotators before beginning the task, as the task was injected into a general purpose image-label validation system used by a professional rater pool to perform a variety of tasks other than this one. For each example, 19 annotators viewed a single image and selected one of three answer options indicating whether a given label applies to that image, does not apply, or they are unsure (Appendix Figure 3).

Human annotation task 2—Model error verification. In Task 2, annotators examined the model error dataset (8,326 image-label pairs from the CATS4ML dataset with human and machine labelers disagreement from Task 1). For each example, 14 annotators saw a machine label produced for an image, and they indicated whether the model was correct or not (Appendix figure 4.A). Guidelines (presented to annotators before starting) included definitions of seven categories of model error that were identified by experts in a qualitative analysis of a subset of the model error dataset (see § 3.5). Annotators answered two questions about each item: (i) whether the machine prediction indicated correctly whether the label was present in the image or not (Appendix figure 4.B), and (ii) in the case of model error, select one out of seven possible error types (Appendix figure 4.C). Annotators could select an additional model error type if it was needed. At any time, annotators could return to previous items and change their responses as needed. Annotators were not given any information on how the “machine prediction” was constructed in order to avoid biasing them towards agreeing or disagreeing with the machine prediction.

3.3. Annotators

Data submitted by challenge participants was validated three times—twice by paid annotators and once by members of the research team. The paid annotators were recruited from professional rater pools and had prior experience in data annotation tasks. To ensure that the annotations on the image-label pairs reflected a range of human perspectives, particularly because we expected that the examples would be especially challenging, we recruited raters from different geographic locales (US, Canada, and India). We selected these locales because they have English as a dominant language and are common locales for recruiting annotators. We did not collect demographic information aside from locale for these annotators. There was no overlap between the raters in Task 1 and Task 2. We summarize the unique number of annotators and the total size of the annotator pools in each task in Table 1.

The annotators in Task 1 consisted of 41 unique raters. Table 1 (left side) shows the number of raters from each locale. We gathered 19 ratings per image-label pair (7 from raters in the US, 7 from raters in India, 5 from raters in Canada), as shown in the right side of Table 1. Each rater labeled an average of 4726 image-label pairs (median 4088). However, 4 annotators (3 from the US, 1 from India) chose to end the task early, providing fewer than 100 annotations each, so the total number of ratings provided by individual raters ranged from 3 to 9932. We ensured that each image-label pair was rated

by the same number of unique annotators from the same locale distributions to ensure that the image-label-pair-level ratings were not imbalanced. Task 1 raters were compensated monetarily in alignment with local norms of the region in which they were working.

Subsequently, two members of the research team performed a qualitative analysis (see § 3.5 for details) to classify the causes of model error in a sample of about 20% (2,035 image-label pairs) of the dataset from Task 1. This validation was performed in order to identify possible model error types (detailed in Appendix Table 9), and qualitatively categorize them for Task 2, described next. The experts each had in-depth experience with machine vision models.

The annotators in Task 2 consisted of 56 raters from two different locales: US and India. Table 1 shows the number of raters from each locale. As in Task 1, example-level annotations were balanced across the locales, as we gathered 14 annotations per image label pair (7 from raters in the US, 7 from raters in India). Each annotator labeled an average of 2080 image-label pairs (median 1652), with the total number of ratings provided by raters ranging from 368 to 8325. Task 2 annotators were compensated monetarily in alignment with local norms of the region in which they were working.

3.4. Scoring

Merging Task 1 and Task 2 human labels.

Tasks 1 and 2 both required annotators to assess if a label was in a given image. In Task 1, this question was direct (“is the label in the image?”); in Task 2, it was indirect (“a machine predicted X, is the machine correct?”). Thus we end up with labels that are not directly comparable, and we need to infer the *intent* of the annotator’s judgment with respect to whether the label is in the image in Task 2. To analyze and directly compare the combined annotations from both tasks, we transformed Task 2 responses to reflect the annotator’s judgment about whether the label was in the image (e.g., if the machine label was “no,” and the annotator marked that “no, that the machine was not correct,” we transform that annotation to “yes, the label is in the image” for comparison with the interpretation of the Task 1 label where the annotators were directly asked if the label is in the image). In cases where the annotator rated an image-label pair “unsure,” we maintain the “unsure” label.

Aggregation of human labels to supermajority vote.

Machine vision datasets often carry only positive or negative annotations for image-label pairs. Though we have a high replication of annotations in the dataset that allows for working with soft-

| | Size of the total rater pool | | | | Unique raters per example | | | |
|--|------------------------------|-----------|-----------|--------------|---------------------------|-----------|-----------|--------------|
| | US raters | IN raters | CA raters | Total raters | US raters | IN raters | CA raters | Total raters |
| Task 1: Is label in image? <i>Annotated 10,668 image-label pairs</i> | 23 | 13 | 5 | 41 | 7 | 7 | 5 | 19 |
| Model error categorization <i>Annotated 2,035 image-label pairs</i> | 2 experts | | | | 2 experts | | | |
| Task 2: Confirm model error <i>Annotated 8,326 image-label pairs</i> | 22 | 34 | – | 56 | 7 | 7 | – | 14 |

Table 1: For each annotation task, (i) the size of rater pools, and (ii) the number of unique raters for each task example.

label distributions, we consider majority vote for comparison with standard machine vision datasets and to assess sources of disagreement that emerge when considering standard majority-vote aggregations. We classify image-label pairs along three dimensions: (i) “clear yes” (positive examples) where at least 66% of annotators indicated the label was in the image, (ii) “clear no” (negative examples) where at least 66% of annotators indicated that the label was not in the image, and (iii) “ambiguous,” for all other examples that did not meet either of the previous two criteria. Image-label pairs may fall into the ambiguous category due to either a high degree of disagreement in terms of “yes”/“no” votes, or because of a high rate of “unsure” answers.

3.5. Data analysis

Annotator agreement metrics: We measure both inter-rater reliability (IRR, Krippendorf’s alpha) and cross-replication reliability (xRR; Wong et al., 2021) to assess the agreement patterns of annotators. We measure Krippendorf’s alpha because this metric is robust to imbalanced data, where different sets of annotators rate different sets of examples. Higher values of alpha indicate greater agreement among annotators. xRR is based on Cohen’s Kappa, and is used to compare different groups of annotators to determine if the agreement between the two annotation distributions is more similar than would be expected by chance. xRR values are interpreted on the same scale as IRR, and higher values indicate greater similarity in responses across the two groups.

Linear modeling: Linear mixed effects models can be used to simultaneously account for random effects related to individual annotators and items, while also taking into account complex interactions between experimental conditions. We construct a null model predicting whether the rater indicated that the label is in the image or not (i.e., “yes” or “not yes”, which collapses together “no” and “unsure” ratings), with random intercepts for raters and

items. We compare this null model to three single-predictor models that add fixed effects of (i) rater locale, (ii) label id, and (iii) task type, and also two models that consider all three fixed effects as (i) additive, (ii) interactive predictors, and we perform model comparisons using ANOVA to compare the three single-predictor models to the null model, and to compare the additive and interactive models to ensure that we are making matched comparisons.

Qualitative analysis: Two members of the research team provided expert annotations for a qualitative analysis of the reasons for model errors. They assessed a 20% sample of the model error dataset, visually comparing the image and the model predictions for the target label on that image. The two experts proposed a taxonomy of error reasons that were then discussed with the larger research team and adapted to be used by human annotators in Task 2 to label a larger dataset. We provide examples of each error reason, with images labeled as that reason, in Appendix Table 9.

4. Results

We classify the image-label pairs from the challenge as either positive or negative examples of the submitted label. We use supermajority vote of human scores to identify which image-label pairs are positive examples (“clear yes”), negative examples (“clear no”), or could not be reliably classified due to rater disagreements or high rates of “unsure” ratings (“ambiguous”). Using the aggregated Task 1 and 2 results, we find 4300 positive examples (40.3%), 2264 negative examples (21.2%), and 4104 ambiguous examples (38.5%); Appendix table 8 breaks down these aggregate values by the target labels.

Model performance and image adversariality. As over one third of image-label pairs from the challenge were ambiguous to human raters, we investigate whether these examples were also ambiguous to machine vision models. To do this, we quantify

the *adversariality of the image-label pairs* using the 61.5% of the dataset (6564 image-label pairs) on which we can compute a high-agreement human label (the “clear yes” and “clear no” examples in Table 8). Adversariality is computed as the number of human-model disagreements observed across the models tested. We identify 710 (10.8%) highly adversarial image-label pairs that none of the 6 models got correct (where “correct” means “agrees with the human consensus”). This method allows us to rank the adversariality of individual images (Table 2), based on how many models made incorrect judgements. We find that 72.8% of images were adversarial to at least one of the state-of-the-art models.

| Adversariality strength: number of models fooled | N. image-label pairs | Percent of 6564 dataset |
|--|----------------------|-------------------------|
| 0 (not adversarial) | 1784 | 27.2 |
| 1 | 1207 | 18.4 |
| 2 | 1426 | 21.7 |
| 3 | 578 | 8.8 |
| 4 | 472 | 2.7 |
| 5 | 387 | 5.9 |
| 6 (very adversarial) | 710 | 10.8 |

Table 2: Image-label pair adversariality across the dataset. To accurately reflect human-model agreement patterns, we exclude items with no human supermajority vote.

Reasons for adversariality. We break down this measure of adversariality by using the qualitative labels assigned by annotators in Task 2 to identify which model error reasons are most associated with high adversariality (Table 3). We observe that visual similarity between the label and the image (e.g., the label is “bird” and the image shows a bat) is the most frequently identified reason for model errors and is most associated with highly adversarial image-label pairs, with 55% of the 710 most adversarial images falling into the category of visual similarity. Annotators also identified misleading background context and atypical depictions of the label as primary causes of model failures, covering 30% and 33% of the most adversarial images, respectively.

Factors in human disagreements. We investigate potential reasons for the disagreement between humans that we observed in the 38% of image-label pairs with no supermajority agreement. For this, we consider the full CATS4ML dataset, and we assess ambiguity from three perspectives: (1) *disagreements due to rater characteristics*, (2) *disagreements due to characteristics of image-label pairs*, and (3) *disagreements due to characteristics*

of rating task. These three perspectives have previously been identified as relevant to understanding crowd labels and rater disagreements (Aroyo and Welty, 2014). We use a linear mixed effects model (see § 3.5), and we compare three single-predictor models to the null model using an ANOVA. Table 4 shows that each of these three models is a significantly better fit for the data compared to the null model, indicating that rater characteristics (as indexed by locale), label name, and task framing all explain a significant amount of variance in the data. To determine whether these three factors interact with each other, we construct both an additive and an interactive model using all three predictors; the interactive model is a significantly better fit for the data compared to the additive model ($p < 0.001$).

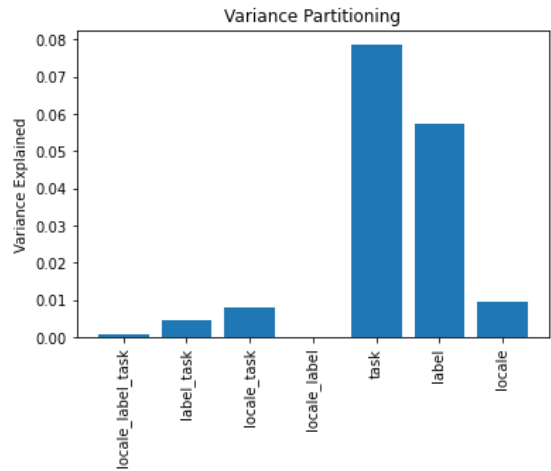


Figure 2: **Variance partitioning on a linear additive model.** First, rater id was regressed out by fitting these features to a multi-class logistic regression model with l2 penalty with raters’ judgments (yes, no, unsure) as the dependent variable. Using log loss as the unit deviance or residuals, we then fit several additive models on those residuals using a combination of locale, label, and task features as independent variables. The figure above shows the shared and unique variance of these different submodels. We observe that the submodels with task followed by label and locale have the highest unique variance.

We used *variance partitioning analysis* to identify which of the three factors (rater locale, label name, task type) had the greatest impact on raters’ judgments. Variance explained by rater id was accounted for first, and then an additive model was fitted to the residuals using features from the three factors ($R^2 = 0.159$). To understand the shared and independent variance of each set of features, several submodels were fitted to these residuals. Figure 2 shows that task type ($R_{uniq}^2 = 0.079$) followed by label ($R_{uniq}^2 = 0.057$) and rater locale

| Adversariality | Total pairs | Ambiguous label | Artistic depiction | Quality issue | Background context | Visual similarity | Out of context | Atypical depiction | Other error reason |
|----------------|-------------|-----------------|--------------------|---------------|--------------------|-------------------|----------------|--------------------|--------------------|
| 1 | 1137 | 7 | 6 | 1 | 10 | 45 | 0 | 15 | 13 |
| 2 | 1404 | 36 | 21 | 22 | 570 | 207 | 6 | 832 | 549 |
| 3 | 562 | 16 | 7 | 5 | 216 | 102 | 0 | 308 | 156 |
| 4 | 471 | 18 | 5 | 6 | 172 | 144 | 4 | 228 | 107 |
| 5 | 387 | 14 | 5 | 1 | 114 | 196 | 2 | 151 | 85 |
| 6 | 710 | 26 | 15 | 10 | 212 | 389 | 0 | 234 | 125 |
| TOTAL | 4671 | 117 | 59 | 45 | 1294 | 1083 | 12 | 1768 | 1035 |

Table 3: Total image-label pairs for which a given error reason was indicated by at least 25% of raters in the Task 2 qualitative labeling task. Totals are different from Table 2 because only a subset of the full CATS4ML dataset was rated in Task 2. “Total pairs” represents the total number of image-label pairs rated in Task 2. Totals across rows may be different than the “total pairs,” as examples can have more than one error reason, and examples can have no error reasons achieving at least the 25% threshold or annotators selecting that reason.

| Model description | Model definition | AIC | BIC | Fit compared to null model |
|------------------------------------|--|----------|----------|----------------------------|
| Null (baseline) | Rating $\sim 1 + (1 rater_id) + (1 item_id)$ | 289711.9 | 289754.4 | N/A |
| Rater locale | Rating $\sim Locale + (1 rater_id) + (1 item_id)$ | 289677.0 | 289740.7 | $p < 0.001$ |
| Task type | Rating $\sim Task_type + (1 rater_id) + (1 item_id)$ | 289669.7 | 289722.8 | $p < 0.001$ |
| Label name | Rating $\sim Label_name + (1 rater_id) + (1 item_id)$ | 282069.8 | 282324.5 | $p < 0.001$ |
| Additive model (all predictors) | Rating $\sim Locale + Label_name + Task_type + (1 rater_id) + (1 item_id)$ | 282007.2 | 282293.8 | $p < 0.001$ |
| Interactive model (all predictors) | Rating $\sim Locale * Label_name * Task_type + (1 rater_id) + (1 item_id)$ | 271579.6 | 272725.9 | $p < 0.001$ |

Table 4: Mixed effect model definitions and fit statistics.

($R_{uniq}^2 = 0.010$) have the highest amount of explained unique variance, with these features’ combined unique variance accounting for 91.57% of observed variance in the original additive model. Shared variance across these features did not impact raters’ judgements as much as each individual factor. While these analyses are useful in understanding how these factors interact and contribute to raters’ judgments, we seek to further understand sources of disagreement *within* each factor by investigating these factors independently in our qualitative analyses.

Understanding disagreements due to rater characteristics. For both Tasks 1 and 2, we investigate rater agreement with Krippendor’s alpha (inter-rater reliability; IRR) and cross-replication reliability (xRR). Overall agreement was only moderate in both tasks. In Task 1, IRR was higher within locale for US and Indian raters than the overall IRR; xRR revealed that the Indian and American raters agreed with each other more than did Indian & Canadian raters or Canadian & American raters (Table 5). In Task 2, agreement was even lower than in Task 1. Taken together, these results show that human labelers did not tend to agree with each other on label judgments, and that a rater’s

| Metric | Rater locale | Agreement |
|--------|----------------|-----------|
| IRR | OVERALL | 0.4737 |
| | India | 0.5739 |
| | USA | 0.5739 |
| | Canada | 0.3794 |
| xRR | India x USA | 0.5429 |
| | India & Canada | 0.4653 |
| | USA & Canada | 0.5088 |

Table 5: Task 1 IRR & xRR scores, by locale.

| Metric | Rater locale | Agreement |
|--------|--------------|-----------|
| IRR | OVERALL | 0.1982 |
| | India | 0.3624 |
| | USA | 0.1299 |
| xRR | India & USA | 0.1846 |

Table 6: Task 2 IRR & xRR scores by locale.

locale impacted how that rater labeled images. Appendix table 10 provides example images where different locales reached different consensus labels. In panel (a), US raters affirmed the label “bird,” Canadian raters rejected the label “bird,” and

Indian raters unanimously indicated “unsure;” in (b), 92% of American raters affirmed that the label “bird” while 86% of Indian raters were unsure. Both examples are artistic depictions of a “bird”—they are drawings that represent a bird (or just the bird’s skeleton), and the different response patterns from raters in different locales highlights the way that a person’s cultural context may influence their judgments in what many would consider an *objective* labeling task.

Understanding disagreements due to the image-label pairs. To identify image-label pairs that are inherently ambiguous, we identify examples where a high number of raters responded that they were “unsure” if the label was in the image. As many image labeling tasks are presented to annotators with only binary labels available (“yes” or “no”), we expect that examples in which the majority vote label is “unsure” would lead to disagreements in a binary task set up. In 2039 examples (21.5% of all image-label pairs), the “unsure” label was the most frequently selected label across Task 1 raters. Appendix Table 11 shows two illustrative examples. In the first case, where the label is “Thanksgiving,” it is genuinely ambiguous whether the meal is a Thanksgiving dinner; in the second it is ambiguous whether the people wearing white coats are “physicians,” as opposed to any other profession that wears a lab coat. In both cases, the label is potentially consistent with the image, but crucially disambiguating background information about the image’s setting, date, or participants is unavailable to the raters. The labels in this study spread across a range of different types of concepts: concrete, abstract, events, roles, and professions. Some of these categories are inherently more difficult to identify in an image-labeling task. Professions and roles (two of the more inherently ambiguous labels in the challenge) can be strongly context-dependent, and identification relies on cultural knowledge and assumptions about the people and event being depicted. Events can be difficult to determine from a single image as well, as many types of events include multiple sub-parts to the whole (e.g., is “Thanksgiving” just a nicely-dressed turkey?). However, we also observe that concrete object labels (e.g., “bird”) can lead to consistent unsure annotations; for example when the image is a painting of a bird, a bird mascot for a sports team, or a whole roasted chicken, annotators disagree on or are unsure about whether the label “bird” should apply.

Understanding disagreements due to the rating task. To identify cases where the task may have affected rater judgments, we analyze examples for which the supermajority vote label on a

given example *changes* between the two tasks. In Table 7, we show a cross-task comparison with the number of examples that fall into each of the nine possible combinations of labels from Task 1 and Task 2. We observe that 35.8% of the image-label pairs switch supermajority vote labels between Tasks 1 and 2. Most flips involve the “ambiguous” label, indicating relatively few cases where raters truly change their vote from “yes” to “no” (or vice versa). We describe observations from these cases in Appendix G and show randomly selected image-label pairs from each of the six different kinds of label flips observed to illustrate these cases.

5. Discussion & recommendations

In this paper, we are concerned with label ambiguity in large label space models, which is typically deleterious to model performance. We identified three key factors contributing to label ambiguity: rater background, label characteristics, and task design. These factors influence whether humans tend to disagree with both model predictions and each other. We demonstrated that it is, in fact, challenging for human raters and machines to agree on label ground truth, even for relatively concrete concepts such as “bird.” We further demonstrated that the geographical location in which a human rater is situated can have an impact on their answers in a labeling task. Finally, we demonstrated that small changes to the way a labeling task is framed can have an impact on how the task is performed. Given these potential complications to performing the bedrock task of machine vision model training (assigning ground truth to images), we conclude with our recommendations as to how developers, annotation guidelines and policymakers can best address label ambiguity.

- **Take a community-driven approach to data labeling.** Make sure that the people doing the labeling are from the communities that are going to be impacted by the model deployment.
- **Assume variance, ambiguity, and subjectivity are always present in any data labeling task,** regardless of how simple it may seem. There is not, and cannot be, one singular “gold standard.” To the extent possible, identify and explore potential sources of ambiguity in any data set, and understand how these sources of ambiguity might be related to the communities impacted by the model.
- **Define and deploy metrics to measure ambiguity in data.** For example, if data is labeled in different sessions, on different interfaces, or by different pools raters, measure and track differences between data subsets. Measure and

| Supermajority vote label | | Number of examples | Percent of total |
|-------------------------------|--------------------------------|--------------------|------------------|
| Task 1: Is label in image? | Task 2: Is machine correct? | | |
| Yes | Yes | 2714 | 32.6 |
| Yes | No | 6 | 0.1 |
| Yes | Ambiguous | 464 | 5.8 |
| No | Yes | 9 | 0.1 |
| No | No | 845 | 10.2 |
| No | Ambiguous | 614 | 7.4 |
| Ambiguous | Yes | 1561 | 18.8 |
| Ambiguous | No | 325 | 3.9 |
| Ambiguous | Ambiguous | 1787 | 21.5 |

Table 7: Cross Task comparison. In **bold** are rows representing image-label pairs that had consistent supermajority labels across tasks. All other rows represent image-label pairs that had inconsistent supermajority labels across tasks. The label for Task 2 represents the transformed label to make it comparable in interpretation to Task 1.

track any differences across data subset by demographic properties of the community that will be impacted by the data (e.g., geographic location, gender, age, ability).

There has been little work that provides specific recommendations for policies pertaining to large label space models. Currently, content moderation strategies recommend employing machine safety filters that comprise several safety classification models (Hao et al., 2023). Although our dataset does not include safety content, our challenge shows that even for categories that are non-controversial, there is ambiguity. Thus, for more subjective labels that pertain to safety (e.g., porn, violence), these ambiguities may be amplified (Homan et al., 2023), which can result in unreliable safety classifications. Adopting these recommendations will ensure that a deployed model has been contributed to by the community it serves, that possible sources of model failure are understood and tracked, and that the way the model is serving different subsets of the community is also tracked. A model deployed under these conditions is on the right track to responsibly serve its community.

Reproducibility Statement

The original CATS4ML data is available on github at github.com/google-research-datasets/cats4ml-dataset. This dataset contains the image-label pairs collected for the challenge along with an aggregation of five human annotations for each example. For the study described in this paper, we collected additional human annotations not part of the original repository; those annotations will be made available as a supplemental dataset, along with the code for the analyses conducted in this paper (descriptive stats,

task score conversions, IRR, xRR, mixed-effects modelling, variance partitioning). To accompany the additional data release, we will also include a datasheet (Gebru et al., 2018).

Bibliographical References

- Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *HCJ* 1, 1 (Sept. 2014).
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *CoRR* abs/1803.09010 (2018). arXiv:1803.09010 <http://arxiv.org/abs/1803.09010>
- Susan Hao, Piyush Kumar, Sarah Laszlo, Shivani Poddar, Bhaktipriya Radharapu, and Renee Shelby. 2023. Safety and Fairness for Content Moderation in Generative Models. (June 2023). arXiv:2306.06135 [cs.LG]
- Christopher M Homan, Greg Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. 2023. Intersectionality in Conversational AI Safety: How Bayesian Multilevel Models Help Understand Diverse Perceptions of Safety. (June 2023). arXiv:2306.11530 [cs.HC]
- Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (Feb. 2015). arXiv:1502.03167 [cs.LG]
- Xu Ji, Andrea Vedaldi, and Joao Henriques. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation.

In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (Seoul, Korea (South)). IEEE.

Sanjay Kairam and Jeffrey Heer. 2016. Parting Crowds: Characterizing Divergent Interpretations in Crowdsourced Annotation Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) (CSCW '16). Association for Computing Machinery, New York, NY, USA, 1637–1648.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. 2020. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Med. Image Anal.* 65 (Oct. 2020), 101759.

Duerig T. Alldrin N. Ferrari V. Abu-El-Hajja S. Kuznetsova A. Rom H. Uijlings J. Popov S. Veit A. Belongie S. Gomes V. Gupta A. Sun C. Chechik G. Cai D. Feng Z. Narayanan D. Murphy K Krasin, I. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from <https://github.com/openimages>.

Ka Wong, Praveen Paritosh, and Lora Aroyo. 2021. Cross-replication Reliability – An Empirical Approach to Interpreting Inter-rater Reliability. (June 2021). arXiv:2106.07393 [stat.AP]

A. Task Interfaces

Figures 3 and 4 show the interfaces that were shown to human raters in the two annotation tasks described in the main text.

B. Label distribution in the CATS4ML dataset

In Figure 5, we show the distribution of raw counts of each label that was submitted in the CATS4ML challenge. Challenge participants were not restricted in terms of which labels they chose in their example submissions, and thus we could not ensure equal distribution across the labels. The skew towards ‘bird’ labels is likely due to multiple factors, including the number of instances of ‘bird’ in the source data, the ease of browsing images for the target object, and participant familiarity with the range of ways the label may be represented in images.

C. By-label supermajority vote results

In Table 8, we show how many images from the challenge were assigned each label (‘yes,’ indicating the label is in the image, or ‘no,’ indicating the label is not in the image), and how many were classified as ‘ambiguous,’ indicating that neither the ‘yes’ or ‘no’ supermajority vote label could be applied.

D. Qualitative labels of model error reasons

Table 9 (spanning three pages to ensure the images are legible) shows an example of each of the qualitative labels used in the Task 2 (“confirm model error”). These labels are derived from expert validation of human-model disagreements from Task 1 (“is label in image”).

E. Examples of disagreements due to the rater locale

Table 10 shows randomly selected examples where raters from different locales gave systematically different ratings on the same image-label pair.

F. Examples of disagreements due to the image-label pair

Table 11 shows randomly selected examples where raters consistently indicated that the image itself was ambiguous with respect to the target label.

G. Examples of disagreements due to the rating task

As reported in the main text, one third of image-label pairs flip their label based on the task phrasing. Most of these flips involve the ‘ambiguous’ supermajority vote label, indicating that there are relatively few cases where raters truly change their vote from “yes” to “no” (or vice versa). To illustrate these cases, we randomly select an image-label pair from each of the six different kinds of label flips observed, and show the examples along with the raters’ labeling patterns in Tables 12, 13 and 14. We observe patterns where the human supermajority vote label switches to align with the machine label shown in Task 2 (12a, 13a, 14b) and to contradict the machine label shown (14a). These images are illustrative of the kinds of difficulties that annotators had in assigning labels, and they show that slight changes in the wording or presentation of the task can lead to different results, even on a task that appears straightforward.

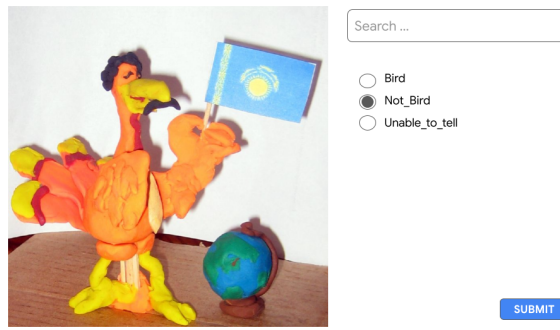



Figure 3: Sample interface for Task 1: Is label in image?

A.

| Link to view image | Label on this image | Machine prediction for this label |
|-------------------------------|---------------------|-----------------------------------|
| link_to_image | Child | no |
| link_to_image | Bird | yes |
| link_to_image | Child | no |
| link_to_image | Lipstick | no |



B.

| Link to view image | Label on this image | Machine prediction for this label | Did the machine correctly predict the label? | Select the <u>main reason</u> you think the machine is wrong (Select N/A if the machine is correct) |
|-------------------------------|---------------------|-----------------------------------|--|---|
| link_to_image | Child | no | <input type="text"/> | |
| link_to_image | Bird | yes | Yes | |
| link_to_image | Child | no | No | |
| link_to_image | Lipstick | no | Unsure | |
| link_to_image | Child | no | | |
| link_to_image | Child | yes | | |
| link_to_image | Athlete | yes | | |
| link_to_image | Child | no | | |

C.

| Link to view image | Label on this image | Machine prediction for this label | Did the machine correctly predict the label? | Select the <u>main reason</u> you think the machine is wrong (Select N/A if the machine is correct) | Styl |
|-------------------------------|---------------------|-----------------------------------|--|---|------|
| link_to_image | Child | no | No | | |
| link_to_image | Bird | yes | | | |
| link_to_image | Child | no | | Machine over-relied on background context | |
| link_to_image | Lipstick | no | | Object is depicted out of typical context (e.g., no background) | |
| link_to_image | Child | no | | Unexpected or atypical depiction of the label | |
| link_to_image | Child | yes | | Artistic depiction of the label | |
| link_to_image | Athlete | yes | | Visually similar shape of the label | |
| link_to_image | Child | no | | Ambiguous meaning of the label (e.g. triggers different interpretation) | |
| link_to_image | Lipstick | no | | Image has quality issue | |
| link_to_image | Selfie | no | | OTHER reason for model error | |
| link_to_image | Child | yes | | N/A (the model was correct) | |
| link_to_image | Child | no | | | |
| link_to_image | Child | yes | | | |


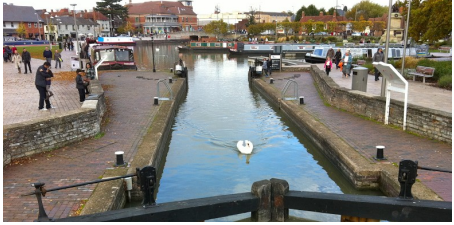
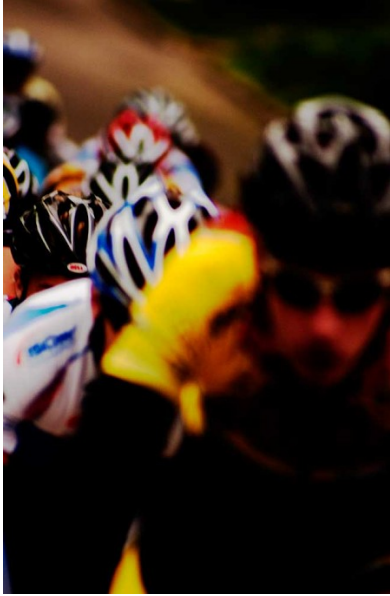
Figure 4: Sample interface for Task 2: Confirm model error.






Figure 5: Histogram of valid image-label pair counts per label name.

| Target Label | clear yes | clear no | ambiguous | TOTAL |
|---------------------|-------------|-------------|-------------|--------------|
| Bird | 1305 | 43 | 1433 | 2781 |
| Smile | 721 | 53 | 261 | 1035 |
| Lipstick | 451 | 20 | 465 | 936 |
| Canoe | 63 | 488 | 382 | 933 |
| Chopsticks | 108 | 702 | 67 | 877 |
| Athlete | 630 | 14 | 123 | 767 |
| Muffin | 19 | 428 | 92 | 539 |
| Child | 387 | 29 | 88 | 504 |
| Chef | 32 | 214 | 138 | 384 |
| Firefighter | 69 | 70 | 160 | 299 |
| Coach | 9 | 19 | 187 | 215 |
| Construction worker | 49 | 60 | 101 | 210 |
| American football | 65 | 27 | 82 | 174 |
| Pizza | 87 | 12 | 65 | 164 |
| Selfie | 49 | 12 | 91 | 152 |
| Funeral | 24 | 23 | 98 | 145 |
| Croissant | 88 | 8 | 41 | 137 |
| Bus driver | 30 | 20 | 50 | 100 |
| Thanksgiving | 9 | 7 | 78 | 94 |
| Physician | 24 | 6 | 35 | 65 |
| Teacher | 13 | 3 | 41 | 57 |
| Graduation | 49 | 3 | 3 | 55 |
| Nurse | 19 | 3 | 23 | 45 |
| TOTAL | 4300 | 2264 | 4104 | 10668 |

Table 8: Counts of how many image-label pairs for each label fell into each supermajority vote category based on aggregated labels from raters in Tasks 1 and 2.

| Error reason | Supermajority vote | Task 2 machine label | Percent of raters | Image |
|---|----------------------------------|----------------------|-------------------|---|
| Artistic depiction of the label | Task 1: Ambiguous Task 2: Yes | No | 78.6 | <p>Label: BIRD</p>  |
| Machine over-relied on background context | Task 1: Ambiguous Task 2: Yes | No | 85.7 | <p>Label: BIRD</p>  |
| Object is depicted out of typical context (e.g., no background) | Task 1: Yes Task 2: Yes | No | 35.7 | <p>Label: ATHLETE</p>  |

| Error reason | Supermajority vote | Task 2 machine label | Percent of raters | Image |
|---|--|----------------------|-------------------|--|
| Unexpected or atypical depiction of the label | Task 1: Ambiguous Task 2: Ambiguous | No | 71.4 | Label: CHILD  |
| Ambiguous meaning of the label (e.g. triggers different interpretation) | Task 1: Ambiguous Task 2: No | Yes | 35.7 | Label: CONSTRUCTION WORKER  |
| Visually similar shape of the label | Task 1: No Task 2: No | Yes | 85.7 | Label: MUFFIN  |



| Error reason | Supermajority vote | Task 2 machine label | Percent of raters | Image |
|------------------------------|----------------------------|----------------------|-------------------|---|
| Image has quality issue | Task 1: No Task 2: No | Yes | 64.3 | <p>Label: SELFIE</p>  |
| OTHER reason for model error | Task 1: Yes Task 2: Yes | No | 64.3 | <p>Label: SMILE</p>  |

Table 9: All error reasons from Task 2. Percent of raters indicates the percentage of Task 2 raters who indicated that the model was wrong for that particular error reason, either as the primary or secondary reason for the model error.



| | | | | | | |
|-----------|---|------------|-----------|--|-----------|------|
| | A) | | | B) | | |
| |  | | |  | | |
| | Label: BIRD Human majority: Unsure | | | Label: BIRD Human majority: Yes | | |
| | Yes % | Unsure % | No % | Yes % | Unsure % | No % |
| US raters | 67 | 17 | 17 | 92 | 0 | 8 |
| CA raters | 20 | 20 | 60 | 40 | 60 | 0 |
| IN raters | 0 | 100 | 0 | 0 | 86 | 14 |

Table 10: Examples of images where the raters in different locales respond differently when asked if the label is in the image.




| | | | | | | |
|-----------|---|-----------|------|--|-----------|------|
| | A) | | | B) | | |
| |  | | |  | | |
| | Label: THANKSGIVING Human majority: Unsure | | | Label: PHYSICIAN Human majority: Unsure | | |
| | Yes % | Unsure % | No % | Yes % | Unsure % | No % |
| US raters | 42 | 50 | 8 | 25 | 50 | 25 |
| CA raters | 40 | 60 | 0 | 20 | 60 | 20 |
| IN raters | 25 | 75 | 0 | 29 | 57 | 14 |

Table 11: Examples of images where the majority of humans indicate they are UNSURE if the label is in the image.

| Label: LIPSTICK | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|---|--|-------------------------------|--------|----|------|-----|------|---|-----|--------|----|------|-----|------|-----|--------|----|------|-----|------|
|  | <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>68.4</td> <td>0.0</td> <td>31.6</td> </tr> </table> | Yes | Unsure | No | 68.4 | 0.0 | 31.6 | <p>Machine Label: No</p> <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>71.4</td> <td>0.0</td> <td>28.6</td> </tr> </table> <p>Transformed responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>28.6</td> <td>0.0</td> <td>71.4</td> </tr> </table> | Yes | Unsure | No | 71.4 | 0.0 | 28.6 | Yes | Unsure | No | 28.6 | 0.0 | 71.4 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| | 68.4 | 0.0 | 31.6 | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 71.4 | 0.0 | 28.6 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 28.6 | 0.0 | 71.4 | | | | | | | | | | | | | | | | | | |
| Final human label: | LIPSTICK in image | LIPSTICK not in image | | | | | | | | | | | | | | | | | | |



| Label: CHOPSTICKS | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|---|--|--------------------------------|--------|----|------|------|-----|---|-----|--------|----|------|-----|------|-----|--------|----|------|-----|------|
|  | <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>84.2</td> <td>10.5</td> <td>5.3</td> </tr> </table> | Yes | Unsure | No | 84.2 | 10.5 | 5.3 | <p>Machine Label: No</p> <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>35.7</td> <td>0.0</td> <td>64.3</td> </tr> </table> <p>Transformed responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>64.3</td> <td>0.0</td> <td>35.7</td> </tr> </table> | Yes | Unsure | No | 35.7 | 0.0 | 64.3 | Yes | Unsure | No | 64.3 | 0.0 | 35.7 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| | 84.2 | 10.5 | 5.3 | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 35.7 | 0.0 | 64.3 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 64.3 | 0.0 | 35.7 | | | | | | | | | | | | | | | | | | |
| Final human label: | CHOPSTICKS in image | CHOPSTICKS ambiguous for image | | | | | | | | | | | | | | | | | | |

Table 12: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original ‘yes’ label in the Label-in-Image task. Note that score transformation is needed when the ‘machine label’ is ‘no,’ in order to align the interpretation of the human label between tasks 1 and 2.

| Label: CANOE | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|---|---|-------------------------------|--------|----|------|------|------|--|-----|--------|----|------|-----|------|-----|--------|----|------|-----|------|
|  | <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>15.8</td> <td>10.5</td> <td>73.7</td> </tr> </table> | Yes | Unsure | No | 15.8 | 10.5 | 73.7 | <p>Machine Label: Yes</p> <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>71.4</td> <td>7.1</td> <td>21.4</td> </tr> </table> <p>Transformed responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>71.4</td> <td>7.1</td> <td>21.4</td> </tr> </table> | Yes | Unsure | No | 71.4 | 7.1 | 21.4 | Yes | Unsure | No | 71.4 | 7.1 | 21.4 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| | 15.8 | 10.5 | 73.7 | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 71.4 | 7.1 | 21.4 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 71.4 | 7.1 | 21.4 | | | | | | | | | | | | | | | | | | |
| Final human label: | CANOE not in image | CANOE in image | | | | | | | | | | | | | | | | | | |


| Label: FIREFIGHTER | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|---|---|---------------------------------|--------|----|------|------|------|--|-----|--------|----|------|------|------|-----|--------|----|------|------|------|
|  | <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>15.8</td> <td>15.8</td> <td>68.4</td> </tr> </table> | Yes | Unsure | No | 15.8 | 15.8 | 68.4 | <p>Machine Label: Yes</p> <p>Participant responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>50.0</td> <td>14.2</td> <td>35.7</td> </tr> </table> <p>Transformed responses (%)</p> <table> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>50.0</td> <td>14.2</td> <td>35.7</td> </tr> </table> | Yes | Unsure | No | 50.0 | 14.2 | 35.7 | Yes | Unsure | No | 50.0 | 14.2 | 35.7 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| | 15.8 | 15.8 | 68.4 | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 50.0 | 14.2 | 35.7 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 50.0 | 14.2 | 35.7 | | | | | | | | | | | | | | | | | | |
| Final human label: | FIREFIGHTER not in image | FIREFIGHTER ambiguous for image | | | | | | | | | | | | | | | | | | |

Table 13: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original ‘no’ label in the label-in-image task. Note that score transformation is only needed when the ‘machine label’ is ‘no,’ in order to align the interpretation of the human label between tasks 1 and 2. As the machine label was ‘yes’ on both examples in this table, the transformation did not alter the labels.



| Label: BIRD | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|---|---|-------------------------------|--------|----|------|------|------|---|-----|--------|----|------|-----|-----|-----|--------|----|-------------|-----|-------------|
|  | <p>Participant responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>31.6</td> <td>63.2</td> <td>5.3</td> </tr> </table> | Yes | Unsure | No | 31.6 | 63.2 | 5.3 | <p>Machine Label: Yes</p> <p>Participant responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>92.9</td> <td>0.0</td> <td>7.1</td> </tr> </table> <p>Transformed responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>92.9</td> <td>0.0</td> <td>7.1</td> </tr> </table> | Yes | Unsure | No | 92.9 | 0.0 | 7.1 | Yes | Unsure | No | 92.9 | 0.0 | 7.1 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| 31.6 | 63.2 | 5.3 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 92.9 | 0.0 | 7.1 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 92.9 | 0.0 | 7.1 | | | | | | | | | | | | | | | | | | |
| Final human label: | BIRD ambiguous for image | CANOE in image | | | | | | | | | | | | | | | | | | |
| Label: SMILE | Task 1: Is the label in the image? | Task 2: Is the model correct? | | | | | | | | | | | | | | | | | | |
|  | <p>Participant responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>0.0</td> <td>52.6</td> <td>47.6</td> </tr> </table> | Yes | Unsure | No | 0.0 | 52.6 | 47.6 | <p>Machine Label: No</p> <p>Participant responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>92.9</td> <td>0.0</td> <td>7.1</td> </tr> </table> <p>Transformed responses (%)</p> <table border="1"> <tr> <td>Yes</td> <td>Unsure</td> <td>No</td> </tr> <tr> <td>7.1</td> <td>0.0</td> <td>92.9</td> </tr> </table> | Yes | Unsure | No | 92.9 | 0.0 | 7.1 | Yes | Unsure | No | 7.1 | 0.0 | 92.9 |
| | Yes | Unsure | No | | | | | | | | | | | | | | | | | |
| 0.0 | 52.6 | 47.6 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 92.9 | 0.0 | 7.1 | | | | | | | | | | | | | | | | | | |
| Yes | Unsure | No | | | | | | | | | | | | | | | | | | |
| 7.1 | 0.0 | 92.9 | | | | | | | | | | | | | | | | | | |
| Final human label: | SMILE ambiguous for image | SMILE not in image | | | | | | | | | | | | | | | | | | |

Table 14: Examples of images where the supermajority vote label was different between the two tasks, focusing on examples that flipped an original ‘no’ label in the label-in-image task. Note that score transformation is only needed when the ‘machine label’ is ‘no,’ in order to align the interpretation of the human label between tasks 1 and 2.

Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation

Flor Miriam Plaza-del-Arco, Debora Nozza, Dirk Hovy

MilaNLP, Bocconi University, Department of Computing Sciences, Milan, Italy
{flor.plaza, debora.nozza, dirk.hovy}@unibocconi.it

Abstract

Large Language Models (LLMs) exhibit remarkable text classification capabilities, excelling in zero- and few-shot learning (ZSL and FSL) scenarios. However, since they are trained on different datasets, performance varies widely across tasks between those models. Recent studies emphasize the importance of considering human label variation in data annotation. However, how this human label variation also applies to LLMs remains unexplored. Given this likely model specialization, we ask: *Do aggregate LLM labels improve over individual models (as for human annotators)?* We evaluate four recent instruction-tuned LLMs as “annotators” on five subjective tasks across four languages. We use ZSL and FSL setups and label aggregation from human annotation. Aggregations are indeed substantially better than any individual model, benefiting from specialization in diverse tasks or languages. Surprisingly, FSL does not surpass ZSL, as it depends on the quality of the selected examples. However, there seems to be no good information-theoretical strategy to select those. We find that no LLM method rivals even simple supervised models. We also discuss the tradeoffs in accuracy, cost, and moral/ethical considerations between LLM and human annotation.

Keywords: model annotation, model label variation, subjective tasks, label aggregation, ethics

1. Introduction

Large Language Models (LLMs) have revolutionized many aspects of Natural Language Processing (NLP). Brown et al. (2020) showed that LLMs have few-shot (FSL) and even zero-shot learning (ZSL) capabilities in text classification due to their extensive pre-training. Subsequent iterations have further improved these capabilities. Those improvements have seemingly obviated one of the most time- and labor-intensive aspects of NLP: annotating enough data to train a supervised classification model. Instead, we can use LLMs to directly predict the labels via prompting. Indeed, various papers tested this hypothesis and found good performance on various NLP tasks (Zhao et al., 2023; Su et al., 2022; Wei et al., 2022; Brown et al., 2020; Plaza-del-Arco et al., 2023). However, upon closer inspection, these claims require some caveats: different models excel at different tasks, datasets, and formulations (Gilardi et al., 2023; Törnberg, 2023).

What if the answer is not to wait for one model to rule them all, though, but to treat their variation as specializations we can exploit, similar to the disagreement among human annotators? Different annotators have different strengths (or levels of reliability), and recent work (Basile et al., 2021; Plank, 2022) has suggested using this human label variation to our advantage. We test whether the same applies to LLMs if we treat them as “annotators”.

We use four state-of-the-art open-source instruction-tuned LLMs to assess their capabilities as “annotators”: Flan-T5 (Chung et al., 2022), Flan-UL2 (Tay et al., 2023), T0 (Sanh et al., 2022)

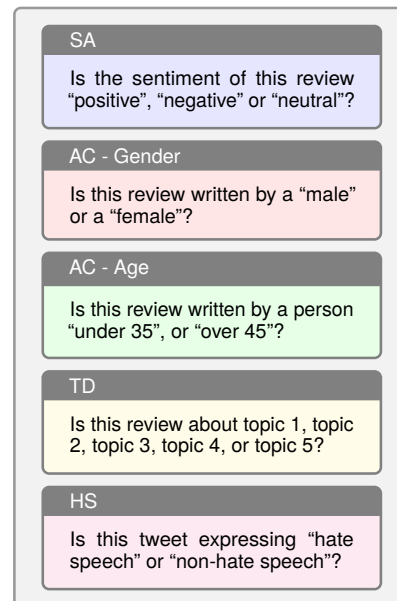


Figure 1: Instructions used to prompt the instruction-tuned LLMs for each classification task.

(alongside its multilingual variant, mT0 (Muenighoff et al., 2022)), and Tk-Instruct (Sanh et al., 2022). (We use open models to mitigate potential concerns regarding data contamination during the evaluation process and to facilitate replication.) We evaluate them across five subjective prediction tasks (age, gender, topic, sentiment prediction, and hate speech detection) in four distinct languages: English, French, German, and Spanish. We use ZSL and FSL prompt instructions, similar to the

ones we would give to human annotators. For FSL, we explore different entropy-based strategies to choose the seed examples. We then aggregate the LLM answers into a single label and evaluate their performance.

We find that different models indeed excel on some tasks or languages, but not on others. Some models even specialize on certain labels in a given task, but perform poorly on the others. **Their behavior thus mimics human expertise in wisdom-of-the-crowd settings.**

We then aggregate the model answers into a single label for each example. The simplest approach is *majority voting*: use the label that most LLMs suggested (ties are split randomly). However, the majority can still be wrong. Instead, we can use *Bayesian models of annotation* (Passonneau and Carpenter, 2014; Paun et al., 2018) to weigh the answers based on the inferred reliability of each annotator. This approach is similar to Bayesian classifier combination, but does not require gold labels to assign the scores. Instead, it is completely unsupervised. That distinction is crucial, as we want to work with unannotated data.

In most cases, the aggregated labels of either method outperform even the best individual LLM. On average, aggregated labels are 4.2 F1-points better than the average LLM. However, even the best-aggregated performance is still well below that of even simple supervised models trained on the same data, and substantially lower than Transformer-based supervised models (by over 10 F1 points on average).

In sum, aggregating several ZSL-prompted LLMs is better than using a single LLM. Surprisingly, FSL-prompting is too varied to consistently improve performance. **However, treating LLMs as annotators cannot rival using human annotators for fine-tuning or supervised learning.** We also discuss what these results mean for the role of human annotation and supervised learning in NLP, with respect to performance, but also time, cost, bias, and ethics.

Contributions (1) We explore the feasibility of four open-source instruction-tuned LLMs as “annotators” via ZSL and FSL prompting on five subjective classification tasks; (2) we compare them across four languages; (3) we analyze the robustness of two label aggregation methods to check whether we can benefit from model label variation in subjective tasks; and (4) we discuss the technical, moral, and ethical ramifications of this development for NLP and annotation.

2. Data

For our experiments, we use two datasets: **Trustpilot** (Hovy et al., 2015) and **HatEval** (Basile et al., 2019). Note that for most models, these datasets are “unseen”, i.e., the data was not part of the LLMs’ training. The one exception is HatEval in EN, which is included in Flan-T5 and Flan-UL2 models. We aim to evaluate their performance in a data contamination scenario, offering insights into models’ generalization capabilities unaffected by such contamination.

Trustpilot (Hovy et al., 2015) is a multilingual dataset with demographic user information from various countries. It uses reviews from the user review website Trustpilot. To test a variety of languages commonly found in LLMs, we select data with English from the United States, German from Germany, and French from France for our experiments. The data includes labels for sentiment, the topic of the review, and two demographic dimensions of the review authors: self-declared gender and age (these two are not available for all data points). We use the same data splits as Hovy (2015) to ensure comparability and consistency. Given our ZSL setup, we evaluate on their evaluation sets for each language, which consists of the joint development and test sets.

HatEval (Basile et al., 2019) is a multilingual dataset for HS detection against immigrants and women on Twitter, part of a SemEval 2019 shared task. The dataset contains Spanish and English tweets manually annotated via crowdsourcing. We use the benchmark test set provided by the HatEval competition for both languages.

2.1. Tasks

We evaluate the performance of LLMs as annotators on five prediction tasks: four from the Trustpilot dataset and one from the HatEval corpus. These tasks involve sentiment analysis, topic detection, and predicting demographic attributes (gender and age). These two **attribute classification (AC)** tasks are binary: the **gender** of the text author (*male* or *female*)¹ and the **age** of the text author (*under 35* or *above 45* years old). In the **sentiment analysis (SA)** task, reviews are classified into *negative*, *neutral*, and *positive* sentiments based on 1, 3, and 5-star ratings, respectively. The **topic detection (TD)** task uses the review categories of the texts to classify them into one of five topics. For this task, the exact topics vary across languages.

¹The data does not allow a more fine-grained classification of gender identities, as the original website only provided users with those two options. See Ethical Considerations for more discussion.

EN: *Car lights, fashion accessories, pets, domestic appliances, and hotels.* **DE:** *Wine, car rental, drugs and pharmacy, flowers, and hotels.* **FR:** *Clothes and fashion, fashion accessories, pets, computer and accessories, and food and beverage.*

For the **hate speech detection (HS)**, the task is to classify a tweet as either hate speech or non-hate speech.

3. Models

We experiment with four state-of-the-art instruction-tuned LLMs from the same model family, the T5 with an encoder-decoder (Raffel et al., 2020) architecture. We specifically select these models because they were fine-tuned on a diverse range of instructions. They use intuitive explanations of the downstream task to respond to natural language prompts, similar to the instructions provided to human annotators. Furthermore, these models are all open-source, letting us inspect the training data and examine data contamination. Our selection represents a realistic LLM annotator pool for a current NLP practitioner. As models evolve rapidly, though, this selection is likely to change. However, the results from using a diverse pool of LLM annotators should hold regardless.

In particular, we use the following models:

- **Flan-T5** (Chung et al., 2022) is a sequence-to-sequence transformer model built on the T5 architecture (Raffel et al., 2020). The model has been pre-trained with standard language modeling objectives and subsequent fine-tuning on the extensive FLAN collection (Longpre et al., 2023). The FLAN collection contains more than 1,800 NLP tasks in over 60 languages. We use the largest version² of this model.
- **Flan-UL2** (Tay et al., 2023) is the Flan version of the T5 and UL2 model. It has a similar architecture to T5, but with an upgraded pre-training procedure known as UL2³.
- **T0** (Sanh et al., 2022) and the multilingual **mT0** (Muennighoff et al., 2023). T0⁴ is an encoder-decoder model based on T5 that is trained on a multi-task mixture of NLP datasets over different tasks. For the non-English languages, we use mT0⁵ since T0 has been trained on English texts. mT0 is based on Google’s mT5 (Xue et al., 2021) and has been fine-tuned on xP3⁶, which covers 13 training tasks across

²<https://huggingface.co/google/flan-t5-xxl>

³<https://huggingface.co/google/flan-ul2>

⁴<https://huggingface.co/bigscience/T0>

⁵<https://huggingface.co/bigscience/mt0-xxl>

⁶<https://huggingface.co/datasets/bigscience/xP3>

46 languages with English prompts.

- **Tk-Instruct** (Wang et al., 2022) is a generative model for transforming task inputs given declarative in-context instructions, like “*Given an utterance and the past 3 utterances, output ‘Yes’ if the utterance contains the small-talk strategy, otherwise output ‘No’. Small-talk is a cooperative negotiation strategy...*” (adapted from Wang et al., 2022). It is also based on T5 but trained on all task instructions in a multi-task setup. It is fine-tuned on the SUPER-NATURALINSTRUCTIONS dataset (Triantafillou et al., 2020), a large benchmark of 1,616 NLP tasks and their natural language instructions. It covers 76 task types across 55 different languages⁷.

Computing Infrastructure We run all experiments on a server with three NVIDIA RTX A6000 and 48GB of RAM.

3.1. Prompting

A prompt is an input that directs an LLM’s text generation, ranging from a single sentence to a paragraph. It guides the model’s comprehension and influences its output. Figure 1 depicts the task formulations (prompt instructions) we give to the LLMs, who act as our annotators, for every considered text classification task. We add “Answer” to mark the output field after the instruction to improve the LLMs’ understanding and output format. For the TD tasks, the list of five topics varies by language. For instance, the prompt for the English TD task is: “I love the earrings I bought,” “Is this review about ‘car lights,’ ‘fashion accessories,’ ‘pets,’ ‘domestic appliances,’ or ‘hotels?’” <Answer>: {LM answer}.

Tk-Instruct requires a prompt template with specific fields: “definition,” “input,” and “output.” The “definition” is used to specify the instruction or guidance, the “input” contains the instance to be classified, and the “output” is the output indicator. For instance, the prompt for the HS task is the following: <Definition> Is this tweet expressing “hate speech” or “non-hate speech?” <Input> “I hate you” <Response>: {LM answer}.

We use task-specific prompts to assess the model’s performance on the resulting outputs for zero- and few-shot prediction. We used default parameters for the models.

If the output does not correspond to a valid class, we assign the most common class for that task. For instance, these out-of-label (OOL) predictions vary between tasks and models for the ZSL setup. Binary or ternary classification tasks (AC, SA, HS)

⁷<https://huggingface.co/allenai/tk-instruct-3b-def>

exhibit a very low OOL percentage (<1%). In contrast, TD shows a significantly higher percentage (13%) due to the larger number of classes and their more descriptive nature (e.g., “fashion accessories”). At the model level, Flan models have a very low OOL percentage (1%), T0 and Tk-Instruct have a low OOL percentage (~2%).

3.2. Baselines

We compare the LLMs across our five tasks to two baselines: *the most frequent class* and *random choice*.

The *most frequent class* baseline does not require any model. It always picks the most frequent label for a task as final prediction. This standard baseline method is very strong in unbalanced datasets. However, it requires knowledge of the label distribution. The *random-choice* method randomly picks a label from the set of labels for a task. It represents a lower bound.

3.3. Aggregation of Labels

For each example, we get four labels: one from each LLM annotator. We use two different methods to aggregate these four labels into a single label: *majority voting*, and a *Bayesian model of annotation*, Multi-Annotator Competence Estimation (MACE, Hovy et al., 2013). These methods use different aggregation methods. Both are common in the literature (Klie et al., 2023).

Majority-voting selects the label returned most by the four models. In case of a tie, it randomly chooses among the top candidates. This approach is common in many annotation projects, but has the drawback that the majority can still be wrong.

MACE is a Bayesian annotation tool that computes two scores: the competence (reliability) of each annotator (i.e., the probability an annotator chooses the “true” label based on their expertise instead of guessing one) and the most likely label. MACE uses variational Bayesian inference to infer both variables, and works on unlabeled data. The aggregated MACE labels are usually more accurate downstream than majority voting, and competence scores correlate strongly with actual annotator proficiency (Paun et al., 2018). Competence scores tend to correlate with annotators’ actual expertise (Hovy et al., 2013), and can therefore be used to directly compare annotator quality in the absence of gold labels. As a probabilistic model, it also computes the entropy of each example, including both annotator competence and agreement. It is therefore a proxy for how “difficult” an example is to label. We use MACE to get the competence of each LLM and the entropy of each example, which we use to select seed examples for FSL.

| Task/Language | | Cohen | Fleiss | Krip. | Raw |
|---------------|----|-------|--------|--------|-------|
| SA | EN | 0.708 | 0.705 | 0.703 | 0.837 |
| | DE | 0.636 | 0.633 | 0.630 | 0.792 |
| | FR | 0.665 | 0.662 | 0.660 | 0.809 |
| AC-Gender | EN | 0.299 | 0.279 | 0.229 | 0.615 |
| | DE | 0.271 | 0.136 | -0.007 | 0.566 |
| | FR | 0.236 | 0.227 | 0.179 | 0.596 |
| AC-Age | EN | 0.101 | 0.044 | -0.154 | 0.428 |
| | DE | 0.068 | 0.040 | -0.124 | 0.596 |
| | FR | 0.099 | 0.093 | 0.014 | 0.679 |
| TD | EN | 0.510 | 0.495 | 0.477 | 0.622 |
| | DE | 0.586 | 0.578 | 0.571 | 0.712 |
| | FR | 0.316 | 0.305 | 0.283 | 0.598 |
| HS | EN | 0.222 | 0.220 | 0.209 | 0.605 |
| | ES | 0.155 | 0.099 | -0.019 | 0.629 |

Table 1: Inter-model agreement scores.

4. Results

We compare the four models as annotators along several dimensions:

How much do models agree with each other?

This assesses the consensus among them and indicates specialization. **How reliable is each model?** This evaluates the consistency and trustworthiness of individual model predictions, key for label aggregation. **How accurate are the predictions of the individual models versus their aggregations?** This last question assesses the prediction quality of individual LLMs vis-a-vis aggregations to determine whether this approach is a viable alternative to supervised learning.

We first report ZSL results and then discuss the FSL setting separately (Section 4.4).

4.1. Inter-model Agreement

To assess the level of specialization among the LLMs as annotators, we evaluate their agreement. We use four common inter-annotator-agreement metrics: Cohen’s κ (Cohen, 1960), Fleiss’ κ (Fleiss, 1971), and Krippendorff’s α (Krippendorff, 2011) (which all correct the observed agreement for expected agreement), as well as the unweighted raw agreement (i.e., the uncorrected level of agreement between LLMs). The results are shown in Table 1. Note that the number of labels does not factor into agreement, and that raw agreement is usually higher than chance-corrected inter-annotator agreement measures.

The results show a wide range of agreement values, but a few takeaways become apparent:

1) **The scores suggest that the different models specialize on different tasks and labels.** As we will see in the performance and reliability analysis, some models perform better on some tasks

than others. Model specialization suggests that aggregation is likely beneficial (as the aggregation hopefully benefits from differing expertise).

2) **Language does not factor into the differences.** The models we test are all multi-lingual, and the languages we test are generally high-resource. The agreement difference between the different languages on the same task is negligible.

3) **Some tasks show higher agreement than others:** SA has higher scores than TD, and the others have little to no agreement. However, we do not know whether high-agreement tasks are inherently easier, or whether the models are all wrong in the same direction.

| Task/Language | T0 | Flan-T5 | Flan-UL2 | Tk-Instruct | |
|---------------|----|--------------|--------------|--------------|-------|
| SA | EN | 0.755 | 0.958 | 0.856 | 0.803 |
| | DE | 0.724 | 0.928 | 0.767 | 0.757 |
| | FR | 0.759 | 0.909 | 0.796 | 0.789 |
| AC-Gender | EN | 0.317 | 0.613 | 0.719 | 0.445 |
| | DE | 0.152 | 0.267 | 0.359 | 0.270 |
| | FR | 0.094 | 0.699 | 0.599 | 0.601 |
| AC-Age | EN | 0.255 | 0.146 | 0.325 | 0.083 |
| | DE | 0.295 | 0.039 | 0.418 | 0.014 |
| | FR | 0.432 | 0.061 | 0.608 | 0.017 |
| TD | EN | 0.388 | 0.737 | 0.933 | 0.511 |
| | DE | 0.660 | 0.793 | 0.735 | 0.712 |
| | FR | 0.556 | 0.327 | 0.492 | 0.210 |
| Mean | EN | 0.429 | 0.614 | 0.708 | 0.461 |
| | DE | 0.458 | 0.507 | 0.569 | 0.438 |
| | FR | 0.460 | 0.499 | 0.624 | 0.404 |
| HS | EN | 0.358 | 0.919 | 0.327 | 0.402 |
| | ES | 0.466 | 0.212 | 0.131 | 0.099 |

Table 2: MACE competence scores of each LLM across tasks and languages on the Trustpilot and HatEval datasets. For non-English languages, we use the multilingual mT0 model.

4.2. Reliability

When aggregating specialized annotators, we might want to trust more specialized ones more. We use the competence scores from MACE to assess the reliability of each model. Table 2 shows the competence scores.

The competence scores support the specialization hypothesis for the different models on different languages and tasks. **No model is dominant in all settings, though the Flan models tend to have higher competence scores than the other models** (reflected in their higher mean competence scores).

4.3. Model Performance and Robustness of Label Aggregation

Ultimately, we care about the predictive performance of the annotator method. Table 3 shows the macro-F1 scores of the LLMs on all tasks and languages. We compute the statistical difference of the individual results over the random-choice baseline, using a bootstrap sampling test with the *bootsa*⁸ Python package. We use 1,000 bootstrap samples, a sample size of 20%, and $p \leq 0.01$.

Most models clearly and significantly outperform the random-choice and even most-frequent-label baselines. Note though that Flan-T5 and Flan-UL2 included the HatEval dataset in their training. Consequently, they perform substantially better than the other models (with Flan-T5 receiving a very high competence score from MACE).

Aggregation When aggregating annotations into a single label, we implicitly assume that a) there is a single correct answer and b) the wisdom of the crowd will find it. The first assumption is up for debate (Basile et al., 2021), but the latter is clearly borne out by the results here. **On average and in most individual cases, majority voting and MACE aggregation predictions are better than most models.** In 6 out of 14 tasks, MACE was the best model. Note that for SA, Tk-Instruct performs better than the aggregation methods in all languages. For AC-Gender in English, Flan-UL2 is better, and in German, no method outperforms random choice (though MACE aggregation is close).

Overall, the two aggregation methods are substantially more robust than any one individual model across all languages and datasets (see the Mean results in Table 3). Presumably, they suffer less from the variance across tasks and languages and instead are able to exploit the specialization of each model as a source of information. The MACE competence score correlates with the actual performance of the models: 0.93 Spearman ρ and 0.83 Pearson ρ . This correlation suggests that MACE identified the model specializations correctly. A custom weighting of each model’s prediction (for example, based on the actual performance) might perform even better. In practice, though, this weighting would of course be unknown.

Comparison to supervised learning ZSL holds a lot of promise for quick predictions, but to assess its worth, we need to compare it to supervised models based on human annotation. For the Trustpilot data, we compare our best ZSL result for each task and language (see Table 3) to two supervised models. 1) a simple Logistic Regression model (the baseline “agnostic” results reported in Hovy, 2015)

⁸<https://github.com/fornaciari/bootsa>

| Task/Lang. | | Models | | | | Baselines | | Aggregate | |
|------------|----|---------------|---------|---------------|---------------|-----------|--------------|-----------|---------------|
| | | T0 | Flan-T5 | Flan-UL2 | Tk-Instruct | Most Freq | Random | Majority | MACE |
| SA | EN | 0.453* | 0.532* | 0.482* | 0.553* | 0.167 | 0.334 | 0.503* | 0.514* |
| | DE | 0.469* | 0.495* | 0.433* | 0.517* | 0.167 | 0.331 | 0.480* | 0.484* |
| | FR | 0.460* | 0.518* | 0.445* | 0.528* | 0.167 | 0.337 | 0.486* | 0.490* |
| AC-Gender | EN | 0.516 | 0.594* | 0.624* | 0.541* | 0.337 | 0.501 | 0.617* | 0.623* |
| | DE | 0.456 | 0.437 | 0.447 | 0.431 | 0.334 | 0.497 | 0.458 | 0.485 |
| | FR | 0.428 | 0.573* | 0.566* | 0.563* | 0.335 | 0.503 | 0.577* | 0.579* |
| AC-Age | EN | 0.495 | 0.442 | 0.516* | 0.397 | 0.336 | 0.497 | 0.569* | 0.572* |
| | DE | 0.458 | 0.366 | 0.503 | 0.344 | 0.334 | 0.500 | 0.422 | 0.499 |
| | FR | 0.497 | 0.375 | 0.550* | 0.343 | 0.335 | 0.500 | 0.443 | 0.542* |
| TD | EN | 0.558* | 0.579* | 0.588* | 0.567* | 0.085 | 0.195 | 0.588* | 0.596* |
| | DE | 0.506* | 0.514* | 0.513* | 0.493* | 0.105 | 0.193 | 0.516* | 0.520* |
| | FR | 0.314* | 0.271* | 0.264* | 0.257* | 0.096 | 0.193 | 0.281* | 0.293* |
| Mean | EN | 0.506 | 0.537 | 0.553 | 0.515 | 0.231 | 0.382 | 0.569 | 0.576 |
| | DE | 0.472 | 0.453 | 0.474 | 0.446 | 0.235 | 0.380 | 0.469 | 0.497 |
| | FR | 0.425 | 0.434 | 0.456 | 0.423 | 0.233 | 0.383 | 0.447 | 0.476 |
| HS | EN | 0.621* | 0.726* | 0.670* | 0.579* | 0.367 | 0.490 | 0.717* | 0.726* |
| | ES | 0.601* | 0.532* | 0.519 | 0.449 | 0.370 | 0.492 | 0.533* | 0.603* |

Table 3: Zero-shot Macro-F1 results obtained by the LLMs on the Trustpilot and HatEval tasks, the baselines and the aggregation methods. Best result per language and task is shown in bold. Significant improvement over Random baseline ($*$: $p \leq 0.01$) with bootstrap sampling. For non-English languages, we use the multilingual mT0 model.

and a recent Transformer-based model (the best results from Hung et al., 2023). Similarly, for HatEval, we compare with 1) a simple linear Support Vector Machine based on a TF-IDF representation (the baseline results reported in Basile et al., 2019) and a fine-tuned multilingual Transformer model (Nozza, 2021). Table 4 shows the results. The two methods approximate an upper and lower bound on supervised learning for these datasets.

| Task/Language | | best | supervised | |
|---------------|----|--------------|-------------|--------------|
| | | ZSL | Standard ML | Transformer |
| SA | EN | 0.553 | 0.610 | 0.680 |
| | DE | 0.517 | 0.610 | 0.677 |
| | FR | 0.528 | 0.612 | 0.706 |
| AC-Gender | EN | 0.624 | 0.601 | 0.638 |
| | DE | 0.497 | 0.540 | 0.629 |
| | FR | 0.579 | 0.546 | 0.650 |
| AC-Age | EN | 0.572 | 0.620 | 0.636 |
| | DE | 0.503 | 0.602 | 0.611 |
| | FR | 0.550 | 0.540 | 0.568 |
| TD | EN | 0.596 | 0.656 | 0.705 |
| | DE | 0.520 | 0.605 | 0.671 |
| | FR | 0.314 | 0.385 | 0.444 |
| HS | EN | 0.726 | 0.451 | 0.416 |
| | ES | 0.603 | 0.701 | 0.752 |

Table 4: Macro-F1 results for best ZSL model (Table 3), compared to previous supervised results on the same datasets.

Except for 4 cases (AC-Gender and AC-Age in French, AC-Gender in English, HS in English), even the simple ML models beat the best ZSL result we achieved, be it from an LLM or aggregation method. Compared to the upper bounds from Hung et al. (2023), we see an average performance gap of 10.5 F1 points. Only for HS in English ZSL achieves better results, likely attributed to the data contamination found in the Flan models.

These results show that while ZSL might be a fast approximation for prediction tasks, it is still far from competitive with supervised learning.

4.4. Few-shot Learning

FSL has the potential to perform better than ZSL, so we ask whether using FSL models as annotators improves over our ZSL experiments. We investigate whether providing a limited set of examples enhances annotation capabilities, similar to instructing human annotators. The short answer is no. We apply this method to the English Trustpilot dataset.

To choose the seed examples, we compare three methods. Random selection and selection based on the MACE entropy scores⁹. Entropy lets us identify two groups of examples: (1) maximum entropy indicates models were less confident or disagreed more, indicating higher difficulty, and

⁹<https://github.com/dirkhovoy/MACE>

(2) low entropy indicates models were more confident or agreed more, indicating lower difficulty. For each task and label, we randomly choose 4,000 instances¹⁰ and use MACE to compute the entropy for each instance. From the initial pool of 4,000, we sample three exemplars per class based on the method (low entropy, max entropy) and use these as few-shot seeds, prepending them to the prompt. We compare these results to the random baseline.

Figure 2 shows a comparison between the ZSL and FSL approaches. **Our analysis reveals that there are no statistically significant differences between the two prompting methods.** In general, though, FSL does not perform as well as ZSL across subjective tasks. Within FSL, a prominent pattern emerges: it exhibits notably higher variance across tasks than ZSL. Presumably, exemplar quality heavily influences performance.

Regarding the two entropy-based selection methods, our results show no consistent trends between them. The choice is somewhat task-dependent: Max entropy seems to perform well for SA and AC-Gender tasks, while low entropy works best for AC-Age and TD tasks. The random strategy is less consistent across tasks. This discrepancy further underlines the inherent challenge in selecting ‘good’ exemplars for FSL. Our findings suggest that using no exemplars (ZSL) is generally more stable and consistent for aggregation.

5. Related Work

Generating human-annotated data is time-consuming and costly, especially for complex or specialized tasks with limited available data. Instead, a possible solution is leveraging automatic annotation models, often using a small subset of labeled data (Smit et al., 2020; Rosenthal et al., 2021), known as ‘weak supervision’ (Stanford AI Lab Blog, 2019). Supervised learning has emerged as the dominant method, driven by the widespread adoption of traditional machine learning models and Transformer-based models like BERT (Devlin et al., 2019).

More recently, LLMs have shown zero-shot and few-shot learning capabilities (Brown et al., 2020). Researchers have further advanced these models with natural language instructions (Chung et al., 2022; Wang et al., 2022; Taori et al., 2023), enabling innovative techniques like prompting (Liu et al., 2023) without the need to train a supervised model. Several papers explored these new techniques with promising performance on various NLP tasks (Brown et al., 2020; Plaza-del-Arco et al., 2022; Su et al., 2022; Sottana et al., 2023). Recent work has focused on exploring the capabilities

¹⁰We use 4,000 instances to make things as comparable as possible.

of LLMs as annotators. For instance, Lee et al. (2023) evaluate the performance and alignment between LLMs and humans, revealing that these models not only fall short in performing natural language inference tasks compared to humans but also struggle to capture the distribution of human disagreements accurately. Other studies have used ChatGPT as an annotator. Some report excellent capabilities (Huang et al., 2023; Gilardi et al., 2023; Törnberg, 2023; He et al., 2024), but Kuzman et al. (2023) found that ChatGPT’s performance notably drops for less-resourced languages. Similarly, Kristensen-McLachlan et al. (2023) show that on two seemingly simple binary classification tasks, the performance of ChatGPT and open-source LLMs varies significantly and often unpredictably, and supervised models systematically outperform both types of models.

For annotation, it remains to be seen whether different instruction-tuned LLMs can generalize to *any* subjective text classification tasks in different languages, especially if these tasks and languages are not well-represented in the training data. Recent studies have shown the importance of considering human label variation (Basile et al., 2021; Plank, 2022), i.e., the disagreement between human annotators, as a source of information rather than a problem. However, how this human label variation also applies to LLMs remains unexplored.

6. Discussion

Our results indicate that treating LLMs as annotators and aggregating their responses is cost-effective and quick. However, we also find that the overall performance is still well below that of even simple supervised models.

Human annotation still has a vital role if we focus on performance. As LLMs become more capable, this edge might diminish to the point where LLM annotation is equivalent to human annotation. As an aside, although all models are likely to improve across the board, we still expect specialization effects, meaning aggregation approaches will likely stay relevant for the foreseeable future.

But what about bias? Human label variation is not only due to different levels of expertise or diligence (Snow et al., 2008). It can also vary due to differing opinions, definitions, and biases (Shah et al., 2020). Specific tasks are subjective by nature (Basile et al., 2019; Rottger et al., 2022), but even seemingly objective tasks like part-of-speech tagging can have different interpretations (Plank et al., 2014). **The current discussion around the moral and ethical alignment of LLMs (Liu et al., 2022) should make us cautious about using these models as annotators in subjective or sensitive tasks.** Aggregation can overcome

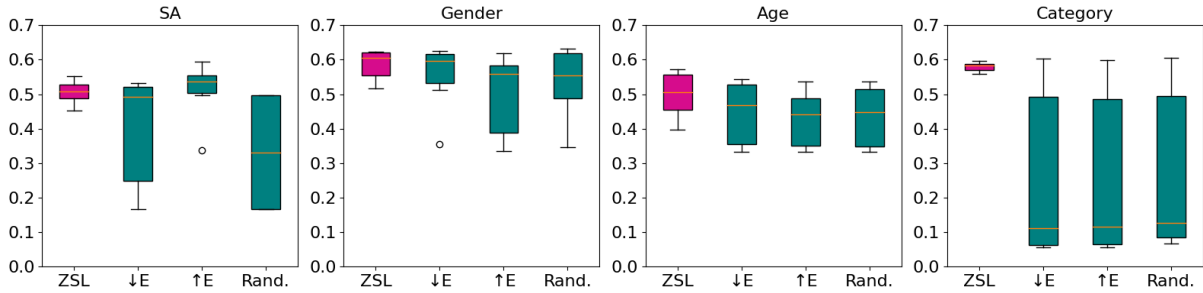


Figure 2: ZSL vs. FSL Macro-F1 scores on English Trustpilot tasks. FSL sample selection strategies: Low Entropy ($\downarrow E$), Max Entropy ($\uparrow E$), and Random (Rand.). All FSL methods show much greater variance than ZSL.

the biases of any one particular model, but it cannot safeguard against widespread biases. Lastly, annotation might be exploratory (the “descriptive” paradigm in Rottger et al., 2022), where the goal is to find the range of human responses.

However, replacing human annotators with LLMs has ramifications beyond performance and bias issues. While crowdsourced annotations can be problematic regarding worker exploitation (Fort et al., 2011), they often provide low to moderate-income earners with a way to supplement their living. Replacing this option with LLMs is a real-life example of automation making human jobs obsolete. Conversely, it may mitigate the mental health risks associated with annotating toxic or sensitive content, such as racist content or tasks related to mental disorders. **Hybrid human and LLM annotation might offer a way forward here.**

7. Conclusion

We use zero- and few-shot prompting to compare four current instruction-tuned LLMs as annotators on five subjective tasks in four languages. We find specialization across models and tasks. We leverage this variance similarly to human label variation by aggregating their predictions into a single label. This approach is, on average, substantially better than any individual model. This suggests that label aggregation consistently enhances performance compared to relying on a single LLM. Despite the rapid development of LLMs enhancing generalization to new tasks, aggregation remains a beneficial strategy.

Our findings suggest that practitioners aiming to label large amounts at minimal cost (both financially and time-wise) can benefit from the outlined aggregation approach. However, we also find that even the best models cannot compete with “traditional” supervised classification approaches. Furthermore, human annotation allows practitioners to encode a specific view or approach in a prescriptive manner or explore the range of responses descrip-

tively (Rottger et al., 2022). Relying on LLMs while alignment and bias still need to be solved (Mökander et al., 2023) makes this approach unsuitable for sensitive applications.

Limitations

We were unable to compare to closed LLMs like GPT-4. While they are often state-of-the-art on many tasks, their training and setup change frequently and are, therefore, not replicable. Their pay-by-use nature also makes them less affordable for many practitioners than free open models. We do suspect, though, that including closed or generally better models will not change the overall conclusions of this paper.

Ethical Considerations

The data we use for AC-gender classification only makes a binary distinction (the Trustpilot website allowed users only to choose from two options). We do not assume this to be representative of gender identities and only use this data to test our hypotheses.

The languages we evaluate all come from the Indo-European branch of languages. The selection was due to data availability and our knowledge of languages. While we do not expect results to systematically differ from other languages, we do note that this is conjecture.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). The research was made possible in part through an unrestricted Google research gift to explore variance in annotation. Flor Miriam Plaza-del-Arco, Debora Nozza, and Dirk Hovy are members

of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis.

8. Bibliographical References

- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *arXiv preprint arXiv:2210.11416*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. [Last words: Amazon Mechanical Turk: Gold mine or coal mine?](#) *Computational Linguistics*, 37(2):413–420.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowdworkers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.
- Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao'Kenneth' Huang. 2024. If in a crowdsourced data annotation pipeline, a gpt-4. *arXiv preprint arXiv:2402.16795*.
- Dirk Hovy. 2015. [Demographic factors improve classification performance](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.
- 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.
- Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. [User review sites as a resource for large-scale sociolinguistic studies](#). In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, page 452–461. International World Wide Web Conferences Steering Committee.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. [Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech](#). In *Companion Proceedings of the ACM Web Conference 2023*. ACM.

- Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. [Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers.](#) In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.
- Jan-Christoph Klie, Ji-Ung Lee, Kevin Stowe, Gözde Şahin, Nafise Sadat Moosavi, Luke Bates, Dominic Petrak, Richard Eckart De Castilho, and Iryna Gurevych. 2023. [Lessons learned from a citizen science project for natural language processing.](#) In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3594–3608, Dubrovnik, Croatia. Association for Computational Linguistics.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. *Computing*, 1:25–2011.
- Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kardos, Mia Jacobsen, and Lene Aarøe. 2023. [Chatbots are not reliable text annotators.](#)
- Taja Kuzman, Igor Mozetič, and Nikola Ljubešić. 2023. [ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification.](#) *arXiv preprint arXiv:2303.03953*.
- Noah Lee, Na Min An, and James Thorne. 2023. [Can large language models infer and disagree like humans?](#) *arXiv preprint arXiv:2305.13788*.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.](#) *ACM Computing Surveys*, 55(9):1–35.
- Ruibao Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. [Aligning generative language models with human values.](#) In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. [The flan collection: Designing data and methods for effective instruction tuning.](#) *arXiv preprint arXiv:2301.13688*.
- Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. [Auditing large language models: a three-layered approach.](#) *AI and Ethics*, pages 1–31.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark.](#) In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. [Crosslingual generalization through multi-task finetuning.](#) *arXiv preprint arXiv:2211.01786*.
- Debora Nozza. 2021. [Exposing the limits of zero-shot cross-lingual hate speech detection.](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.
- Rebecca J. Passonneau and Bob Carpenter. 2014. [The benefits of a model of annotation.](#) *Transactions of the Association for Computational Linguistics*, 2:311–326.
- Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. [Comparing Bayesian models of annotation.](#) *Transactions of the Association for Computational Linguistics*, 6:571–585.
- 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, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. [Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora.](#) In *Proceedings of the 29th International Conference*

- on *Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Flor Miriam Plaza-del-Arco, Debora Nozza, and Dirk Hovy. 2023. [Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *The Journal of Machine Learning Research*, 21(1).
- Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. [SOLID: A large-scale semi-supervised dataset for offensive language identification](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 915–928, Online. Association for Computational Linguistics.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. [Multitask prompted training enables zero-shot task generalization](#). *arXiv preprint arXiv:2110.08207*.
- Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. [Predictive biases in natural language processing models: A conceptual framework and overview](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. [Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. [Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.
- Stanford AI Lab Blog. 2019. [Weak supervision](http://ai.stanford.edu/blog/weak-supervision/). <http://ai.stanford.edu/blog/weak-supervision/>.
- Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Tao Yu. 2022. [Selective annotation makes language models better few-shot learners](#). *arXiv preprint arXiv:2209.01975*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [Stanford Alpaca: An Instruction-following LLaMA model](https://github.com/tatsu-lab/stanford_alpaca). https://github.com/tatsu-lab/stanford_alpaca.
- Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. [UI2: Unifying language learning paradigms](#). *arXiv preprint arXiv:2205.05131*.
- Petter Törnberg. 2023. [ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning](#). *arXiv preprint arXiv:2304.06588*.

Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. 2020. [Meta-dataset: A dataset of datasets for learning to learn from few examples](#). In *International Conference on Learning Representations*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoor-molabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. [Emergent abilities of large language models](#). *arXiv preprint arXiv:2206.07682*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. [A survey of large language models](#). *arXiv preprint arXiv:2303.18223*.

Revisiting Annotation of Online Gender-Based Violence

Gavin Abercrombie[♡], Nikolas Vitsakis[♡], Aiqi Jiang^{♡♣}, Ioannis Konstas^{♡♠}

[♡]Interaction Lab, Heriot-Watt University, Edinburgh

[♣]Computational Linguistics Lab, Queen Mary University of London

[♠]Alana AI

{g.abercrombie, nv2006, a.jiang, i.konstas}@hw.ac.uk

Abstract

Online Gender-Based Violence (GBV) is an increasing problem, but existing datasets fail to capture the plurality of possible annotator perspectives or ensure representation of affected groups. In a pilot study, we revisit the annotation of a widely used dataset to investigate the relationship between annotator identities and underlying attitudes and the responses they give to a sexism labelling task. We collect demographic and attitudinal information about crowd-sourced annotators using two validated surveys from Social Psychology. While we do not find any correlation between underlying attitudes and annotation behaviour, ethnicity does appear to be related to annotator responses for this pool of crowd-workers. We also conduct initial classification experiments using Large Language Models, finding that a state-of-the-art model trained with human feedback benefits from our broad data collection to perform better on the new labels. This study represents the initial stages of a wider data collection project, in which we aim to develop a taxonomy of GBV in partnership with affected stakeholders.

Keywords: Gender-Based Violence, Misogyny, Sexism, Abusive language, Hate speech, Annotation, LLMs

1. Introduction

Gender-Based Violence (GBV) is an increasing problem in online spaces, affecting around half of all women and targeting those from marginalised groups in particular (Glitch UK and ERAW, 2020).

To counter this, there have been attempts to facilitate moderation of such content using natural language processing (NLP) methods to automatically identify misogynistic language. As a result, there now exist a number of datasets designed for supervised classification of various forms of GBV.

However, Abercrombie et al. (2023) identified a number of weaknesses in approaches to the creation of corpora for this task. One prominent shortcoming has been the lack of representation in the labelled data of people’s different points of view, and particularly of people with the minoritised identities who are best placed to recognise GBV.

To fill this gap, we aim to revisit the task of annotating online text following *strongly perspectivist* data practices (Abercrombie et al., 2022; Basile et al., 2023; Cabitza et al., 2023) in the collection, modeling, and distribution of datasets, preserving the labels provided by multiple annotators. In this pilot study, we re-annotate a recently collected dataset, this time with (1) multiple ratings per item; and (2) demographic and attitudinal information about the annotators.

We make the following **research contributions**: (1) we collect a corpus of the responses of multiple annotators to each item in a subset of a widely used English language GBV dataset, along with demographic and attitudinal information about the

annotators. We make this resource available to the research community at <https://github.com/GavinAbercrombie/EquallySafeOnline>.

(2) We analyse this data to investigate the relationship between annotator demographics and attitudes and the labels that they apply to items. (3) We conduct benchmark experiments to investigate the capabilities of current state-of-the-art systems in identifying GBV in text.

2. Background

The GBV framework encompasses phenomena such as sexism, misogyny, and violence against women and girls—although it also recognises that people of all genders are affected by GBV.¹ It was first introduced by the United Nations (UN General Assembly, 1993; United Nations, 2021). For further details of the theoretical foundation of this framework and motivation for its application to the field of NLP, see Abercrombie et al. (2023).

Annotator Variability and Perspectivist Data Practices While labels collected for supervised classification have traditionally been aggregated to a single ‘gold’ or ‘ground truth’ label for each item, recent work has recognised that this can lead to the erasure of minoritised voices, and can subsequently hinder the ability of classifiers to recognise subtle and implicit forms of abuse. *Standpoint theory* (Harding, 1991) contends that only people with

¹For example, men face pressure to conform to masculine gender role norms (European Institute for Gender Equality, 2021).

relevant lived experience are capable of recognising subtle, implicit abuse such as stereotypes and micro-aggressions. According to the *matrix of domination* Collins (2002), this experience likely results from sharing intersectional social categorisations with the intended targets of the abuse. With label aggregation, the labels provided by people with such identities and experiences are often erased.

There is now a growing recognition of the need to collect, retain, and distribute labels provided by multiple annotators, and this has been adopted across a range of NLP tasks (Plank, 2022). This is particularly so for controversial tasks such as identification of abusive or toxic language, in which annotator variation may be caused by differences of opinion or ideology (e.g. Akhtar et al., 2021; Almanea and Poesio, 2022; Cercas Curry et al., 2021; Leonardelli et al., 2021). *Strong Perspectivism* aims to preserve this variation through modelling, classification, and evaluation (Cabitza et al., 2023). For further background, see the Perspectivist Data Manifesto at <https://pdai.info/>.

Beliefs and attitudes We ground our theoretical approach in the Dual Process Motivational Model of Ideology and Prejudice (Duckitt and Sibley, 2009; Duckitt, 2001), specifically, the differential effect hypothesis aspect of the model. This hypothesis explains that sociopolitical and ideological attitudes linked to prejudice can be adequately captured by two distinct but often related constructs, Right Wing Authoritarianism (RWA) and Social Dominance Orientation (SDO) related attitudes. The former explains propensity towards cultural conservatism and traditionalism related beliefs (Altemeyer, 1983; Feather and McKee, 2012; Van Assche et al., 2019), while the latter explains favourable views towards social hierarchies of power, where inequality between groups is seen as inevitable or even natural (Christopher and Wojda, 2008; Pratto et al., 1994; Jagayat and Choma, 2021).

Both of these constructs have been extensively assessed and found to be strongly related and to explain different forms of sexism and gender based discrimination. RWA has been found to be a good predictor of ‘benevolent sexism’, that is attitudes that force women into traditional predefined roles (i.e., being a mother) that seem subjectively advantageous but are, in reality, marginalising and disempowering (De Geus et al., 2022). SDO pertains towards beliefs towards deterministic gender imbalances justifies male dominance through a disparaging characterisation of women (La Macchia and Radke, 2020; De Geus et al., 2022).

Taken as a whole, these constructs have been widely used to explain gender based discrimination, through both offline (Perez-Arche and Miller, 2021; Christopher and Wojda, 2008; Patev et al., 2019)

and online (Jagayat and Choma, 2021) contexts, and have been validated across cultures (Çetiner and Van Assche, 2021; De Geus et al., 2022), while also being used to explain that such beliefs transcend demographic identities (Renström, 2023).

3. Related Work

Annotator Characteristics A number of NLP studies have attempted to group annotators according to their demographic characteristics and use these factors as predictors of their responses to items (e.g. Akhtar et al., 2021; Gordon et al., 2022; Goyal et al., 2022). However, it has repeatedly been shown that demographic characteristics do not predict annotator behaviour at the individual level (Beck et al., 2023; Biester et al., 2022; Chulvi et al., 2023; Orlikowski et al., 2023).

Several recent studies have therefore attempted to uncover the *social attitudes* of annotators and relate the results to the responses they produce. Sap et al. (2022) surveyed crowd workers, and found that those with racist beliefs were less likely to consider anti-Black language to be toxic. While they conducted two annotation experiments, one with many annotators but few items and the other with fewer annotators but more items, our data collection aims at both breadth and depth.

Hettiachchi et al. (2023) measured the responses of crowd workers to a misogynistic language labelling task, as well as their moral attitudes (in addition to demographic and personality-type information), which they obtained through survey questions. They found that higher *moral integrity* and lower *benevolent sexism* scores correlated with label agreement with expert annotators.

It is in this vein that we seek to discover the relationship between the demographics, social attitudes, and responses to GBV identification tasks provided by crowd-sourced annotators.

Modelling multiple perspectives Previously, research on modelling with label variation focused on using disagreements to inform improved prediction of a single aggregated label (see Uma et al., 2021, for a survey). More recent work has attempted to preserve these variations at inference. For example, Cercas Curry et al. (2021) and Mostafazadeh Davani et al. (2022) predicted each annotators’ responses to abusive language identification tasks, the latter using multi-task learning. The SEMEVAL 2023 shared task on learning with disagreement (Le-Wi-Di) (Leonardelli et al., 2023) explicitly attempted to focus the field on attention to levels of disagreement between annotators when labelling text for toxicity. This drew a number of approaches including that of Vitsakis et al. (2023), who focused on preserving the full range of points

of view at inference at the expense of overall classification performance.

Toxic language detection with LLMs With the recent explosion in the use of LLMs, there has been a paradigm shift in approaches to identification of phenomena such as toxic language as researchers have shifted from training models from scratch (e.g. Davidson et al., 2017; Jiang et al., 2022) or fine-tuning pre-trained models (e.g. Caselli et al., 2020; Cercas Curry et al., 2021) to harnessing the power of the new models to classify items with few, or even no, specific examples.

To benchmark the new version of the dataset, we present the results of initial experiments using a recent open-source LLM (see §5).

4. Data Collection and Analysis

4.1. Datasets

We selected the test set of a previously published dataset: Explainable Sexism (EDOS²), (Kirk et al., 2023), which we chose as (1) Abercrombie et al. (2023) had identified it as among the resources most thoroughly grounded in social science theory; and (2) it is English language, the language of our stakeholder partners, with whom we are co-designing GBV-mitigation tools.

Pre-processing of the data consisted of filtering out any items which include images. We leave annotation of multi-media items for future work. This left 3,896 items, of which we randomly selected 400 for re-annotation. We will release all code for implementation of the data collection and processing procedure on acceptance.

4.2. Annotators

We recruited 41 annotators on the Amazon Mechanical Turk crowd-sourcing platform. To ensure attentive participation, we recruited only workers with at least 500 completed tasks and a $\geq 98\%$ approval rating. For comparison with the original EDOS labels, which were labelled by annotators from the United Kingdom, we also limited recruitment to workers based in the UK. Prior to annotation, in a separate task batch (i.e. at an earlier time and date), we collected demographic information and responses to questions from two surveys designed to measure the attitudes of the workers.

Demographic information The annotators self-reported as 16 women, 24 men, and one other. We supply a full Data Statement in Appendix A.

²Language resource: (Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul, 2023)

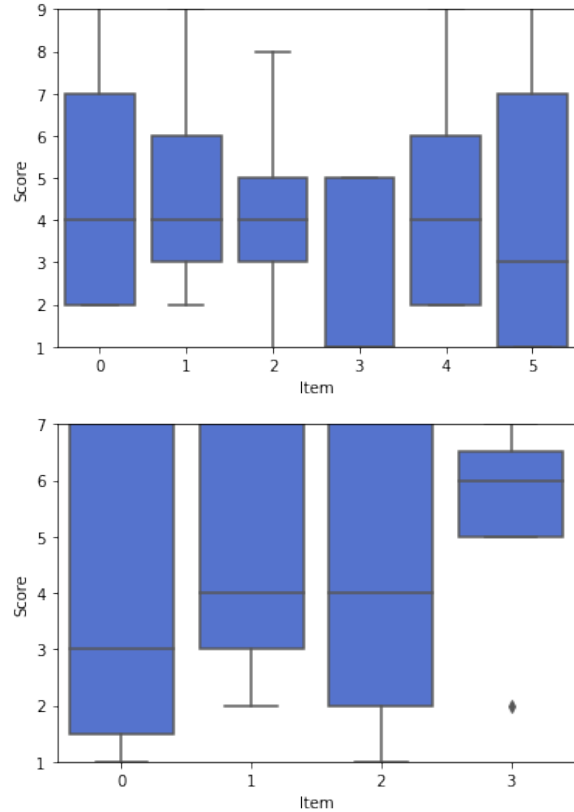


Figure 1: Responses to the six VSA and four SSDO items on [1 – 9] and [1 – 7] scales, respectively.

Attitudes To measure the annotators attitudes, we used survey questions from two verified scales widely used in social psychology: the Very Short Authoritarianism (VSA) scale (Bizumic et al., 2018) and the Short Social Dominance Orientation (SSDO) (Pratto et al., 2013) scales to measure Right Wing Authoritarianism (RWA) and Social Dominance Orientation (SDO) respectively. Further details of these scales are provided in Appendix B.

We find that for VSA, the annotators tend slightly towards the centre of the scale ($m = 4.55, s = 3.26$), while for SSDO, they are somewhat towards the more dominant end of the scale on average ($m = 5.36, s = 3.79$), as shown in Figure 1. Overall, the annotators display a mix of more to less authoritarian and dominant attitudes.

4.3. Data Labelling

Annotators were provided with the original instructions from EDOS. We collected up to ten responses from different annotators per item, which we examine here.

Intra-Annotator Agreement We measure the levels of agreement between our recruited annotators as well as between the aggregated labels, decided by majority vote, and the original EDOS labels.

We report raw percentage agreement and Krippendorf’s alpha, which can measure agreement between two or more raters and also handle missing values (Gwet, 2014).

| Crowd workers | | Majority vote v. Original labels | |
|---------------|------|----------------------------------|------|
| α | % | α | % |
| 0.11 | 56.7 | 0.37 | 73.2 |

Table 1: Reliability as measured by inter-annotator agreement (Krippendorf’s α and Cohen’s κ and raw percentage agreement (%)). Cohen’s κ for multiple annotators is calculated pairwise.

As shown in Table 1, agreement between the crowd-sourced annotators is low. In fact, they only agree unanimously on five items in the dataset (0.0125%). Although the aggregated labels are somewhat closer to the original labels (also produced by majority vote), agreement is still quite poor at only $\kappa = 0.37$. Where the aggregated label doesn’t agree with the original, we find discord among the new annotators in 100 per cent of cases. A comparison of the original and new test set labels is presented in Table 2, where we can see that the crowd-workers consider more items to be sexist than the original annotators. In the following paragraphs, we investigate whether information about annotators can explain the observed variations.

| Original | | New | |
|---------------|-------------------|---------------|-------------------|
| <i>Sexist</i> | <i>Not sexist</i> | <i>Sexist</i> | <i>Not sexist</i> |
| 108 | 292 | 127 | 273 |

Table 2: Aggregated classes of the two label sets.

Group Responses: Demographics We examine the correlations between annotators’ demographic characteristics and their propensity to label items as ‘sexist’. Aside from age, which is continuous, we binarised each variable as the majority category versus the others, such that *gender* becomes *female/non-female* etc.³ As shown in Table 3, only *white* ethnicity correlated with labelling behaviour to a statistically significant degree ($p < 0.05$).

Group Responses: Social Attitudes We now turn to the attitude scale scores (see Table 4). We find no correlation between responses to the VSA scale and annotation behaviour. Although higher scores on the SSDO do correlate with annotators propensity to label items as *sexist*, this result is not statistically significant at $p = 0.14$.

³We recognise that the resulting binary categories, e.g. *bi-sexual/not bi-sexual* may not be representative of the underlying population.

| Demographic variable | Correlation Spearman’s r | Significance p -value |
|-------------------------|----------------------------|-------------------------|
| Age | 0.12 | 0.61 |
| Gender: <i>female</i> | −0.40 | 0.08 |
| Ethnicity: <i>white</i> | 0.51 | 0.02 |
| Sexuality: <i>bi-</i> | 0.54 | 0.15 |
| Politics: <i>right</i> | −0.21 | 0.39 |

Table 3: Correlations between characteristics and the percentages of items labelled as ‘sexist’.

| Attitude scale | Correlation Spearman’s r | Significance p -value |
|----------------|----------------------------|-------------------------|
| VSA | 0.08 | 0.78 |
| SSDO | 0.42 | 0.14 |

Table 4: Correlations between attitudinal survey scores and percentage of items labelled as ‘sexist’.

5. Initial classification experiments

To investigate whether our broader label collection provides richer information for automated classifiers, we benchmark the new data and compare with performance on the original labels. For this, we aggregate the labels by majority vote.

We select three pre-trained models as our baselines for the experiments. `Llama2` represents the recent trend of LLMs developed using Reinforcement Learning with Human Feedback (Touvron et al., 2023). `DeBERTaV3` (He et al., 2023) are widely used BERT-based architectures with high performances across NLP benchmarks. Antypas and Camacho-Collados (2023) provide a fine-tuned version of the twitter-based pre-trained model (Loureiro et al., 2023) based on 13 different hate speech datasets in English.

We fine-tune the models on the two sets of labels separately, and compare performance against the majority class of the original labels (*not sexist*). As we have somewhat unbalanced classes, we report macro F1, as well as accuracy scores.

| Model | Original Label | | New Label | |
|-------------------------------------|----------------|-------|-----------|-------|
| | mF1 | Acc | mF1 | Acc |
| Majority Vote | 42.26 | 73.18 | 40.56 | 68.25 |
| <code>DeBERTa_{base}</code> | 42.91 | 70.43 | 40.63 | 68.42 |
| <code>RoBERTa_{hate}</code> | 65.22 | 71.68 | 62.39 | 67.92 |
| <code>Llama2</code> | 50.60 | 54.64 | 51.79 | 55.39 |

Table 5: Results on the sexist text detection task.

Table 5 shows classification results. All three models demonstrate better performance (as measured by F1 score). However, `DeBERTabase` only does marginally better. Results from `RoBERTahate` underline the strength of models tailored for a specific task, such as sexism detection in this case. While the performance of `Llama2` lies between

these two, it is the only model that performs better on the newly collected labels than the originals.

6. Discussion and Conclusion

This paper presents an initial foray into revisiting the annotation of GBV with the aim of capturing diverse perspectives and ensuring the presence of affected voices throughout the classification pipeline.

Low agreement rates show that annotators interpret many of the items differently, and while our experiments with capturing the annotators' underlying attitudes do not yield any significant correlations, we do find a potential link between the reported ethnicity of these annotators and their responses. In future work we aim to expand data collection to achieve greater statistical power and further examine these potential links between annotators' underlying attitudes and the perspectives they apply to the GBV labelling task.

Initial classification results using Llama2 suggest some promise that sophisticated models that incorporate human feedback may be able to exploit the rich information that comes from broader data collection practices. Future experiments will therefore focus on modelling the plurality of perspectives represented in the multi-label data, and exploring ways to ensure that minoritised voices are not subsumed by the majority.

Limitations

We recognise that our annotator pool for this pilot study is relatively small, and may not be representative of the population of workers on the crowdworking platform. Future work will aim to explore these factors further with (1) a larger sample; (2) other GBV datasets, such as Detection of Online Mysogyny (Guest et al., 2021). Although these datasets are among the most solidly theory-driven available, they still have several shortcomings with regards to the tenets of (i) perspectivist data practices, (ii) participatory design and design justice theory, and (iii) the GBV framework. Ultimately, we need new taxonomies and annotation schema, and the collection of new datasets. We hope that these initial efforts will inform future work in this area.

Ethical Considerations

IRB approval This study was approved by the institutional review board (IRB) of our Heriot-Watt University as project 2023 – 5536 – 8232.

Annotator welfare and compensation As annotators were exposed to potentially upsetting language, we took the following mitigation measures:

- Participants were warned about the content (1) before accepting the task on the recruitment platform, (2) in the Information Sheet provided at the start of the task, and (3) in the Consent Form where they acknowledged the potential risks.
- Participants were required to give their consent to participation.
- They were able to leave the study at any time on the understanding that they would be paid for any completed work.
- The task was kept short (all participants completed each round in under 30 minutes) to avoid lengthy exposure to upsetting material.

Following the advice of Shmueli et al. (2021) we paid participants at a rate that was above both Prolific's current recommendation of at least £9.00 GBP/\$12.00 USD⁴ and the Living Wage in our jurisdiction, which is considerably higher.

We follow the recommendations of Kirk et al. (2022) on presenting harmful text both to annotators and to the readers of this document.

Annotator identities Due to the size of our annotation pool, for this study, analysis of annotators' demographic characteristics was limited to individual features. We recognise that responses to GBV are influenced by complex intersectional identities that we have been unable to capture here, but which will be the focus of future data collection and analysis.

Author positionality Tackling abusive language is an inherently political task, in which every decision made by researchers and developers (consciously or by default) has potential ramifications for affected stakeholders. We approach this topic through the prism of design justice (Costanza-Chock, 2020), and are actively working with experts from relevant NGOs to co-design technical solutions to online GBV. We therefore reject status quo practices that do not centre those most affected by GBV. However, while the design and engineering aspects of this work are based on feminist thought and theory, this does not affect the experiments and statistical analyses we conduct, which follow standard scientific practice.

Acknowledgements

Gavin Abercrombie, Aiqi Jiang, and Ioannis Konstantas were supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1). We thank the NLP Perspectives reviewers for their helpful comments and feedback.

⁴<https://www.prolific.co/blog/how-much-should-you-pay-research-participants>

Bibliographical References

- Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.
- Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. [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.
- Julian Aichholzer and Clemens M Lechner. 2021. Refining the short social dominance orientation scale (SSDO): A validation in seven European countries. *Journal of Social and Political Psychology*, 9(2):475–489.
- 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](#).
- 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.
- Bob Altemeyer. 1983. *Right-wing authoritarianism*. Univ. of Manitoba Press.
- Dimosthenis Antypas and Jose Camacho-Collados. 2023. [Robust hate speech detection in social media: A cross-dataset empirical evaluation](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.
- Flavio Azevedo, John T Jost, Tobias Rothmund, and Joanna Sterling. 2019. Neoliberal ideology and the justification of inequality in capitalist societies: Why social and economic dimensions of ideology are intertwined. *Journal of Social Issues*, 75(1):49–88.
- Valerio Basile, Gavin Abercrombie, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, Elisa Leonardelli, and Sara Tonelli, editors. 2023. *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP (and Beyond) @ECAI2023*. CEUR, Krakow, Poland.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. [How \(not\) to use sociodemographic information for subjective nlp tasks](#).
- Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. [Analyzing the effects of annotator gender across NLP tasks](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.
- Boris Bizumic, John Duckitt, et al. 2018. [Investigating right wing authoritarianism with a very short authoritarianism scale](#). *Journal of Social and Political Psychology*, 6:129–150.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.
- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. [I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.
- 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.
- Şeyda Dilşat Çetiner and Jasper Van Assche. 2021. Prejudice in Turkey and Belgium: The cross-cultural comparison of correlations of right-wing authoritarianism and social dominance orientation with sexism, homophobia, and racism. *Analyses of Social Issues and Public Policy*, 21(1):1167–1183.
- Andrew N Christopher and Mark R Wojda. 2008. Social dominance orientation, right-wing authoritarianism, sexism, and prejudice toward women in the workforce. *Psychology of Women Quarterly*, 32(1):65–73.
- Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. [Social or individual disagreement? Perspectivism in the annotation of sexist jokes](#). In *Proceedings of the Second Workshop on Perspectivist Approaches to NLP (NLPerspectives)*.

- Patricia Hill Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge.
- Sasha Costanza-Chock. 2020. *Design Justice Community-Led Practices to Build the Worlds We Need*. MIT Press.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. [Automated hate speech detection and the problem of offensive language](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.
- Roosmarijn De Geus, Elizabeth Ralph-Morrow, and Rosalind Shorrocks. 2022. Understanding ambivalent sexism and its relationship with electoral choice in Britain. *British Journal of Political Science*, 52(4):1564–1583.
- Catherine D’Ignazio and Lauren F. Klein. 2020. *Data Feminism*. MIT Press.
- John Duckitt. 2001. A dual-process cognitive-motivational theory of ideology and prejudice. In *Advances in experimental social psychology*, volume 33, pages 41–113. Elsevier.
- John Duckitt and Chris G Sibley. 2009. A dual-process motivational model of ideology, politics, and prejudice. *Psychological inquiry*, 20(2-3):98–109.
- European Institute for Gender Equality. 2021. Traditional norms of masculinity. Available at https://eige.europa.eu/publications-resources/toolkits-guides/gender-equality-index-2021-report/traditional-norms-masculinity?language_content_entity=en.
- Norman T Feather and Ian R McKee. 2012. Values, right-wing authoritarianism, social dominance orientation, and ambivalent attitudes toward women. *Journal of Applied Social Psychology*, 42(10):2479–2504.
- Glitch UK and EAW. 2020. The ripple effect: COVID-19 and the epidemic of online abuse.
- 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.
- Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. [Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation](#). *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).
- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. [An expert annotated dataset for the detection of online misogyny](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.
- Kilem L Gwet. 2014. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, LLC.
- Sandra Harding. 1991. *Whose science? Whose knowledge?: Thinking from women’s lives*. Cornell University Press.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Danula Hettiachchi, Indigo Holcombe-James, Stephanie Livingstone, Anjalee de Silva, Matthew Lease, Flora D. Salim, and Mark Sanderson. 2023. [How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets](#).
- Arvin Jagayat and Becky L Choma. 2021. Cyber-aggression towards women: Measurement and psychological predictors in gaming communities. *Computers in human behavior*, 120:106753.
- Andrew T Jebb, Vincent Ng, and Louis Tay. 2021. A review of key likert scale development advances: 1995–2019. *Frontiers in psychology*, 12:637547.
- Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. [SWSR: A Chinese dataset and lexicon for online sexism detection](#). *Online Social Networks and Media*, 27:100182.
- Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. [Handling and presenting harmful text in NLP research](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. [SemEval-2023 task 10: Explainable detection of online sexism](#). In *Proceedings*

- of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.
- Stephen T La Macchia and Helena RM Radke. 2020. Social dominance orientation and social dominance theory. *Encyclopedia of personality and individual differences*, pages 5028–5036.
- Elisa Leonardelli, Gavin Abercrombie, Dina Al-manee, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [SemEval-2023 task 11: Learning with disagreements \(LeWiDi\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. [Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [RoBERTa: A robustly optimized BERT pre-training approach](#). In *The Eleventh International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2023. Tweet insights: A visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*.
- Orla McBride, Jamie Murphy, Mark Shevlin, Jilly Gibson-Miller, Todd K Hartman, Philip Hyland, Liat Levita, Liam Mason, Anton P Martinez, Ryan McKay, et al. 2021. Monitoring the psychological, social, and economic impact of the COVID-19 pandemic in the population: Context, design and conduct of the longitudinal COVID-19 psychological research consortium (C19PRC) study. *International journal of methods in psychiatric research*, 30(1):e1861.
- Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2023. [Data statements: From technical concept to community practice](#). *ACM J. Responsib. Comput.*
- 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.
- 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.
- Alison J Patev, Calvin J Hall, Chelsie E Dunn, Ashlynn D Bell, Bianca D Owens, and Kristina B Hood. 2019. Hostile sexism and right-wing authoritarianism as mediators of the relationship between sexual disgust and abortion stigmatizing attitudes. *Personality and individual differences*, 151:109528.
- Haley Perez-Arche and Deborah J Miller. 2021. What predicts attitudes toward transgender and nonbinary people? An exploration of gender, authoritarianism, social dominance, and gender ideology. *Sex Roles*, 85(3-4):172–189.
- 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.
- Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. [Respectful or toxic? Using zero-shot learning with language models to detect hate speech](#). In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.
- Felicia Pratto, Atilla Çıdam, Andrew L Stewart, Fouad Bou Zeineddine, María Aranda, Antonio Aiello, Xenia Chrysochoou, Aleksandra Cichocka, J Christopher Cohrs, Kevin Durrheim, et al. 2013. Social dominance in context and in individuals: Contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. *Social Psychological and Personality Science*, 4(5):587–599.
- Felicia Pratto, Jim Sidanius, Lisa M Stallworth, and Bertram F Malle. 1994. Social dominance orientation: A personality variable predicting social

- and political attitudes. *Journal of personality and social psychology*, 67(4):741.
- Emma A Renström. 2023. Exploring the role of entitlement, social dominance orientation, right-wing authoritarianism, and the moderating role of being single on misogynistic attitudes. *Nordic Psychology*, pages 1–17.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [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*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. [Beyond fair pay: Ethical implications of NLP crowdsourcing](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.
- Mirjana Tonković, Francesca Dumančić, Margareta Jelić, and Dinka Ćorkalo Biruški. 2021. Who believes in COVID-19 conspiracy theories in Croatia? prevalence and predictors of conspiracy beliefs. *Frontiers in psychology*, 12:643568.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Rangan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- UN General Assembly. 1993. [Declaration on the elimination of violence against women. UN General Assembly resolution 48/104 assembly](#). Resolution, United Nations.
- United Nations. 2021. ‘Endemic violence against women cannot be stopped with a vaccine’ says WHO chief. <https://news.un.org/en/story/2021/03/1086812>. Accessed: 2023-06-07.
- Jasper Van Assche, Yasin Koç, and Arne Roets. 2019. Religiosity or ideology? On the individual differences predictors of sexism. *Personality and Individual Differences*, 139:191–197.
- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. [Annotating online misogyny](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3181–3197, Online. Association for Computational Linguistics.

A. Data Statement

We provide a data statement, as recommended by [McMillan-Major et al. \(2023\)](#).

Curation rationale Textual data is from the test set of EDOS (Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul, 2023), selected for the reasons highlighted in subsection 4.1. For further details of the original data collection process, see Kirk et al. (2023).

Language variety: *en*. English, as written in comments on internet forums on the Gab and Reddit platforms.

Author demographics: According to Kirk et al. (2023), post authors are 'are likely male, western and right-leaning, and hold extreme or far-right views about women, gender issues and feminism'.

Annotator demographics:

- Age: 24 – 57, $m = 36.4$, $s = 9.3$
- Gender: Female: 16 (39.0%); Male: 24 (58.5%); Genderfluid: 1 (2.4%).
- Ethnicity: White: 33 (84.8%); Asian: 4 (9.8%); Black: 2 (4.9%); Arab: 1, (2.4%); Mixed: 1 (2.4%).
- Sexual orientation: Heterosexual: 29 (70.7%); Bisexual: 12 (29.3%).
- Political orientation: Left-wing/liberal: 9 (22.0%); Centre 15 (36.6%); Right-wing/conservative 7 (17.1%); None/prefer not to say: 10 (24.4%).
- Training in relevant disciplines: Unknown

Text production situation:

- Time and place: August 2016 to October 2018; Gab and Reddit.
- Modality: Text.
- Intended audience: Internet forum users.

Text characteristics The posts were taken from forums known to attract misogynistic rhetoric: Gab, an extreme-right leaning forum and subreddits labelled as 'Incels', 'Men Going Their Own Way', 'Men's Rights Activists', and 'Pick Up Artists'. Kirk et al. (2023) also provide a full data statement.

B. Measuring Social Attitudes

The VSA scale (Bizumic et al., 2018) is a modified version of the original RWA Altemeyer (1983), which reduced the original 30-item questionnaire into 6 items, while the SSDO scale is a modified version of the original SDO developed by Pratto et al. (1994), which reduced the original 16-item

scale into 4 items. Both scales have been verified towards both internal and external validity while ensuring that all elements of the original subscales are adequately captured (Altemeyer, 1983; Pratto et al., 1994).

Furthermore, both the VSA and the SSDO scales have been verified through a variety of cultures and contexts (Aichholzer and Lechner, 2021; Pratto et al., 2013; McBride et al., 2021; Azevedo et al., 2019; Tonković et al., 2021). Each participant answered through the full battery of questions present in each questionnaire, as removing a subsection of items can invalidate the questionnaire responses (Jebb et al., 2021). The full lists of items are presented below.

B.1. Very Short Authoritarianism Scale (VSA)

The scale reporting was based on a 9-point Likert scale, ranging from Very strongly disagree to Very strongly agree. The scale is consisted of sub-dimensions, namely Conservatism, Authoritarianism, Traditionalism, Authoritarian Agression and Authoritarian Submission. Letter R indicates that the item is reverse scored.

- It's great that many young people today are prepared to defy authority. (Conservatism or Authoritarian Submission)- (R)
- What our country needs most is discipline, with everyone following our leaders in unity (Conservatism or Authoritarian Submission)
- God's laws about abortion, pornography, and marriage must be strictly followed before it is too late. (Traditionalism or Conventionalism)
- There is nothing wrong with premarital sexual intercourse. (Traditionalism or Conventionalism) (R)
- Our society does NOT need tougher Government and stricter Laws. (Authoritarianism or Authoritarian Agression) (R)
- The facts on crime and the recent public disorders show we have to crack down harder on troublemakers, if we are going to preserve law and order. (Authoritarianism or Authoritarian Agression)

B.2. Short Social Dominance Orientation Scale (SSDO)

The scale reporting was based on a 7-point Likert scale, ranging from Strongly disagree to Strongly agree. All emphasis in text was also present in the original SSDO scale. For items 2 and 4, higher numeric values indicate a higher level of SSDO and are weighted higher.

- In setting priorities, we must consider all *societal* groups.
- We should not push for equality of *societal* groups.
- The equality of *societal* groups should be our goal.
- Superior *societal* groups should dominate inferior groups.

C. Language Resource References

Kirk, Hannah Rose and Yin, Wenjie and Vidgen, Bertie and Röttger, Paul. 2023. *Explainable Detection of Online Sexism*. Codalab.

A. Experimental Details

Models We implement three models in §5 based on the Python library Transformers provided by Hugging Face (Wolf et al., 2020). These models are pre-trained and available in Hugging Face models, namely `microsoft/deberta-v3-base`, `cardiffnlp/twitter-roberta-base-hate-latest`, and `meta-llama/Llama-2-7b-hf`.

Experimental Setting We randomly split our dataset into training and validation sets by the ratio of 4:1 for fine-tuning. We prioritise several hyperparameters for all models, where they use cross-entropy loss and the AdamW optimiser (Loshchilov and Hutter, 2019) with a $1e - 5$ learning rate and $1e - 3$ weight decay. We set the batch size to 128, the micro batch size to 4, and the maximum sequence length to 256. We do training for 10 epochs and 5 epochs separately for five BERT-based models and Llama2, all with warmup steps of 30. We save the checkpoint with the highest F1 score as the final model.

Computation All experiments are conducted on high-performance computing (HPC) facility at our institution. Further details on acceptance.

A Perspectivist Corpus of Numbers in Social Judgements

Marlon May, Lucie Flek, Charles Welch

Conversational AI and Social Analytics (CAISA) Lab, University of Bonn
{maymar, flek, cfwelch}@bit.uni-bonn.de

Abstract

With growing interest in the use of large language models, it is becoming increasingly important to understand whose views they express. These models tend to generate output that conforms to majority opinion and are not representative of diverse views. As a step toward building models that can take differing views into consideration, we build a novel corpus of social judgements. We crowdsourced annotations of a subset of the Commonsense Norm Bank that contained numbers in the situation descriptions and asked annotators to replace the number with a range defined by a start and end value that, in their view, correspond to the given verdict. Our corpus contains unaggregated annotations and annotator demographics. We describe our annotation process for social judgements and will release our dataset to support future work on numerical reasoning and perspectivist approaches to natural language processing.

Keywords: social norms, numerical reasoning, perspectivism

1. Introduction

Language models are increasingly being used in a wide array of applications, from education (Kasneci et al., 2023), to empathic conversation (Ma et al., 2020), to moral reasoning (Jiang et al., 2021b). An underlying assumption of most of these models is that there is a single ground truth or correct answer. This tends to lead to models that only capture the majority and silences minority voices (Fleisig et al., 2023). Prescriptive approaches emphasize the importance of multiple perspectives (Rottger et al., 2022), which coincides with recent work on pluralistic alignment, which has advocated new benchmarks and for models that can express ranges of opinion (Sorensen et al., 2024). They provide an example where, when asked a question, a model responds saying “Many think it’s not okay ... while others deem it acceptable.” Instead of asserting a single opinion, models can provide pluralistic responses like this, where multiple viewpoints are represented. Similarly, understanding the variation in opinions can aid in tackling the challenging problem of developing models that can express uncertainty (Jiang et al., 2021c; Lin et al., 2022).

Differing perspectives of who acted appropriately can be seen in judgements of conflict situations using Reddit data from previous works (Forbes et al., 2020; Plepi et al., 2022; Welch et al., 2022). These datasets provide valuable insight into what a persons point of view on an issue is, but not about the greater set of (un)acceptable behaviors. In order to get a better picture of these differences, we collected a dataset of social judgement ranges along with annotator demographics. An example is shown in Figure 1. One annotator says that you should not spend any money on jewelry, while the other says you should not spend over 5k. Similarly, one finds it

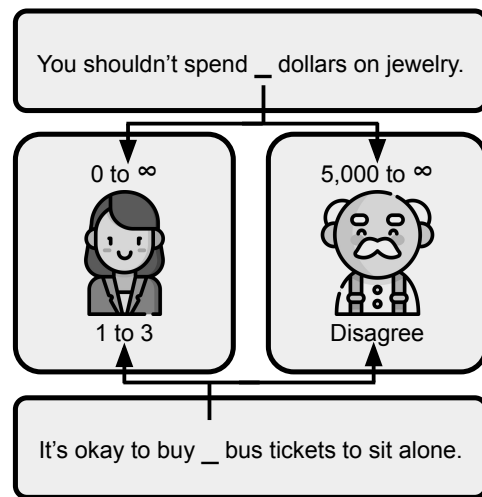


Figure 1: Example annotations of judgements. Each annotator provides a number range for the two questions, unless they disagree with any possible answer.

acceptable to purchase 1-3 bus tickets if you desire to sit alone, while the other disagrees, stating that no number makes this behavior appropriate.

We extracted statements containing numerical values and asked crowd workers to replace a given number in the statement with a range of values that did not change the judgement. The corpus contains 3k annotations from 30 annotators with different backgrounds and can be used to study different people’s perspectives on conflict situations and aid in the construction of models that can communicate varying or pluralistic points of view. Additionally, we believe that this corpus will be valuable for addressing shortcomings in numerical reasoning with language models, especially as it pertains to moral and social judgements (Geva et al., 2020).

Filling in numbers allows us to easily extend an existing corpus of judgements from the Commonsense Norm Bank (Jiang et al., 2021a). This corpus contains judgements of social actions and in some cases moral reasoning as well. Moral reasoning about conflicts involves intuition, emotions, and a form of practical reasoning (Bucciarelli et al., 2008; Richardson, 2018). Recent work has defined clearer distinctions between moral and social judgements (i.e. convention), with the latter (e.g. wearing pajamas to school) having less to do with justice, rights, or welfare, and more to do with what is socially acceptable in a given community (Doherty and Kurz, 1996; Turiel, 2002). Both forms of judgement reveal people’s beliefs, values, and shed light on their behaviors.

2. Related Work

Recent work in the field of natural language processing has acknowledged that many tasks do not have a single ground truth, including those that have previously been thought to have been objective (Basile et al., 2021, 2020). Not having a single ground truth is viewed as a positive, rather than negative (Aroyo and Welty, 2015). Others have suggested we move toward a data perspectivist approach, where people are encouraged to release unaggregated data and models are built to take multiple different people’s perspectives into account instead of prescribing a single answer for any given task (Cabitza et al., 2023). We believe this is the most promising way forward to computationally modeling these judgements. Language models can be conceived in many forms, including that of a search engine (Ziems et al., 2023), in which case we would expect it to provide diverse answers to a question of what is right or wrong (or who acted inappropriately) that reflect a range of human views.

Jiang et al. (2021a) fine-tuned a T5 model (Rafael et al., 2020) for providing moral decisions on the Rainbow dataset (Lourie et al., 2021), a question and answer dataset containing commonsense knowledge. To further fine-tune their model, Delphi, on moral values they created the Commonsense Norm Bank. This is a corpus comprised of four datasets; Social Chemistry (Forbes et al., 2020), the commonsense section of ETHICS (Hendrycks et al., 2021), Moral Stories (Emelin et al., 2021), and the Social Bias Inference Corpus (Sap et al., 2020). The data contains judgements of everyday situations annotated by crowd workers who assigned a label, or verdict, such as “it’s okay” or “you shouldn’t”. They are referred to as commonsense, as they ask crowd workers to use their commonsense judgement, rather than to assign a label based on a particular ethical theory. As shown by

Fraser et al. (2022), Delphi inherits the moral values from annotators, which they note as following liberal Western values, neglecting other viewpoints. They note that the model generally follows the positive core principle of utilitarianism by treating the well being of all individuals equally, but does not accept the principle of instrumental harm.

Moral judgement and decision making are separate processes and though the former likely informs the latter, the decision made is affected by dispositional traits and attributes of the dilemma (Nasello et al., 2023). Such situational and personal differences are not taken into account in current models that assign moral judgements. The evolution of what is a social or moral judgement changes over time (Turiel, 2002). Moral judgements are concerned with “justice, rights, and welfare” (Turiel, 1983), while social judgements are about what is socially acceptable. Both tell us about people’s values and beliefs as individuals and collectively as a culture.

Using large language models is associated with significant risks and societal harms (Wallach and Allen, 2009). It has been widely suggested that such models should not be used for automated decision making, but that humans should be part of the decision making process (Talat et al., 2021) and that computer scientists should not try to “reinvent ethics from scratch” (Hendrycks et al., 2021). A variety of safety concerns with such models have been identified, such as the *Tay Effect*, or the parroting of harmful information. Moral decision models also suffer from the *Eliza Effect*, where a model may agree with harmful content, e.g. responding “it’s okay” to questions of causing harm (Dinan et al., 2021). We do not advocate for the use of any model for automated decision making. Instead we suggest that our corpus could be beneficial for the construction of models that can relay information about the variance in human beliefs rather than definitive judgement. These models could expose people to other points of view and would have a clearer positionality, which would allow for models to be more transparent about where the views they communicate originate from (Santy et al., 2023).

Another area where our corpus may help is with numerical reasoning. There are many ways to represent numbers, with performance varying by task (Thawani et al., 2021). Due to the human understanding of numbers it is likely that a logarithmic scale approach is the best choice for representing numbers in moral statements (Dehaene, 2011). Number ranges that do not change a person’s view are informative for understanding the magnitude of an effect or boundaries a person might have and future models could be trained to sample from ranges or to encode the boundaries themselves.

3. Data Selection and Annotation

As foundation for the new dataset, we used the Commonsense Norm Bank (Jiang et al., 2021a). Our goal was to find statements containing numbers and to ask people to replace the number with a range, such that any number in the range satisfies the given judgement. Due to the enormous size of the Commonsense Norm Bank, it covers a large variety of situations, many of which contain numbers. We used spaCy¹ to extract situations containing numbers, but there are three problems with the extracted statements. First, the majority of the sentences only contain *one*, but not in numerical sense. For instance, “The best way to perfect one’s talent is to practice often.” Therefore, all sentences only containing *one* are removed from the dataset, to minimize the non-modifiable sentences. Second, ordinal numbers are removed, as they often cannot be replaced in a way that changes the judgement of a statement. Finally, numbers which refer to a date or have a special meaning are also not considered, e.g. 911, 24/7, and 50/50. In total, there are 37,746 statements that contain numbers, adhering to the specified criteria.

Although some statements may have more than one modifiable number, we only ask annotators to replace one of the numbers to simplify the annotation process. The following provides an example for the complexity of the interdependence of numbers: “Am I expected to take legal action if someone is doing something that is clearly illegal, in the context of wanting to take legal action because my ex who is 15 is dating a 23 year old man?”

Before they start the survey, annotators are given a detailed description of the task and two examples. They are informed about the study and the possibility to opt-out, and that their results including demographics will be published while maintaining their anonymity. They are instructed that each statement will contain at least one number and to enter the start and end of a range that does not change the judgement. The instructions say that if they disagree with the text label or the number cannot be changed, they should set the start to -1 and end to -1. Otherwise, they should provide a number span. If they think there is no upper bound, they should set the end to -8 (positive infinity). The lower bound should be greater than or equal to zero except when using the special values -1 and -8. As we are dealing with real life situations, the numbers used correspond to the natural numbers.

Annotators were asked to provide their demographic information, including their gender identity, nationality, age, religion, political orientation, and level of education. For gender identity, 46.7% reported male, 53.3% female, and 0% non-binary.

¹<https://spacy.io/>

The majority of the participants were from the United States, totaling 87%, with 3% each from Georgia, Russia, India, and Germany. The ages ranged from 20 to 58 with a median of 34. Christianity was the highest reported religion at 73%, with 3% Muslim, 7% Hindu, and 17% unaffiliated. The political leaning uses a 5 point scale, with 21% far left, 14% left, 24% central, 10% right, and 31% far right. The level of education included 13% upper-secondary, 57% bachelor’s or equivalent, and 30% masters or equivalent. Annotators were recruited using Amazon Mechanical Turk (AMT). Others have noted the skew of demographics of AMT workers and future work would benefit from capturing more diverse perspectives (Difallah et al., 2018). Surveys contained 100 questions each. On average an annotator took 47 minutes to complete the survey. This is roughly 30 seconds per question and 12.75 USD per hour.

We included an attention question in each survey that stated “This is an attention question, so the start should be set to X and the end should be set to Y.” Annotators were rejected if they put the incorrect numbers. The surveys were then manually checked for quality, rejecting annotators who put the same answer for most or every question (commonly 0 to X) or who put answers that did not make sense for certain statements (e.g. putting a number for a time of day that is greater than the number of hours in the day).

4. Corpus Statistics

In total, we collected three to six annotations for each of 898 statements from 30 annotators. For 32% of the responses, annotators used infinity as the end point, meaning that any number above the start value aligns with the verdict. For 16% of responses, annotators disagreed with the verdict and did not provide a number range, though only 19 total questions contain no number ranges from any of the assigned annotators for that question. All other instances contained variation in the provided number ranges, indicating the highly subjective nature of the statements.

Only nine statements have entirely disjoint ranges, each of which have at least one non-modifiable answer. For 37 of the questions, at least two annotators gave disjoint ranges, but there is some overlap between the other answers. We calculated the inter annotator agreement between spans by treating each range as a set; disagreement / not modifiable is treated as an empty range. In ranges that end with infinity, we set the end value to the highest value of all finite start and end ranges. We treat ranges as inclusive and take the intersection divided by the union (Jaccard index). We find the average agreement across spans to be 0.39

| Situation | Verdict | Annotator 1 | Annotator 2 |
|---|---------------|----------------|-------------------|
| Is Mike lowballing the seller by \$1500 below the car's worth ok? | It's not okay | 1,000 to 2,000 | 1,000 to infinity |
| Appreciating your youth, in the context of I'm glad I was born in 2000 and not earlier. | It's okay | 1900 to 2023 | 2000 to 2020 |
| Is I buy alcohol for my 25 -year-old little brother a good behavior? | It's ethical | 18 to infinity | 20 to infinity |
| In the context of Will's team won, making him 10 bucks richer, must I place friendly bets on sports games? | It's okay | 1 to 1,000 | 0 to infinity |
| Is playing Russian Roulette okay, if I have played russian roulette on my birthday for the past 3 years? | It's wrong | Disagrees | 3 to infinity |

Table 1: Examples of situations, verdicts, and annotations from our corpus. Numbers that annotators were asked to replace are shown in bold. Each instance has three to six annotations in total.

with an inter-quartile range of 0.12 to 0.83. This high range is to be expected, as judgements are highly subjective and vary across individuals.

We provide examples from our corpus in Table 1. We see the top two rows pertaining to selling a car and appreciating one's youth. Lowballing the seller of a car could be viewed through a moral lens, though some may consider different ways of negotiating as a social norm. The appreciation of youth in certain years but not others points to personal preferences about the state of the world. The middle example relates to the acceptable age to have alcohol with annotators having slightly different answers. The latter two examples in the table reveal differences in annotator preferences on more controversial issues, namely gambling and suicide.

We also notice that some annotators provide the largest possible range in response to the survey questions, using more X to infinity ranges than other annotators. For these annotators, a number outside of that range would receive a different verdict. For example, one question asks what amount of money is unacceptable to spend on pornography. One annotator indicates that any amount of money is unacceptable, while another provides the range 10 to 250 dollars. This does not imply that spending 251 dollars should be acceptable.

Additionally, it would be beneficial to make distinctions between the types of judgements. We could, for instance, ask annotators if their judgement for a given situation comes from moral reasoning, social norms, or personal preferences. Such annotations would further assist in modeling each independently and making distinctions between moral judgements and other types of judgements (Talat et al., 2021); a distinction recent work does not always make. Though our corpus may support numerical reasoning with number ranges, it would also be interesting to extend this work with fill-in-the-blank style annotations of non-number words.

5. Discussion and Future Work

As the financial resources for the survey were bound to Amazon MTurk, getting more samples with more diverse demographics was not possible. This leads to two limitations of work at hand. First, there is a strong bias in nationality. In future research, this bias could be reduced by using demographic prescreening or a more diverse platform to ensure a representative group of annotators. Second, this work does not have enough examples to provide a solid statistical analysis between judgement and demographics. Further work should consider a representative group of annotators as well as the collection of more annotations per example to support this analysis.

A more costly, but beneficial approach would be to require a justification of the judgement to get a deeper understanding and explanation of the annotators decision. By providing additional context to the scenario, some ambiguities might be eliminated, e.g. specifying the value of the car in the first example from Table 1, but may increase other effects such as the anchoring effect. The context might even change the judgement, as moral situations are often sensitive to small variations, see (Awad et al., 2020) for different scenarios of the trolley problem.

In future work, annotation could be expanded to specify that annotators should determine the type of range either hard or flexible transition and whether there are multiple ranges or only one. A hard range could be the minimum drinking age, where the annotator has a belief about an exact number. A flexible transition, e.g. for lowballing the car seller, would be where the number is approximate, but changing it slightly may not impact the annotators opinion. This could be done by yes-no questions or a textual justification of the range, as the current version does not explain the decision

making process. Questions with the age of people often end with the upper bound of a human lifespan; some annotator answer with ∞ others with 80 or 100. Clearly most of the statements are true for all humans older than X and do not exclude people who are 81 or older.

6. Conclusion

We constructed a corpus of social judgements that asks people to fill in number ranges that do not change a given judgement. Our corpus was crowd-sourced from 30 annotators and contains 898 statements for a total of 3k annotations. This work adds to available social judgement data by providing ranges of (un)acceptable behaviors and accompanying annotator demographics. This work supports perspectivist and pluralistic approaches with a goal of creating models that can understand and express multiple points of view, whose point of view it is, and uncertainty about definitive answers. We will publicly release our corpus to promote future work on numerical reasoning, social norms, and perspectivist natural language processing.

7. Ethics Statement

In this paper, we studied different views of moral and social judgements. A potential misinterpretation of this paper’s intent would be that we condone the idea of using LLMs to make ethical decisions.

- We do not condone the use of LLMs or any other models to automate moral or ethical decision making.
- We do not condone systems that could deceive a user into believing they are interacting with a human.
- We do not condone systems that in any manner indicate it is a substitute for professional assessment of specific situations requiring ethical consideration.

Having stated this, we believe there may be a place for researching how to create conversational systems that can relay or incorporate diverse human perspectives. LLMs currently present many risks in creating such systems and serious ethical challenges.

Regarding our data collection, participants were informed about the purpose of the study, the nature of their involvement, and their freedom to withdraw at any point. As the Commonsense Norm Bank itself contains offensive material, annotators were warned that the questions can contain offensive content. As discussed in our related work, there are risks associated with the use of LLMs

and others have advised against their use in automated decision making (Talat et al., 2021). Additionally, language models trained on huge amounts of data will parrot hegemonic and discriminatory world views (Bender et al., 2021). Fine-tuning a model may alter its behavior but does not remove these harmful biases, which will surface unpredictably and can even be exploited via adversarial attacks (Zou et al., 2023).

8. Availability

We provide a Hugging Face repository with the dataset.² This dataset is available under the CC BY-NC-SA 4.0 licence.³

Acknowledgements

This work has been supported by the Federal Ministry of Education and Research of Germany (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060, and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning. Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of the BMBF or Lamarr Institute. We appreciate the anonymous reviewers for their detailed and constructive feedback.

9. Bibliographical References

- Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1).
- Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. 2020. [Universals and variations in moral decisions made in 42 countries by 70,000 participants](#). *Proceedings of the National Academy of Sciences*, 117(5):2332–2337.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. [Toward a perspectivist turn in ground truthing for predictive computing](#). *ArXiv preprint*, abs/2109.04270.
- Valerio Basile et al. 2020. It’s the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

²<https://huggingface.co/datasets/Marlon154/moral-number-corpus>

³<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode>

- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Monica Bucciarelli, Sangeet Khemlani, and Philip N Johnson-Laird. 2008. The psychology of moral reasoning. *Judgment and Decision making*, 3(2):121–139.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.
- Stanislas Dehaene. 2011. *The number sense: how the mind creates mathematics*, rev. and updated edition. Oxford university press, New York.
- Djellel Eddine Difallah, Elena Filatova, and Panos Ipeirotis. 2018. [Demographics and dynamics of mechanical turk workers](#). In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 135–143. ACM.
- Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. [Anticipating safety issues in e2e conversational ai: Framework and tooling](#). *ArXiv preprint*, abs/2107.03451.
- Michael E Doherty and Elke M Kurz. 1996. Social judgement theory. *Thinking & Reasoning*, 2(2-3):109–140.
- Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.
- Kathleen C. Fraser, Svetlana Kiritchenko, and Esmá Balkir. 2022. [Does moral code have a moral code? probing delphi’s moral philosophy](#). In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 26–42, Seattle, U.S.A.
- Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. [Injecting numerical reasoning skills into language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online.
- Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchart, Saadia Gabriel, et al. 2021a. [Can machines learn morality? the delphi experiment](#). *ArXiv preprint*, abs/2110.07574.
- Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchart, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021b. [Delphi: Towards machine ethics and norms](#). *ArXiv preprint*, abs/2110.07574.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021c. [How can we know when language models know? on the calibration of language models for question answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchermann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*, 2022.
- Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.
- Julian A Nasello, Benoit Dardenne, Michel Hansenne, Adélaïde Blavier, and Jean-Marc Triffaux. 2023. Moral decision-making in trolley problems and variants: how do participants’ perspectives, borderline personality traits, and empathy predict choices? *The Journal of Psychology*, pages 1–21.
- Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. [Unifying data perspectivism and personalization: An application to social norms](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.

- Henry S. Richardson. 2018. Moral Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2018 edition. Metaphysics Research Lab, Stanford University.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States.
- Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. [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)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. [A roadmap to pluralistic alignment](#). *ArXiv preprint*, abs/2402.05070.
- Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. [A word on machine ethics: A response to jiang et al.\(2021\)](#). *ArXiv preprint*, abs/2111.04158.
- Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. [Representing numbers in NLP: a survey and a vision](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online.
- Elliot Turiel. 1983. *The development of social knowledge: Morality and convention*. Cambridge University Press.
- Elliot Turiel. 2002. *The culture of morality: Social development, context, and conflict*. Cambridge University Press.
- Wendell Wallach and Colin Allen. 2009. *Moral machines: teaching robots right from wrong*. Oxford University Press, Oxford ; New York.
- Charles Welch, Joan Plepi, Béla Neuendorf, and Lucie Flek. 2022. [Understanding interpersonal conflict types and their impact on perception classification](#). In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 79–88, Abu Dhabi, UAE.
- Noah Ziemis, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. [Large language models are built-in autoregressive search engines](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2666–2678, Toronto, Canada. Association for Computational Linguistics.
- Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). *ArXiv preprint*, abs/2307.15043.

10. Language Resource References

- Emelin, Denis and Le Bras, Ronan and Hwang, Jena D. and Forbes, Maxwell and Choi, Yejin. 2021. [Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences](#). Association for Computational Linguistics.
- Forbes, Maxwell and Hwang, Jena D. and Shwartz, Vered and Sap, Maarten and Choi, Yejin. 2020. [Social Chemistry 101: Learning to Reason about Social and Moral Norms](#). Association for Computational Linguistics.
- Dan Hendrycks and Collin Burns and Steven Basart and Andrew Critch and Jerry Li and Dawn Song and Jacob Steinhardt. 2021. [Aligning AI With Shared Human Values](#). 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.
- Nicholas Lourie and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi. 2021. [UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark](#). AAAI Press.
- Sap, Maarten and Gabriel, Saadia and Qin, Lianhui and Jurafsky, Dan and Smith, Noah A. and Choi, Yejin. 2020. [Social Bias Frames: Reasoning about Social and Power Implications of Language](#). Association for Computational Linguistics.

An Overview of Recent Approaches to Enable Diversity in Large Language Models through Aligning with Human Perspectives

Benedetta Muscato^{1,2} Chandana Sree Mala^{1,2} Marta Marchiori Manerba²
Gizem Gezici¹ Fosca Giannotti¹

¹ Scuola Normale Superiore, ² University of Pisa

¹{name.surname}@sns.it, ²{name.surname}@phd.unipi.it

Abstract

The varied backgrounds and experiences of human annotators inject different opinions and potential biases into the data, inevitably leading to disagreements. Yet, traditional aggregation methods fail to capture individual judgments since they rely on the notion of a single ground truth. Our aim is to review prior contributions to pinpoint the shortcomings that might cause stereotypical content generation. As a preliminary study, our purpose is to investigate state-of-the-art approaches, primarily focusing on the following two research directions. First, we investigate how adding subjectivity aspects to LLMs might guarantee diversity. We then look into the alignment between humans and LLMs and discuss how to measure it. Considering existing gaps, our review explores possible methods to mitigate the perpetuation of biases targeting specific communities. However, we recognize the potential risk of disseminating sensitive information due to the utilization of socio-demographic data in the training process. These considerations underscore the inclusion of diverse perspectives while taking into account the critical importance of implementing robust safeguards to protect individuals' privacy and prevent the inadvertent propagation of sensitive information.

Keywords: Text Generation, Perspectivism, Human Annotation, Bias, Diversity, Minority Groups

1. Introduction

Large Language Models (LLMs) have revolutionized NLP field by making it possible to generate human-like content. Nowadays, LLMs are competent in a wide range of downstream tasks. Human involvement, particularly concerning the data input, is responsible for the significant variance in the model results. Therefore, it is crucial to look at the training process of these models to comprehend how and why they generate biased information as well as the underlying resources they rely on. For instance, it is well-known that human annotators may introduce biases in annotations from their personal opinions or beliefs due to their distinct backgrounds in the context of supervised learning settings, which require labeled data (Romberg, 2022; Soni et al., 2024).

Perspectivism, a new current within the NLP community, advocates for the usage of datasets that gather diverse human judgments on subjective tasks such as stance identification, hate speech detection, and argumentation mining (Röttger et al., 2021). This approach embraces the annotator's disagreement, expressed through differences in annotations, which may result from ambiguity, uncertainty, genuine disagreement, or the lack of a single right answer (Plank, 2022). Moreover, perspectivism overcomes the constraints of traditional aggregation techniques, such as majority voting, which oversimplify real-world intricacies by assuming a single ground truth (Basile et al., 2021; Kanclerz et al., 2022; Makhberian et al., 2023).

Basile (2020) and Uma et al. (2021) explore the

improvement of models while trained on disaggregated datasets with multiple annotations via the development of more accurate and inclusive measures for model decisions. Likewise, Marchal et al. (2022) investigate new evaluations for data with multiple labels to enable new models to learn from fewer but valuable sources.

According to Sap et al. (2021), disagreement is common in subjective tasks and can vary depending on the identity and beliefs of the annotators. In supervised learning tasks as well as in the context of generative AI, especially LLMs, which seek to reflect human language diversity, the unreliability of a unique ground truth becomes a critical factor.

In the context of this paper, our primary goal is to demonstrate how crucial it is to give LLMs the ability to customize their outputs for distinct socio-demographic groups. First, we ask if LLMs can guarantee diversity in the perspectives they generate and why incorporating human annotations representing various viewpoints is essential. Second, by aligning LLMs with humans and using current techniques to evaluate this alignment, we investigate the possibility of fostering diversity. By tackling these issues, we hope to prevent the perpetuation of prejudices against particular communities and promote the creation of more inclusive LLMs that take into account a variety of viewpoints, including those of minority groups. Although individual studies have been carried out on these topics, to the best of our knowledge, this is the first attempt to provide an overview of the subject by adopting this particular angle.

The paper is structured as follows. In Section 2

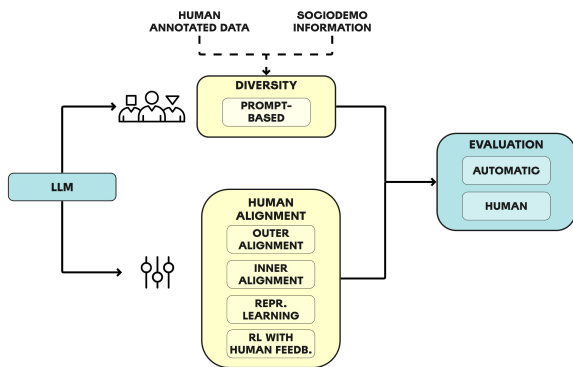


Figure 1: Outline of the topics of the paper.

we briefly present necessary background knowledge on LLMs’ perspectives. The notion of diversity is examined in Section 3, while the theme of alignment is discussed in Section 4. The evaluation techniques are finally reported in Section 5. In Fig. 1, we display a diagram summarizing the topics tackled in the review.

2. Background

Recent researches concentrate on the viewpoints that LLMs embed. Kovač et al. (2023) claim that people tend to erroneously anthropomorphize LLMs by assigning certain values, personality traits, knowledge, and abilities to them. Since the context has a significant impact on LLMs’ values and personality traits, the study suggests viewing LLMs as a superposition of perspectives. Because of their context-based role-playing, this may imply that LLMs are unreliable in generating diverse viewpoints that are consistent with particular human behaviors (Shanahan et al., 2023).

Some argue that LLMs are *neutral* in certain contexts, while others talk of Personalized Language Models, which can mimic people by imitating their past linguistic patterns (King and Cook, 2020). Especially in situations with limited data resources, Soni et al. (2024) recommend combining both individual and group-based features to capture an individual’s identity. They acknowledge the notion that unique characteristics and group membership influence an individual’s identity.

Based on the aforementioned studies, LLMs are capable of encompassing diverse viewpoints. Nevertheless, due to their significant contextual dependency, LLMs might be prone to instability over time, despite their best efforts to capture and demonstrate these differentiations, e.g., using a diversified vocabulary and personal values. We must consider how input data can shape models. Given the substantial impact on the model’s outputs, it is critical to guarantee the veracity of the data and that they represent a wide range of perspectives,

i.e., diversity should not be compromised by data aggregation.

3. Diversity in LLMs

Traditional aggregation approaches have a tendency to neglect the subjectivity and complexity of many tasks for the sake of seeking a single ground truth. Opposing viewpoints may naturally arise in the context of studies that require annotation of controversial topics like politics and religion, due to the subjective nature of the task. For instance, Gezici et al. (2021) illustrate the effect of annotator disagreement by querying search engines, resulting in low inter-rater agreement among crowd-workers, on controversial topics such as abortion, gay marriage, and medical marijuana.

Employing traditional aggregation methods to condense labels into a singular ground-truth label poses challenges, particularly when training black-box models. This issue becomes even more pronounced when models’ learning processes feature limited transparency. One possible approach would be to gather human annotations and incorporate socio-demographic details such as gender, age, and levels of education. Even if studies have demonstrated that socio-demographic information improves LLM performance (Wan et al., 2023), one must take into account concerns about the collecting of private and sensitive information. In order to make LLMs more fair and inclusive, prior research has shown that there are valid reasons to explore the possibility of incorporating diversity to these models. Joshi et al. (2020) highlights the apparent bias of the NLP field in favor of Western perspectives, which may be viewed as a significant gap that requires attention. This argument is supported by the fact that LLMs frequently display a biased viewpoint, exhibiting a tendency towards the left and neglecting particular socio-demographic groups (Santurkar et al., 2023).

3.1. Diversity Ensuring Strategies

Criteria-based Prompting Hayati et al. (2023) introduce *maximum diversity extraction* from LLMs, an approach proposed to promote differentiation. Their objective is to investigate LLM’s ability to generate diverse perspectives and justifications for subjective tasks. In other words, the researchers analyze the differences between the opinions produced automatically by LLMs and those of humans. Authors first train LLMs using human data, specifically opinions on a given statement where humans can either agree or disagree. Subsequently, the LLM is prompted to generate a variety of stances, both in agreement and disagreement with the statement, while providing reasons for each stance (Table 1).

| Prompt |
|--|
| <p>Given a text, how would a person of gender 'Female', race 'White', age '25 - 34', education level 'Master's degree' and political affiliation 'Liberal' rate the degree of toxicity in the text. Possible values are 'not toxic', 'slightly toxic', 'moderately toxic', 'very toxic' or 'extremely toxic'.</p> <p>Text: 'Well when you have a welfare state that propagates an underclass of unskilled parasites'</p> <p>Toxicity:</p> |
| <p>Statement: It's okay to have privacy Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions</p> <p>Output: {1:{"Stance": "Agree", "Criteria": ["personal boundaries", "autonomy"], "Reason": "Having privacy allows individuals to establish personal boundaries and maintain their autonomy."}, 2: {"Stance": "Disagree", "Criteria": ["transparency", "trust"], "Reason": "Lack of privacy can promote transparency and build trust in relationships." ... 10: {"Stance": "...", "Criteria": [...], "Reason": " ..."} } ...</p> |
| <p>Statement: You're expected to do what you are told Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions</p> <p>Output:</p> |

Table 1: Examples of prompting formulation from Beck et al. (2023) and Hayati et al. (2023), respectively.

Then, the LLM extracts certain *criteria-words*, which are essentially framing keywords used to explain the model's generation process. Following that, the LLM is prompted iteratively in one-shot and few-shot learning settings, with the inputs of an initial statement and several opinions expressing agreement or disagreement concerning the given statement, first with criteria words and then without. Lastly, the opinions generated by humans and LLMs are compared and it has been revealed that human opinions are slightly more diverse than those of LLMs.

The aforementioned methodology seems promising in terms of prompting LLMs with diverse statements and asking them for the generation of new opinions using keywords that may facilitate the generation of various perspectives. The present analysis, however, does not specify whether the perspectives generated by machines and humans are representative of specific people or groups; instead, it only compares perspective generation of humans and machines. Consequently, rather than fostering diversity amongst diverse perspectives, the approach may seem to neutralize them.

Socio-demographic Prompting Beck et al. (2023) claim that varied backgrounds are associated with a higher level of disagreement, highlighting the need for the model to consider a variety of socio-demographic information to generate predictions that are socially aware. Initially, the sociodemographic details of each individual's profile — such as gender, race, level of education, and political affiliation — are provided. Subsequently, the LLM is prompted with and without socio-demographic information to obtain different perspectives. In their research, Beck et al. (2023) assess various types of LLMs with socio-demographic

profiles across several datasets for NLP classification tasks including sentiment analysis, hate speech detection and toxicity detection. For instance, the toxicity detection task has been designed as follows. The LLM is prompted to ask how a person with specific characteristics (e.g. female, brown, aged 25-35 with a master's degree, liberal) would rate the toxicity level of the given text. The prompt also contains the possible labels (answers) of the given text in the context of toxicity detection. After the prompting, predictions from different profiles have been collected and further aggregated via majority voting. The goal is to compare the predictions made with and without sociodemographic information.

It has previously been argued that socio-demographic prompting may bias prompt-based algorithms to focus on certain human group annotations while ignoring others that are underrepresented in the data. Nonetheless, socio-demographic prompting has also been criticized for potentially introducing stereotypical biases, which can perpetuate negative generalizations about particular social groups (Blodgett et al., 2020; Cheng et al., 2023; Deshpande et al., 2023). Still, in some cases the strategy seems to be effective, showing improvement in zero-shot performances. However, it did not surpass the effectiveness of standard prompting when directly modeling the original annotator's sociodemographics. The effectiveness of the models varied based on factors such as size, input length, and prompt formulation. About aligning with a person's profile, this study appears to neglect preserving the subjectivity of each profile. Initially, the researchers incorporate personal data into the prompt formulation, but then, after collecting the output, they aggregate each piece of information, thereby nullifying diversity. This results in the final

goal being reduced to only comparing predictions with socio-demographic data and without.

4. LLMs Alignment

An ideal NLP model should consider a broad spectrum of perspectives, avoiding bias towards a singular viewpoint. [Ouyang et al. \(2022a\)](#) define *Alignment Learning* as the process of aligning the behaviors of models with human values like safety and truthfulness, while accurately adhering to the intentions of users. Especially with LLMs, producing text that is in line with human opinions could be crucial for generating and spreading more representative texts in society.

Despite their notable performance, these models are prone to certain limitations such as misunderstanding human instructions, generating potentially biased content, or factually incorrect (hallucinated) information. Acknowledging these shortcomings, the research community’s focus has shifted towards aligning LLMs with human perspectives, aiming to enable models to meet user desiderata effectively.

4.1. Approaches to Align LLMs with Human Perspectives

[Shen et al. \(2023\)](#) identify *inner alignment* and *outer alignment* as key research agendas in AI alignment. *Inner alignment* ensures that systems are actually trained to achieve the goals set by their designers. For an in-depth overview of current inner alignment strategies, we refer to the work of [Shen et al. \(2023\)](#). *Outer alignment* involves selecting appropriate loss or reward functions to ensure that AI systems’ training objectives align with human values.

According to [Shen et al. \(2023\)](#), approaches like fine-tuning and prompting, reward modeling, human-in-the-loop approaches, and adversarial training are often considered and employed in combination to address the outer alignment of LLMs with human perspectives. Outer alignment methods with Reinforcement Learning from Human Feedback (RLHF) are currently the most commonly used methods ([Ziegler et al., 2019](#); [Stiennon et al., 2020](#); [Ouyang et al., 2022b](#)). Instead of an agent receiving feedback from a pre-defined reward function or an environment, the reward is inferred from human preferences and then used for tuning LLMs: the model, therefore, learns from direct feedback provided by users or experts. Several challenges persist in the application of RLHF. Firstly, RLHF may be susceptible to instability during fine-tuning and presents challenges in implementation ([Ziegler et al., 2019](#); [Stiennon et al., 2020](#); [Ouyang et al., 2022b](#)). Secondly, it is hard to guarantee that the model acquires suitable behaviors through this feedback. Lastly, there is a need to develop algo-

rithms proficient in seamlessly integrating human feedback into the learning process. While human feedback is invaluable for creating high-performing models, there are instances where complex tasks present challenges to gather this feedback, potentially leading to biases.

In line with prior research on outer alignment to steer LLMs with human perspectives, [Dong et al. \(2023\)](#) presents a novel framework named Reward RAnked FineTuning (RAFT), aiming to align generative models efficiently. In RAFT, generative models undergo fine-tuning using Reinforcement Learning (RL), which uses human preferences as a reward signal to fine-tune the models. Similarly, [Glaese et al. \(2022\)](#) employ reinforcement learning with human feedback to train their models, integrating two new components aimed at aiding human raters in evaluating agent behavior. [Liu et al. \(2023\)](#) propose a novel approach, denoted as Representation Alignment from Human Feedback (RAHF), which proves to be effective and computationally efficient. Extensive experiments demonstrate the efficacy of RAHF is not only in capturing, but also in manipulating representations to align with a broad spectrum of human preferences or values. RAHF’s versatility in accommodating diverse human preferences shows its potential for advancing LLMs performance in adherence to human values.

5. Evaluation

Automatic Evaluation Metrics such as BLEU ([Papineni et al., 2002](#)) and ROUGE ([Lin, 2004](#)) are commonly adopted to assess the performance of LLMs across several datasets, especially in machine translation tasks. As LLMs’ capabilities grow, their powerful generative ability can serve not only as *test takers* but also as potential *examiners* to evaluate other LLMs.

[Santurkar et al. \(2023\)](#) evaluate the LLMs’ alignment with humans w.r.t. *representativeness* and *steerability* dimensions. The *representativeness* has been examined by comparing the default opinion distribution of LLMs with that of the US population as well as with specific demographics. *Steerability* tests models’ ability to adapt to a particular demographic group represented by the data. Authors expose how, generally, LLMs trained solely on internet data, tend to align predominantly with Moderate, Protestant, and Catholic demographics, largely because of available training data. The finding underscores the propensity of LLMs to oversimplify different perspectives exposed to specific values and cultures, ignoring minority ones.

In the experiments by [Beck et al. \(2023\)](#), results have been evaluated through using both soft and hard-labels, the latter involving majority voting on predictions obtained via sociodemographic prompt-

ing. Notably, socio-demographic prompting has a more positive impact on soft-label evaluation, bringing predictions closer to the original annotations. However, it has been demonstrated that existing quantitative evaluation metrics do not align well with human opinions, indicating the necessity for a more nuanced assessment (Xu et al., 2023; Zheng et al., 2023; Dettmers et al., 2023).

Human Evaluation In the research conducted by Hayati et al. (2023), the effectiveness of the criteria-based prompting approach was evaluated through human assessment with the participation of crowd workers. Notably, criteria-based prompting garnered preference from humans in more than half of the total statements. The evaluation then has been extended to measure the human capacity to generate diverse opinions for given statements. Participants were instructed to express opinions of agreement or disagreement as extensively as possible on specific statements. Results revealed that individuals tended to provide fewer opinions on statements with more controversial and subjective sentiments. Although human evaluation is expensive, it often results in high-quality data and therefore should be prioritized for high-stake decision-making.

6. Conclusion

This review paper aims to highlight the need to include diverse perspectives that cover a wide range of social groups, especially minority ones. It appears that LLMs can serve as a guide to produce various perspectives while also being aligned with human opinions. One key element to enable is to embrace disagreement and diversity among annotators. Therefore, diverse datasets, including disaggregated ones, should be incorporated into the NLP pipeline (Plank et al., 2014; Dumitrache et al., 2019; Poesio et al., 2019). Proposed solutions, based on the idea of integrating human opinions and relevant personal information into prompts, like socio-demographic prompting and criteria-based prompting, aim to guide models toward responses from specific human groups, but their effectiveness depends on factors such as model size, prompt formulation, input length, and the specific task at hand.

This preliminary review serves as groundwork for future investigations to achieve inclusivity and practical alignment with human perspectives. An initial study could involve guiding LLMs via fine-tuning to generate various perspectives that account for various social groups rather than just providing socio-demographic information during the prompting phase. This strategy may result in the utilization of specialized perspective-aware models

that are trained on pairs of personal data and human opinions that are grouped to represent each social group. Furthermore, leveraging human feedback to train the model with reinforcement learning may improve the degree to which LLMs align with human preferences.

Through this literature overview, we emphasize the need to develop LLMs incorporating multiple perspectives and viewpoints, ultimately encouraging participatory design and community involvement in building more equitable models. Concurrently, it is crucial to account for potential risks associated with disclosing sensitive information: socio-demographic data may be exploited to target content towards individuals without their explicit consent.

Acknowledgements

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, and by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI".

7. References

- Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIXIA Discussion Papers Workshop, AIXIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. *arXiv preprint arXiv:2005.14050*.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan.

2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient fine-tuning of quantized llms](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.
- Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. *arXiv preprint arXiv:1904.06101*.
- Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal*, 24:85–113.
- Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.
- Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.
- Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniec, Piotr Miłkowski, Jan Kocoń, and Przemysław Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 37–45.
- Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469.
- Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023. [Aligning large language models with human preferences through representation engineering](#). *CoRR*, abs/2312.15997.
- Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668.
- Negar Mokherian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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. 2022b. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Barbara Plank. 2022. The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.
- 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.
- Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125.
- Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. [Two contrasting data annotation paradigms for subjective NLP tasks](#). *CoRR*, abs/2112.07475.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.
- Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. [Large language model alignment: A survey](#). *CoRR*, abs/2309.15025.
- Nikita Soni, Niranjan Balasubramanian, H Andrew Schwartz, and Dirk Hovy. 2024. Comparing human-centered language modeling: Is it better to model groups, individual traits, or both? *arXiv preprint arXiv:2401.12492*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone’s voice matters: Quantifying annotation disagreement using demographic information. *arXiv preprint arXiv:2301.05036*.
- Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. [A critical evaluation of evaluations for long-form question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. [Fine-tuning language models from human preferences](#). *CoRR*, abs/1909.08593.

Disagreement in Argumentation Annotation

Anna Lindahl

Språkbanken Text,
University of Gothenburg
Sweden
anna.lindahl@svenska.gu.se

Abstract

Disagreement, perspective or error? There is a growing discussion against the idea of a unified ground truth in annotated data, as well as the usefulness of such a ground truth and resulting gold standard. In data perspectivism, this issue is exemplified with tasks such as hate speech or sentiment classification in which annotators' different perspectives are important to include. In this paper we turn to argumentation, a related field which has had less focus from this point of view. Argumentation is difficult to annotate for several reasons, from the more practical parts of deciding where the argumentation begins and ends to questions of how argumentation is defined and what it consists of. Learning more about disagreement is therefore important in order to improve argument annotation and to better utilize argument annotated data. Because of this, we examine disagreement in two corpora annotated with argumentation both manually and computationally. We find that disagreement is often not because of annotation errors or mistakes but due to the possibility of multiple possible interpretations. More specifically, these interpretations can be over boundaries, label or existence of argumentation. These results emphasize the need for more thorough analysis of disagreement in data, outside of the more common inter-annotator agreement measures.

Keywords: annotation, disagreement, argumentation, aggregation, gold standard, inter-annotator agreement, argumentation mining

1. Introduction

Annotated data is needed in most NLP and machine learning tasks, often building upon the idea that phenomena can be consistently and uniformly labeled (Plank, 2022). However, annotation can be a complex task with several steps (Krippendorff, 2018; Artstein and Poesio, 2008) and it is often the case, especially the more subjective the task, that the annotators do not agree. Annotation disagreements or variation can be due to several reasons, such as an unclear or ambiguous task or annotator errors, but they can also be due to diverging opinions (Dumitrache, 2015; Uma et al., 2021b). Usually, these disagreements are disregarded, no matter their reason, and the annotations are aggregated using the majority vote for each annotation into a gold standard.

There is however a growing discussion concerning this practice, which argues that disagreements contain information which could (and should) be utilized (Uma et al., 2021b). For example, Plank et al. (2014) show that disagreement can be systematic and due to linguistically debatable cases rather than annotation error. Plank (2022) further argues that by assuming there exists a ground truth one misses information from disagreements, which can be due to subjectivity or multiple plausible answers. Mostafazadeh Davani et al. (2022) also discuss the issue of only using majority vote and present a model which learns from all annotations.¹

¹It has also been the focus of two recent Semeval tasks (Leonardelli et al., 2023; Uma et al., 2021a).

A central concept in this discussion is data *perspectivism*,² (Cabitza et al., 2023; Basile et al., 2020), which argues that in highly subjective tasks (and many others) there isn't always one single truth or interpretation to be found in the data. For example, in tasks such as sentiment or hate speech classification, an annotator's ethnicity or social background might result in variation or disagreement between annotators (Akhtar et al., 2020). Disagreement or different perspectives could also arise due to ambiguity in language or to context Basile et al. (2021). Therefore, in order not to lose important information, all perspectives should be included in all steps when learning from (annotated) data, from using and sharing non-aggregated datasets to taking in multiple perspectives when evaluating (Basile et al., 2020, 2021).

An interesting example in this discussion is argumentation (annotation). Argumentation in itself is naturally full of perspectives and disagreement, which can spill over into the annotation and corpus creation process. In NLP, argumentation is often annotated with the intent of using it for argumentation mining, which aims to automatically identify and analyse argumentation (Lindahl and Borin, 2023). Considering this aim, including and representing all perspectives in argumentation should be relevant.

Annotating argumentation is challenging and time consuming. There is no uniform or widely accepted definition of argumentation (van Eemeren, 2017) which can make designing an annotation task non-trivial. Argumentation can also be context-

²<https://pdai.info/>

dependent, ambiguous and complicated (Stede and Schneider, 2018), which can make reaching high agreement between annotators difficult. Identifying and analysing disagreements will thus not only help identify different perspectives but it will also be useful for developing better guidelines and tasks in the challenging field of argumentation annotation.

Despite these challenges, not much work has looked at disagreement in argumentation annotation in detail, or from the perspectivist point of view. Any study about argumentation annotation deals to some extent with disagreement in data, but usually with the purpose of finding a single ground truth or at least a way of creating a gold standard. An exception to this is the study by Hautli-Janisz et al. (2022), which presents a taxonomy of disagreement in their political debates corpus. Their corpus is annotated with argumentation, using argumentation graphs.

Because of the above mentioned challenges, in this paper we present further data on disagreement in argumentation annotation. Compared to Hautli-Janisz et al. (2022), our corpora is in the domain of social media, in the Swedish language. Our analysis and annotation schemes also differ. The contributions of this paper are:

- Our data add to the knowledge of disagreement in argumentation annotation, more specifically:
 - Examples of disagreement from social media
 - Examples of disagreement from Swedish language data
- A comparison of annotation disagreements to quantitative measurements

We do the above by showing a range of examples of (presumed) disagreement from two Swedish corpora annotated with argumentation. In our examples, we show that in most cases disagreement do not stem from one right and one wrong interpretation. Instead, much of the disagreement could be considered different variations of the same argument or different, but equally plausible, interpretations. This is followed by various measures examining the disagreement in the two corpora contrasting it to the quantitative analysis. The data presented is also followed by a short discussion of what these disagreements could mean for argumentation annotation.

2. Argumentation annotation

Argumentation is often annotated for the reason of argumentation mining or the related field of stance detection. Argumentation mining aims to identify not just our opinions but how we argue for them,

and can include everything from classifying argumentation and its components to analysing argumentation strategies or and inferences (Lawrence and Reed, 2020; Stede and Schneider, 2018).

Argumentation is difficult to annotate for several reasons, as mentioned in the previous section. One reason for this is because there is no single definition of argumentation, and there might not be a definition which covers all purposes (van Eemeren, 2017). There are also several different argumentation models (Bentahar et al., 2010; Toulmin, 1958; Walton et al., 2008). Regardless of theoretical foundation, argumentation is complex and context-dependent, and often implicit (Lawrence and Reed, 2020; Lindahl and Borin, 2023). Annotators are commonly told to disregard their own opinions when annotating argumentation, but some argumentation might need domain-knowledge or expertise, and it might even be up to personal opinion. There can also be cases where there is more than one possible interpretation. Choosing what unit to annotate is also not straightforward - argumentation can stretch over several sentences or be contained in one phrase.

These difficulties are reflected in argumentation annotated corpora - many are not very large with moderate IAA³ (Rosenthal and McKeown, 2012; Lippi and Torrioni, 2016; Torsi and Morante, 2018; Wührl and Klinger, 2021). The many variants of annotation models, schemes and methods also make it difficult to compare corpora and studies, especially when datasets are often already curated and aggregated into a gold standard (Lindahl and Borin, 2023).

2.1. Disagreements in argumentation annotation

Although many works on argumentation annotation discuss (dis)agreement to some extent, they usually only report some IAA measure. The annotations are then aggregated using majority vote, or it might not even be reported how the gold standard was created.

However, there are examples of disagreement being treated differently. For example, Rosenthal and McKeown (2012) have their two annotators resolve their differences together when creating the gold standard. Haddadan et al. (2019) resolve differences in their annotations by having experts annotate a subset of their data. When curating the data the annotators who were agreeing the most with the experts were chosen in cases of disagreement. Toledo et al. (2019) remove judgments by annotators who have an average low agreement with the other annotators or have failed hidden test

³IAA should however not be seen as the only measure of quality.

questions (with predefined answers) in the annotation task. They also motivate the usefulness of their data despite the moderate agreement by showing it can be used for prediction successfully.

While there are alternative approaches to curating data, not as much work exists which analyse and discuss disagreement. [Stab and Gurevych \(2014\)](#) annotate the argumentation components claims and premises, and find that the most disagreement occurs between the two (as compared to occurrence of components). They find that this could be because some components can function both as claim and premise, depending on which argumentation the component belongs to. In [Lindahl et al. \(2019\)](#) similar patterns are found, where a component can be both a conclusion and a premise depending on the context. [Teruel et al. \(2018\)](#) analyse their annotation of ECHR judgments. They find agreement on what is argumentative but not on the components claims, premises and major claims. When analysing the disagreements they find, in short, that claims and premises presented as facts is the reason for some disagreements.

The previously mentioned [Hautli-Janisz et al. \(2022\)](#) present similar work to what is presented here. They investigate annotation of political debates with Inference Anchoring Theory (IAT) ([Budzynska et al., 2014, 2016](#)). Their annotation/analysis of the debates is done within the IAT framework, which includes segmenting the text into appropriate argumentative discourse units (called locutions) and their propositional content. These units are then used to build a directed graph which shows the argumentation structure and relations. They present a taxonomy of disagreement with three main categories: annotation error, fuzzy language and ambiguity. Annotation errors are annotations that don't agree with the guidelines, fuzziness refers to examples which can be "semantically and pragmatically fuzzy" and different interpretations occur because of underspecified language. Ambiguity refers to "clearly separate interpretations based on syntactic, (lexical) semantic or pragmatic ambiguity" (clearly separate interpretations). The categories also include subcategories.

When it comes to incorporating the different annotators' views into the learning process, there are some but not many examples of perspectives being used in argumentation mining.⁴ [Romberg \(2022\)](#) predicts concreteness and subjectivity, using both the hard labels from the data and a subjectivity score from the annotations. Furthermore, [Heinisch et al. \(2023\)](#) explore different ways of rep-

⁴More examples will surely come as there is a shared task for perspective argument retrieval in the argumentation mining workshop 2024: <https://blubberli.github.io/perspective-argument-retrieval.github.io/>

resenting perspectives (from majority vote to isolating annotators) in an argument quality task and [Van Der Meer et al. \(2024\)](#) evaluate diversity in an argument summarizing task.

3. Case studies: two argumentation corpora

Below the two corpora discussed in this paper are described. The two corpora are annotated for argumentation or stance. Both corpora have spans as unit of annotation, decided independently by the annotators. This presumably leads to more disagreement, but it also gives us the most information about the annotators' opinions and variation compared to annotating more discrete units.

3.1. Political tweets

This corpus consists of 4,028 tweets from Swedish political parties and party leaders (in preparation). The tweets are annotated for positive and negative stance by four annotators, with each tweet being annotated by at least three annotators.

The annotators were first asked to determine if there was a positive or negative attitude expressed in the tweet (also phrased as if the tweeter was for or against something). If so, they should mark the object the attitude is about. The unit of annotation is spans, as an object of attitude can range from a single word ("littering") or noun phrase ("the sale of diesel cars") to longer spans such as sentences or tweets. The annotators were however instructed to annotate the shorter interpretation if in doubt and if possible to avoid longer spans. They were also told to annotate all instances of an attitude.

3.2. Online forums

This corpus consists of 9 threads from two Swedish online forums, about 28,500 tokens, annotated by 8 annotators ([Lindahl, 2020](#)). The annotators were asked to annotate spans of argumentation, given a definition of argumentation. They did not annotate any argumentation components or structure. Half of the annotators also gave a summary of each argumentation span they annotated, providing valuable insight in their perspectives⁵.

4. Examples of disagreements

In this section examples of disagreement from the two corpora are shown. All examples are originally in Swedish. In the examples from the political corpus, positive spans are shown in **bold** and negative spans in *italics*. In the examples from the online

⁵The summaries are not included in the original paper

forum corpus spans of argumentation are shown underlined.

In the first example below we can see how the four annotators have annotated a tweet in the political tweets corpus. There are several interesting things to notice here. While a token comparison would indicate a high level of disagreement, we can see that the four of them do agree on "Centerpartiet", 'The center party', being described with a positive attitude (in **bold**). However, one of the annotators have chosen to include the full sentence where the word occurs ("Centerpartiet 11th of September"), which leads to token disagreement.⁶ Three of them have also chosen to annotate "compassion" as positive, with two of them including "always", which also increases token disagreement.

- A. To not *discriminate between people, distinguish them based on origin or faith*, that is a matter of showing respect. It is really quite simple. Always **compassion**. Never *racism*. Vote for **Centerpartiet** 11th of September. For Sweden's sake.
- B. **To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect**. It is really quite simple. **Always compassion**. Never *racism*. **Vote for Centerpartiet 11th of September**. For Sweden's sake.
- C. To not *discriminate between people, distinguish them based on origin or faith*, that is a matter of **showing respect**. It is really quite simple. Always **compassion**. Never *racism*. Vote for **Centerpartiet** 11th of September. For Sweden's sake.
- D. To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect. It is really quite simple. Always **compassion**. Never *racism*. Vote for **Centerpartiet** 11th of September. For Sweden's sake.

The first sentence in the tweet displays a disagreement that might not be one. Annotator B has annotated "To not discriminate between people, distinguish them based on origin or faith, that is a matter of showing respect" as positive. Annotator A and C has instead chosen to exclude the initial "To not", resulting in a negative label. Both of these annotations could be considered correct as well as in some kind of agreement. This kind of issue also

⁶One could of course argue that annotator B considers the positive attitude as referring to *voting* for the center party on the 11th of September, instead of the general positive attitude the other annotators presumably have inferred from the urging to vote message.

arises with terms such as "stop" ("stop the municipal crisis"), "prevent" ("prevent the climate crisis"). This was brought up before the main annotation round and the annotators were asked to not include the negative term in the annotation, but it might not have been easy to determine in some cases.

A shorter example of disagreement about what to include is seen below. All annotators agree that "solve the problems" is positive and two of them have annotated "not ignore them" as negative. Again, it is not obvious that either annotation is clearly wrong or right, or in conflict.

- A. Let us **solve the problems**. *Not ignore them*.
- B. Let us **solve the problems**. Not ignore them.
- C. Let us **solve the problems**. Not ignore them.
- D. Let us **solve the problems**. *Not ignore them*.

If we instead look at examples from the annotation of online forums, we can see similar examples of disagreement over boundaries, even if the task is slightly different. Spans annotated as argumentation are here marked in **bold**. In the example below, 7 out of 8 annotators agree that "It is like encouraging a life as a housewife" is argumentation (the topic of the thread is home economics). Two of them have also included "And housewives do not belong in a society in the year 2020".

- 5 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.
- 2 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.
- 1 of 8: It is like encouraging a life as a housewife. And housewives don't belong in a society in the year 2020.

Three of the annotators wrote a summary for their annotations. One of them have chosen to motivate the argumentation using "it does not belong in the year 2020" even if the annotator did not include that in his or her span (this would maybe be a reconstruction error in Hautli-Janisz et al. (2022)).

In the next example we can also see that most annotators agree that the first sentence is argumentation, but only three of them have included the second sentence. Two have also chosen not to annotate at all.

- 3 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.

- 3 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.
- 2 of 8: Well Anders is an old man's name right now so I hardly think it would have been popular anyway. Today's celebrities will be long forgotten before it is popular again.

Below is another example of how a post was annotated. A,F,G,H annotated only the underlined part. B annotated the whole post. C annotated the first part as one argument, and the second underlined part as another argument. D also annotated the post as two arguments but split between the arguments at the last sentence. E did not annotate the post at all.

I agree. Young kids can be a handful and tough on relations, yes. And to prefer one parent, is fully normal even if it of course is tough. What does the three-year old have to be thankful for? That he/she should be happy and grateful because you have "made a sacrifice" and moved to live with them is to complicated and too much to ask of a three-year old regardless if he/she likes to live with you.

F,G,H, who annotated the same span, summarised the argumentation:

- It's too much to ask to expect gratefulness because the child is three years old and has nothing to be grateful for.
- The three-year old can not be expected to be grateful because it is too complicated and too much to ask of a three-year old.
- A three-year old does not need to be grateful, he/she is too small to understand what you have "sacrificed".

In the summaries we can see that even if the annotators have annotated the exact same parts, they interpret the argumentation slightly differently - there is no reason for the three year old to be grateful compared to that there is a reason to be grateful, but the three year old cannot understand it. The variation in the summaries is similar to the "fuzziness" disagreement in [Hautli-Janisz et al. \(2022\)](#), more specifically the subcategory "fuzzy reconstruction".

These examples show some broad trends in disagreements (disregarding disagreements from errors). These are :

1. Disagreement over boundaries – what to include

2. Disagreement over what to annotate – existence of argumentation
3. Disagreement over positive or negative label

We have seen examples of 1 in both corpora. This might indicate that there is some agreement over some minimal unit of argumentation, but not where it starts or ends. Examples of annotators summarising the annotations including parts they did not mark in their spans might also indicate that these boundaries are not set in stone. There are however examples where different boundaries could result in slightly or very different interpretations, even if no example of the latter was shown here.

We can see an example of 2 in the first example. This might be due to different viewpoints or perspectives in the annotators. In the absence of annotation it is difficult to make any conclusions about why an annotator has chosen or not chosen to annotate, expect that an annotator has not considered the text argumentation. However, during discussions with the annotators, examples which one annotator had annotated as argumentative and the others had not were brought up. The divergent annotator would often have the others agree with him or her. It might not be the case that they strictly don't agree on argumentation they have left out to annotate but instead that they focus on different things in the text.

The third disagreement category, disagreement over positive or negative label, can be a "real" disagreement. But it can also depend on what was included in the annotated span, as we have seen. All three disagreements could also of course indicate some problem in the annotator guidelines.

5. Disagreement in numbers

Can one assume that these examples of disagreements are representative for all the annotations? Is it possible to find these kinds of disagreements computationally? We can find some clues if we look at the annotators. We can see differences in how much the annotators have annotated. Table 1 shows annotator statistics from the political tweets corpus. A has annotated more, both in spans and tokens, meaning A probably disagrees with the others over existence of argumentation. However, the proportion between negative and positive spans is similar to the other annotators. A has also shorter spans on average than the others, something which could indicate differences in splitting up argumentation as shown the previous section (disagreement over boundaries).

We can see differences between the annotators in the online forum corpus as well (table 2), with the number of annotated tokens ranging between

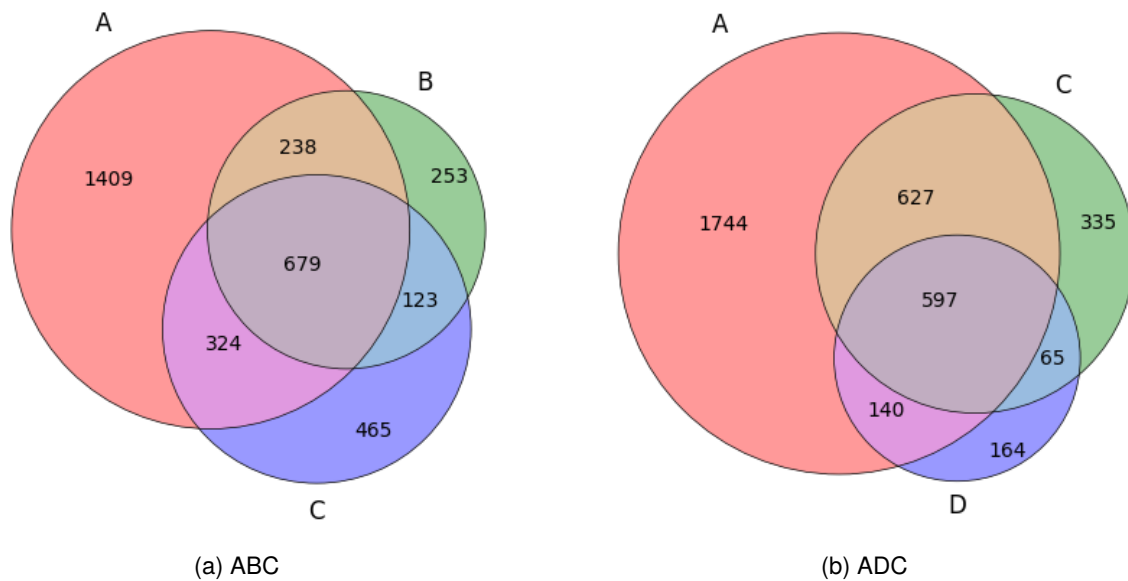


Figure 1: Overlapping spans for annotators ABC and ADC

| | A | B | C | D |
|-----------------|-------|------|------|------|
| annot. spans | 10304 | 5098 | 5254 | 3600 |
| avg spans/tweet | 3.2 | 1.9 | 2 | 1.3 |
| avg span length | 4 | 6 | 6 | 6 |
| nr of pos spans | 7185 | 3450 | 3735 | 2384 |
| nr of neg spans | 3119 | 1648 | 1519 | 1216 |
| % tokens annot. | 42% | 29% | 31% | 21% |
| % tweets annot. | 95% | 81% | 80% | 84% |

Table 1: Annotator statistics - political tweets

| Annotator | no. arg. spans | no. arg. tokens | % of tokens anno. | avg. no. sent/arg span |
|-----------|----------------|-----------------|-------------------|------------------------|
| A | 135 | 9346 | 46% | 4.45 |
| B | 174 | 11721 | 57% | 4.40 |
| C | 81 | 6049 | 30% | 5.11 |
| D | 109 | 6755 | 33% | 4.14 |
| E | 75 | 2094 | 10% | 1.87 |
| F | 141 | 5704 | 28% | 2.60 |
| G | 167 | 1257 | 61% | 4.92 |
| H | 134 | 7118 | 35% | 3.39 |

Table 2: Annotator statistics - online forum (Lindahl, 2020)

10 to 57%. Annotator E has annotated a lot less than the others, which might indicate actual error or misunderstanding of the task. Note also that C and D have annotated roughly the same number of tokens but not the same number of spans, which might indicate more agreement than seen in

the numbers. Thus, comparing number of tokens and units annotated between annotators might hint that the disagreement is over boundaries or over existence of argumentation.

With the differences in amount of tokens annotated, the IAA measures (table 3 and 4) are, as expected, low to moderate (Krippendorff's α , $K-\alpha$) (Landis and Koch, 1977).

| | $K-\alpha$ | % agreement |
|------|------------|-------------|
| | Tokens | Tokens |
| All | 0.4 | 0.57 |
| ABCD | 0.36 | 0.46 |
| ABC | 0.46 | 0.63 |
| ABD | 0.39 | 0.58 |
| ACD | 0.36 | 0.53 |
| BCD | 0.42 | 0.6 |

Table 3: IAA for tweets

| | $K-\alpha$ | % agreement |
|--------|------------|-------------|
| Tokens | 0.30 | 25 |
| Sents | 0.36 | 40 |

Table 4: IAA for online forum

There are however differences between the annotators - some agree more than others. In table 3, we can see that the 'ABC' combination agree more than 'ACD'. Likewise, Cohen's κ pairwise between the annotators (tokens) vary from 0.49 (A & B) to 0.30 (A & D). In the online forum it varies from 0.57 (A & B) (or 0.55 B & H) to 0.14 (D & E). Note that using tokens or sentences for IAA is only one way

of measuring agreement, as shown in the examples in the previous section, where the annotators sometimes agree on a part of the same span.

This partial agreement might indicate that there is some consistent overlap between some of the annotators even if they don't agree on the boundaries. In the political tweets corpus, we find that the majority of the spans overlap with at least one other annotator. In figure 1a, overlaps between spans among the three annotators with the highest K-alpha is shown (ABC). Annotator A has annotated the most spans, and most of the spans from the other two annotators overlap with A's. B and C do not overlap as much with each other. The overlaps between the annotator combination (ACD) with the lowest k-alpha is shown in figure 1b. Although the number of overlapping spans between all annotators is greater in figure 1a than in figure 1b, annotator A's spans overlap with more spans individually in figure 2. The other annotator combinations show similar patterns (see appendix A).

| Tag combination | % of total tokens |
|-----------------|-------------------|
| O,O,O | 48 |
| O,O,POS | 16 |
| O,POS,POS | 11 |
| POS,POS,POS | 8 |
| O,O,NEG | 8 |
| NEG,NEG,O | 5 |
| NEG,NEG,NEG | 3 |
| NEG,POS,O | 1 |
| NEG,POS,POS | 1 |
| NEG,NEG,POS | 0.3 |

Table 5: Distribution of tag combinations

If we instead look at the labels in the political tweets corpus, we can see that despite the example of the label changing depending on span length, the agreement is high. About 10% of tokens were annotated with either a positive or negative label, and the observed agreement is 92% and $K-\alpha$ is 0.86. This indicates that the annotators agree on what is negative and positive. The most common disagreement is instead between no label and the positive label, followed by no label and negative. Disagreement over existence or boundaries of positive spans seems to be more difficult than negative spans. This can be seen in table 5. This table shows the distribution of the tag combinations for all tweets which has been seen by three annotators, regardless of annotator identity.

6. Discussion

In comparing our disagreement categories to the categories in Hautli-Janisz et al. (2022) we can find both similarities and differences. Their first

category, annotation errors w.r.t. the guidelines is difficult to compare against since our annotation schemes differ (annotation of spans compared to construction of argumentation graphs). As our guidelines allowed for any span length, we can't consider boundary disagreement as errors. While we do find some annotation errors in our data, they do not seem to be behind the disagreement examined so far. Annotation errors make up most of the disagreement in Hautli-Janisz et al. (2022). Our manual analysis does not look at as many examples as theirs, but it seems like disagreement over boundaries are more frequent.

Our first disagreement category, boundary disagreement, is similar both to the 'fuzziness' and 'ambiguity' category. Hautli-Janisz et al. (2022) distinguishes between the two by defining ambiguity as "those instances where a string yields two fully discrete discourse or argumentative structures" whereas fuzziness relates to language patterns common in natural language such as vagueness which "therefore result in different analyses which themselves are valid, but illustrate the uncertainty in representing partially underspecified or vague language." A disagreement in boundary could result in both separate and similar interpretations. Looking at the reformulations made by the annotators in the online forums corpus, it seems that they do interpret the argumentation similar but slightly different. This would mean that we found more fuzziness than ambiguity.

No matter the type of disagreement, dealing with disagreements require some kind of strategy. As mentioned in section 2.1, analysing and utilizing disagreements in argumentation corpora is usually disregarded in favor for majority vote, or some other aggregation method is used. It would perhaps make more sense, that in order to deal with disagreements one must first know what kind of disagreements there are. If the disagreements are actual annotation errors these should be dealt with accordingly. For example, there are methods for finding unreliable annotators (Hovy et al., 2013; Simpson and Gurevych, 2019).

However, as we have shown examples of here, disagreement in argumentation annotation is not always because of annotation errors but can be due to the possibility of several interpretations or boundaries. A more thorough analysis of the annotations, including both quantitative and qualitative aspects, instead of only relying on standard IAA measures could help identify disagreements. For example, the manual analyses we have shown here found that boundary disagreement wasn't necessarily wrong. A more liberal matching approach in combination with agreement measures could help with resolving and measuring such disagreement. Manual analysis could also identify specific

disagreements like the effect inclusion of negation in a span has on disagreement. This possibly could be solved (or identified) by automatically inverting the negation in the text.

This still leaves the cases where there are different interpretations of the same argumentation, or cases where annotators have annotated varying number of argumentation. Assuming we want to keep all perspectives, we could resolve this by either *weak perspectivism*: creating a gold standard combining all voices in some way, or *strong perspectivism*: using the data from the annotators individually (Cabitza et al., 2023).

7. Conclusion and Outlook

In our examples, we have shown that not all disagreements in argumentation corpora are the same, and that not all of them should be considered disagreements but rather variation or perspectives. In order to determine what kinds of disagreement there are, IAA measures are not enough and a thorough look at the data is needed. This requires methodologies and research about disagreement in argumentation annotation. The development of taxonomies of disagreement specific to argumentation annotation, as in Hautli-Janisz et al. (2022), will also help categorizing disagreement. More research is needed on disagreement in argumentation corpora in order to find further patterns of disagreement or perspectives. An important part of this would be access to more non-aggregated datasets, which would enable more studies across argumentation domains and models. And finally, methods for learning from disagreement, such as soft loss (Uma et al., 2020) or labels (Fornaciari et al., 2021; Wu et al., 2023), is as far as we know a relatively unexplored area for argumentation annotated data and will surely give interesting results when applied.

8. Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

9. Bibliographical References

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.

Ron Artstein and Massimo Poesio. 2008. 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.

Valerio Basile et al. 2020. It’s the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR WORKSHOP PROCEEDINGS*, volume 2776, pages 31–40. CEUR-WS.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. [A taxonomy of argumentation models used for knowledge representation](#). *Artificial Intelligence Review*, 33(3):211–259.

Katarzyna Budzynska, Mathilde Janier, Juyeon Kang, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. Towards argument mining from dialogue. In *Fifth International Conference on Computational Models of Argument*, pages 185–196. IOS Press.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. 2016. [Theoretical foundations for illocutionary structure parsing](#). *Argument & Computation*, (1):91–108.

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

Anca Dumitrache. 2015. Crowdsourcing disagreement for collecting semantic annotation. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 701–710. Springer.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

- Shohreh Haddadan, Elena Cabrio, and Serena Vilata. 2019. [Yes, we can! mining arguments in 50 years of US presidential campaign debates](#). In *Proceedings of ACL 2019*, pages 4684–4690, Florence. ACL.
- Annette Hautli-Janisz, Ella Schad, and Chris Reed. 2022. [Disagreement space in argument analysis](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 1–9, Marseille, France. European Language Resources Association.
- Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. [Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives!](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.
- 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.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- John Lawrence and Chris Reed. 2020. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- 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.
- Anna Lindahl and Lars Borin. 2023. [Annotation for computational argumentation analysis: Issues and perspectives](#). *Language and Linguistics Compass*, 18(1):e12505.
- Anna Lindahl, Lars Borin, and Jacobo Rouces. 2019. [Towards assessing argumentation annotation - a first step](#). In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence. ACL.
- Marco Lippi and Paolo Torroni. 2016. [Argumentation mining: State of the art and emerging trends](#). *ACM Trans. Internet Technol.*, 16(2):10:1–10:25.
- 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.
- 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.
- Julia Romberg. 2022. [Is your perspective also my perspective? enriching prediction with subjectivity](#). In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Sara Rosenthal and Kathy McKeown. 2012. [Detecting opinionated claims in online discussions](#). In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37.
- Edwin Simpson and Iryna Gurevych. 2019. [A Bayesian approach for sequence tagging with crowds](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1093–1104, Hong Kong, China. Association for Computational Linguistics.
- Christian Stab and Iryna Gurevych. 2014. [Annotating argument components and relations in persuasive essays](#). In *Proceedings of COLING 2014*, pages 1501–1510. ACL.
- Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Morgan & Claypool, San Rafael.
- Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena

Villata. 2018. [Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines](#). In *Proceedings of LREC 2018*, pages 4061–4064, Miyazaki. ELRA.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635.

Benedetta Torsi and Roser Morante. 2018. [Annotating claims in the vaccination debate](#). In *Proceedings of the 5th Workshop on Argument Mining*, pages 47–56, Brussels. ACL.

Stephen Edelston Toulmin. 1958. *The use of argument*. Cambridge University Press, Cambridge.

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.

Michiel Van Der Meer, Piek Vossen, Catholijn Jonker, and Pradeep Murukannaiah. 2024. [An empirical analysis of diversity in argument summarization](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2028–2045, St. Julian’s, Malta. Association for Computational Linguistics.

Frans H. van Eemeren. 2017. Rhetoric and argumentation. In Michael J. MacDonald, editor, *The Oxford Handbook of rhetorical Studies*, pages 661–672. Oxford University Press, Oxford.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge.

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. [Don’t waste a single annotation: improving single-label classifiers through soft labels](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Amelie Wüthrl and Roman Klinger. 2021. [Claim detection in biomedical Twitter posts](#). In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. ACL.

10. Language Resource References

Lindahl, Anna. 2020. [Annotating argumentation in Swedish social media](#). ACL.

A. Overlap between annotators

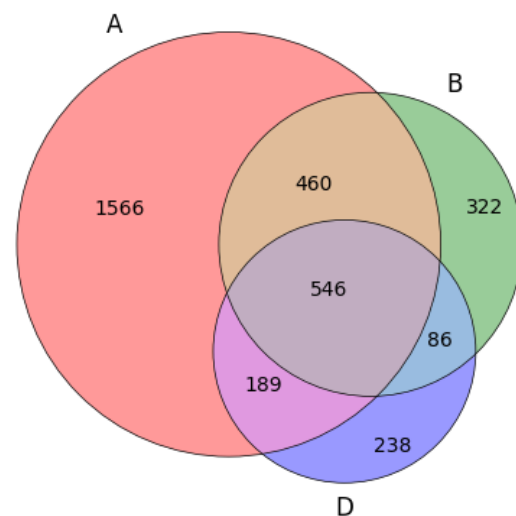


Figure 2: Overlapping spans for annotators ABD

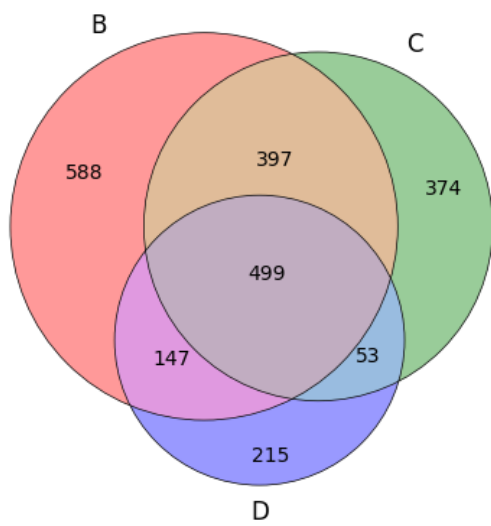


Figure 3: Overlapping spans for annotators BCD

Moral Disagreement over Serious Matters: Discovering the Knowledge Hidden in the Perspectives

Anny Álvarez, Oscar Araque

Universidad Politécnica de Madrid, ETSI Telecomunicación, Intelligent Systems Group
Avenida Complutense, 30, Madrid, 28040, Spain
a.anogales@alumnos.upm.es, o.araque@upm.es

Abstract

Moral values significantly define decision-making processes, notably on contentious issues like global warming. The Moral Foundations Theory (MFT) delineates morality and aims to reconcile moral expressions across cultures, yet different interpretations arise, posing challenges for computational modeling. This paper addresses the need to incorporate diverse moral perspectives into the learning systems used to estimate morality in text. To do so, it explores how training language models with varied annotator perspectives affects the performance of the learners. Building on top of this, this work also proposes an ensemble method that exploits the diverse perspectives of annotators to construct a more robust moral estimation model. Additionally, we investigate the automated identification of texts that pose annotation challenges, enhancing the understanding of linguistic cues towards annotator disagreement. To evaluate the proposed models we use the Moral Foundations Twitter Corpus (MFTC), a resource that is currently the reference for modeling moral values in computational social sciences. We observe that incorporating the diverse perspectives of annotators into an ensemble model benefits the learning process, showing large improvements in the classification performance. Finally, the results also indicate that instances that convey strong moral meaning are more challenging to annotate.

Keywords: moral foundations theory, language models, perspectivism

1. Introduction

The language we use mirrors our thoughts, emotions, values, and cultural background, shaping our interactions with others. The proliferation of online communication platforms and social media has empowered individuals to voice and disseminate their opinions on contentious issues rapidly and to a larger audience. Under these circumstances it is relevant to assess the attitude of individuals towards certain topics of interest. Moral values play an essential role in shaping our decision-making process, particularly when addressing contentious subjects. When dealing with issues such as global warming or political regulations, individuals reference their moral value system, consciously or subconsciously. The Moral Foundations Theory (MFT) has been developed to interpret the concept of morality across diverse cultures (Haidt and Joseph, 2004), outlining five core foundations: *care*, *fairness*, *loyalty*, *authority*, and *sanctity*. The MFT has benefited from refinement with the addition of a sixth foundation: *liberty* (Haidt, 2012).

Despite being recent, the MFT is currently a well-established theory in psychology and the social sciences. Besides, it has found broad acceptance in the field of computational social science due to the creation of a clear taxonomy of values and the development of several computational resources, such as the Moral Foundations Dictionary (MFD) (Graham et al., 2009), which serves

as a central resource for natural language processing applications. The creators of the MFD report some challenges involved in the construction process of such a resource since linguistic, cultural and historical contexts influence language usage.

Attending to the nature of moral values, the MFT has been designed with the idea of harmonizing the variety of moral expressions across different cultures. That is, the MFT models innate foundations that are common to different cultures. Of course, this also means that different cultures and thus, individuals will instantiate the moral foundations differently under the same circumstances. This shows one of the key challenges of generating computing models of the MFT: considering different moral perspectives on the same topic.

While the current datasets and lexicons (Hoover et al., 2020; Trager et al., 2022) do consider the annotations of different individuals, ultimately these annotations are treated in an aggregated manner (i.e., using a voting mechanism) and do not explore the richness introduced by a diverse set of annotators. This lack of understanding of morality computational models introduces a severe bias that can influence individuals (Krügel et al., 2023). Moreover, recent works highlight the necessity of considering a diverse set of annotations simultaneously, without recurring to aggregations that lose relevant information (Cabitza et al., 2023). In light of this, this work explores the information contained within a set of annotators when modeling morality

in an attempt to shed light on such a relevant issue.

Thus, we explore the effect of considering the views from several annotators in an already annotated moral dataset, the Moral Foundations Twitter Corpus (MFTC) (Hoover et al., 2020). In doing so, this paper investigates the impact of training different language models with the perspective of each annotator and then combining these models in an ensemble fashion. Additionally, the task of assessing whether an instance is particularly challenging to annotate is considered, providing further insight into the language usage of this type of text.

To frame the contributions of the paper, we explore the following research questions (RQs). **RQ1: To what extent can the diversity of views in moral annotations be useful for automated moral assessments?** This work examines the variance of the annotations of the MFTC, training different language models with different annotations. Using these trained models, we explore the effect of this additional knowledge in the framework of automatically estimating morality in text.

Following, we also inspect **RQ2: Is it possible to automatically assess whether a text is challenging to annotate?** This question reflects on the characteristics of texts where annotators diverge in their ratings, offering a basis on which we can understand the difficulties of evaluating moral foundations. In this sense, this paper evaluates the performance of several models in the task of predicting whether a text is challenging to annotate, using the disagreement that the annotators of the MFTC have shown.

The rest of the paper is structured as follows. Section 2 describes the fundamentals of the Moral Foundations Theory (MFT) and how it has been previously addressed from a computational perspective. Section 3 presents the data and methodology used in this work. Next, the experimentation is detailed in Section 4. Finally, the conclusion and future work is delineated in Section 5.

2. Background

In this section, we summarize key concepts and methodologies for our research. First, we explore the Moral Foundations Theory (MFT), which represents the underlying principles that influence human moral judgments in diverse cultural contexts and resources such as the Moral Foundations Dictionary (MFD). We also discuss the application of these resources in computational models, including the use of prompts, which has demonstrated the potential to enhance the comprehension and generation of texts.

2.1. Moral Foundations Theory

Previously, it has been mentioned that the Moral Foundations Theory (MFT) describes, through the definition of several foundations, common axes to measure morality across diverse cultures and sensibilities. In this work, we study the five basic foundations (Haidt and Joseph, 2004). *Care/harm*: This foundation relates to our capacity to empathize with and perceive the pain of others. It encompasses virtues such as kindness, gentleness, and nurturance. *Fairness/cheating*: This foundation underscores the virtues of justice and rights. *Royalty/betrayal*: manifests the principles of solidarity. It embodies virtues like patriotism and willingness for group-oriented self-sacrifice. *Authority/subversion*: This foundation emphasizes virtues associated with leadership and followership. It entails deference to esteemed authority figures and reverence for traditional norms. *Purity/degradation*: This foundation emphasizes aspirations for elevated living, often found in religious narratives. It encompasses virtues of self-discipline, self-improvement, naturalness, and spirituality.

We have already covered that one of the main reasons the MFT has become so popular in computational social sciences is the development of the Moral Foundations Dictionary (MFD) (Graham et al., 2009). This lexical resource, based on the known Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), covers a basic annotation of lemmas and how they convey meanings toward the moral foundations. While this resource has been crucial for the development of computational models of morality, it is not without limitations. Among the notable limitations of the MFD are: (i) a limited number of tokens; (ii) inclusion of “radical” lemmas seldom encountered in everyday language, such as “homologous” and “apostasy”; and (iii) classification based on a moral bipolar scale denoting vice and virtue, lacking any indication of “strength.”

Concerning the dataset we use in this work, the Moral Foundations Twitter Corpus has been used in several scientific studies and natural processing tasks whose work is based on the MFT (Graham et al., 2013) and the MFTC as a reference to evaluate distinct moral narratives in natural language texts. On one hand, this was used to study how a moral lexicon (Araque et al., 2020) can be exploited at the document level using different machine learning and engineering techniques, obtaining better results in the detection of morality in text. On the other hand, Guo et al. (2023) propose a refinement model that uses Sentence-BERT embeddings to capture moral information, investigating the performance, generalisation and transferability of moral embeddings with a specific focus on how these

embeddings can improve the accuracy of moral classifiers. Finally, [Liscio et al. \(2022\)](#) perform an extensive investigation on the effects of cross-classification of moral values in text, comparing a deep learning model on seven different domains.

2.2. Prompts for inserting knowledge

The utility of pre-trained language models for a large variety of natural language processing applications is clear due to its success and popularity ([Han et al., 2021](#)). In this regard, language models show characteristics in their internal representations and behaviors that indicate that they are capable of generating a depiction of moral concepts ([Scherrer et al., 2023](#); [Fitz, 2023](#)). For example, it has been found that language models' internal representations induce a moral dimension that, in principle, could be utilized by the model ([Fitz, 2023](#)). We argue that this kind of morality knowledge can be exploited to assess moral values in text.

Following on the previous, one common method to control the output of a language model is to steer their generation process through *prompts*. Prompts are instructions or fragments designed to guide the model during the performance of a specific task. Although this approach has not been previously used in the context of moral values assessment, we build on the evidence of positive results obtained in other tasks using pre-trained models such as BERT ([Luo et al., 2022](#)).

For an comprehensive review on the use of prompts, please consult the work of [Liu et al. \(2023\)](#).

3. Data and Methods

As described, this work pursues to gain insights into how an already annotated dataset can be used to characterize different perspectives in the process of annotating moral values in text. This section describes the dataset used in the experimentation (Sect. 3.1) and the methods designed to explore the knowledge of the annotations (Sect. 3.2).

3.1. Dataset

To perform the experiments detailed in Section 4, we have used the Moral Foundations Twitter Corpus dataset ([Hoover et al., 2020](#)) and its corresponding annotations. It is structured into seven subsets of data, each addressing distinct and socially relevant discursive topics. The corpus has been labelled by various annotators; it is composed of a considerable size of tweets (approximately 35 thousand) and a diversity of ideas in various social movements, from politics and human rights to natural disasters. These aspects provide a comprehensive

view of how morality is reflected in different social media, thus making it a benchmark for machine learning tasks such as multi-labelled morals.

Originally, the dataset consists of seven different subsets that contain Twitter messages pertaining to different societal issues: All Lives Matter (ALM), Black Lives Matter (BLM), Baltimore, Davidson, Election, MeToo Movement (MT) and Sandy. We work with 6 of them, which are available online¹. These are the following: All Lives Matter (ALM), related to 'All Lives Matter' Movement; Black Lives Matter (BLM), related to 'Black Lives Matter' Movement; Baltimore, related to the Baltimore protest following the death of Freddie Gray in US; Davidson, texts collected by [Davidson et al. \(2017\)](#) for hate speech and offensive language research; Election, tweets about the 2016 US presidential election; and Sandy, related to Hurricane Sandy in 2012.

This set of human-annotated English tweets has labels of moral foundations in 10 classes distinguishing between vice and virtue for each moral trait, including a 'non-moral' class. Tweets were tagged following the MFT, described in Section 2.1, and each domain was evaluated by at least three trained annotators as set out in the original labeling guide ([Hoover et al., 2017](#)), which has been designed as a comprehensive manual that establishes common practices and clear guidelines for the identification of moral sentiments expressed in texts. Despite the training given to annotators, the authors put emphasis on the use of personal views even if they diverged from common values, increasing the variety in the annotations. Each tweet was therefore labelled with an indication of the presence or absence of each virtue and vice or using a 'non-moral' label.

In this study, a basic pre-processing and subsequent tokenization has been carried out to the data, as required by this type of transformer model. Numbers, punctuation marks, symbols, usernames, URLs, and emoticons were removed, and stop-words were preserved. The final label for each text was obtained by aggregating the labels of several annotators using the majority vote as the true class, resulting in the distribution of morality found in each dataset and reflected in Table 1.

To assess the overview of different annotators, we set each annotator's label to the corresponding text the person had annotated. Table 5 shows the final distribution of labels per annotator.

One observable concern is the imbalance towards the 'non-moral' class, where in Davidson and Baltimore cases, they are approximately 90% of the total. Although we use the original data to take advantage of the largest dataset, these limitations were taken into account when analyzing the results and reflecting on the conclusion.

¹<https://osf.io/k5n7y/>

| Dataset | C/H | F/C | L/B | A/S | P/D | NM |
|-----------|-------|-----|-------|-------|-----|-------|
| ALM | 1,314 | 723 | 408 | 274 | 182 | 585 |
| BLM | 1,048 | 934 | 528 | 491 | 253 | 1,040 |
| Baltimore | 434 | 292 | 895 | 120 | 37 | 2,366 |
| Davidson | 447 | 130 | 319 | 1,039 | 118 | 2,784 |
| Election | 798 | 736 | 286 | 177 | 349 | 2,019 |
| Sandy | 708 | 708 | 1,010 | 519 | 560 | 291 |

Table 1: Distribution of foundations presence in all data domains. The column names are encoded as follows. C/H: care/harm, F/C: fairness/cheating, L/B: loyalty/betrayal, A/S: authority/subversion, P/D: purity/subversion, NM: non-moral.

3.2. Methodology

To satisfy the research questions previously raised (see Sect. 1), this work studies (i) how the information of the disagreement among annotators can be exploited, as well as (ii) the characteristics of what constitutes an instance prone to be subject to disagreement.

For all experiments, we have used Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018)² as the base model. Given the unbalance of the data labels and its effect on the performance in the classification tasks, all results are reported using the macro-averaged F-score.

Regarding the model specifications, it was used the pre-trained BERT model `bert-base-uncased` along with its corresponding tokenizer. Each model was trained for 15 epochs, using a batch size of 32 and learning rates of 0.01 and 2e5 respectively.

Diversity exploitation. Regarding the first challenge, this work proposes an evaluation that probes the utility of understanding the moral views of the different annotators. In this regard, we first assess the variety of the annotators by training a model that predicts the moral of the text as judged by each annotator. As Figure 1 illustrates, we fine-tune a different instance of the same model using as training labels the annotations expressed by each annotator. In this way, we intend that each captures the particularities and views of each annotator. Additionally, we evaluate the classification performance of each of these models, which can offer further insights into the consistency of the annotations.

Following, a supplemental evaluation is done. To predict the aggregated label of each data instance, we use the previously fine-tuned models trained on the specific annotations of each annotator and the corresponding text.

²<https://huggingface.co/google-bert/bert-base-uncased>

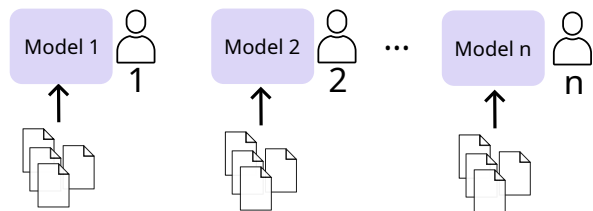


Figure 1: Fine-tuning procedure where models are trained with the specific annotations of n different annotators.

To carry out this evaluation, we decided to explore the use of a prompt-based approach adding the predictions of each fine-tuned model as additional information alongside the original text. Then, a second training was performed using the enriched dataset to analyse how it contributed to the performance of the model in the classification task. This is shown in Figure 2.

The choice of this strategy is based on the proven effectiveness of these models in natural language understanding and leveraging the ability to capture semantics, incorporating multiple perspectives through the predictions of morals provided by different annotators. We believe that this approach could provide a more complete and refined view of the moral dimensions present in the data, which in turn could improve the performance of models on the moral classification task.

Feeding the model in a consistent way with the perspectives of each annotator enriches the dataset by providing it with additional information about each text, especially about the different model perspectives it may contain. Taking into account the limitation of choosing the prompt template manually due to the numerous possibilities and choosing the one that maximises the performance of the model, a structure has been used that reflects as clearly as possible that the additional information conveys the view of different annotators.

During the evaluation of diversity explained above, the predictions of each model were used for each data instance and annotator. These predictions were added to the standardised prompt at the input, following the structure: *'The text {...} has been annotated by different annotators with the following moral values $\{m_1, m_2, \dots, m_n\}$ '*, where {...} is the original input and $\{m_1, m_2, \dots, m_n\}$ is a concatenation of the annotations for the text.

By providing these, we can better align the predictions with the characteristics and evaluation features of each text, improving the accuracy and consistency in the prediction of the aggregated labels. Once the new inputs were obtained, the training of the BERT model was performed, and the results were compared with the base training.

Thereby, we propose that having an overview

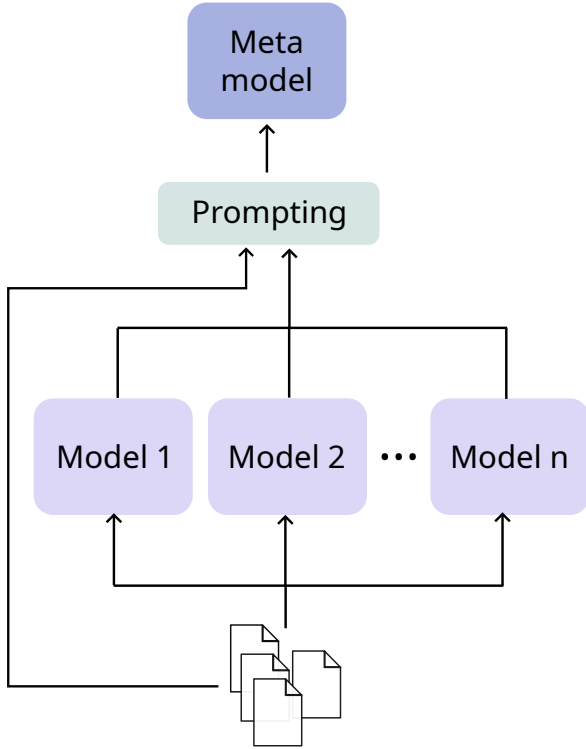


Figure 2: Proposed ensemble method that combined the predictions on the different perspectives of the base models with the textual input through a prompt approach.

of each of the annotator’s subjective perspectives can aid in the overall estimation of morality in text. To assess this, this work compares the ensemble method to the baseline of estimating morality using solely the text. These models, as shown in Figures 1 and 2, are thoroughly evaluated in Section 4.1.

Disagreement estimation. To address the second research question, we propose the use of a learning model to assess whether a text is challenging to annotate. This task is oriented to exploit the information inherent in the disagreement among annotators, where some instances will show a high agreement and other instances’ content will be harder to annotate. This proposition follows the ideas presented by (Basile, 2020) in the sense that it is an attempt to consider all different perspectives contained within the original annotations in a machine learning setting.

In this way, we fine-tune a different instance of a BERT model for each dataset, having as label the level of disagreement of the data instance. To facilitate the analysis, we have considered a binary approach so that each instance can be considered as either challenging to annotate (i.e., that shows a high disagreement among annotators) or not. Thus, in this proposal, the learning models perform a bi-

nary classification task: given a document, predict whether the instance is challenging to annotate.

To assess if a given instance is positive or negative under the mentioned distinction, we define a *divergence* metric that allows us to encode this idea of agreement among annotators. More formally, consider a set of annotations for a given data instance $A = \{a_1, a_2 \dots a_N\}$; we then define a measure of agreement among annotators. Thus, the agreement for annotator i is defined as:

$$g_i = \frac{1}{N} \sum_{i \neq j} a_i == a_j \quad (1)$$

where N is the number of annotations. The $==$ operation returns a value of 1 if $a_i = a_j$, and a value 0 otherwise. Naturally, g_i encodes the number of times that annotator i agrees with the rest of the annotators for that data instance. Thus, $g = \{g_1, g_2, \dots, g_N\}$ Following, we define the divergence metric as the opposite of the previous:

$$d = 1 - \frac{g}{\max(g)} \quad (2)$$

where $d \in [0, 1]$. The closer d is to 0, the less divergent the instance (i.e., the more agreement among annotators); conversely, the closer d is to 1, the higher disagreement in the annotations we observe. Therefore, we utilize the divergence metric d as a measure to identify instances that are challenging to annotate.

Since we are modeling the problem through a binary approach, a threshold concerning the divergence metric has been defined. Thus, we consider an instance to be challenging to annotate if $d \geq d_{th}$. Section 4.2 describes how this threshold has been estimated.

Finally, to study the characteristics of the language in documents that have diverging annotations, we use the SHapley Additive exPlanations (SHAP) method (Lundberg and Lee, 2017). Such a method assigns an importance score to each of the features considered for an specific prediction. These SHAP values allow us to inspect the learning models trained, inspecting how the language affects the decision on the disagreement of a document.

To perform this analysis, we extract the SHAP values of all models trained, aggregating them to obtain a whole overview of the classification process. To do so, we extract the SHAP values for all words in all documents, aggregating them into a set of values for each word considered.

These evaluations, which address the estimation of disagreement, are described in Section 4.2s.

4. Experimentation

In this section, we present the results obtained. Concretely, Section 4.1 focuses on both the individual performance of the fine-tuned models according to different annotators and the impact of using these predictions as additional knowledge for morality prediction. Following, Section 4.1 describes the analysis done on the modeling of agreement among annotators.

4.1. Annotation diversity exploitation

Firstly, Table 2 presents the results of the performance evaluation of the models on each dataset and for each annotator, comparing them to the baseline results.

| Dataset | Annot. | Baseline | F1-Score |
|-----------|--------|----------|----------------|
| ALM | 00 | 64.71 | 13.46 (-51.24) |
| | 01 | | 60.52 (-04.18) |
| | 02 | | 26.22 (-38.48) |
| | 03 | | 79.35 (+14.64) |
| BLM | 00 | 85.46 | 79.44 (-06.01) |
| | 01 | | 79.38 (-06.07) |
| | 02 | | 37.94 (-47.51) |
| | 03 | | 83.05 (-02.40) |
| | 04 | | 83.55 (-01.90) |
| Baltimore | 02 | 42.58 | 40.17 (-02.40) |
| | 13 | | 39.59 (-02.98) |
| | 14 | | 49.54 (+06.96) |
| Davidson | 05 | 15.84 | 15.21 (-00.62) |
| | 06 | | 15.50 (-00.33) |
| | 07 | | 14.64 (-01.19) |
| Election | 00 | 61.11 | 58.40 (-02.70) |
| | 02 | | 32.24 (-28.86) |
| | 03 | | 65.57 (04.46) |
| | 04 | | 70.81 (09.70) |
| Sandy | 09 | 55.73 | 56.58 (00.85) |
| | 10 | | 52.73 (-02.99) |
| | 11 | | 48.49 (-07.23) |

Table 2: Results of the classification performance in predicting the moral as judged by the different annotators.

It can be observed that there is significant variability in performance across different datasets and between different annotators. One relevant observation is that better results are found in the cases where there is a more balanced distribution of classes. Additionally, we argue that the interpretation of moral values may depend significantly on the context and domain of the text, which can influence the consistency and accuracy of different annotator’s labels.

In general, the results only diverge slightly from the baseline results, except for annotator 00 in the ALM dataset, where it performs much worse. The

pronounced disparities observed in some cases are mainly due to the amount of data labelled by these annotators. An insufficient number of examples prevents the model from accurately learning and predicting the labels assigned by these annotators.

The lowest metric values are observed in the Davidson dataset. This is likely due to class imbalance and subjectivity in the interpretation of moral values in this specific context. In the Davidson case, approximately 60% of the labelled data was identified as ‘non-moral’. For more details on the class distributions for each annotator, see Table 5.

Finally, as reflected in Table 3, in terms of the model’s performance when using prompts, a significant improvement in the classification performance was obtained in all domains compared to the baseline model without prompts. This suggests that the choice of prompt and additional information on different perspectives can influence and improve the results.

The incorporation of this additional information has effectively provided more contextual cues, allowing the model to better understand and classify morality in different texts across various domains. Moreover, the observed improvements in F1 scores highlight the effectiveness of leveraging diverse perspectives from annotators. By adding these into the training process, the model becomes more efficient at recognizing moral nuances present in texts. However, it’s remarkable that while the prompt-based approach has led to considerable enhancements, certain domains, like Davidson, still present challenges for accurate classification. Overall, the success of using prompts underscores the significance of contextual information and diverse perspectives in morality estimation tasks.

| | F1-score | |
|-----------|----------|-----------|
| | Baseline | Prompting |
| ALM | 64.71 | 88.74 |
| BLM | 85.46 | 95.82 |
| Baltimore | 42.58 | 76.32 |
| Davidson | 15.84 | 66.03 |
| Election | 61.11 | 88.22 |
| Sandy | 55.73 | 86.44 |

Table 3: Evaluation of the addition of different perspectives in training. The F1-Score results are compared with baseline results in all domains.

4.2. Disagreement estimation

As described in Section 3.2, we study the nature of the disagreement among annotators by training a learning model to predict whether a given text is challenging to annotate. By approaching the issue in this manner, we are operating on the basis that

annotators diverge in their annotations driven by certain characteristics of the texts they are annotating.

Firstly, we have defined a threshold d_{th} on the divergence metric that allows us to distinguish whether a text is challenging to annotate. Figure 3 shows the evolution of the percentage of positive instances, that is, instances that show a divergence metric where $d > d_{th}$. Based on the distributions of the d metric along all datasets, we manually set this threshold to $d_{th} = 0.7$. As can be seen, the majority of the distributions in the figure suffer an abrupt decline when the threshold is at the indicated number.

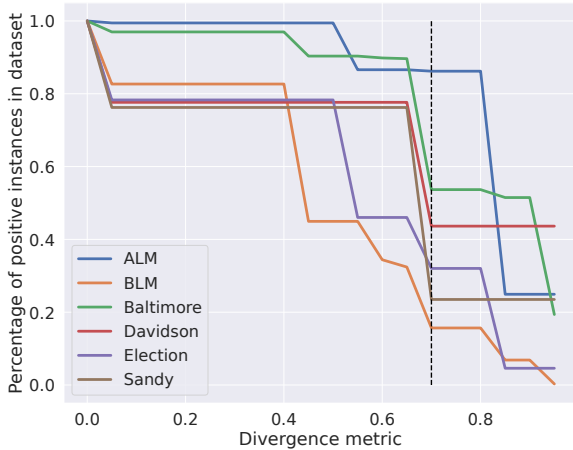


Figure 3: Percentage of instances considered to be challenging (vertical axis) to annotate with the divergence metric (horizontal axis).

Following this, fine-tuning and evaluation of the learning models has been performed. We have trained a different instance of the same model for each of the data domains in an attempt to capture the specific characteristics of each domain. To avoid the negative effect of the imbalance of the two classes considered, we balanced the resulting data by randomly sampling the majority class. The results of such an experiment are shown in Table 4, including the number of derived instances for each data domain.

It is clear, attending to the results, that in some cases the classifier is able to distinguish the divergent instances (i.e., the instances where annotators show a higher divergence metric). These cases include the BLM, Baltimore, and Election domains, with the highest performance metrics. In contrast, in the ALM, Davidson, and Sandy domains, the classifiers are not able to properly discern the divergence of the data instances, although for ALM and Sandy the f-score reaches 58%. This dissimilar behaviour among domains is a consistent result: as studied previously by [Liscio et al. \(2023\)](#), the differences in the domains of the Moral Founda-

| | Acc. | F1-score | Neg. inst. | Pos. inst. |
|-----------|-------|----------|------------|------------|
| ALM | 58.55 | 58.36 | 94 | 99 |
| BLM | 68.94 | 68.49 | 120 | 115 |
| Baltimore | 79.26 | 79.25 | 402 | 355 |
| Davidson | 48.17 | 47.75 | 442 | 403 |
| Election | 71.61 | 71.39 | 272 | 288 |
| Sandy | 58.82 | 58.81 | 180 | 177 |

Table 4: Evaluation in the task of predicting whether a text is challenging to annotate with morality. Accuracy, macro averaged F-score, and the number of negative and positive instances are reported.

tions Twitter Corpus (MFTC) do affect the quality of prediction tasks.

Overall, these positive results are a clear indication that there are language cues that indicate to the learners whether a text is prone to be challenging to annotate. Since these language signals are sure to vary with the domain of annotations, we seek to gain a better understanding of this process. To do so, as described in Section 3.2, we use SHAP to inspect how the learners analyse the text in terms of divergent annotations. In this study, we have aggregated the SHAP values from all data domains, as we aim to obtain a general view of this process rather than a specific examination of each domain’s particularities.

Figure 4 shows the results obtained from a selection of the tokens that have the highest relevance for either the negative or positive classes. Tokens with negative SHAP values are relevant for detecting the negative class (i.e., instances that show low disagreement), while tokens with positive SHAP values are related to detecting the positive class, where the disagreement is higher.

We observe that the tokens with negative SHAP values are generally words with semantics not pertinent to morality and innocuous in terms of societal or cultural issues. Interesting examples of these terms are *photo*, *wonderful*, *green*, *internet* or *babies*. This is an intuitive result since annotators will generally agree within texts that do not convey a strong moral or cultural position. In contrast, tokens with positive SHAP values tend to express strong moral significance. Some examples of these words are *democrats*, *evil*, *god*, *duty*, *racism*, *homo* (from homosexuality), and *respect*. Again, this can be explained if we consider that annotators will disagree more frequently when assessing documents that include morally and culturally stronger positions. Interestingly, some tokens with higher positive SHAP values revolve around polemic or even harmful matters such as religion, sexual practices, and racism.

To better understand the insights obtained by this last study, we include some interesting exam-

ples of texts that show the characteristics found through the SHAP analysis. For instance, the following text, contained in the All Lives Matter (ALM) dataset, “*#blacklivesmatter is for unity equality respect between races all lives matter ignores the truth of injustice to claim reverse racism*” has been annotated with the foundations care, loyalty and fairness by the different annotators, which indicate that the annotators have identified different foundations in the text, although all of them are virtues as defined in the MFT.

Another instance, extracted from the Sandy domain, is as follows: “*Sandy is god’s way of saying ignoring climate change is equal to saying you are willing to destroy my creation*”. This text has been annotated with the foundations of authority, purity, and fairness. While the purity and authority annotations probably reference the religious content, the debatable fairness annotation may relate to a sense of divine justice, alluded to in the original message.

5. Conclusion

This paper explores the effect of diverse human annotations in the context of computationally modeling moral foundations through the Moral Foundations Theory. Under the lenses of perspectivism (Cabitza et al., 2023)³, we explore a known dataset in the field of moral value estimation, the Moral Foundations Twitter Corpus. This dataset contains annotations from different annotators that are commonly aggregated. This work investigates the effect of separately considering the perspectives of the annotators toward morality.

Concretely, we raise two research questions (RQs) that are thoroughly studied in this work. Firstly, RQ1 inspects the effect of exploiting the diversity of annotators’ perspectives for automated moral estimation. In this regard, we have shown that the different annotators do highly impact the quality of the predictions if taken in isolation. Attending to this, it is clear that the diversity of annotators and domains are variables to take into account when generating new data repositories. In contrast, the experiments show notable and consistent improvements in the classification performance when adding the predictions of models trained to estimate individual annotators’ perspectives into an ensemble model. Such a positive result motivates future research on harnessing diverse perspectives into learning systems.

Secondly, RQ2 proposes the task of estimating whether a data instance is challenging to annotate. That is, if an instance generates disagreement among annotators. Through this task, we intend to

³The perspectivist data manifesto: <https://pdai.info/>.

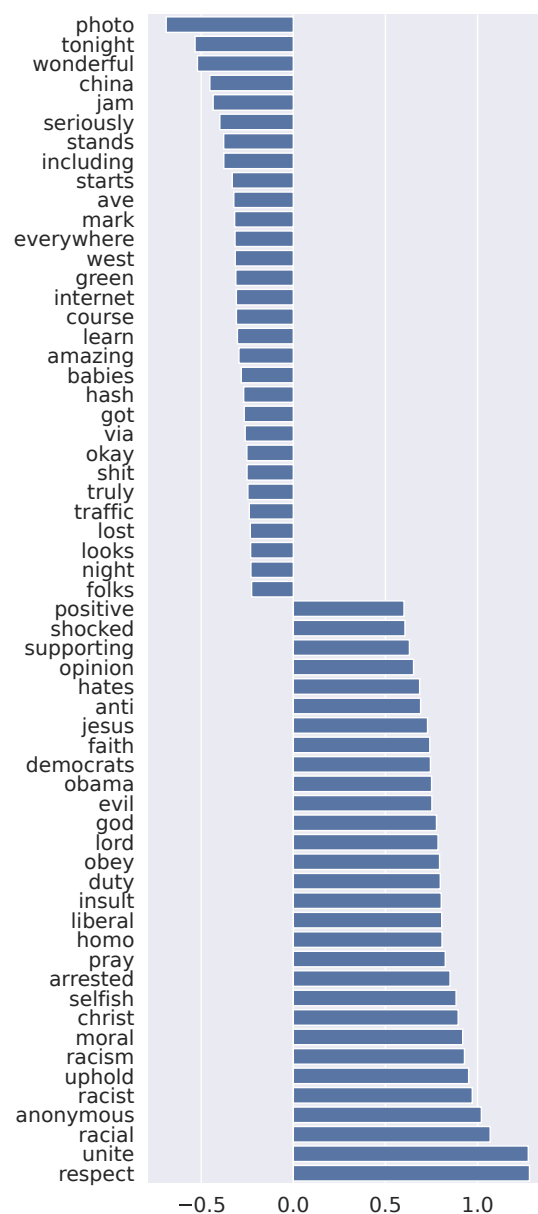


Figure 4: SHAP values of interesting tokens. Positive values indicate relevance towards the positive class, while negative values indicate otherwise.

analyse the linguistic cues that indicate disagreement factors. The experiments show that the ability to estimate disagreement can achieve high performance scores but varies across domains, indicating considerable variance. By doing a subsequent analysis using SHAP values, we have discovered that the disagreement instances tend to contain strong moral, political, or cultural meanings. On the contrary, instances where annotators typically agree normally contain more neutral language.

Addressing the limitations of the work, we evaluate the ensemble method using an aggregated label for moral values. Oddly, this challenges one of the principles of the perspectivism movement,

which states that traditional *golden* labels should be avoided, thus taking into account the diversity of views from annotators. This part of the proposed evaluation does simplify the challenge of moral estimation for the ensemble method due to the large complexity involved in designing a model that predicts over such a substantial set of target labels (i.e., all possible combinations of moral foundations for each of the annotators). Future work should tackle this issue by modeling the prediction objective more tractable.

Another limitation of the work is related to our definition of what constitutes a divergent instance. We have defined a straightforward metric that aids in defining a learning problem related to disagreement. In this regard, future work should investigate this direction, further defining the divergence of annotated documents and how we can handle them.

6. Acknowledgements

This project has been funded by the project UNICO I+D Cloud - AMOR, financed by Ministry of Economic Affairs and Digital Transformation, and the European Union through Next Generation EU.

7. Appendix

Table 5 describes the distributions of moral annotations for each annotator and data domain.

After a first analysis of the different annotations in the texts, it was observed that there was a disparity in the amount of data labelled by each annotator. In order to ensure a correct comparison, we initially used the intersection of instances annotated by all the annotators. However, this strategy faced the challenge of dealing with very small datasets due to annotators with minimal contributions. Thus, to overcome this problem four annotators from different domains were removed.

For ALM dataset, annotator00 was excluded because only 94 instances were labelled, which is a considerable lower proportion in comparison to the 3486 instances from the original dataset. In the case of Baltimore dataset, annotator12 and annotator15 were also discarded for their low contribution. Finally, in Davidson dataset, annotator08 was removed because their annotations consisted in 1 instance.

Removing these annotators was done to prevent the datasets from being too small and negatively impacting the training process. In the case of Baltimore dataset, when all the annotators were considered, the data was reduced from 4144 examples to 402, resulting in significant missing data and poor metrics in performance. Excluding annotator 12 and 15 results in a large dataset formed by the

intersection of 3528 examples, leading to better model performance.

8. Bibliographical References

- 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.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Stephen Fitz. 2023. Do large gpt models discover moral dimensions in language representations? a topological study of sentence embeddings. *arXiv preprint arXiv:2309.09397*.
- Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.
- Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 281–291.
- Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.
- Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

| Dataset | Annotator | C/H | F/C | L/B | A/S | P/D | NM |
|-----------|-----------|------|-----|-----|------|-----|------|
| ALM | 00 | 48 | 26 | 7 | 4 | 4 | 5 |
| | 01 | 1252 | 714 | 403 | 266 | 181 | 524 |
| | 02 | 1299 | 718 | 404 | 273 | 179 | 579 |
| | 03 | 1312 | 722 | 408 | 273 | 182 | 582 |
| BLM | 00 | 881 | 774 | 448 | 435 | 222 | 858 |
| | 01 | 906 | 784 | 458 | 441 | 236 | 856 |
| | 02 | 910 | 787 | 452 | 443 | 228 | 870 |
| | 03 | 924 | 795 | 457 | 446 | 236 | 874 |
| | 04 | 923 | 793 | 456 | 447 | 234 | 864 |
| Baltimore | 02 | 390 | 267 | 806 | 115 | 35 | 2054 |
| | 12 | 102 | 84 | 130 | 33 | 4 | 599 |
| | 13 | 423 | 281 | 855 | 113 | 34 | 2313 |
| | 14 | 426 | 286 | 884 | 114 | 36 | 2280 |
| | 15 | 61 | 60 | 100 | 30 | 3 | 276 |
| Davidson | 05 | 435 | 113 | 290 | 957 | 118 | 2705 |
| | 06 | 396 | 121 | 294 | 1036 | 95 | 2121 |
| | 07 | 150 | 38 | 161 | 174 | 74 | 1880 |
| | 08 | 1 | 0 | 0 | 0 | 0 | 0 |
| Election | 00 | 728 | 667 | 263 | 161 | 335 | 1991 |
| | 02 | 736 | 673 | 265 | 165 | 327 | 1922 |
| | 03 | 791 | 733 | 283 | 175 | 345 | 2004 |
| | 04 | 791 | 734 | 285 | 177 | 347 | 1951 |
| Sandy | 09 | 701 | 702 | 990 | 517 | 554 | 285 |
| | 10 | 706 | 701 | 989 | 514 | 553 | 289 |
| | 11 | 703 | 705 | 992 | 509 | 540 | 290 |

Table 5: Distribution of foundations by annotator. The column names are encoded as follows. C/H: care/harm, F/C: fairness/cheating, L/B: loyalty/betrayal, A/S: authority/subversion, P/D: purity/subversion, NM: non-moral.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. 2023. Chatgpt’s inconsistent moral advice influences users’ judgment. *Scientific Reports*, 13(1):4569.

Enrico Liscio, Oscar Araque, Lorenzo Gatti, Ionut Constantinescu, Catholijn Jonker, Kyriaki Kalimeri, and Pradeep Kumar Murukanniah. 2023. [What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14113–14132, Toronto, Canada. Association for Computational Linguistics.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukanniah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural lan-

guage processing. *ACM Computing Surveys*, 55(9):1–35.

Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Xianchang Luo, Yinxing Xue, Zhenchang Xing, and Jiamou Sun. 2022. Prcbert: Prompt learning for requirement classification using bert-based pretrained language models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. [Evaluating the moral beliefs encoded in llms](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 51778–51809. Curran Associates, Inc.

Yla R. Tausczik and James W. Pennebaker. 2010. [The psychological meaning of words: Liwc and computerized text analysis methods](#). *Journal of Language and Social Psychology*, 29(1):24 – 54.

9. Language Resource References

- Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. [Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction](#). *Knowledge-Based Systems*, 191:105184.
- Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.
- Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.
- Joseph Hoover, Kate Johnson-Grey, Morteza Dehghani, and Jesse Graham. 2017. Moral values coding guide. *PsyArXiv*.
- Jackson Trager, Alireza S Ziabari, Aida Mostafazadeh Davani, Preni Golazizian, Farzan Karimi-Malekabadi, Ali Omrani, Zhihe Li, Brendan Kennedy, Nils Karl Reimer, Melissa Reyes, et al. 2022. The moral foundations reddit corpus. *arXiv preprint arXiv:2208.05545*.

Perspectives on Hate: General vs. Domain-Specific Models

Giulia Rizzi*[†], Michele Fontana*, Elisabetta Fersini*

* University of Milano-Bicocca, Milan, Italy,

[†] Universitat Politècnica de València, Valencia, Spain

{g.rizzi10, m.fontana36}@campus.unimib.it, elisabetta.fersini@unimib.it

Abstract

The rise of online hostility, combined with broad social media use, leads to the necessity of the comprehension of its human impact. However, the process of hate identification is challenging because, on the one hand, the line between healthy disagreement and poisonous speech is not well defined, and, on the other hand, multiple socio-cultural factors or prior beliefs shape people's perceptions of potentially harmful text. To address disagreements in hate speech identification, Natural Language Processing (NLP) models must capture several perspectives. This paper introduces a strategy based on the Contrastive Learning paradigm for detecting disagreements in hate speech using pre-trained language models. Two approaches are proposed: the General Model, a comprehensive framework, and the Domain-Specific Model, which focuses on more specific hate-related tasks. The source code is available at <https://github.com/MIND-Lab/Perspectives-on-Hate>.

Keywords: Hate Speech, Disagreement, Contrastive Learning

1. Introduction

With the widespread use of social media, the opportunities to share people's experiences and opinions have grown rapidly. As a consequence, hatred on social media is growing accordingly, with people sharing hateful content towards various targets and minorities. To ensure the continued shared of knowledge and ideas and improve individual and social well-being in the online environment, it is critical to understand the potential harm that hate content can cause on a human level. However, as people use online forums and Social Media to express themselves and engage in debate, the distinction between healthy disagreement and toxic speech becomes increasingly blurred. Moreover, individuals' susceptibility to objectionable content is substantially influenced by their cultural beliefs and origins, emphasizing the importance of considering various perceptions (Sang and Stanton, 2022; LaFrance and Roberts, 2019; Sap et al., 2021). Addressing disagreement, especially in the context of hate speech identification has received more attention in recent years. Nevertheless, the development of Natural Language Processing (NLP) models capable of completely capturing and representing diverse perspectives is critical. Various approaches have been proposed to address disagreements in hate speech identification, and explored the area of perspectivism (Akhtar et al., 2021; Sachdeva et al., 2022; Uma et al., 2021). According to recent studies, it may be beneficial to consider the exploration of more elaborate and established techniques, such as integrated gradients or uncertainty quantification (Astorino et al., 2023; Davani et al., 2022; Rizzi et al., 2023). The identification

of disagreements among hateful statements and the identification of disagreement-related aspects would lead to more reliable benchmarks. Moreover, it would allow the definition of specific annotation policies (e.g., adding more annotators, removing samples from the dataset that need annotation, etc.) to be adopted for contents that are likely to cause disagreement among readers. In this paper, we exploit the Contrastive Learning paradigm to predict Disagreement in hateful content. In particular, we exploit pre-trained large language models for hate speech detection and leverage the embedding representation derived from this model to accurately predict disagreement among annotators. We propose two different approaches with distinct characteristics:

- **General Model:** a comprehensive approach, combining multiple tasks (e.g. aggressive, offensive, and abusive language detection) under the umbrella of hate speech identification (Poletto et al., 2021). This inclusive viewpoint enables the model to effectively capture the subtle manifestations of hate across multiple linguistic dimensions and different languages, resulting in a more robust and versatile solution for identifying and treating various forms of harmful text.
- **Domain-Specific Model:** The Domain-Specific Model represent a more refined approach, focusing solely on elements that share specific characteristics. This approach focuses on instances of the same hate-related task that share homogenous aspects such as language, type of text, and hate target, recognizing the close relationship between those

characteristics and annotator disagreement on hate speech.

The paper is organized as follows: Section 2 provides an overview of the state of the art. Section 3 describes the adopted datasets. Section 4 digs into the specifics of the proposed approach. The obtained results are presented in Section 5. Finally, Section 6 summarizes the findings of this study and outlines future investigations.

2. Related Works

Over the years, significant progress has been made in the development of automatic hate content detection systems, exploiting advances in Natural Language Processing (NLP), machine learning, large language models, and deep learning technologies (Mozafari et al., 2020; Alatawi et al., 2021; Saleh et al., 2023). However, hate speech detection, like many natural language tasks, is characterized by intrinsic ambiguity or subjectivity (Uma et al., 2021). These characteristics have led to datasets with multiple annotations that incorporate varied annotator perspectives and understandings or with confidence levels associated with labels. The representation of annotators’ disagreement has found utility in three ways: (i) to enhance the quality of the dataset by removing instances marked by annotator disagreement (Beigman Klebanov and Beigman, 2009), (ii) to weight instances during training aiming at prioritizing those with higher confidence levels (Dumitrache et al., 2019), or (iii) to directly train a machine learning model from disagreement without considering aggregated labels (Uma et al., 2021; Fornaciari et al., 2021). While prior research focused on utilizing disagreement information, limited attention has been given to predicting and explaining annotators’ disagreement. An important contribution in the field is represented by the SemEval 2023 Task 11 (Leonardelli et al., 2023) where the main goal is to model the disagreement between annotators on different types of textual messages. A first insight in explaining disagreement sources is represented by (Astorino et al., 2023). The authors leverage integrated gradients to detect both disagreement and hate speech and introduce a *filtering strategy* for textual constituents that aids in explaining hateful messages. In this paper, we investigate whether is possible to grasp disagreement from pre-trained language models fine-tuned for the hate-detection task, exploiting Contrastive Learning strategies.

3. Dataset

We employ four benchmark datasets from SemEval 2023 Task 11 focused on Learning With

Disagreement (LWD) (Leonardelli et al., 2023), each exhibiting diverse characteristics such as types (social media posts and conversations), languages (English and Arabic), goals (misogyny, hate speech, offensiveness detection), and annotation methods (experts, specific demographic groups, and general crowd). In particular, we used Hate Speech on Brexit (HS-Brexit) (Akhtar et al., 2021), Arabic Misogyny and Sexism (ArMIS) (Almanea and Poesio, 2022), ConvAbuse (Cercas Curry et al., 2021) and Multi-Domain Agreement (MD-Agreement) (Leonardelli et al., 2021). A summary of the datasets is presented in Table 1.

All datasets feature hard-labels (hateful/non-hateful) and soft-labels (disagreement) for each instance. The purpose of this work is to discern agreement and disagreement rather than different levels of disagreement, therefore the number of annotators is not taken into account. The disagreement prediction is treated as a binary task. Therefore, an *agreement label* was derived from the soft-label by setting the value to (+) when there is 100% agreement between the annotators, regardless of the value of the hard label; it is set to (-) otherwise.

4. Disagreement Estimation

The proposed approach exploits Contrastive Learning techniques that allow the comparison among multiple instances (in contrast with the pairwise comparison of the previous approach). The proposed approach includes an initial fine-tuning on hate detection task and a subsequent Disagreement predictions based on the extracted embeddings. The main phases can be summarized as follows:

1. **Fine-tuning of a pre-trained LM:** The *bert-base-multilingual-cased* has been fine-tuned to distinguish hateful content from non-hateful ones (considering the provided hard labels), proposing a loss function that is grounded on the Binary Cross Entropy and InfoNCE¹ (Khosla et al., 2020) specifically adapted for the considered problem:

$$\begin{aligned} \mathcal{L} &= \lambda L_{bce} + (1 - \lambda) L_{InfoNCE} = \\ &= -\lambda \sum_s t(s) \log(p(s)) + \\ &+ (1 - \lambda) \left(-\log \frac{e^{s \cdot k^{pos}} / \tau}{\sum_{k^{neg} \in K} e^{s \cdot k^{neg}} / \tau} \right) \end{aligned} \quad (1)$$

¹In order to reinforce the impact of the Contrastive Loss InfoNCE, the hyperparameter λ has been set to 0.3. The fine-tuning has been performed for 4 epochs, adopting a learning rate of 3e-5

| Dataset | Language | Task | Annotators | Pool Ann. | % of items with full agr. | Agreement Distribution (Test Set) |
|--|----------|-------------------------------|------------|-----------|---------------------------|-----------------------------------|
| HS-Brexit (Akhtar et al., 2021) | En | Hate Speech | 6 | 6 | 69% | 116/168 |
| ArMis (Almanea and Poesio, 2022) | Ar | Misogyny and sexism detection | 3 | 3 | 86% | 92/145 |
| ConvAbuse (Cercas Curry et al., 2021) | En | Abusive Language detection | 2-7 | 7 | 65% | 727/840 |
| MD-Agreement (Leonardelli et al., 2021) | En | Offensiveness detection | 5 | >800 | 42% | 1292/3057 |

Table 1: Datasets characteristics.

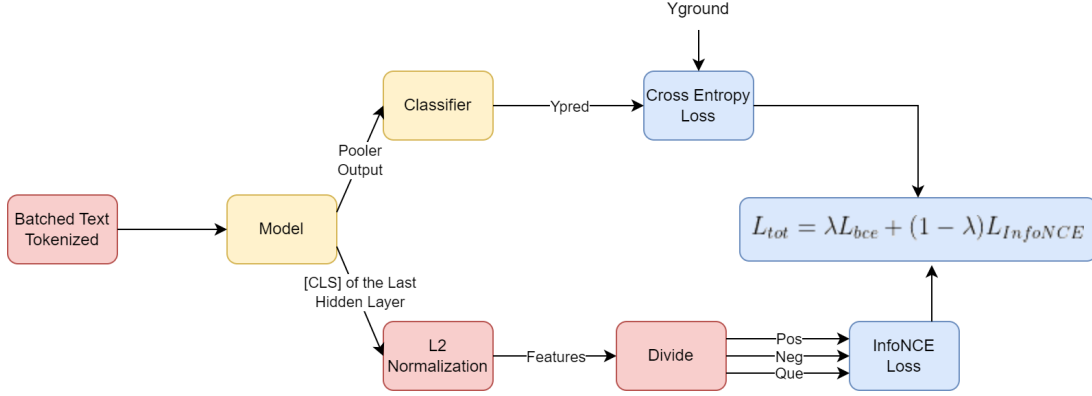


Figure 1: Schematic representation of the Fine-Tuning step.

where s indicates a given sample in the dataset, $t(s)$ denotes the target distribution, $p(s)$ represents the prediction probability distribution, k^{pos} is an instance in the dataset that has the same ground truth label of s , k^{neg} denotes an instance in the dataset that has the opposite ground truth label with respect to s , K is the set of instances in the dataset that have label opposite to s , and τ is the temperature.

The procedure is, in fact, composed of two parts. The first part allows to compute the Binary Crossentropy Loss while the second part exploits information derived from the representation of the [CLS] token in the last model layer. The Binary Crossentropy Loss has the main goal of minimizing the difference between the prediction probabilities and truth values, while the InfoNCE is aimed at maximising the agreement between positive samples and minimizing the agreement between the negative ones in the learned representation. In this way, The derived features are then normalized with L2 regularization to extract *query*, *positive* and *negative* features, used for computing the InfoNCE. The the fine-tuning phase is summarized in Figure 1.

2. **Similarity Matrix definition:** The fine-tuned model has been used to generate embeddings² for the samples in the training and test set in order to define a similarity matrix. The

²The embedding representation has been obtained merging the last seven layers of the model.

last contains embedding distances computed towards cosine similarity.

3. **Disagreement prediction:** For each instant in the test set, disagreement is predicted starting from the distribution of samples with agreement and with disagreement in the closer neighborhood. Two different strategies have been proposed, distinguishing the definition of the neighborhood:

General Model. The General Model takes a comprehensive approach, combining multiple activities under the umbrella of hate speech identification. This framework incorporates tasks linked to aggressive, offensive, and abusive language, relying on the idea that these behaviors frequently share a common foundation in manifestations of hatred, disregarding the targeted minority. This inclusive viewpoint enables the model to effectively capture the subtle manifestations of hate across multiple linguistic dimensions, different languages, and towards several targets, resulting in a more robust and versatile solution for identifying and treating various forms of harmful speech. According to this rationale, for each instance in the test set, the corresponding neighbor is computed in order to include instances that appear in the overall training set (i.e. achieved via the union of the four training datasets).

Domain-Specific Model. The Domain-Specific Model takes a more refined approach, focusing solely on elements that share specific characteristics. This approach focuses

| Dataset | Approach | P+ | R+ | F+ | P- | R- | F- | Macro F |
|--------------|-------------------------|------|------|------|------|------|------|-------------|
| HS-Brexit | m-BERT | 0.85 | 0.69 | 0.76 | 0.51 | 0.73 | 0.60 | 0.68 |
| | General Model | 0.78 | 0.83 | 0.80 | 0.56 | 0.48 | 0.52 | 0.66 |
| | Domain-Specific Model | 0.80 | 0.94 | 0.86 | 0.77 | 0.46 | 0.58 | 0.72 |
| | (Astorino et al., 2023) | 0.84 | 0.78 | 0.81 | 0.57 | 0.67 | 0.62 | 0.71 |
| ArMIS | m-BERT | 0.60 | 0.27 | 0.37 | 0.32 | 0.65 | 0.43 | 0.40 |
| | General Model | 0.63 | 0.95 | 0.75 | 0.17 | 0.02 | 0.03 | 0.39 |
| | Domain-Specific Model | 0.65 | 0.88 | 0.75 | 0.48 | 0.19 | 0.27 | 0.51 |
| | (Astorino et al., 2023) | 0.67 | 0.75 | 0.71 | 0.47 | 0.38 | 0.42 | 0.56 |
| ConvAbuse | m-BERT | 0.87 | 0.99 | 0.93 | 0.33 | 0.03 | 0.05 | 0.49 |
| | General Model | 0.87 | 0.97 | 0.92 | 0.21 | 0.04 | 0.07 | 0.50 |
| | Domain-Specific Model | 0.71 | 0.13 | 0.22 | 0.88 | 0.99 | 0.93 | 0.58 |
| | (Astorino et al., 2023) | 0.94 | 0.70 | 0.80 | 0.27 | 0.72 | 0.40 | 0.60 |
| MD-Agreement | m-BERT | 0.43 | 0.34 | 0.38 | 0.58 | 0.68 | 0.63 | 0.50 |
| | General Model | 0.66 | 0.53 | 0.59 | 0.70 | 0.80 | 0.74 | 0.67 |
| | Domain-Specific Model | 0.66 | 0.53 | 0.59 | 0.70 | 0.80 | 0.75 | 0.67 |
| | (Astorino et al., 2023) | 0.54 | 0.52 | 0.53 | 0.66 | 0.68 | 0.67 | 0.60 |

Table 2: Comparison of the different approaches on the test set. **Bold** denotes the best approach according to the F1-Score.

on instances of the same hate-related task (i.e. aggressiveness, general hatred, or abusive language identification). Furthermore, the Domain-Specific Model focuses on data that shares homogenous aspects such as language, type of text (e.g. Tweets, discussion, etc.), and target, recognizing the close relationship between those characteristics and annotator disagreement on hate speech. A given term can be, in fact, interpreted as controversial and generate disagreement on a dataset that focuses on hate towards a specific task (e.g. misogyny identification) and neutral in different datasets with different characteristics (e.g. racism detection). As a result, when developing this strategy, the datasets have not been combined. For each instance in the test set, the corresponding neighborhood is computed in order to include only instances that appear in the respective training set in order to guarantee the comparison with samples that share similar characteristics (i.e., topic, type, language, etc.). In both cases, the hyperparameter n that defines the numerosity of the selected neighborhood has been estimated towards a grid search approach.

The estimated configurations are summarized in Table 3.

| dataset | n |
|------------------|-----|
| ArMIS | 22 |
| HS-Brexit | 50 |
| ConvAbuse | 19 |
| MD_Agreement | 105 |
| Overall Datasets | 59 |

Table 3: Estimated Hyperparameter

Once the neighbor has been selected, the final disagreement label is predicted evaluating the number of samples with agreement and the number of samples with disagreement in the selected neighborhood. In particular, if the difference between the number of samples

with agreement and the number of samples with disagreement in the selected neighbor is smaller than τ^3 , then the predicted label is set to disagreement. On the other hand, if the difference between samples with agreement and samples with disagreement in the selected neighbor is bigger than τ the prediction is computed toward majority voting (i.e., Agreement if the majority of samples in the selected neighbor are labeled as agreement, Disagreement otherwise).

5. Results

In this section, the results obtained by the proposed approaches are reported. We measured Precision (P), Recall (R) and F-Measure (F), distinguishing between Agreement (+) and Disagreement (-) labels and reporting also the Macro F-Measure.

Table 2 summarized the achieved results. We also report results achieved by (Astorino et al., 2023) for a state-of-the-art comparison. This last approach exploits integrated gradients from pre-trained language models in the recognition of disagreements' causes and hate speech contents. One of the main contribution is given by the introduction of a filtering strategy that contributes to explain hateful messages via textual constituents. It can be easily noted that, in the majority of the considered datasets, the proposed approach "Domain-Specific Model" outperforms the considered baseline m-BERT and achieves competitive results with (Astorino et al., 2023). It is also interesting to highlight that the *Domain-Specific Model* outperforms the *General* one in all the proposed datasets. The *Domain-Specific Model* is designed to concentrate on a single dataset, allowing it to define its representation based on its unique characteristics, such as

³ n has been estimated via Grid Search. It has been set to 7 for the General approach and to 2 for the Domain-Specific approach.

the type of text, target of hate, language, and more. This leads to a better understanding of the terms in relation to the hate task at hand, and therefore to higher performance with respect with the *General* approach. More important, although the proposed approach is comparable or in some cases even better than (Astorino et al., 2023), it has the great advantage of being computationally less complex than (Astorino et al., 2023) thanks to presence of a simpler objective function compared to the two fine-tuning losses in the considered baseline model.

6. Conclusions and Future works

The proposed paper introduces a novel approach for detecting disagreement in hateful content. The method exploits contrastive learning techniques applied to pre-trained language models to predict both hate speech and potential disagreement arising from different readers. The proposed approach outperforms m-BERT and achieve competitive results on four benchmark datasets from the Learning With Disagreement (LeWiDi) task at SemEval (Leonardelli et al., 2021). Overall, the proposed approach demonstrates the potential to encapsulate Contrastive Learning technique in Natural Language tasks. Future work could focus on exploring the applicability of the proposed approach to other datasets in different domain and expanding the scope to include multimodal data analysis.

Acknowledgments

The work of Elisabetta Fersini has been partially funded by MUR under the grant REGAINS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca and by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR)

7. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. [Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection](#).

Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

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.

Alessandro Astorino, Giulia Rizzi, and Elisabetta Fersini. 2023. Integrated gradients as proxy of disagreement in hateful content. In *CEUR WORKSHOP PROCEEDINGS*, volume 3596. CEUR-WS. org.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

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.

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.

Anca Dumitrache, FD Mediagroep, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of NAACL-HLT*, pages 2164–2170.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Marianne D LaFrance and Sarah J Roberts. 2019. The role of bias in hate speech detection. *Journal of Language Aggression and Conflict*, 7(1):1–20.

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.
- Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. [Semeval-2023 task 11: Learning with disagreements \(lewid\)](#).
- Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.
- Giulia Rizzi, Alessandro Astorino, Paolo Rosso, and Elisabetta Fersini. 2024. Unraveling disagreement constituents in hateful speech. In *Advances in Information Retrieval*, pages 21–29, Cham. Springer Nature Switzerland.
- Giulia Rizzi, Alessandro Astorino, Daniel Scalena, Paolo Rosso, and Elisabetta Fersini. 2023. Mind at semeval-2023 task 11: From uncertain predictions to subjective disagreement. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 556–564.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia Von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@LREC2022*, pages 83–94.
- Hind Saleh, Areej Alhothali, and Kawthar Moria. 2023. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1):2166719.
- Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future: 17th International Conference, iConference 2022, Virtual Event, February 28–March 4, 2022, Proceedings, Part I*, pages 425–444. Springer.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.
- Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Soft metrics for evaluation with disagreements: an assessment

Giulia Rizzi^{1,2}, Elisa Leonardelli³, Massimo Poesio^{4,5}, Alexandra Uma,
Maja Pavlovic⁴, Silviu Paun, Paolo Rosso^{2,6}, Elisabetta Fersini¹

¹University of Milano-Bicocca, ²Universitat Politècnica de València, ³Fondazione Bruno Kessler

⁴Queen Mary University of London, ⁵University of Utrecht

⁶ValgrAI - Valencian Graduate School and Research Network of Artificial Intelligence

g.rizzi10@campus.unimib.it, eleonardelli@fbk.eu, {m.poesio, m.pavlovic}@qmul.ac.uk,

{alexandra.uma2, spaun3691}@gmail.com, proso@dsic.upv.es, elisabetta.fersini@unimib.it

Abstract

The move towards preserving judgement disagreements in NLP requires the identification of adequate evaluation metrics. We identify a set of key properties that such metrics should have, and assess the extent to which natural candidates for soft evaluation such as Cross Entropy satisfy such properties. We employ a theoretical framework, supported by a visual approach, by practical examples, and by the analysis of a real case scenario. Our results indicate that Cross Entropy can result in fairly paradoxical results in some cases, whereas other measures Manhattan distance and Euclidean distance exhibit a more intuitive behavior, at least for the case of binary classification.

1. Introduction

As the realization grows that disagreement between subjects in many natural language tasks may be the result of genuine differences in interpretation rather than of unclear guidelines or poor quality annotators (Poesio and Artstein, 2005; Passonneau et al., 2012; Plank et al., 2014; Aroyo and Welty, 2015; Akhtar et al., 2019; Basile et al., 2021; Uma et al., 2021b,a; Davani et al., 2022; Sap et al., 2022; Leonardelli et al., 2023), many researchers have started investigating methods for learning and evaluating models from datasets in which such differences in interpretation are preserved, particularly for subjective tasks (Basile et al., 2021; Uma et al., 2021b,a; Leonardelli et al., 2023). However, our understanding of this form of evaluation is still only at the beginning.

In this paper, we argue that soft evaluation metrics – metrics to evaluate the ability of NLP models to predict not just the preferred interpretation of an item, but also its probability and the probability of alternative interpretations according to human judgements, that Uma et al. called soft label (Uma et al., 2021b) – should satisfy a number of properties, that we define within a theoretical framework.

We then analyze four candidate metrics with respect of this set of formal properties. The metrics analysed include Cross Entropy, possibly the most widely used among such metrics, and which was also the main soft evaluation metric in the two recent Learning With Disagreements (LeWiDi) SemEval shared tasks (Uma et al., 2021a; Leonardelli et al., 2023). The other considered candidates are Manhattan Distance, Euclidean Distance and the Jensen-Shannon Divergence. For the binary label case, we also provide empirical examples and

graphical visualizations of the metrics' behavior. Moreover we analyze how the metrics behave in a real case scenario, namely the LeWiDi shared task. Finally we discuss the case of multi-class labels.

One key result is that the widely used Cross Entropy metric has several counterintuitive properties, which other metrics considered do not suffer from, at least for the binary classification case. The situation is more complex for multi-label classification.

2. Soft Evaluation Metrics

The fundamental characteristic required of a soft evaluation metric is the ability to compare two probability distributions: the target distribution obtained from annotator judgments, and the distribution predicted by a model. In this Section, we introduce four metrics that have been used or could be used for such soft evaluation (Uma et al., 2021b; Basile et al., 2021; Uma et al., 2021a; Leonardelli et al., 2023).

Cross Entropy Cross Entropy is a common measure used in information theory and machine learning to quantify the difference between two probability distributions.

Given two distributions p , and q , their Cross Entropy is defined as:

$$\mathbb{H}(p, q) = \mathbb{E}_p [\log q] = - \sum_k p(k) \log(q(k)) \quad (1)$$

Where \mathbb{E}_p is the expected value operator with respect to the distribution p .

In the binary classification case, Cross Entropy simplifies to:

$$\mathbb{H}(p, q) = -[p \log(q) + (1 - p) \log(1 - q)] \quad (2)$$

Manhattan Distance The Manhattan distance, also known as \mathbb{L}_1 distance measures the absolute differences between corresponding elements of two distributions. Given two distributions p and q , the Manhattan distance is defined as:

$$\mathbb{L}_1(p, q) = \sum_k |p(k) - q(k)| \quad (3)$$

Euclidean Distance The Euclidean distance, also known as \mathbb{L}_2 distance measures the the straight-line distance between two points in Euclidean space. Given two distributions p and q , the Euclidean distance is defined as:

$$\mathbb{L}_2(p, q) = \sqrt{\sum_k (p(k) - q(k))^2} \quad (4)$$

Jensen-Shannon Divergence (JSD) The Jensen-Shannon Divergence is a symmetrized and smoothed version of the Kullback-Leibler Divergence (KL Divergence). Given two distributions p and q , the Jensen-Shannon Divergence is defined as:

$$\text{JSD}(p, q) = \frac{1}{2} (D_{KL}(p \parallel m) + D_{KL}(q \parallel m)) \quad (5)$$

Where D_{KL} is the Kullback-Leibler Divergence and $m = \frac{1}{2}(p + q)$. That corresponds to:

$$\text{JSD}(p, q) = \frac{1}{2} \left(\sum_k p(k) \log \left(\frac{p(k)}{m(k)} \right) + \sum_k q(k) \log \left(\frac{q(k)}{m(k)} \right) \right) \quad (6)$$

where $m(k) = \frac{1}{2}(p(k) + q(k))$.

Although the Wasserstein distance is commonly used to quantify the difference between two probability distributions, it was not included in our analysis: it is crucial to highlight that, in the specific case of two binary distributions that are not rearranged, the Wasserstein distance reduces to the Manhattan distance.

3. Desirable properties

In this Section, we identify a set of properties that soft evaluation metrics should satisfy. We will use $q(k)$ to indicate the probability of an item k having the positive label according to the model, and $p(k)$ to indicate the real probability of k having the positive label according to the gold (soft) standard. Finally, we use \mathbb{M} to indicate the general measure to quantify the difference between two probability distributions.

Property 1 [Symmetry] Given two probability distributions $q(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label and the corresponding real value associated with k in the golden standard,

$$\mathbb{M}(p(k), q(k)) = \mathbb{M}(q(k), p(k))$$

Property 2 [Boundedness] Given two probability distributions $q(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label and the corresponding real value associated with k in the golden standard, there exist constants a and b such that, for every item k ,

$$a \leq \mathbb{M}(p(k), q(k)) \leq b$$

Property 3 [Triangle Inequality] Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$),

$$\mathbb{M}(p(k), q(k)) + \mathbb{M}(q(k), r(k)) \geq \mathbb{M}(p(k), r(k))$$

Property 4 [Transitivity] Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$),

$$\begin{aligned} \mathbb{M}(p(k), q(k)) &< \mathbb{M}(p(k), r(k)) \\ \Rightarrow \mathbb{M}(q(k), r(k)) &< \mathbb{M}(p(k), r(k)) \end{aligned}$$

Property 5 [Sum invariant] Given two probability distributions $q(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by a model ($q(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$). A divergence \mathbb{M} is sum invariant if whenever c is independent from p, q

$$\mathbb{M}(c + p(k), c + q(k)) \leq \mathbb{M}(p(k), q(k))$$

This property is strictly related to the following three subproperties:

Property 5.a [Minimum penalization at perfect match] Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), if $p(k) = q(k)$ and $r(k) \neq p(k)$, then

$$\mathbb{M}(p(k), q(k)) < \mathbb{M}(p(k), r(k))$$

Property 5.b [Fair penalization] *Given three probability distributions $q^{(k)}$, $r^{(k)}$ and $p^{(k)}$ representing the probability of an item k being classified with the positive label by two different models ($q^{(k)}$ and $r^{(k)}$) and the corresponding real value associated with k in the golden standard ($p^{(k)}$), if $|p^{(k)} - q^{(k)}| < |p^{(k)} - r^{(k)}|$, then*

$$\mathbb{M}(p^{(k)}, q^{(k)}) < \mathbb{M}(p^{(k)}, r^{(k)})$$

Property 5.c [Fair penalization on perfect match] *Given two probability distributions $q^{(k)}$, and $p^{(k)}$ representing the probability of an item k being classified with the positive label by two different models ($q^{(k)}$ and $r^{(k)}$) and the corresponding real value associated with k in the golden standard ($p^{(k)}$); given two items k_i and k_j , if $p^{(k_i)} = q^{(k_i)}$, $p^{(k_j)} = q^{(k_j)}$ and $p^{(k_i)} \neq p^{(k_j)}$, then*

$$\mathbb{M}(p^{(k_i)}, q^{(k_i)}) = \mathbb{M}(p^{(k_j)}, q^{(k_j)})$$

Property 6 [Scale sensitivity] *Given two probability distributions $p^{(k)}$ and $q^{(k)}$ representing the probability of an item k being classified with the positive label ($q^{(k)}$) and the corresponding real value associated with k in the golden standard, ($p^{(k)}$), We say that \mathbb{M} is scale sensitive (of order β), if there exists a $\beta > 0$, and a real value $c > 0$, such that for all k*

$$\mathbb{M}(cp^{(k)}, cq^{(k)}) \leq |c|^\beta \mathbb{M}(p^{(k)}, q^{(k)})$$

If \mathbb{M} is scale sensitive of order $\beta = 1$ then the divergence $\mathbb{M}(\delta, \delta_{1/2})$ can be no more than half the divergence $\mathbb{M}(\delta_0, \delta_1)$. If \mathbb{M} is sum invariant, then the divergence of δ_0 to δ_0 is equal to the divergence of the same distributions shifted by a constant c , i.e., of δ_c to δ_{1+c} .

The above mentions set of perproperties are desired when evaluating soft metrics in order to ensure the fairness and the consistency of the evaluation process. In particular the *Symmetry property* ensure an objective evaluation, independent by the arrangement of the input data (i.e. regardless of whether we evaluate predictions against ground truth or vice versa). The *Boundedness* property guarantee that the evaluation values remains in a defined range, allowing for comparison among different models and facilitating the identification of outliers. The *Triangle Inequality* property is essential for a consistent evaluation since it guarantee that composed metrics remain coherent and does not leads to contradictory results. Similarly, the *sum invariant* property ensuer the consistency of the metric when combined or aggregated. The *Transitivity* property guarantees consistency in comparisons across different instances or groups. It ensures the consistency of a model performances when comparing

across different tasks, datasets, or experimental conditions. Finally, the *scale sensitivity* property guarantee that the metric correctly capture the magnitude of the differences among models. In other words, it ensure that sligh variations in the model's performance are reflected as a minor change in the metric score, while big changes in performance lead to a significant change in the metric score.

4. Metric properties assessment in the binary case

In this Section, we analyze the extent to which the evaluation metrics under consideration and presented in Section 2, satisfy the properties we deem desirable and presented in Section 3, in the case of binary labels. An analysis is performed (Section 4.1), focusing on the selected properties, providing theoretical background and practical examples when the defined properties are not fulfilled. In Section 4.2 a graphical representation of the metrics behaviour at different target distributions is shown. Furthermore, the figure is used as a visual support to discuss some metrics' properties. Finally, in Section 4.3 we compare metrics behaviour in the real case scenario of the LeWiDi competition.

4.1. Properties assessment and examples

In this section, properties are discussed with respect to selected metrics. Table 1 summarizes the properties satisfied by the metrics.

Property 1 All the selected metrics satisfy the *symmetry* property (P1), i.e. inverting target and prediction does not affect the result, except for Cross Entropy.

Cross Entropy, in fact, is *asymmetric*, given its relation to Kullback-Leibler Divergence. Cross Entropy is related to KL-Divergence as follows:

$$\mathbb{H}(p, q) = D_{KL}(p||q) + \mathbb{E}(p) \quad (7)$$

where \mathbb{H} is the Cross Entropy of distribution p and q , $D_{KL}(p||q)$ is the KL-Divergence and $\mathbb{E}(p)$ is the Entropy of the distribution p . Since $\mathbb{E}(p)$ can be considered as a constant, Cross Entropy follows the same asymmetry of KL-Divergence. The definition of Cross Entropy reported in equation (1) leads to the following inequality:

$$-\sum_i p_i \log q_i \neq -\sum_i q_i \log p_i \quad (8)$$

Example 1 shows two distributions for which the symmetry property is not fulfilled by Cross Entropy: in the proposed example, $\mathbb{H}(p(k_1), q(k_1)) \neq \mathbb{H}(p(k_2), q(k_2))$, although $p(k_1) = q(k_2)$ and $q(k_1) = p(k_2)$.

Table 1: Properties of Evaluation Metrics (Binary Case)

| Metric | Properties | | | | | | | | | |
|---------------------------|------------|----|----|----|----|-----|-----|-----|----|--|
| | P1 | P2 | P3 | P4 | P5 | P5a | P5b | P5c | P6 | |
| Cross Entropy | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | |
| Manhattan Distance | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Euclidean Distance | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Jensen-Shannon Divergence | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | |

Example 1 [Symmetry violation]

| k | Target p(k) | Prediction q(k) |
|----------------|--------------|-----------------|
| k ₁ | [0.83; 0.17] | [0.5; 0.5] |
| k ₂ | [0.5; 0.5] | [0.83; 0.17] |

Cross Entropy values:

$$\mathbb{H}(p(k_1), q(k_1)) = 0.6931$$

$$\mathbb{H}(p(k_2), q(k_2)) = 0.9791$$

Property 2 In the binary case, all the selected metrics satisfy the *Boundedness* property (P2), i.e. they can only assume finite values. To note that Cross Entropy is left-bounded by definition, also given its relationship with the KL-Divergence. Commonly, it is bounded by introducing a smoothing that affects the extremants. But the scaling technique that is chosen to make the metric *bounded* has a great effect on the interval where the \mathbb{H} values are distributed.

Property 3 The *Triangle inequality* property (P3) is satisfied by all metrics except Cross Entropy. When comparing two binary distributions, the asymmetry and sensitivity to specific distribution values of Cross Entropy can lead to instances where the distance between two distributions is not guaranteed to be less than or equal to the sum of their distances to a third distribution. Therefore, Cross Entropy does not consistently satisfy the *triangle inequality* property (P3).

Example 2 reports an example in which the Triangle Inequality property is not fulfilled by the Cross Entropy. Triangle Inequality property implies that the sum of the Cross Entropies for two *consecutive* predictions should be greater than or equal to the Cross Entropy between the target distribution and the direct prediction. However, in the proposed example, $\mathbb{H}(p(k), q(k)) + \mathbb{H}(q(k), r(k))$ is less than $\mathbb{H}(p(k), r(k))$ and contradicts the Triangle Inequality property for Cross Entropy.

Example 2 [Triangle Inequality violation]

| Target p(k) | Prediction q(k) | Prediction r(k) |
|-------------|-----------------|-----------------|
| [0.7, 0.3] | [0.95, 0.05] | [1, 0] |

Cross Entropy values:

$$\mathbb{H}(p(k), r(k)) = 8.2893$$

$$\mathbb{H}(p(k), q(k)) + \mathbb{H}(q(k), r(k)) = 2.3162$$

Property 4 The *Transitivity* property is satisfied by all metrics, except the Cross Entropy and the Jensen-Shannon divergence that do not consistently satisfy it.

Example 3 shows how despite $\mathbb{H}(p(k), q(k)) < \mathbb{H}(p(k), r(k))$, the expected transitivity property ($\mathbb{H}(q(k), r(k)) < \mathbb{H}(p(k), r(k))$) is not satisfied by Cross Entropy. Similarly, for Jensen-Shannon Divergence: despite $\mathbb{JSD}(p(k), q(k)) < \mathbb{JSD}(p(k), r(k))$, the expected transitivity property ($\mathbb{JSD}(q(k), r(k)) < \mathbb{JSD}(p(k), r(k))$) is not satisfied.

Example 3 (P4)[Transitivity violation]

| Target p(k) | Prediction q(k) | Prediction r(k) |
|-------------|-----------------|-----------------|
| [0.9, 0.1] | [0.7, 0.3] | [1, 0] |

Cross Entropy values:

$$\mathbb{H}(p(k), q(k)) = 0.4414$$

$$\mathbb{H}(p(k), r(k)) = 2.7631$$

$$\mathbb{H}(q(k), r(k)) = 8.2893$$

Jansen-Shannon values:

$$\mathbb{JSD}(p(k), q(k)) = 0.1801$$

$$\mathbb{JSD}(p(k), r(k)) = 0.1897$$

$$\mathbb{JSD}(q(k), r(k)) = 0.3425$$

Property 5.a In the binary case, all the selected metrics satisfy the *Minimum penalization at perfect match* property. Indeed for each possible target, the perfect match (the exact prediction of the target) assumes the minimum values possible for the target considered. (See also Figure 1 and relative discussion in Section 4.2)

Property P5.b Cross Entropy tends to penalize predictions that perfectly match the target distribution when the target distribution itself is characterized by a large entropy, resulting in an *unfair penalization*. This is because, as shown in Equation 7, when p is ‘highly entropic’, $\mathbb{E}(p)$ is large.

Example 4 shows how Cross Entropy tends to unfairly penalize probability distributions close to the boundaries. Despite the error performed in the prediction $q(k)$ is smaller than the one performed by the prediction $r(k)$, $\mathbb{H}(p(k), q(k))$ is bigger than $\mathbb{H}(p(k), r(k))$ and contradicts the *Fair penalization* property.

The Jensen-Shannon tends to penalize those cases that are less entropic (disregarding which distribution, target, or prediction, is more entropic than the other). Therefore, the Jensen-Shannon measure does not fulfill the *fair penalization* property (P5b). An example of distributions for which the property is not fulfilled is reported in Example 4. In fact, despite the error performed in the prediction $q(k)$ is smaller than the one performed by the prediction $r(k)$, $\mathbb{JSD}(p(k), q(k))$ is bigger than $\mathbb{JSD}(p(k), r(k))$ and contradicts the *Fair penalization* property.

Example 4 (P5b)

Unfair penalization

| Target $p(k)$ | Prediction $q(k)$ | Prediction $r(k)$ |
|---------------|-------------------|-------------------|
| [0.9, 0.1] | [1, 0] | [0.7, 0.3] |

Cross Entropy values:

$$\mathbb{H}(p(k), q(k)) = 2.7631$$

$$\mathbb{H}(p(k), r(k)) = 0.4414$$

Jensen-Shannon values:

$$\mathbb{JSD}(p(k), q(k)) = 0.1897$$

$$\mathbb{JSD}(p(k), r(k)) = 0.1801$$

Property P5.c Another effect of the entropy in the distribution on the Cross Entropy emerges when comparing the scores associated to different distributions that correctly predict the target. Example 5 reports an example showing that the *Fair penalization on perfect match* property is not satisfied: despite both distributions correctly predict the target, $\mathbb{H}(p(k_1), q(k_1))$ is not equal to $\mathbb{H}(p(k_2), q(k_2))$, due to the corresponding entropy in the distributions.

Example 5 (P5c)

Unfair penalization on perfect match

| k | Target $p(k)$ | Prediction $q(k)$ |
|-------|---------------|-------------------|
| k_1 | [0.5, 0.5] | [0.5; 0.5] |
| k_2 | [0.9; 0.1] | [0.9; 0.1] |

Cross Entropy values:

$$\mathbb{H}(p(k_1), q(k_1)) = 0.6932$$

$$\mathbb{H}(p(k_2), q(k_2)) = 0.3251$$

Property 6 Considering two binary distributions p and q , and a positive real value c ; let p' and q' be scaled versions of p and q by the constant factor c : $p' = c \cdot p$ and $q' = c \cdot q$.

Cross Entropy: In the binary classification scenario, the Cross Entropy distance does not fulfill the scale sensitivity property. Substituting the scaled distributions into the Cross Entropy distance formula (Eq. 1), we obtain:

$$\begin{aligned} \mathbb{H}(p', q') &= - \sum_k c \cdot p(k) \log(c \cdot q(k)) \\ &= -c \cdot \sum_k p(k) \log(c \cdot q(k)) \end{aligned} \quad (9)$$

By comparing this with $|c| \cdot \mathbb{H}(p, q)$ we obtain:

$$|c| \cdot \mathbb{H}(p, q) = |c| \cdot - \sum_k p(k) \log(c \cdot q(k)) \quad (10)$$

The two expressions are not directly proportional. Therefore, the Cross Entropy distance does not satisfy the scale sensitivity property.

Manhattan distance: In the binary classification scenario, the Manhattan distance satisfies the scale sensitivity property.

Considering two binary distributions p and q , the Manhattan distance between them is defined as shown in Eq. 3.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 3), we obtain:

$$\begin{aligned} \mathbb{L}_1(p', q') &= \sum_i |c \cdot p(k) - c \cdot q(k)| \\ &= c \cdot \sum_k |p(k) - q(k)| \end{aligned} \quad (11)$$

By comparing this with $|c| \cdot \mathbb{L}_1(p, q)$ we obtain:

$$|c| \cdot \mathbb{L}_1(p, q) = |c| \cdot \sum_k |p(k) - q(k)|$$

Indicating that the Manhattan distance scales linearly with the constant factor c , fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

Euclidean Distance: In the binary classification scenario, the Euclidean distance fulfills the scale sensitivity property.

Considering two binary distributions p and q , the Euclidean distance between them is defined as shown in Eq. 4.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 4), we obtain:

$$\begin{aligned} \mathbb{L}_2(p', q') &= \sqrt{\sum_k (c \cdot p(k) - c \cdot q(k))^2} \\ &= c \cdot \sqrt{\sum_k (p(k) - q(k))^2} \end{aligned} \quad (12)$$

By comparing this with $|c| \cdot \mathbb{L}_2(p, q)$ we obtain:

$$|c| \cdot \mathbb{L}_2(p, q) = |c| \cdot \sqrt{\sum_k (p(k) - q(k))^2}$$

Indicating that the Euclidean distance scales linearly with the constant factor c , fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

Jensen-Shannon: In the binary classification scenario, the Jensen-Shannon distance fulfills the scale sensitivity property.

Considering two binary distributions p and q , the Jensen-Shannon distance between them is defined as shown in Eq. 5.

Substituting the scaled distributions into the Manhattan distance formula (Eq. 5), we obtain:

$$\begin{aligned} \mathbb{JSD}(p', q') &= \frac{1}{2} (D_{KL}(p' \parallel m') + D_{KL}(q' \parallel m')) \\ &= \frac{1}{2} (D_{KL}(c \cdot p \parallel c \cdot m) + \\ &\quad D_{KL}(c \cdot q \parallel c \cdot m)) \\ &= \frac{1}{2} (c \cdot D_{KL}(p \parallel m) + c \cdot D_{KL}(q \parallel m)) \end{aligned} \quad (13)$$

where $m'(k) = \frac{1}{2}(p'(k) + q'(k))$ and $m(k) = \frac{1}{2}(p(k) + q(k))$.

By comparing this with $|c| \cdot \mathbb{JSD}(p, q)$ we obtain:

$$|c| \cdot \mathbb{JSD}(p, q) = \frac{1}{2} (|c| \cdot D_{KL}(p \parallel m) + |c| \cdot D_{KL}(q \parallel m)) \quad (14)$$

Indicating that the Jensen-Shannon distance scales linearly with the constant factor c , fulfilling the scale sensitivity property with a sensitivity order (β) of 1.

4.2. Metrics graphical representation

Figure 1 shows distinct plots for each metric, with the x-axes representing the prediction values and the y-axes representing the corresponding distance values (or score) based on the metric under consideration. These plots provide a detailed visual representation of the metrics behaviors at different target values. Moreover, we can visually explore the properties, and in the following we discuss P5.a, P5.b and P5.c.

In Figure 1 we can observe how all the selected metrics satisfy the *Minimum penalization at perfect match property* (P5.a): for each target's curve plotted, the minimum values of the curve corresponds to the perfect match, i.e. the exact prediction of the target.

To demonstrate the influence of prediction errors on metric performance (P5.b), distance values when a nominal error of 0.2 appears in the forecast are highlighted with points within the same plots. This intentional perturbation enables an investigation of the metric's robustness in the presence of slight prediction mistakes, evaluating the ability of fair evaluation across a range of targets. Horizontal alignment between two prediction points that are equally distant from the target, indicate that property P5.b is respected (see the cases of Manhattan Distance and Euclidean Distance, Figure 1 b and c). Conversely, deviations from this horizontal alignment implies unfair penalizations (see the cases of Cross Entropy and Jensen Shannon Divergence, Figure 1 a and d).

Finally, to contribute to a more detailed understanding of the *Fair penalization on perfect match property* (P5.c), within each plot, dots are used to highlight the resulting score, when the target is correctly predicted. The alignment of all dots along a horizontal axis (such as in the case of Figure 1 b, c and d), indicates a fair penalty for perfect matches across targets. Deviations from this horizontal alignment, such in the case of Figure 1 a (Cross Entropy) imply diverse penalization levels for perfect matches on diverse targets, revealing disparities in the metric's treatment of different target values.

4.3. Impact on a Leaderboard: The LeWiDi Case Study

In this section, we aim to investigate the application of some of the discussed evaluation metrics to a real case scenario. To this end, we exploit the data from a recent shared task, the Learning With Disagreements task (LeWiDi) (Leonardelli et al., 2023) proposed at the 2023 edition of SemEval¹. The challenge proposed by the task foresees to

¹<https://semeval.github.io/>

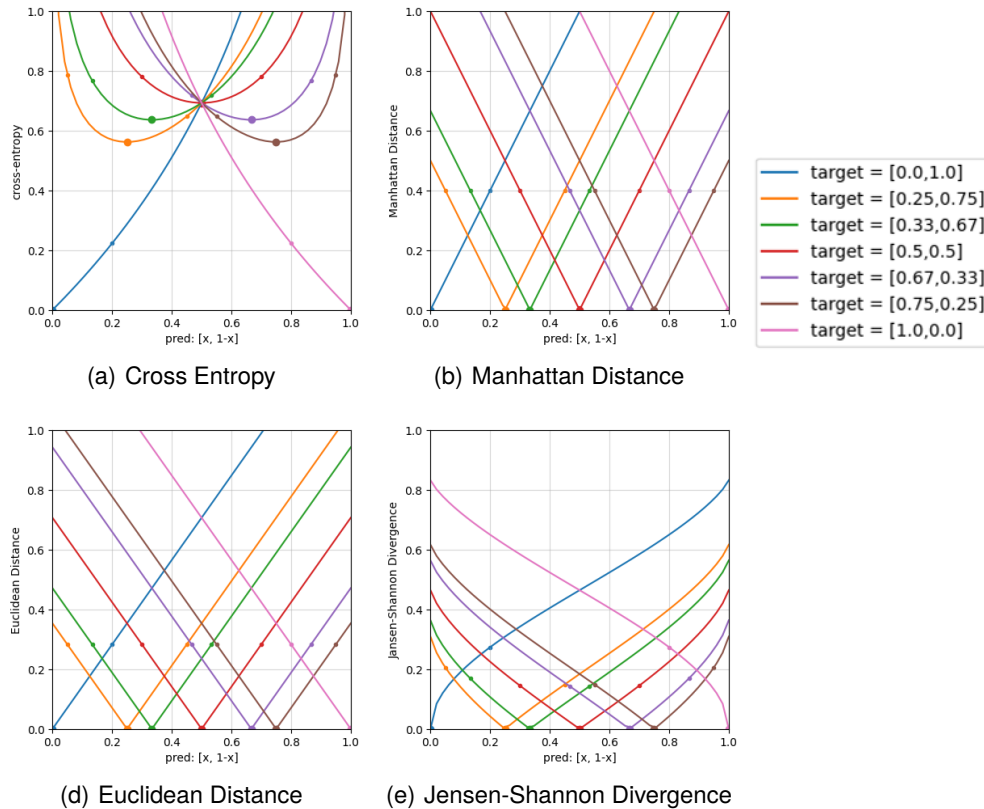


Figure 1: Metric Visualization: Each plot demonstrates the sensitivity of the analyzed metric to varying target values. X-axes represent prediction values, while Y-axes depict corresponding distance values according to the metric. Dots highlight distance values for accurately predicted targets, while points represent distance values when a nominal error of 0.2 appears in the prediction.

model the disagreements among annotators in four textual datasets that encompass different binary classification tasks (e.g. hate speech, offensive language, sexism detection). Teams competing in the shared task were asked to model annotators agreement/disagreement, represented in the form of soft labels: the probability of each item to be assigned to one class or the other is given by the agreement among annotators on the label. In the official competition to evaluate the performance of participants, Cross Entropy was considered the main evaluation metric. Here, for each of the four datasets that were part the LeWiDi task, rankings were recalculated for the evaluation metrics considered, and statistical difference was assessed from top to bottom using the Wilcoxon signed-rank test. Pairwise comparison among the different evaluation metrics, in terms of percentage of teams for which rank changed and the mean rank's position change, are summarized in Table 2. Results shown report the average value across the four datasets of the LeWiDi challenge.

From Table 2 we can observe how Cross Entropy rankings are substantially different from all the other metrics considered (although the mean position

Table 2: Pairwise comparison of evaluation metrics rankings: percentage of teams ranked differently and mean position's change across the LeWiDi datasets

| Evaluation metrics compared | % of teams re-ranked | Mean position change $\pm std$ |
|--|----------------------|--------------------------------|
| Cross Entropy vs Manhattan distance | 79% | 2.1 \pm 2 |
| Cross Entropy vs Euclidean distance | 73% | 2 \pm 2.1 |
| Cross Entropy vs J-S Divergence | 75% | 2 \pm 2 |
| Manhattan distance vs Euclidean distance | 2% | 0.1 \pm 0.2 |
| Manhattan distance vs J-S Divergence | 21% | 0.4 \pm 0.6 |
| J-S Divergence vs Euclidean distance | 23% | 0.4 \pm 0.6 |

change is relatively small). On the contrary, the other metrics produce more homogeneous results, with Manhattan distance and Euclidean distance exhibiting almost no difference. This confirms that

the metrics' differences in adhering to the properties outlined above, exert a certain influence on the application of the metrics in the real cases.

5. Multiclass Classification

Additional analyses have been performed considering the Multiclass Classification scenario. The most promising metrics, selected through the binary classification analysis (i.e., Manhattan distance and Euclidean distance), have been evaluated with respect to further desirable properties defined in the scope of multiclass classification.

Property 7 [*Non-Invariance with respect to the Most Probable Label*] Given three probability distributions $q(k)$, $r(k)$, and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), let $\mathbb{M}(p, q)$ and $\mathbb{M}(p, r)$ denote the distance measure between the two probability distributions and the golden standard if the most probable label in q corresponds to the target distribution p , and the most probable label in r does not correspond to the target distribution p , then $\mathbb{M}(p(k), q(k)) < \mathbb{M}(p(k), r(k))$.

The proposed property is not fulfilled by the selected metrics. For instance, Example 6 reports an example in which two different predictions lead to the same value, according to the Manhattan distance. However, $r(k)$ leads to a wrong classification, while $q(k)$ still preserves the ground truth of the target distribution.

| Example 6 | | |
|--------------------|----------------------|--------------------|
| Target p(k) | Prediction q(k) | Prediction r(k) |
| [0, 0.1, 0.1, 0.8] | [0.1, 0.3, 0.2, 0.4] | [0, 0.1, 0.5, 0.4] |

Manhattan Distance values:

$$\mathbb{L}_1(p(k), q(k)) = 0.8$$

$$\mathbb{L}_1(p(k), r(k)) = 0.8$$

Similarly, Example 7 reports an example in which, despite the most probable label in the second prediction ($r(k)$) does not correspond to the most probable label in the target prediction ($p(k)$), it is considered closer, according to the Euclidean distance, with respect to the other prediction ($q(k)$). However, in the last prediction, the most probable label corresponds to the most probable label in the target prediction.

| Example 7 | | |
|--------------------|----------------------|--------------------|
| Target p(k) | Prediction q(k) | Prediction r(k) |
| [0, 0.1, 0.4, 0.5] | [0.1, 0.2, 0.3, 0.4] | [0, 0.1, 0.5, 0.4] |

Euclidean Distance values:

$$\mathbb{L}_2(p(k), q(k)) = 0.2$$

$$\mathbb{L}_2(p(k), r(k)) = 0.1414$$

Property 8 [*Positional Error Sensitivity for Multiple Labels*] Given three probability distributions $q(k)$, $r(k)$ and $p(k)$ representing the probability of an item k being classified with the positive label by two different models ($q(k)$ and $r(k)$) and the corresponding real value associated with k in the golden standard ($p(k)$), if $\sum_i |p_i(k) - q_i(k)| \leq \sum_i |p_i(k) - r_i(k)|$, then $\mathbb{M}(p(k), q(k)) \leq \mathbb{M}(p(k), r(k))$.

The Manhattan distance confers equivalent significance to a substantial error on a single label and to minor distributed errors across multiple labels relative to the target distribution. In other words, even if a prediction leads to performing the smallest number of errors (implying a more realistic prediction that is close to the target one) it has the same distance of a probability distribution that spreads the wrong prediction across the remaining labels (having a distribution that is characterized by a higher entropy). On the other hand, the Euclidean distance penalizes more a single large error on a given label than small distributed errors on multiple labels.

An example of these behaviors is shown in Example 8. Even if a prediction results in the fewest number of errors (implying a more realistic prediction that is close to the target one), it achieves an equal or lower distance score (according to the Manhattan and the Euclidean distance respectively), than a probability distribution that spreads the incorrect prediction across the remaining labels. This indicates that the largest-scaled probability value will outperform the rest.

| Example 8 | | |
|-----------------|---------------------|----------------------------|
| Target p(k) | Prediction q(k) | Prediction r(k) |
| [0, 0, 0, 0, 1] | [0, 0, 0, 0.2, 0.8] | [0, 0.05, 0.05, 0.05, 0.8] |

Manhattan Distance values:

$$\mathbb{L}_1(p(k), q(k)) = 0.4$$

$$\mathbb{L}_1(p(k), r(k)) = 0.4$$

Euclidean Distance values:

$$\mathbb{L}_2(p(k), q(k)) = 0.2828$$

$$\mathbb{L}_2(p(k), r(k)) = 0.2236$$

The unfulfillment of this property can lead to some cases in which the Euclidean distance penalizes

less completely misclassified distributions than partial (erroneously) label distributions, as shown in Example 9.

| Example 9 | | |
|-------------------|---------------------------|-------------------|
| Target $p(k)$ | Prediction $q(k)$ | Prediction $r(k)$ |
| [0,0,0,0,0.3,0.7] | [0.25,0.25,0.25,0.25,0,0] | [0,0,0,0,1,0] |

Euclidean Distance values:

$$\mathbb{L}_2(p(k), q(k)) = 0.911$$

$$\mathbb{L}_2(p(k), r(k)) = 0.99$$

The identification of unique properties for Multiclass Classification problems is crucial due to the intricate nature of multiclass categorization itself. Multiclass settings frequently have hierarchical structures or allow for potential label relationships. The complexity of multiclass issues is further increased in multilabel classification, where multiple labels for instance are allowed. Specific properties for each classification problem might be defined, for instance, to deal with the concept of label similarity, to attribute a lower penalization for failures in predicting similar labels with respect to errors in predicting dissimilar labels. The proposed property offers a preliminary insight into the study of multiclass classification, highlighting the need for a more sophisticated understanding.

6. Related Work

We are grateful to one of the reviewers of this paper for directing us towards (Geng, 2016), which we had never previously encountered and appears to come from an entirely different research community. The objectives of that paper are, however, very different from ours, and closer to those of (Uma et al., 2021b). Geng considers six approaches to what he calls Label Distribution Learning and we would call Learning from Disagreement, and compares their performance on 16 datasets, none of which are of NLP tasks (1 is artificial, 11 are biological datasets, 3 are image understanding datasets, and 1 is movie ratings). To do this, he selects six metrics supporting a comparison between label distributions, chosen among 41 (!) measures proposed in previous literature—this selection is made in order to maximize diversity between the metrics. There is essentially no overlap between the metrics considered in the paper, and no proposal regarding the properties such metrics should satisfy, or analysis of the extent to which they satisfy them. This said, that paper does point out to the existence of an extensive literature on soft evaluation metrics we should investigate in the future.

7. Conclusion and Future Directions

In this paper, we propose a set of properties that soft evaluation metrics should have in order to allow for a fair comparison of computational models, and

assess the extent to which plausible candidate metrics satisfy these properties. Our analysis suggests that Manhattan distance and Euclidean distance are the most suitable metrics for a robust and fair soft evaluation for binary classification problems, since they adhere to all the desired properties. Our investigation of the LeWiDi real case scenario gave us some indication as to the impact of the adoption of different metrics in a real-case scenario, showing differences in the rankings definitions and thus implying the importance of selecting the best evaluation metric for ensuring a fair evaluation. Further preliminary analysis in the Multiclass Classification domain demonstrated however the unsuitability of the analyzed metrics to provide a fair comparison of models in this scenario. Future works will concentrate on Multiclass Classification and will include the definition of properties in accordance with the different task specifics (e.g. hierarchical, multilabel, etc). The performed analysis suggests the need for a novel metric that overcomes the limitations that arise in Multiclass Classification evaluation.

Ethical issues

This study analyzes the impact of metrics on a real-case scenario. Data from Learning With Disagreements task (LeWiDi) have been exploited. However, no sensitive information is used nor reported within the paper.

Limitations

The investigation of the application of the explored metrics limits to one real-case scenario (Learning With Disagreements task (LeWiDi) at SemEval 2023). The achieved results highlight a relationship among the entropy of the dataset and the impact of a variation of the evaluation metric on the leaderboard. The four LeWiDi datasets exhibit diverse characteristics such as types, languages, goals (misogyny, hate speech, offensiveness detection), and annotation methods and represent therefore a solid case-study. However, additional analysis on real-case scenarios would provide a deepen understanding of the studied phenomena.

Acknowledgments

The work of Paolo Rosso was in the framework of the FairTransNLP-Stereotypes research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. The work of Massimo Poesio is supported in part by the AINED Fellowship Grant *Dealing with Meaning Variation in NLP*, NGF.1607.22.002. Elisa Leonardelli's work has been partly supported by the Precrisis EU project (GA 101100539 - ISF-2022-TF1-AG-PROTECT).

The work of Elisabetta Fersini has been partially funded by MUR under the grant REGAINS, *Dipartimenti di Eccellenza 2023-2027* of the Department of Informatics, Systems and Communication at the University of Milano-Bicocca and by the European Union – NextGenerationEU under the National Research Centre For HPC, Big Data and Quantum Computing - Spoke 9 - Digital Society and Smart Cities (PNRR-MUR)

Bibliographical References

- Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. [A new measure of polarization in the annotation of hate speech](#). In *AI*IA - XVIIIth International Conference of the Italian Association for Artificial Intelligence*, Lecture Notes in Computer Science, page 588–603. Springer.
- Anthony McEnery and others. 2004. *The EMILLE/CiIL Corpus*. EMILLE (Enabling Minority Language Engineering) Project. distributed via ELRA: ELRA-Id W0037, ISLRN 039-846-040-604-0.
- 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.
- 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 *Proc. of the ACL-IJCNLP Workshop on Benchmarking: Past, Present and Future*.
- Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. [Dealing with disagreements: Looking beyond the majority vote in subjective annotations](#). *Transactions of the ACL*, 10:92–110.
- Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.
- Khalid Choukri and Niklas Paullson. 2004. *The OrientTel Moroccan MCA (Modern Colloquial Arabic) database*. distributed via ELRA: ELRA-Id ELRA-S0183, ISLRN 613-578-868-832-2.
- Elisa Leonardelli, Gavin Abercrombie, Dina Al-Manea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [SemEval-2023 task 11: Learning with disagreements \(LeWiDi\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement. In *Proc. of EMNLP*.
- Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salieb-Aouissi, and Nancy Ide. 2012. [Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations](#). *Language Resources and Evaluation*, 46(2):219–252.
- 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 Sogaard. 2014. Linguistically debatable or just plain wrong? In *Proc. EACL*.
- Massimo Poesio and Ron Artstein. 2005. [The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account](#). In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.
- Roventini, Adriana and Marinelli, Rita and Bertagna, Francesca. 2016. *ItalWordNet v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics “A. Zampolli”, National Research Council, in Pisa, ISLRN 532-206-426-067-2. PID <http://hdl.handle.net/20.500.11752/ILC-62>. Note: You don’t really need both an ISLRN and another PID, but it can’t hurt.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [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*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Speecon Consortium. 2011. *Catalan Speecon database*. SpeeCon. Speecon Project, distributed via ELRA: ELRA-Id S0327, Speecon resources, 1.0, ISLRN 935-211-147-357-5.
- Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, and Massimo Poesio. 2021a. Semeval-2021 task 12: Learning with disagreements. In *Proc. of SEMEVAL*. Association for Computational Linguistics.

Alexandra 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.

Designing NLP Systems That Adapt to Diverse Worldviews

Claudiu Creangă^{2,3}, Liviu P. Dinu^{1,3}

¹ Faculty of Mathematics and Computer Science

² Interdisciplinary School of Doctoral Studies, ³ HLT Research Center

University of Bucharest, Romania

claudiu.creanga@s.unibuc.ro, ldinu@fmi.unibuc.ro

Abstract

Natural Language Inference (NLI) is foundational for evaluating language understanding in AI. However, progress has plateaued, with models failing on ambiguous examples and exhibiting poor generalization. We argue that this stems from disregarding the subjective nature of meaning, which is intrinsically tied to an individual's *weltanschauung* (which roughly translates to worldview). Existing NLP datasets often obscure this by aggregating labels or filtering out disagreement. We propose a perspectivist approach: building datasets that capture annotator demographics, values, and justifications for their labels. Such datasets would explicitly model diverse worldviews. Our initial experiments with a subset of the SBIC dataset demonstrate that even limited annotator metadata can improve model performance.

Keywords: weltanschauung, perspectivism, alignment

1. Introduction

Natural Language Inference (NLI) lies at the heart of developing and evaluating language understanding in AI models. As Montague stated, entailment is "the basic aim of semantics" (Montague, 1970) and a lot of focus has been put in building models that score high on NLI datasets. Recently, two big issues emerged. Firstly, the research has reached a plateau on these datasets where the models perform almost as well as humans on samples where there is high human agreement, but perform poorly on samples with high entropy (from 0.9 to 0.5 accuracy) (Nie et al., 2020a). Secondly, models have poor generalisation abilities and their high score on in-distribution samples doesn't translate to a high score on out-of-distribution samples (Bras et al., 2020), (Zhou and Bansal, 2020), a sign that they actually use shallow heuristics rather than understanding. We will argue that these models lack a worldview and that using perspectivist approaches we can improve our solutions to both problems. A significant obstacle in this research is the scarcity of datasets that preserve annotator demographics and reveal their socio-political values, which are crucial for understanding their worldviews. Our code is open source and available to use on [Anonymous GitHub](#). Our present work focuses on worldviews which are shaped culture, values, beliefs and less by demographic data. Several studies (Orlikowski et al., 2023) have shown that demographic data alone is not a good predictor of annotator's views.

2. Related Work

Basile et al. (2021) offer a valuable synthesis of prior research in perspectivist machine learning,

clearly delineating two distinct approaches within the field: weak and strong perspectivism. Historically, the illusion of ground truth was ingrained in every NLP dataset. At start, NLP datasets were built only with a single annotator per label. That label was taken as the truth, even if the language was ambiguous or the annotator made a mistake. Then, a step forward was made when weak perspectivist research acknowledged the potential for disagreement and errors. It therefore adopted a multi-annotator approach to capture diverse perspectives. However, where there was human disagreement, this was solved either by aggregation (majority voting) or by filtering (removing low agreement samples). But, by removing the low agreement samples, we remove a big part of human speech which is by its virtue ambiguous and, by aggregation, we obscure the rich diversity of valid interpretations inherent in human communication. Strong perspectivist research embraces linguistic diversity, recognizing that even when aiming for a single target label, models benefit from exposure to non-aggregated data (manifesto¹). In the literature we've found three main ways to embrace variation:

- **Multilabel categorical classification:** where a single model predicts multiple labels for each sample (Ferracane et al., 2021), (Jiang and de Marneffe, 2022) and others;
- **Soft label classification:** where a single model is trained on the distribution of labels for a given sample and the model predicts that distribution (Peterson et al., 2019), (Uma, 2020) and others.
- **Radical perspectivist:** where we train a model for each annotator so that a model

¹<https://pdai.info/>

learns the behaviour of one particular annotator (Akhtar et al., 2021). This requires the identification of each annotator, which few datasets have.

Our research continues in the line of the radical perspectivist approach and takes it one step further, by including not only the demographics of the annotators, but also the values and beliefs which make the annotator's worldview.

3. The goal of an NLP system

NLP is a large field and uses of NLP systems can vary from translations to information extraction, summarization and others. But at its core, we can say that an NLP system aims to enable computers to understand and generate human language in a meaningful way. We argue that meaning is subjective and cannot be separated from a worldview. Our language influences how we perceive and understand the world and the other way around. The subjective nature of meaning doesn't entail that each annotator position is equally valid under the current paradigm (an extreme relativistic view that can have negative implications in some tasks like hate speech detection (Curry et al., 2024)), but that each position is part of a worldview that should be dealt with, not erased.

3.1. Meaning and Worldview

Quine's radical indeterminacy (Quine, 1980) shows that no sentence has only one meaning and there is no way to determine the one correct translation of a sentence in another language. This happens because meaning is embedded within the speaker's entire web of beliefs, culture, and how they experience the world. The term that better describes this concept is **Weltanschauung**, roughly translated by worldview. The main problem of NLP systems today is that it doesn't take into account the different worldviews in which a sentence can be interpreted. We cannot know the right label for large category of utterances if we don't contextualise it in a worldview. If we accept Quine's position that there's no such thing as **purely mental meanings** (mental dictionaries from which we take the definition of every word we use) and that what words mean is inextricably linked to how speakers behave and look at the world (*weltanschauung*), then we can agree that the legacy NLP datasets do not offer the information we need. By **legacy datasets** we mean datasets that only have aggregated labels. But non-aggregated labels are not enough and we should go further and include those that do not offer labels by annotator id, demographic metadata and worldview metadata about each annotator.

Here we need to recognize that while many inferences are influenced by worldview, certain types – such as those grounded in formal logic or mathematical reasoning – hold regardless of the annotator (**deductive inferences**). If we were to use logic to establish what inference means, we would denote it by this formula:

$$\forall w \in W : (P(w) \rightarrow H(w))$$

Which means that for every world w , if the premise P is true in world w , then the hypothesis H is also true in world w . Only deductive inferences can pass this type of rigour:

Premise: All men are mortal and Socrate is a man.

Hypothesis: Socrate is mortal.

Label: Entailment.

Here, no matter what we mean by Socrate and men, the inference is still valid. Compared to deductive inferences, **inductive inferences** need to pass a lower bar. Traditionally the standard was what a "common man" would assume to be true or false about an utterance. The creators of datasets sometimes give annotators instructions on how to evaluate utterances (Bowman et al., 2015). In case of SNLI, entailment meant a "definitely true description" and neutral "might be a true description". In case of MNLI the instructions for entailment was "definitely correct" and "might be correct" for neutral. In Gubelmann 2024 the threshold is much more lax, a "good reason" is sufficient for entailment and it is given the following example:

Premise: The streets are wet.

Hypothesis: It has rained.

Label: Entailment.

Firstly, we argue that even a simple inference like this cannot be made without considering the annotator's context. For instance, an annotator living in a town where streets are washed every morning might attribute wet streets to cleaning rather than rain as his first thought. Secondly, providing annotators with instructions biases the results, as it attempts to override their individual understanding of entailment and contradiction. This prevents the study from authentically reflecting common language use which comes in different varieties. And NLP systems should reflect how humans naturally speak.

3.2. Building a worldview-annotated dataset

Consider the following basic pair of sentences:

Premise: It is dark outside.

Hypothesis: It is dangerous to go outside.

There is no right label for this pair and the annotator would use his life experiences to annotate it. A woman would be more likely to annotate it as an entailment. Many men who live in high crime areas would potentially do the same, while others would see no good reason to be dangerous outside if it is dark. Recently built perspectivist datasets, such as ChaosNLI (Nie et al., 2020b) which collects 100 annotations for each label, but then proceeds to remove the worldviews of annotators and provide only a distribution (i.e. 30% E, 50% N, 20% C) make the same mistake of weak perspectivist research. Relying solely on a distribution, one cannot learn in which worldview this statement is entailment or not and if we don't know the annotator backgrounds and how diverse they are, we risk to capture only a limited range of potential interpretations.

In Figure 1 we explain how we can build worldview-annotated datasets. It is necessary to have a diverse pool of annotators aligned to the task. For instance, a BioNLP task would benefit from annotators with expertise in health sciences. Metadata should be collected about each of these annotators: demographics and values. Worldviews are part of a paradigm, a set of principles shared by every annotator at a certain point in time. Each annotator should label items according to their worldview, while being mindful of potential "noise" – errors caused by factors like lack of attention, which are unrelated to their perspective. To mitigate noise and preserve valid interpretations, annotators should provide justifications for their labels. They will then self-review their labels based on their own explanation and reconsider the label or not. This self-review step should remove some noise. The **justifications** are important for a second reason. How can we address the scenario where two annotators assign the same label, but their justifications diverge so significantly that the apparent agreement is misleading? This case suggests that reasons must be kept in the dataset and if reasons diverge significantly there should be different categories in each label. A model that learns a worldview should arrive at a label through the same reasoning as the annotator. Unfortunately there is no public dataset where reasons are available, but we aim to build one and we hypothesize that it would help the models in learning a worldview.

Training a model in this way means modelling a worldview in which the model doesn't pretend universality and is by essence local. We hypothesize that this grounding in a worldview will make

the model more consistent in its judgements and generalise better.

4. Modelling a worldview

Unfortunately, the demographic metadata of annotators is removed in most datasets and there are very few datasets that collected annotators' social and political values in order to create a worldview. However, this is changing, with at least four datasets that we are aware of now including more information: Chulvi 2023, Sachdeva et al. 2022, Sap et al. 2019 and Hettiachchi 2023. In (Sap et al., 2022) we see that annotators with stronger racist beliefs demonstrate a tendency to mislabel African American English (AAE) as toxic, while being less likely to identify anti-Black language as harmful, suggesting that even only demographic metadata can be helpful. To the best of our knowledge, SBIC (Sap et al., 2019) is the most complete dataset in regards to including both annotator demographics and the political beliefs of the annotator (liberal or conservative). Their annotators are from U.S. and Canada and, although they are gender and age-balanced, the ethnicity is not diverse, there are a lot more white (82%) than any other group (4% Asian, 4% Hispanic, 4% Black). They annotated posts from Reddit and Twitter for offensiveness, intent to offend, sexual references and, if it was offensive, which groups did it target.

We used this dataset for our goal of building models with demographic and political metadata. Because the dataset was built with detecting Social Bias Frames in mind, they had 263 different annotators and at most 3 annotators per post. For our use case we had to find annotators from different backgrounds that annotated the same posts. The biggest subset that we could find from this dataset was for 2 annotators, man (liberal, white) and woman (mod-conservative, white), which together annotated 290 samples (worker ids are hidden due privacy concerns).

Given it is a small dataset, we are aware of the limitations of our results. We used K-Fold validation (10) to reduce the risk of overfitting and we used a pre-trained DeBERTaV3 model (He et al., 2021) followed by a fully connected layer for predictions. We used a learning rate of 5e-5 for 3 epochs where we trained only the output layer and then a small learning rate of 2e-5 for one epoch where we trained the output layer and the last layer from the pre-trained model. The target label is if the post is offensive or not.

As shown in (Table 1), leveraging annotator metadata provides a boost in performance. Training on aggregated data without metadata yields a test F1 score of 0.3. While splitting the data by annotator, training 2 separated models and ag-

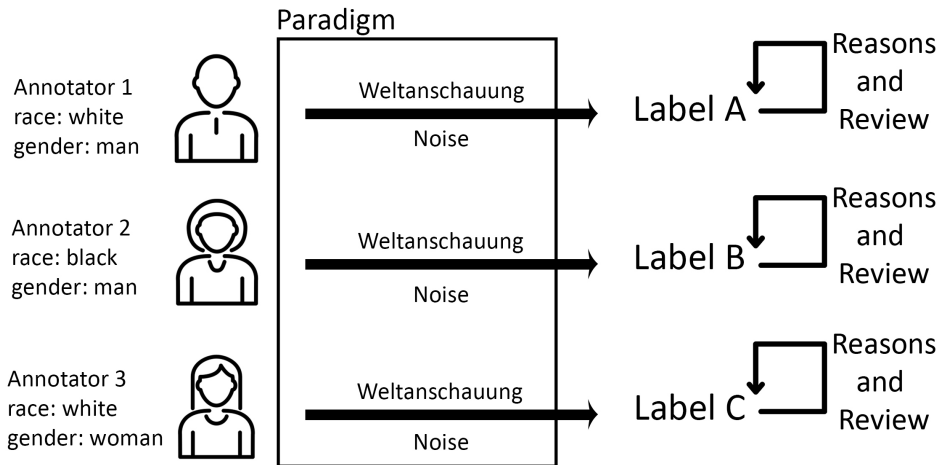


Figure 1: Building a worldview-annotated dataset. It is necessary to have a diverse pool of annotators aligned to the task. Metadata should be collected about each of these annotators: demographic and values. Each annotator should label items according to their worldview, while being mindful of potential "noise" – errors caused by factors like inattention, which are unrelated to their perspective. To mitigate noise and preserve valid interpretations, annotators should provide justifications for their labels. They will then self-review their labels based on their own explanation.

Table 1: F1 scores of the three types of training: training on aggregated data, using 2 models for each annotator with and without metadata.

| Training Type | Val score | Test score |
|-----------------------|-----------|------------|
| Aggregated data | 0.3 | 0.3 |
| 2 models no metadata | 0.37 | 0.36 |
| 2 models and metadata | 0.38 | 0.38 |

gregating the outputs afterwards improves the F1 score to 0.36. Instead, if we also concatenate the annotators' metadata to the input before the encoding layers, we increase the F1 score to 0.38. Even though the dataset is small, there is an improvement when we consider the annotators' background.

5. Conclusion

Traditional NLP approaches have exhibited limitations in both generalization and performance on challenging examples within NLI datasets. Part of these shortcomings stem from the omission of the subjective nature of meaning and the diverse worldviews that shape individual interpretations of language. For inductive inferences, the notion of a universally "correct" label detached from a worldview is misleading.

We propose that embracing perspectivist approaches and building worldview-annotated datasets is crucial for advancement in NLP. Such datasets must capture annotator justifications, demographic information, and the values that com-

prise their worldview, offering a richer understanding of linguistic variation. Such datasets are missing at the moment, but initial experiments with a subset of the SBIC data support our hypothesis: even with limited demographics and a basic political orientation label, incorporating annotator metadata improves model performance.

Acknowledgements

This work was partially supported by a grant on Machine Reading Comprehension from Accenture Labs and by the POCIDIF project in Action 1.2. "Romanian Hub for Artificial Intelligence".

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.](#)

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. [Toward a perspectivist turn in ground truthing for predictive computing.](#)

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference.](#) In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. [Adversarial filters of dataset biases](#).
- Fontanella-L. Labadie-Tamayo R. Rosso P. Chulvi, B. 2023. [Social or individual disagreement? perspectivism in the annotation of sexist jokes](#).
- Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. [Subjective *Isms*? on the danger of conflating hate and offence in abusive language detection](#).
- Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. [Did they answer? subjective acts and intents in conversational discourse](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.
- Katis-I. Niklaus-C. et al Gubelmann, R. 2024. [Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks](#). page 21–48.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). In *International Conference on Learning Representations*.
- Holcombe-James I.-Livingstone S. de Silva A. Lease M. Salim F. D. Sanderson M. Hettiachchi, D. 2023. [How crowd worker factors influence subjective annotations: A study of tagging misogynistic hate speech in tweets](#).
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. [Investigating reasons for disagreement in natural language inference](#). *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020a. [What can we learn from collective human opinions on natural language inference data?](#)
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. [What can we learn from collective human opinions on natural language inference data?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. [The ecological fallacy in annotation: Modelling human label variation goes beyond sociodemographics](#).
- Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. [Human uncertainty makes classification more robust](#).
- W. V. O. Quine. 1980. Two dogmas of empiricism. from a logical point of view. page 20–46.
- Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. [The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2019. [Social bias frames: Reasoning about social and power implications of language](#).
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [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*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.
- Fornaciari T. Hovy-D. Paun S. Plank B. Poesio M. Uma, A. 2020. [A case for soft loss functions](#).
- Xiang Zhou and Mohit Bansal. 2020. [Towards robustifying NLI models against lexical dataset biases](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.

The Effectiveness of LLMs as Annotators: A Comparative Overview and Empirical Analysis of Direct Representation

Maja Pavlovic¹ and Massimo Poesio^{1,2}

¹Queen Mary University of London, ²University of Utrecht
m.pavlovic@qmul.ac.uk, m.poesio@{qmul.ac.uk,uu.nl}

Abstract

Large Language Models (LLMs) have emerged as powerful support tools across various natural language tasks and a range of application domains. Recent studies focus on exploring their capabilities for data annotation. This paper provides a comparative overview of twelve studies investigating the potential of LLMs in labelling data. While the models demonstrate promising cost and time-saving benefits, there exist considerable limitations, such as representativeness, bias, sensitivity to prompt variations and English language preference. Leveraging insights from these studies, our empirical analysis further examines the alignment between human and GPT-generated opinion distributions across four subjective datasets. In contrast to the studies examining representation, our methodology directly obtains the opinion distribution from GPT. Our analysis thereby supports the minority of studies that are considering diverse perspectives when evaluating data annotation tasks and highlights the need for further research in this direction.

Keywords: large language model (llm), annotation/labelling, representation

1. Introduction

Large Language Models (LLMs) have shown impressive abilities in a variety of natural language related tasks (Brown et al., 2020; Touvron et al., 2023). Brown et al. (2020) demonstrate their ability as few-shot learners and Wei et al. (2022); Kojima et al. (2022) evidence their zero-shot capabilities. Recognising the significance and costliness of annotated data across various research domains, recent work explores the potential of LLMs as data annotators, encompassing both zero- and few-shot learning approaches (Lee et al., 2023; Ziems et al., 2024; Törnberg, 2023; Zhu et al., 2023; Gilardi et al., 2023; Mohta et al., 2023; Ding et al., 2023; He et al., 2023). Considering that LLMs are trained to adhere to instructions guided by human preference (Ouyang et al., 2022; Rafailov et al., 2023), studies examine the extent to which human disagreement is captured (Lee et al., 2023) and whether or not such disagreement aligns with that of humans (Santurkar et al., 2023).

Our work, firstly, offers a comparative overview of twelve previous studies that investigate the capabilities of LLMs as annotators, concentrating on classification tasks and considering whether disagreement is captured by the studies. Secondly, we present an empirical analysis concentrating more specifically on the perspectivist question. We compare the top-performing LLM from the first section (GPT) against human annotators, by examining the degree of alignment between their opinion distributions, for the case of the four subjective datasets recently used for the 2023 SEMEVAL Task on Learning With Disagreement (Leonardelli et al., 2023).

2. Comparative Overview

Labelled data forms the foundation for training supervised models across diverse machine learning tasks. Much recent research has focused on exploring the use of LLMs as a quicker and more cost-effective alternative to traditional data annotation. In this first Section we review the research in this area. Due to rapid developments in this space, we concentrate on works from the past year which leverage recent models with a focus on classification tasks. Our approach to selecting relevant papers followed a combination of keyword searches, monitoring relevant workshops and conferences, and examining citations.

Studies: Wang et al. (2021) employ GPT-3 for the annotation of datasets, which are subsequently used in the training of smaller models. Huang et al. (2023) explore the capability of ChatGPT to accurately label implicit hate speech and provide good explanations for its annotations. Zhu et al. (2023) also investigate the capability of GPT for labelling and He et al. (2023) introduce a two step approach in which they first prompt the LLM to generate explanations and then annotate a sample to improve the annotation quality of LLMs. Both Törnberg (2023); Gilardi et al. (2023) contrast the performance of GPT with that of crowd-workers. Whereas, Goel et al. (2023) introduce a two-stage semi-automated approach employing LLMs and human experts to accelerate annotation for the extraction of medical information. Ziems et al. (2024) conduct a large scale empirical analysis to understand the zero-shot performance of GPT and Flan on 25 computational social science (CSS) benchmarks.

| Paper | model families | # of model versions | # of data-sets | # of metrics | Zero/few shot | Lang. | Dis-agree. | LLM as Anno. |
|--------------------------|--------------------------|---------------------|----------------|--------------|---------------|----------|------------|--------------|
| (Lee et al., 2023) | GPT,Vicuna, Flan,OPT-IML | 9 | 6 | 4 | z&f | en | ✓ | ✗ |
| (Santurkar et al., 2023) | GPT,Jurassic | 9 | 1 | 3 | z | en | ✓ | ✗ |
| (Ziems et al., 2024) | GPT,Flan | 14 | 20 | 2 | z&f | en | ✗ | ✓ |
| (Zhu et al., 2023) | GPT | 1 | 5 | 5 | z&f | en | ✗ | (✓) |
| (Gilardi et al., 2023) | GPT | 1 | 4 | 2 | z | en+ | ✗ | ✓ |
| (Törnberg, 2023) | GPT | 1 | 1 | 3 | z | en | ✗ | ✓ |
| (Mohta et al., 2023) | Vicuna, Flan,Llama | 9 | 5 | 3 | z | en,fr,nl | ✗ | ✗ |
| (Ding et al., 2023) | GPT | 1 | 4 | 4 | z&f | en+ | ✗ | ✓ |
| (He et al., 2023) | GPT | 1 | 3 | 1 | z&f | en | ✗ | ✓ |
| (Huang et al., 2023) | GPT | 1 | 1 | 2 | z | en | ✗ | ✓ |
| (Goel et al., 2023) | Palm | 1 | 1 | 3 | f | en | ✗ | ✓ |
| (Wang et al., 2021) | GPT | 1 | 9 | 2 | f | en | ✗ | ✓ |

Table 1: Overview on LLM’s as Annotators (Language codes follow ISO 639, en+: predominantly English, with some additional language explorations)

Language: The majority of these studies measure LLM performance on English corpora (see Table 1). However, Ding et al. (2023) conduct tests to understand the possibility of using GPT on non-English corpora and Mohta et al. (2023) investigate the performance of open source LLMs on French, Dutch and English natural language inference (NLI) tasks. Thus far, models have shown better performance on English related tasks and performed notably poorly on low-resource languages Srivastava et al. (2023). While Ding et al. (2023) see potential for GPT on languages other than English, Mohta et al. (2023) observe a considerable decline in performance with non-English languages.

Annotator Disagreement: All studies referenced thus far assume the existence of a singular ground truth label for a given sample. There has, however, been a shift in thinking across machine learning towards a collectivist approach, meaning the inclusion of all annotator perspectives rather than having a majority voted ground truth (Uma et al., 2021; Prabhakaran et al., 2021; Cabitza et al., 2023; Rottger et al., 2022; Nie et al., 2020; Pavlick and Kwiatkowski, 2019). In this context, Lee et al. (2023) explore whether LLMs can capture the human opinion distribution. Additionally, Santurkar et al. (2023) investigate the alignment between LLMs and human annotators with respect to the

opinions and perspectives reflected in response to subjective questions. From Table (1) we can see that the latter two studies which investigate the performance of LLMs on opinion distributions don’t yet deem them ready as annotators. However, all studies that investigate the capabilities of GPT as an annotator within the traditional framework of majority voted labels agree with varying degrees that LLMs have the potential to disrupt the annotation process. Within this paradigm of majority voting, the sole exception to the consensus is expressed by Mohta et al. (2023) who conclude that LLMs have not yet attained a sufficient level for the annotation of datasets. Notably, amongst the cited studies, they are the sole study to only use open source LLMs and not consider best performing closed source alternatives (see Table 1).

Models: As mentioned in the previous paragraph, the predominant focus across all studies lies on models belonging to the GPT series. The remaining models under consideration are mostly open-source options, with Flan being the second most investigated, succeeded by Vicuna. Table 1 highlights that only four studies explored model families beyond GPT. Notably, these same studies explored multiple versions of a given model (“# of model versions”). In contrast, the remaining studies exclusively assessed a singular model. More details

on the exact versions can be found in table 7 (Appendix A).

Temperature Parameter: Not all studies mention the settings of their temperature parameter. However, both Törnberg (2023); Gilardi et al. (2023) investigate the variability in responses by experimenting with lower (0.2) and high (1.0) temperature settings. They find that LLMs have higher consistency with lower temperatures without sacrifices in accuracy and thus recommend lower values for annotation tasks. Ziems et al. (2024) and Goel et al. (2023) opt for a temperature of 0 throughout their study, aiming to ensure consistent and reproducible results across their LLM analysis.

Prompting: Wang et al. (2021) and Goel et al. (2023) investigate the efficacy of LLMs as annotators using only few-shot prompting. In contrast, five of the subsequent studies experiment with both zero- and few-shot prompting. Additionally, five other studies employ zero-shot prompting for their annotation tasks (see Table 1). The outcomes of the experiments comparing zero-shot and few-shot prompting show inconsistency. Mohta et al. (2023) experience superior performance using few-shot prompting, while Ding et al. (2023) find that few-shot prompting does not yield superior results across all their approaches. He et al. (2023) report a decrease in performance with few-shot prompting for their specific task. Ziems et al. (2024) conclude that improvements from few-shot prompting are inconsistent across their experiments, suggesting that achieving more substantial gains would require increased efforts in refining the prompting process.

| Paper | Accuracy | F1 | Precision | Recall | Reliability | Other |
|--------------------------|----------|----|-----------|--------|-------------|-------|
| (Lee et al., 2023) | ✓ | - | - | - | - | ✓ |
| (Santurkar et al., 2023) | - | - | - | - | - | ✓ |
| (Ziems et al., 2024) | - | ✓ | - | - | ✓ | - |
| (Zhu et al., 2023) | ✓ | ✓ | ✓ | ✓ | - | ✓ |
| (Gilardi et al., 2023) | ✓ | - | - | - | ✓ | - |
| (Törnberg, 2023) | ✓ | - | - | - | ✓ | ✓ |
| (Mohta et al., 2023) | ✓ | ✓ | - | - | - | ✓ |
| (Ding et al., 2023) | ✓ | ✓ | ✓ | ✓ | - | - |
| (He et al., 2023) | ✓ | - | - | - | - | - |
| (Huang et al., 2023) | ✓ | - | - | - | - | ✓ |
| (Goel et al., 2023) | - | ✓ | ✓ | ✓ | - | ✓ |
| (Wang et al., 2021) | ✓ | - | - | - | - | ✓ |

Table 2: Evaluation Metrics across Papers

Evaluation: Nearly all studies assess their outcomes using metrics such as accuracy or F1. Santurkar et al. (2023) deviate from these conventional performance metrics as, their primary focus lies in evaluating representation. This emphasis leads them to assess LLMs responses based on metrics measuring representativeness, steerability, and consistency (Santurkar et al., 2023). In addition to accuracy and F1, three studies utilise metrics such as precision and recall, while three other studies employ different reliability measures to evaluate inter-coder agreement. Törnberg (2023); Santurkar et al. (2023) specifically investigate model bias, whereas Huang et al. (2023) evaluate the natural language explanations (NLE) that LLMs can provide for their predictions. For the evaluation of LLM and human opinion distributions, Lee et al. (2023) use entropy, Jensen-Shannon divergence (JSD), and the Human Distribution Calibration Error (DistCE) introduced by Baan et al. (2022). Two studies have conducted error analyses. Huang et al. (2023) observe that the instances of disagreement, comprising 20% in their study, align more closely with lay-people’s perspectives. Similarly, Ziems et al. (2024) conclude that in their error analysis, the LLM tends to default to more common label stereotypes. Given the reported accuracy-based performance of LLMs on labelling tasks, it is important to broaden metrics to include more representational measures. For example, Ziems et al. (2024) omit measuring bias in their study, concluding that larger, instruction-tuned models demonstrate superior performance. However, Srivastava et al. (2023) caution that larger models tend to amplify bias.

2.1. Benefits

Törnberg (2023) finds that gpt-4 consistently surpasses the performance of both crowd-workers and expert coders, and the cost associated with labeling a sample is orders of magnitude lower for LLMs compared to humans. Wang et al. (2021) provide a detailed explanation that, in their experiments, utilising labels generated by the LLM resulted in a cost reduction ranging from 50% to 96%, while maintaining equivalent performance in downstream models. Similarly, Goel et al. (2023) determine that the LLM reduces the total time of labelling by 58% while maintaining a comparable baseline performance to medically trained annotators. Gilardi et al. (2023) demonstrate that the LLM shows superior quality compared to annotations obtained through Amazon Mechanical Turk (MTurk), while being approximately 30 times more cost-effective. Ding et al. (2023) find that their approach attains nearly equivalent performance when labeling the same number of samples. However, when they double the amount of data labeled by the LLM, superior performance is achieved at only 10% of the

cost associated with human annotation (Ding et al., 2023). LLMs not only entail lower costs than human annotators but also demonstrate significantly higher speeds in the labeling process (Törnberg, 2023; Wang et al., 2021; Ding et al., 2023).

In addition to diminished cost and time requirements, LLMs demonstrate the capability to provide explanations for their annotation (Mohta et al., 2023). Huang et al. (2023) find that ChatGPT generates explanations comparable, if not superior in clarity, to those produced by human annotators.

2.2. Limitations

As mentioned in Section 2, one limitation lies in the predominant development and testing of LLMs within the confines of the English language. An additional constraint associated with using LLMs as annotators is the challenge in formulating prompts and obtaining meaningful responses. Models might generate unconstrained responses (Goel et al., 2023) or might refrain from providing responses altogether as a result of the implementation of safeguarding measures. Ziems et al. (2024) observed that models tended to predict beyond the presented labels and exhibited a tendency to abstain from responding to tasks deemed offensive. In the event that a model does provide a response, potential issues may arise in the form of bias. Srivastava et al. (2023) show that bias in LLMs increases in with scale and ambiguous contexts. Santurkar et al. (2023) identify that LLMs demonstrate a singular perspective characterised by left-leaning tendencies. Törnberg (2023) notes the absence of substantial disparities between expert annotators and LLMs, while underscoring the notable bias observed among annotators from MTurk. However, Goel et al. (2023) underscore the importance of expert human annotators in attaining high-quality labels. Lee et al. (2023) express concerns regarding the population representation capabilities of current LLMs, whereas Ziems et al. (2024) caution researchers to consider and mitigate the potential risks of bias in their applications through human-in-the-loop methods.

An additional noteworthy limitation in employing LLMs as annotators is their sensitivity to minor alterations in prompting (Loya et al., 2023; Sclar et al., 2024). Both Huang et al. (2023) and Ziems et al. (2024) assert the need for further research to comprehensively investigate the effects of prompting and determine optimal strategies for effective prompting. Lastly, it is important to note that these models show sub-optimal performance as annotators in tasks such as NLI, implicit hate classification, empathy or dialect detection (Lee et al., 2023; Ziems et al., 2024).

3. Results with the SEMEVAL 2023 Subjective Tasks Benchmark

As discussed above, most studies of LLMs as annotators still adopt a majority vote perspective, which is becoming increasingly questionable particularly for subjective tasks (Akhtar et al., 2021; Leonardelli et al., 2021; Uma et al., 2021; Plank, 2022; Cabitza et al., 2023). We decided therefore to carry out a preliminary exploration of the alignment between LLM and human judgment distributions on the datasets used in the recent SEMEVAL 2023 Shared Task on Learning with Disagreement (Leonardelli et al., 2023). Our analysis is centered on the extent to which the most frequently used model (GPT) matches human distribution on datasets for inherently subjective tasks. This was done by extracting opinion distributions in the simplest and most straightforward manner possible: we directly prompt GPT to provide its estimation of the human opinion distribution and compare it against the baseline and optimal results from SemEval-2023.

| Dataset | Task | Lang. | # items train dev test | % full agree. |
|-----------|-------------------------------------|-------|---------------------------------|------------------|
| MD-Agree. | Offensiveness detection | en | 6592 1104 3057 | 42% |
| HS-Brexit | Offensiveness detection | en | 784 168 168 | 69% |
| ConvAbuse | Abusiveness detection | en | 2398 812 840 | 86% |
| ArMIS | Misogyny and sexism detection | ar | 657 141 145 | 65% |

Table 3: Dataset statistics (Leonardelli et al., 2023) (Language codes follow ISO 639)

3.1. Datasets

We leverage four datasets from SemEval2023 on "Learning with Disagreements" for the empirical analysis. All four datasets focus on subjective tasks and contain human annotated target distributions that we compare to the LLM predictions. Table 3 contains key statistics on the datasets (Leonardelli et al., 2023).

Multi-Domain Agreement: MD-Agreement (Leonardelli et al., 2021) is the dataset with the lowest amount of annotator agreement amongst these subjective tasks. Each example was labelled by 5 annotators and was created using English tweets from three domains (BLM, Election2020 and Covid-19).

Hate Speech on Brexit: HS-Brexit (Akhtar et al., 2021) was constructed from English tweets using keywords related to immigration and Brexit. Each example was labelled by 6 annotators with 69% of items having total annotator agreement.

ConvAbuse: ConvAbuse (Cercas Curry et al., 2021) consists of English conversational text collected from dialogue between users and two conversational AI systems. Each example was labelled by between 3 and 8 annotators. 86% of items have total annotator agreement.

Arabic Misogyny and Sexism: ArMIS (Almanea and Poesio, 2022) is the only non-English language task and serves to study the effect on sexism judgements particularly with respect to the annotators leanings towards conservatism or liberalism. Each example was labelled by 3 annotators with 65% of items having total annotator agreement.

3.2. Experimental Parameters

We explore the capability of `gpt-3.5-turbo` to generate opinion distributions for the test data of each SemEval2023 task. Given the sensitivity of LLMs to minor changes in input (Loya et al., 2023; Sclar et al., 2024), we maintain a uniform prompt structure across various tasks and let the LLM assume the role of an expert annotator who considers multiple worldviews and cultural nuances. Modifications are made only on the words related to the respective task under consideration. For instance, in the case of HS-Brexit, the LLM specialises in "hate speech detection," whereas in the ConvAbuse dataset its specialisation lies in "abusiveness detection." ArMIS is approached with slight variation due to the presence of Arabic text. In this instance, we explore two approaches: one involves prompting the models in English and providing them with the Arabic text that requires labelling, while the second approach uses an Arabic prompt (a translated version of the English prompt).

As mentioned in Section 2 there is some variability both among and within studies regarding the preferred prompting approach for LLM annotation. However, given that the multiple studies indicate limited benefits from few-shot prompting, we opt for zero-shot prompting in our tasks. The expectation of a model’s output on a labelling task is to be consistent. In order to achieve such consistent and reproducible results we set the temperature parameter across our models to zero such as Ziems et al. (2024). Gilardi et al. (2023) suggest that a lower temperature value might be preferable for annotation task as it increases consistency without decreasing accuracy across their empirical analysis.

3.3. Evaluation Metrics

We compare the performance of GPT to both the Semeval2023 baseline model as well as the top-performing model on each task. Leonardelli et al. (2023) evaluate point predictions using the F1 measure (1) and distribution similarity using Cross-Entropy (CE) (2). To ensure comparability we use both of these in our analysis.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN} \quad (1)$$

$$CE(y_n, \hat{y}_n) = - \sum_{n=1}^N y_n \log(\hat{y}_n), \quad (2)$$

where y_n is a sample opinion distribution annotated by humans and \hat{y}_n the LLMs predicted distribution for that sample. In addition to the above, we also use Shannon’s entropy to visualise human and LLM uncertainties.

3.4. Results

Figures 1, 2, 3 and 4 contrast the frequency of opinion distributions of human annotators with those predicted by GPT for each SemEval task. We observe that when prompted directly for opinion distributions, the model shows a tendency towards bimodal predictions, with a notable preference for the following opinion distributions: {"0":0.2, "1":0.8} and {"0":0.8, "1":0.2}.

Another notable observation is evident in Figure 1, where we observe a bias towards assigning greater weight to the sexist class ('1') when prompting the LLM with Arabic text. In fact, when these distributions are simplified to a majority-based label, all test samples are categorised as sexist, a pattern not observed when the LLM was prompted with English text. The difference is also evident in the F1 performance (Table 4). The LLM prompted in Arabic only achieves an F1 score of 0.256, whereas prompting the LLM in English results in a score of 0.448, suggesting that LLMs perhaps understand the English prompt better than the Arabic one. The overall performance, however, remains significantly lower compared to other datasets, both in terms of F1 and CE metrics. This finding aligns with Moha et al. (2023) who find that LLMs perform better on English datasets.

Table 4 highlights that while the simplistic baseline performance can be matched, it consistently falls short of the performance achieved by a specifically fine-tuned model on both F1 and CE scores (SE best).

A further examination of the errors when using the final majority voted labels reveals a higher tendency for false positive errors (see Table 5). This indicates that GPT is biased towards annotating samples as offensive, abusive, and misogynistic.

| | MD-Agree. | | | HS-Brexit | | | ConvAbuse | | | ArMIS | | | |
|-----------------|--------------|---------------|-----------|-----------|---------------|-----------|--------------|---------------|-----------|---------------|--------------|---------------|-----------|
| | gpt | SE (baseline) | SE (best) | gpt | SE (baseline) | SE (best) | gpt | SE (baseline) | SE (best) | gpt (english) | gpt (arabic) | SE (baseline) | SE (best) |
| $F1 \uparrow$ | 0.520 | 0.534 | 0.846 | 0.696 | 0.842 | 0.929 | 0.902 | 0.741 | 0.942 | 0.448 | 0.256 | 0.417 | 0.832 |
| $CE \downarrow$ | 3.829 | 7.385 | 0.472 | 5.037 | 2.715 | 0.235 | 3.746 | 3.484 | 0.185 | 5.828 | 6.667 | 8.908 | 0.469 |

Table 4: Prompting gpt-3.5-turbo directly vs. baseline & best results from SemEval2023 (SE)

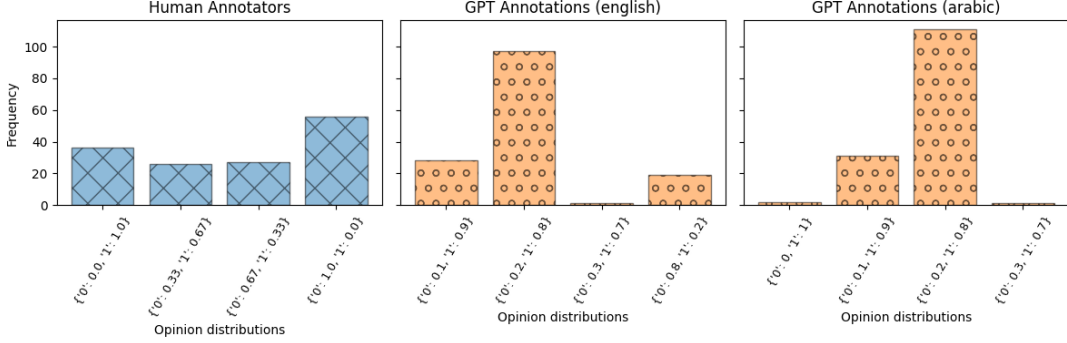


Figure 1: ArMIS opinion distributions

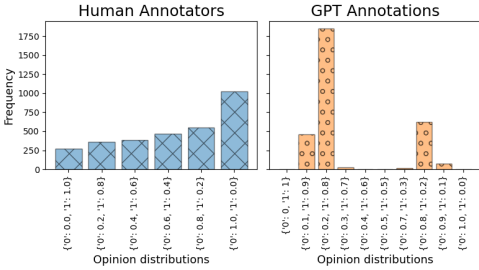


Figure 2: MD-Agreement opinion distributions

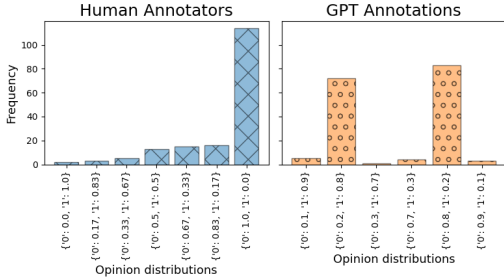


Figure 3: HS-Brexit opinion distributions

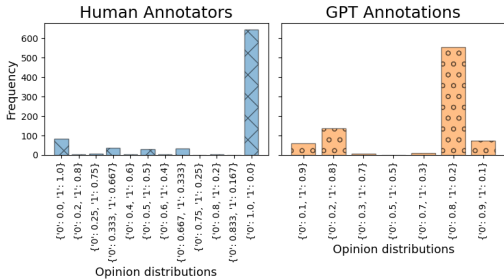


Figure 4: ConvAbuse opinion distributions

| Categorisation of Errors | | |
|--------------------------|---------|-------|
| Dataset | FP | FN |
| MD-Agree | 96.87% | 3.13% |
| HS-Brexit | 100.00% | 0.00% |
| ConvAbuse | 91.11% | 8.89% |
| ArMIS (english) | 95.71% | 4.29% |
| ArMIS (arabic) | 100.00% | 0.00% |

Table 5: Categorisation of errors into percentage that are False Positive vs. False Negative. *GPT 3.5-turbo* across different SemEval2023 tasks

Prompting the LLM to directly return opinion distributions results in higher average entropy values across all four datasets when compared to the average human entropy values (Figure 5). This stems from the observations made in the initial four figures. With the exception of the Arabic prompt, GPT consistently provides opinion distributions that allocate a small proportion to both classes rather than assigning 100 percent to one class. This leads to increased per sample entropy and thereby overall higher average entropy.

4. Conclusion

The overview section is not intended to provide an exhaustive review; however, the variety of tasks, datasets and approaches within the surveyed papers offers first insight into the efficacy of using LLMs to annotate data. Despite the mentioned limitations, the overall findings show a degree of consensus and positive outlook towards the use of

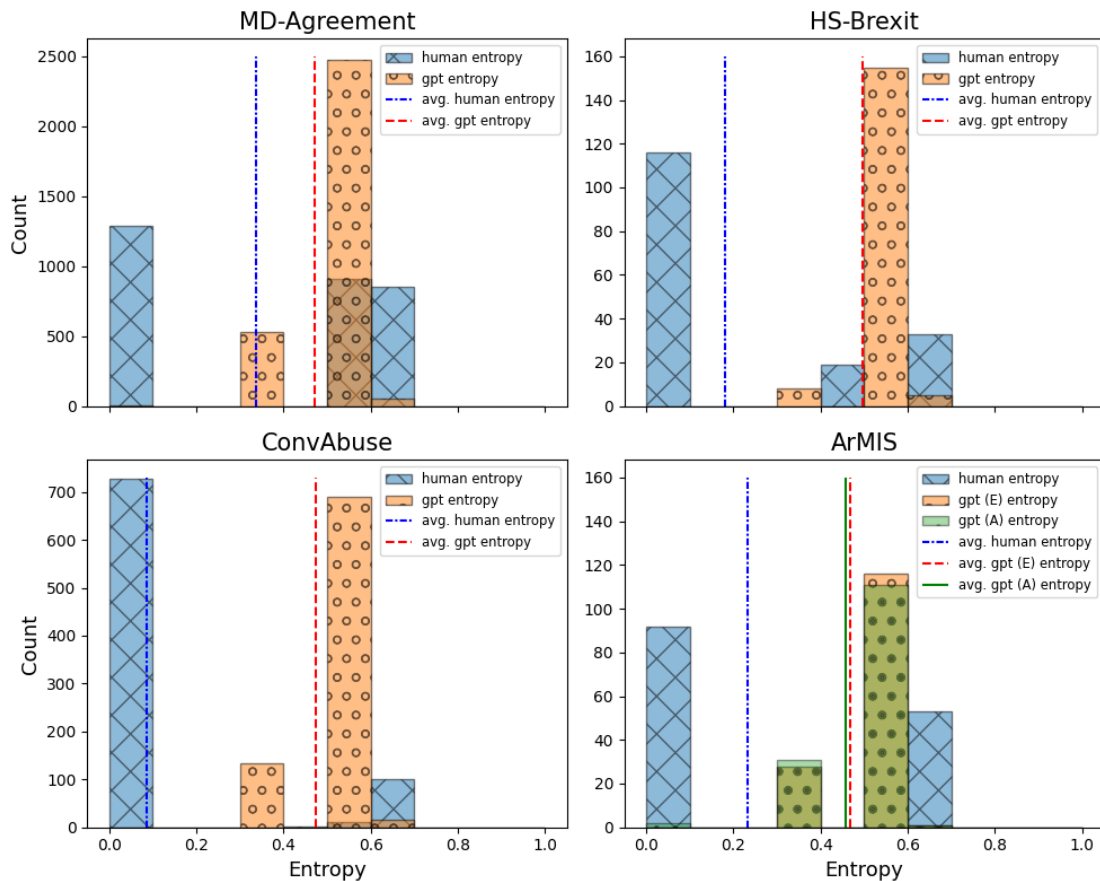


Figure 5: Histogram showing human and GPT entropy

LLMs as data annotators within the majority voting paradigm.

Our initial observations suggest that, when directly prompted, GPT tends to produce label distributions that are not strongly aligned with human opinion distributions. Furthermore, also consistent with prior research, the LLM shows superior performance on English language tasks compared to non-English text, while also showing potential bias in its responses. However, given that LLMs are trained to predict next tokens, directly obtaining opinion distributions from them has inherent limitations. Hence, in future work, we aim to explore further approaches to extracting the probability distributions such as through normalising the log probabilities (Santurkar et al., 2023) or through Monte Carlo estimation (Lee et al., 2023).

Ethical statement

Our study exclusively used pre-existing datasets for experimentation purposes. While the datasets contain instances of offensive language, our approach involved handling this content without direct human involvement.

Acknowledgments

Maja Pavlovic is supported by a Deep Mind PhD studentship to Queen Mary University. The work of Massimo Poesio is supported in part by the AINED Fellowship Grant *Dealing with Meaning Variation in NLP*, NGF.1607.22.002.

5. Bibliographical References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. [Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection.](#)

Dina Almana 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.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel

- Fernandez. 2022. [Stop measuring calibration when humans disagree](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. [Toward a perspectivist turn in ground truthing for predictive computing](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.
- 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.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. [Is GPT-3 a Good Data Annotator?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#).
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [ChatGPT outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences*, 120(30). Publisher: Proceedings of the National Academy of Sciences.
- Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. 2023. [LLMs Accelerate Annotation for Medical Information Extraction](#). In *Proceedings of the 3rd Machine Learning for Health Symposium*, pages 82–100. PMLR. ISSN: 2640-3498.
- Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. [Anollm: Making large language models to be better crowdsourced annotators](#). *arXiv preprint arXiv:2303.16854*.
- Fan Huang, Haewoon Kwak, and Jisun An. 2023. [Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech](#). In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pages 294–297, New York, NY, USA. Association for Computing Machinery.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Noah Lee, Na Min An, and James Thorne. 2023. [Can Large Language Models Capture Dissenting Human Voices?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.
- Elisa Leonardelli, Gavin Abercrombie, Dina Al-Manea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [SemEval-2023 Task 11: Learning with Disagreements \(LeWiDi\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. [Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. [Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.
- Jay Mohta, Kenan Emir Ak, Yan Xu, and Mingwei Shen. 2023. [Are large language models good annotators?](#) In *NeurIPS 2023 Workshop on I Can't Believe It's Not Better (ICBINB): Failure Modes in the Age of Foundation Models*.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. [What can we learn from collective human opinions on natural language inference data?](#) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- 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.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent disagreements in human textual inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- 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.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. [On releasing annotator-level labels and information in datasets](#). In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. [Two contrasting data annotation paradigms for subjective NLP tasks](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 29971–30004. JMLR.org.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. [Do llms exhibit human-like response biases? a case study in survey design](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Petter Törnberg. 2023. [Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning](#).
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Shuohang Wang, Yang Liu, Yichong Xu, Chengguang Zhu, and Michael Zeng. 2021. [Want To Reduce Labeling Cost? GPT-3 Can Help](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. [LLMaAA: Making Large Language Models as Active Annotators](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-UI Haq, Pan Hui, and Gareth Tyson. 2023. [Can chatgpt reproduce human-generated labels? a study of social computing tasks](#).

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, pages 1–55.

Appendix A - Additional tables

| Paper | Datasets |
|--------------------------|--|
| (Lee et al., 2023) | ANLI-R3, QNLI, ChaosNLI, PK2019 |
| (Santurkar et al., 2023) | OpinionQA |
| (Ziems et al., 2024) | Indian English dialect feature detection, Twitter Emotion detection, FLUTE, Latent Hatred, Reddit/Kaggle Humor data, Ideological Books Corpus, Misinfo Reaction Frames Corpus, Random Acts of Pizza, Semeval2016 Stance Dataset, Temporal Word-in-Context benchmark, Coarse Discourse Sequence Corpus, TalkLife dataset, Winning Arguments Corpus, Wikipedia Talk Pages dataset, Conversations Gone Awry Corpus, Stanford Politeness Corpus, Hippocorpus, WikiEvents Article Bias Corpus, CMU Movie corpus dataset |
| (Zhu et al., 2023) | Stance Detection, Hate Speech, Sentiment Analysis, Bot Detection, Russo-Ukrainian Sentiment |
| (Gilardi et al., 2023) | Twitter Content moderation, US Congress, Newspaper article content moderation |
| (Törnberg, 2023) | Twitter Parliamentary Database |
| (Mohta et al., 2023) | MM-IMDB, XNLI, Hateful memes, 2 proprietary datasets |
| (Ding et al., 2023) | SST2, CrossNER, FewRel, ASTEData-V2 |
| (He et al., 2023) | QK (user query & keyword relevance assessment), Word-inContext WiC, BoolQ |
| (Huang et al., 2023) | LatentHatred |
| (Goel et al., 2023) | Mimic-iv-note |
| (Wang et al., 2021) | XSum, Gigaword, SQuAD, SST-2, CB TREC, AGNews, DBPedia, RTE |

Table 6: Datasets used across different studies

| Paper | Model Versions |
|--------------------------|---|
| (Lee et al., 2023) | GPT (text-davinci-002&003); FlanT5 (large,xl,xxl), Flan UL2; Stable Vicuna; OPT-IML-M-S(1.3B)&(30B) |
| (Santurkar et al., 2023) | GPT(ada,davinci, text-ada-001,text-davinci-001&002&003); Jurassic (j1-Grande, j1-Jumbo, j1-Grande-v2 beta) |
| (Ziems et al., 2024) | GPT (text-ada-001, text-babbage-001, text-curie-001, text-davinci-001&002&003, gpt-3.5-turbo, gpt-4); FlanT5 (small, base large, xl, xxl), Flan UL2 |
| (Zhu et al., 2023) | gpt-3.5-turbo |
| (Gilardi et al., 2023) | gpt-3.5-turbo |
| (Törnberg, 2023) | gpt-4 |
| (Mohta et al., 2023) | Instruct-BLIP-Flan-T5; Vicuna(13b v1.3&v1.5), Vicuna (7b v1.5) InstructBLIP (7b & 13b), Llama (7b, 13b), OpenLlama (13b) |
| (Ding et al., 2023) | text-davinci-003 |
| (He et al., 2023) | gpt-3.5 |
| (Huang et al., 2023) | ChatGPT |
| (Goel et al., 2023) | PaLM 2 |
| (Wang et al., 2021) | text-davinci |

Table 7: Model versions used across studies

Appendix B - Prompts

Multi-Domain Agreement: *You are an expert annotator in "offensiveness detection", who considers multiple worldviews and cultural nuances. Now assign probabilities for the text below, which is delimited by three dashes (-), for "offensive speech" or "not offensive speech". Return your probabilities in a json format with the key being 1 for 'offensive' and 0 for 'not offensive'. You obviously know that probabilities need to sum to 1.*

Hate Speech on Brexit: *You are an expert annotator in "hate speech detection", who considers multiple worldviews and cultural nuances. Now assign probabilities for the text below, which is delimited by three dashes (-), for "hateful speech" or "no hateful speech". Return your probabilities in a json format with the key being 1 for 'hate speech' and 0 for 'no hate speech'. You obviously know that probabilities need to sum to 1.*

ConvAbuse: *You are an expert annotator in "abusive speech detection", who considers multiple worldviews and cultural nuances. Now assign probabilities for the text below, which is delimited by three dashes (-), for "abusive speech" or "no abusive speech". Return your probabilities in a json format with the key being 1 for 'abusive speech' and 0 for 'no abusive speech'. You obviously know that probabilities need to sum to 1.*

Arabic Misogyny and Sexism: *You are an expert annotator in "sexism detection", who considers multiple worldviews and cultural nuances. Now assign probabilities for the text below, which is delimited by three dashes (-), for "sexist speech" or "no sexist speech". Return your probabilities in a json format with the key being 1 for 'sexist speech' and 0 for 'no sexist speech'. You obviously know that probabilities need to sum to 1.*

Arabic prompt:

"أنت محلل خبير في "كشف التحيز الجنسي"، وتأخذ في اعتبارك العديد من وجهات النظر العالمية والفروقات الثقافية".
 "(-) الآن قم بتحديد احتمالات للنص أدناه، الذي يتم تحديده بثلاثة شروطات
 "الخطاب التحيز الجنسي" أو "خطاب خالٍ من التحيز الجنسي".
 "مع المفتاح 1 لـ'خطاب التحيز جنسي' و0 لـ'خطاب خالٍ من التحيز الجنسي' JSON، قم بإعادة احتمالاتك بتنسيق
 "من الواضح أنك تعلم أن مجموع الاحتمالات يجب أن يكون 1"

What Does Perspectivism Mean? An Ethical and Methodological Countercriticism

Mathieu Valette

ERTIM, Institut National des Langues et Civilisations Orientales (Inalco, Paris)
2 rue de Lille, 75007 Paris, France
mvalette@inalco.fr

Abstract

In this paper, we address the epistemological and ethical break of perspectivism in NLP. First, we propose to consider data annotation from the point of view of the scientific management of annotation work - which is part of the automation process inherent in NLP, in order to ideologically situate the perspectivist paradigm. We then analyze some of the concepts of perspectivism (in particular, truth). Finally, based on this analysis, we formulate a set of proposals aimed at overcoming the observed limitations of corpus annotation in general and perspectivism in particular.

Keywords: perspectivism, postmodernism, ethics, annotated corpus, corpus linguistics, statistical analysis of textual data

1. Introduction

The dynamism of perspectivist work (Basile et al., 2021; Cabitza et al., 2023) attests to the fact that the construction of datasets in NLP, as in AI, is a research challenge that is more topical than ever. The inclusion of variation and *situated* interpretation in a field of research still strongly marked by a referentialist approach is good news. Of course, the days are long gone when all semantic phenomena were represented by means of ontologies, often general, stable and unchanging, mostly constructed in a top-down or onomasiological manner (Fellbaum, 1998). In this old paradigm, abundance and variety were a *problem* (polysemy problem for linguists, disambiguation problem for NLP scientists), whereas they are inherent in semiotic activity, i.e. the production of signs. Probabilistic approaches have won out over the dominant ontological paradigm.

The NLP community is increasingly asking itself about the political and sociological biases present in the datasets it processes (Feng et al., 2023), especially as these datasets are, as they grow, increasingly opaque. By connecting the issue with dataset human annotation, the perspectivist paradigm links a general scientific question (the process of qualifying an object) to an ethical question (the under-representativeness of minority sensibilities, in particular), paving the way for a more general discussion on the ethics of NLP and AI.

We aim to address the epistemological and ethical break of perspectivism in NLP. First, we propose to consider data annotation from the theoretical point of view of *scientific management* (Taylor, 1911) of annotation work which participates in the automation process inherent in NLP, in order to ideologically situate the perspectivist paradigm. We then analyze some of the concepts of perspectivism (in particular, truth). Finally, based on this analysis, we make a set of proposals aimed at overcoming the observed limits of perspectivism.

2. Annotation as Scientific Management of Work

The aim of NLP is automation, i.e. the elimination of the human component in the processing of textual data. NLP consists in setting up processes to obtain an output result from a set of input data. Among the proposed system improvements, reducing human intervention appears to be almost as important as improving raw performance, computed on the basis of well-known metrics (f-score, accuracy, etc.). Tasks performed by humans are traditionally described as "manual". However, "manual" has two antonyms: automatic, of course, but also intellectual, in which case it's not out of the question for "manual" to have a depreciatory connotation to describe a non-intellectual or low-intellectual task.

Manual work is indeed the stumbling block of NLP. A manual task par excellence is, paradoxically, reading and interpreting datasets. The NLP scientist must neither read nor interpret their dataset, firstly for practical reasons (the dataset is too large) but also for methodological reasons: they must keep their distance from it, and this distancing constitutes a strategy of objectification. In most cases, ethical standards for corpus annotation require that annotators do not participate in algorithm design. Knowledge of the corpus would induce a bias in algorithmic choices and bias the results. Interpretation tasks are therefore outsourced for methodological reasons.

Manual annotation (or partially automated annotation for the most robust tasks, such as POS tagging and named entity recognition) is the only form of text interpretation available in NLP (as we'll see in paragraph 4, it's not the only one possible). NLP scientists entrust annotation tasks to various third parties. Depending on the resources deployed, the stakes involved and the type of task, these annotation (and therefore reading) tasks may be delegated to experts (linguists, doctors, lawyers), or to less qualified individuals (interns, students) or to subcontractors (such as the most famous of them all:

Amazon Mechanical Turk) who provide little or no guarantee of the annotators' expertise - in short, they deskill the annotation work (Cohen et al., 2016).

Annotation campaigns are based on NLP's industrial concepts of automation, optimization and rationalization. Thus, the annotation task must be understood in terms of the scientific management of work also known as Taylorism, and thus the division of labor. Without going into the details of this division, we can distinguish three roles: (i) the project manager selects the corpus, defines the sequence of operations, controls the process, is in charge of the annotation guide, and the methodological and logistical choices, (ii) the curator supervises the annotation tasks and is the interface between the project manager and the annotators. He trains them in the annotation guidelines, checks the quality of annotations and acts as an expert; (iii) the annotators, who have no expert status, comply with the annotation guidelines. Finally, annotators work on data samples and do not have an overall view of the dataset or even of the processing line, which they may even be completely unaware of (Bontcheva, 2010).

In another industrial context, the annotator would correspond to the unskilled laborer working on a production-line. Until perspectivist proposals overturned the paradigm, the annotator was suspected of being, structurally, the weak link in the processing chain (Bontcheva, 2010): he or she may make careless mistakes or fail to understand guidelines. The whole evaluation system (inter-annotator agreement, Cohen's Kappa...) of the annotation task is based on this structural weakness, and determines annotator recruitment processes. Admitting that the annotator has an intrinsically low confidence rating deskills their work and forces the project manager to compensate by multiplying the number of annotators, who are recruited at low cost (which further lowers the level of requirement) or without pay, following practices such as incidental crowdsourcing (Park et al., 2019) or gamification.

A Marxian reading of these annotation campaigns is necessary: it shows a process entirely controlled by computer scientists who own the technological production apparatus (the ability to build up large-scale datasets, machine learning algorithms) and who buy the labor power of annotators-proletarians. This vision in terms of class rule may seem inappropriate when we think of IT, but on the one hand, we can only observe the incredible rise in power of the industrial players in the digital sector, proportional to their technological domination and the human costs it generates. On the other hand, the production of resources, i.e. the creation of value through annotation, is necessarily linked to the production means.

¹ "Aggregation and harmonization destroy any personal opinion, nuance, and rich linguistic knowledge that come as a result of the different cultural and demographic background of the annotators" (<https://pdai.info/>).

3. The Ethics of Perspectivism in Debate

We will not discuss here the scientific value-adding virtues of perspectivism, which consists in considering noise as information and therefore error as a positive value (Basile 2021; Cabitza et al., 2019; Cabitza et al., 2023; Sachdeva et al., 2022; Kralj Novak et al., 2022). We wish to raise two issues, one ethical, the other methodological, which can be linked in terms of solutions.

3.1 Perspectivism Has no Impact on Work Management and Labor Conditions

We could be content to see only the technical and scientific contribution of perspectivism, i.e. the enrichment of data, but the Perspectivist Manifesto explicitly adopts an ethical stance¹, both in its general argument and in the datasets available (Measuring Hate Speech, Pejorative Language in Social Media, Work and Job-Related Well-Being), so it's legitimate to discuss the ethics of perspectivism itself, since technical means are never neutral and, as we've seen, organize work.

The Manifesto forcefully denounces aggregation and harmonization as forms of obliteration of annotators' personal opinions and disregard for cultural background. The authors rightly observe that harmonization, in particular, can take place at the end of deliberative phases, where relationships of domination can be established to the prejudice of minority opinions (Noble, 2012). What we note, however, is that in no case does the Manifesto denounce the alienation of the task. *Crowdsourcing is still production-line work*². Not only do these working conditions potentially expose annotators to abnormally repeated harmful content (Steiger et al., 2021), but also, we are not sure that many annotators declare that corpus annotation is fulfilling work and leads to well-being at work (contrary to what academic computer scientists might say about their high-qualified work). If we exclude the incidental crowdsourcing or gamification techniques already mentioned, annotation is a tedious, repetitive task, socially unrewarded, solitary by method and, finally, poorly paid (Fort et al., 2011; Gray and Suri, 2019). This last point is soberly mentioned - but not discussed - in the Manifesto in a footnote.

3.2 Perspectivism Mistakes Sincerity for Truth

It's tempting to draw a parallel between the perspectivist paradigm in AI and its philosophical homonym (Leibniz, Nietzsche, but especially Deleuze) (Astor, 2020). The subjectivist relativism of the perspectivist paradigm is akin to the postmodern assumption that there is not just one truth, but many truths, and that all truths are equal. Many authors see this as a worrying drift, particularly in science and politics, as it leads to *post-truth politics* (Holzem, ed. 2019) and to pseudoscientific or negationist positions,

² In fact, the term "annotator" is sometimes replaced by "worker" (Aroyo and Welty, 2015).

which in turn provide the breeding ground for totalitarianism (Rastier, 2019).

Of course, the perspectivist paradigm does not claim to offer alternative scientific truths, but several interpretations that reflect the opinions of the annotators, and it is not uncommon for the concept of "truth", sometimes renamed "ground truth" to be relativized, whether in the Manifesto or in *position papers* (Basile et al. 2021; Planck, 2022; Cabitza et al, 2023). For instance, Aroyo and Welthy (2015), pervert the traditional concept of truth with a leitmotiv that systematically contradicts the rational thinking that underpins modern science ("truth is a lie", the "antiquated ideal of truth"...). They even propose a "new theory of truth": "Crowd truth is the embodiment of a new theory of truth that rejects the fallacy of a single truth for semantic interpretation, based on the intuition that human interpretation is subjective and that measuring annotations on the same objects of interpretation (in our examples, sentences) across a crowd will provide a useful representation of their subjectivity and the range of reasonable interpretations" (Aroyo and Welthy 2015: 21). The label "theory of truth" is a misuse of language, firstly because their purpose is limited to corpus annotation and not to a scientific definition of truth (only the 1st of the 7 myths they identify calls into doubt the concept of truth, the other 6 myths are purely methodological criticisms), and secondly, because the authors confuse "truth" with opinion, or some concept we could design as "sincerity".

The confusion maintained between sincerity and truth in almost all the scientific production of the perspectivist paradigm is not just a problem of terminological inaccuracy. It is an epistemological confusion with methodological consequences. What is valued in the perspectivist paradigm is the sincerity of annotations. Truthfulness can be evoked, for example, in the annotation of linguistic norms, or in the perspective of establishing linguistic norms (e.g. POS tagging of poorly endowed languages leads to the identification of variations that are not variations of personal sensitivities but of norm perception). Sincerity is a psychological concept and is not a matter for the annotated texts, but for the annotators themselves, their subjectivity and the resulting interpretation.

Interpretations are not just a matter of the annotator and the sample to be annotated going head to head. Numerous interpretative biases could be cited and, in order to inventory and model them, we could draw on millennia of bibliography, from Aristotle's Rhetoric (what is the ethos of the speaker?) to R. Jakobson's functions of language. We could also ask about the material conditions of the device (how was the corpus constituted? for what purpose? how was the task described? What is the intertext of the sample to be annotated?) as well as the *Dasein* of the annotator (What is his or her psychological state at the moment of annotation? is he or she happy? unhappy? worried about the future? How long has he been working? is he tired? hungry? etc. etc.).

All these questions seem trivial, but from the moment we address the sincerity of the annotator – a fortiori if we aim to make it a truth – it seems necessary to ask the question of their psychological condition. From the point of view of a psychologist, this condition could be as relevant as the sampling recommended by (Cabitza et al. 2023: 6885) "both in regard to their origin and culture as well as to their expertise and skills" – a necessary condition that could involve sociologists capable of correctly sampling the team of annotators. The more extra-linguistic criteria we include in the constitution of the dataset, the more we have to mobilize the corresponding sciences.

To sum up: noting the methodological and ethical limits to immanent corpus annotation, the perspectivist paradigm proposes to abandon corpus annotation in favor of annotation of annotators (by sampling) and their sincere perceptions of the dataset. But the perspectivist literature only partially solves the ethical problems it raises, it only notes some of these problems and, rather than fighting discrimination, it makes it visible, measurable and computable by tagging variations. The real ethical problem with NLP is the management of work.

4. Constructing Rather Than Annotating

Let's change perspective.

What if the problem wasn't the inclusion of annotators, but the very principle of corpus annotation? Let's try to justify our change of perspective by 3 main proposals:

A. Give a philological value to the dataset, i.e. turn it into a corpus

With the widespread use of machine learning, the concept of the corpus as a built set of texts designed for a specific task has been greatly devalued in NLP, in favor of the dataset, i.e. a collection of data often "scraped" from the Internet with little preliminary characterization (mainly some sources and keywords), and whose main characterization comes from the annotation itself. Moreover, with deep learning, datasets are now quantified in terms of gigabytes rather than linguistic units (words, sentences, texts). However, the construction of a corpus is the first scientific act involved in the definition and establishment of the object of science. Consequently, building a corpus is a high-level activity (so much so that it can take years for a linguistics PhD student!). Our first proposal for a change of perspective is this: we need to give value to the construction of the corpus (Dusserre and Padró, 2017). Constructing a corpus requires setting up a task, identifying and examining sources, selecting texts, verifying them, characterizing them: author, type of discourse, textual genre, etc. (Biber and Conrad, 2009). Generally speaking, the reuse of datasets for new tasks, commonly accepted as good scientific practice, is not always acceptable when we're talking about corpora (a problem that also arises for gold standard corpora). Indeed, if we think in terms of corpus rather than data, sharing is only acceptable

if the new task aims to pursue or verify the objectives of the previous task. If the aim is to recycle a dataset that is more or less suitable for a new task, the corpus is downgraded to a dataset, because the intention is inherent in the corpus.

Just as the concept of corpus is underused in NLP, that of text is also poorly understood. A text is not just a collection of sentences or a bag of words, it's a semiotic object produced in a particular social and cultural context, corresponding to an enunciative project and containing interpretative rules, which are conditioned by its intertextuality (Mayaffe, 2002) (i.e. the set of texts linked by a text, for instance, in the case of an interlocution, usual on the social web). Not taking intertextuality into account when collecting texts is to deprive future annotators of their interpretative clues, because the corpus is not a resource but a multi-scale contextualization of observable phenomena (Mayaffe and Viprey, 2008). Text is the first interpretable semantic unit if we have to select a first level of annotation.

B. Focus on coarse-grained annotations or on intrinsically annotated corpus

The more fine-grained the annotation, the greater the number of annotations, the greater the risk of variation. If variation is a quality in the perspectivist paradigm, this is without considering the ethical biases we discussed earlier: variation can indeed be an effect of arduous working conditions. A coarse-grained annotation is one that requires a longer, more reflective - less reflex - less manual, more objective intellectual work of interpretation. For example, when it concerns hate speech, it can be interesting to collect all the tweets of an author identified as a habitual hater, rather than a sample of his or her explicitly hateful tweets, which allows annotators to safely distance themselves from hateful content. Moreover, hateful sentiments are not necessarily expressed in hateful words (Eensoo et al., 2015). Fine-grained annotation is unfortunately confused in NLP with word-level annotation, but the word is only a minimal semiotic unit carrying lexical meaning, not text sense.

An ultimate coarse-grained annotation would be to use corpora that are intrinsically annotated, i.e. corpora that, by their very constitution, already contain metadata that can be used as annotations (e.g. gender declared, age, opinion, city, etc.), as is the case, for example, with corpora of polarized comments with ratings. One of the most widely cited NLP articles in sentiment analysis, (Pang and Lee, 2002), uses this type of corpus. The use of intrinsically annotated corpora makes it possible to concentrate efforts on higher-level (and therefore more skilled, better-paid) annotation tasks.

C. Use computer-aided corpus analysis

"One could determine the different ages of a science by the technique of its measuring tools " said the philosopher Bachelard (1938: 216, our translation). It seems astonishing that corpus annotation is still today an irreducibly manual task when numerous "distant

reading" tools exist and have been used for over 40 years now, first in the Statistical Analysis of Textual Data and then Digital Humanities community (e.g. Compagno, eds., 2018; Iezzi, eds., 2018; Lebart et al. 2019). This is indicative of a stubbornly marked divide between the NLP scientists and the humanities. Yet, humanities provide several bottom-up analytic methods, and tools and heuristics for corpus description. Textometrics tools (Heiden, 2010), combined with ad hoc text semantic theories (Pincemin, 2010; Rastier, 2018), can be used to generate annotations that can then feed philologically built datasets, for instance for opinion mining (Eensoo et al., 2015; Valette, 2018; Baiocchi, 2019). Finally, work combining semiotic theory and word embeddings aims to formally model the 'reader' (Sanna and Compagno, 2020) – a relevant idea in the context of perspectivism.

By combining statistical measurement, corpus analysis and semantic theorization, computer-aided corpus analysis creates an (objectifying) distance between annotator and *empirie*. What's more, this distance protects annotators from prejudicial content to which they might be brutally subjected, without the necessary reflexivity to withstand it. What's more, the proposed corpus analysis reclassifies the creation of value as intellectual work. This twofold distancing, by theory and by tool, leads to an objectification of the phenomena studied. Objectivity could wrongly be seen as contrary to the perspectivist paradigm. On the contrary, it is a construct that emanates from experiences, not subjectivities: As the philosopher Bitbol (2014) says: "*Objectification means focusing attention on what, in experience, can be shared. This presupposes an education, which begins with the transmission of a common language and an ethic of truth*" (our translation, our emphasis).

5. Conclusion

The perspectivist paradigm renews work on the production of learning data for machine learning. It is based on both the methodological obstacles observed in the community and an ethical position that appears to be inspired by the DEI (Diversity, Equity, and Inclusion) ethical management framework. In brief, the aim of this short position paper was to sketch a discussion about the ethical and methodological proposals of perspectivism which is mainly based on a managerial background, by using both a Marxian-based ethical and a methodological criticism inspired by Humanities proposals.

Our conclusion is that an ethical approach to manual annotation of corpora, today organized according to the principles of production-line work, would be to reclassify it as scientific corpus analysis, in order to revalue this necessary interpretative work.

6. References

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

- Astor, D. (2020). *Perspectivisme. Leibniz, Nietzsche, Whitehead, Deleuze*. Phd Thesis, Institut Polytechnique de Paris: 471 p.
- Bachelard, G. (1938). *La formation de l'esprit scientifique*, Paris: Vrin.
- Baiocchi, M. (2019). *Diversité et recommandation : une investigation sur l'apport de la fouille d'opinions pour la distinction d'articles d'opinion dans une controverse médiatique*, PhD thesis, Université de Montréal: 509 p.
- Biber, D. Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.
- Bitbol, M. (2014). L'expérience d'objectiver Ou comment vivre en première personne la possibilité de la troisième. In N. Depraz (eds), *Première, deuxième, et troisième personne*, Presses Universitaires de Rouen.
- Cohen, K.B., Fort, K., Adda, G., Zhou, S., Farri, D. (2016) Ethical issues in corpus linguistics and annotation: pay per hit does not affect effective hourly rate for linguistic resource development on Amazon Mechanical Turk. In *LREC Int Conf Lang Resour Eval*. W40: 8-12.
- Compagno, D. eds. (2018). *Quantitative Semiotic Analysis*, Springer International Publishing: Berlin. 189 p.
- Dusserre, E., Padró, M. (2017). Bigger does not mean better! We prefer specificity. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.
- Fellbaum, Chr. eds. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Eensoo, E., Valette, M. (2015). Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité. *TALN'2015*. Caen (France).
- Fort, K. Adda, G., Bretonnel Cohen, K. (2011). Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, 37 (2): 413–420.
- Gray, M. L., Suri, S., (2019). *Ghost Work. How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt, 288 p.
- Heiden Serge. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*. Sendai, Japan.
- Holzem, M. (ed) (2019). *Les sciences contre la post-vérité : Vérités citoyennes*. Vulaines-sur-Seine: Éditions du Croquant. 174 p
- Iezzi, D. F. Celardo, L., Misuraca, M., eds. (2018). *Proceedings of the 14th International Conference on Statistical Analyses of textual Data (JADT'18)*, UniversItalia, 2 vols.
- Lebart, L., Pincemin, B., Poudat, C. (2019). *Analyse des données textuelles*. Presses de l'Université du Québec. Mesure et évaluation. 510 p.
- Mayaffre, D. (2002). Les corpus réflexifs : entre architextualité et hypertextualité. In *Corpus*, 1: 51-69.
- Mayaffre, D., Viprey, J.-M., eds. (2012). *La cooccurrence, du fait statistique au fait textuel, Corpus*, 11.
- Noble, J. A. (2012). Minority voices of crowdsourcing: Why we should pay attention to every member of the crowd. In *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*: 179-182.
- Planck, B. (2022) The “problem” fo human label variation: on ground truth in data, modeling and evaluation. Planck, 2022, In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*: 10671-10682.
- Park, J., Krishna, R., Khadpe, P., Fei-Fei, L., & Bernstein, M. (2019). AI-based request augmentation to increase crowdsourcing participation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7(1): 115-124).
- Pincemin B. (2010). Semántica interpretativa y textometría. In Duteil-Mougel C. and Cárdenas V. eds., *Semántica e interpretación, Tópicos del Seminario*, 23, Enero-junio 2010: 15-55.
- Rastier, F. (2018). Computer-Assisted Interpretation of Semiotic Corpora. In Compagno, D. eds., *Quantitative Semiotic Analysis*. Springer International Publishing: Berlin: 123-139.
- Rastier, F. (2019). Autour de la « post-vérité », de menaçantes convergences, in Holzem, M., eds. (2019). *Les sciences contre la post-vérité : Vérités citoyennes*. Vulaines-sur-Seine: Éditions du Croquant: 23-57.
- Sanna, L., Compagno, D. (2020). Implementing Eco's Model Reader with Word Embeddings. An Experiment on Facebook Ideological Bots. *Proceedings of the 15th International Conference on Statistical Analyses of textual Data (JADT'20)*, Toulouse, France.
- Steiger, M., Bharucha, T., Venkatagiri, S., Riedl, M.J., Lease, M. (2021). The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing System (CHI '21)*: 1-14.
- Taylor, F. W. (1919). *The Principles of Scientific Management*. Harper & Brothers – via Internet Archive (Prelinger Library)
- Valette, M. (2018). Elements of a Corpus Semantics for Humanities. Application to the Classification of Subjective Texts, In D. Compagno, eds., *Quantitative Semiotic Analysis*, Springer International Publishing: Berlin: 141-150.

OrigamIM: A Dataset of Ambiguous Sentence Interpretations for Social Grounding and Implicit Language Understanding

Liesbeth Allein, Marie-Francine Moens

Department of Computer Science, KU Leuven

Leuven, Belgium

{liesbeth.allein, sien.moens}@kuleuven.be

Abstract

Sentences elicit different interpretations and reactions among readers, especially when there is ambiguity in their implicit layers. We present a first-of-its kind dataset of sentences from Reddit, where each sentence is annotated with multiple interpretations of its meanings, understandings of implicit moral judgments about mentioned people, and reader impressions of its author. Scrutiny of the dataset proves the evoked variability and polarity in reactions. It further shows that readers strongly disagree on both the presence of implied judgments and the social acceptability of the behaviors they evaluate. In all, the dataset offers a valuable resource for socially grounding language and modeling the intricacies of implicit language understanding from multiple reader perspectives.

Keywords: implicit language, interpretation, ambiguity, social grounding, moral reasoning, resource

1. Introduction

A sentence frequently evokes diverse and disagreeing interpretations. Disagreement in interpretation can arise from explicit cues, such as the choice and order of words, triggering phonological, lexical, and structural ambiguities (Kennedy, 2019). This disagreement is further amplified by a diversity among readers, each guided by their unique experiences, knowledge, and viewpoints. Despite extensive exploration of ambiguity within computational linguistics (Bevilacqua et al., 2021; Haber and Poesio, 2023), little attention has been devoted to *ambiguity in the implicit layers of sentences* and the resulting *disagreement in interpretation*.

This underexposure of the implicit is surprising considering a substantial portion of human communication is inherently non-verbal. Even when using language, we convey information between the lines. Implicit communication is efficient since it obviates the need to reiterate common sense or common ground information (Stalnaker, 2002), and it is social as it can prevent a loss of face when sharing social evaluations (Dunbar, 2004). Some people also reside to the implicit layers of communication when targeting a specific audience and deceiving all others (e.g., dogwhistles (Henderson and McCready, 2017)). Achieving such human-like communication skills in computational models therefore necessitates a *transition to multi-perspective language production and understanding*, in which models are equipped with the ability to reason over implicit content from multiple angles.

To facilitate the development of such models, we curate a *first-of-its-kind dataset* of sentences, where each sentence is annotated with multiple interpretations, detailed descriptions of underlying

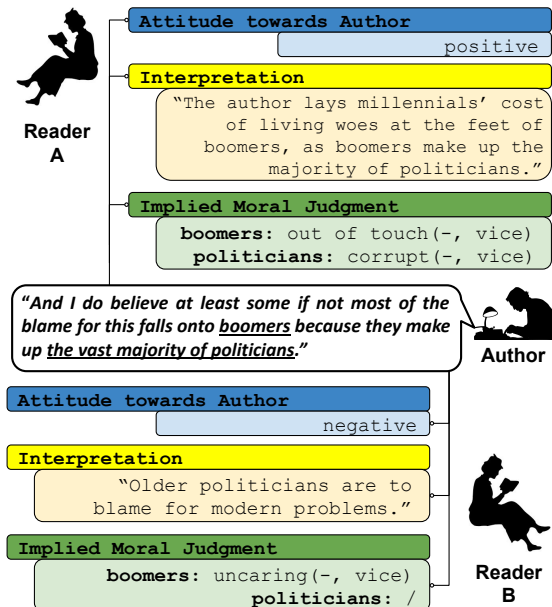


Figure 1: Sample taken from the origamIM dataset, demonstrating the diverging reader attitudes towards the author, slightly different interpretations, and disagreeing understandings of implicit moral judgments a sentence can trigger.

moral judgments of people mentioned in the sentence, and measures of reader attitude describing a reader's first impression of the author upon reading the sentence¹ (Figure 1). The latter two information types socially ground the sentences from multiple perspectives. The name of the dataset, origamIM, refers to the analogy between the Japanese art of paper folding and the diversity of

¹The dataset is publicly available: <https://github.com/laallein/origamIM>.

| Context | Appropriateness | | |
|---|--------------------|--------------------|----------------|
| | Sphere of Action | Vice of Deficiency | Virtue of Mean |
| <i>Confidence, fear, uncertainty</i> | Cowardice | Courage | Rashness |
| <i>Pleasures of the body</i> | Insensibility | Temperance | Profligacy |
| <i>Giving & taking: Small money</i> | Stinginess | Liberality | Prodigality |
| <i>Giving & taking: Added value</i> | Meanness | Magnificence | Vulgarity |
| <i>Pride, honor as cause</i> | Little-mindedness | High-mindedness | Vanity |
| <i>Ambition, honor as goal</i> | Lack of ambition | Proper ambition | Over-ambition |
| <i>Anger</i> | Spiritlessness | Gentleness | Wrathfulness |
| <i>Pleasure and pain of others</i> | Cross, contentious | Agreeableness | Flattery |
| <i>Truth, honesty about oneself</i> | Irony | Truthfulness | Boastfulness |
| <i>Amusing conversation</i> | Boorishness | Wittiness | Buffoonery |

Table 1: Overview of spheres of actions and the degrees of appropriateness (Hursthouse, 1999).

interpretations and attitudes that could be obtained when presented with the same sentence.

2. A Moral Framework for Grounding

Moral judgments offer an interesting case for examining and modeling disagreement. Individuals namely look through their own lenses when judging people and interpreting judgments made by others, despite a shared understanding of moral norms and values. The judgments annotated in the dataset are grounded in Virtue Ethics (Hursthouse, 1999). The moral theory introduced by Aristotle poses that a person’s moral character can be evaluated by the contextual appropriateness of their voluntary behavior within a sphere of action (see Table 1). A virtuous behavior is characterized by moderation and appropriateness within its context (e.g., considering the people involved and the severity of the situation) while contextually deficient or excessive behaviors are not celebrated in society.

The axis of appropriateness in Virtue Ethics provides a distinct advantage over other popular moral frameworks (e.g., Moral Foundation Theory (Haidt and Joseph, 2004)) as it enables individuals not only to differentiate between negative behavior based on its context, but also to annotate their understanding of the implied moral judgments given their cultural and social backgrounds.

3. Dataset Creation

3.1. Data Collection

We automatically retrieve blog posts in English from the Subreddit /r/ChangeMyView that were posted between 13 July 2020 and 3 March 2022. These

posts typically present views on often controversial and polarizing topics, such as abortion and racism. We anticipate that a considerably large portion of the posts pass judgments about people given the human tendency to gossip (Dunbar, 2004; Baumeister et al., 2004; Feinberg et al., 2012). Moreover, negative judgments are expected to be conveyed implicitly due to the subreddit’s moderation policies². We remove duplicated and deleted blog posts and extract the title, body text, and additional metadata³ for each post. Lastly, the body text is segmented into sentences using SpaCy.

3.2. Data Annotation

We recruit crowd workers on Amazon Mechanical Turk⁴ and let them annotate the sentences in two rounds. An annotator never annotates the same sentence in both rounds. The first round distinguishes sentences that mention people and imply a character trait of at least one of them from those that either lack explicit mentions people or do not imply any character trait. A character trait presents a voluntary aspect of a person’s attitude or behavior, e.g., *lazy* and *charitable*. The second round takes the first set of sentences and gathers multiple reader attitudes, interpretations, and entity-level moral judgments for each sentence.

3.2.1. First Round: People Entities

Two annotators mark all entities referring to people other than the author (i.e., ‘I’) in a sentence and indicate whether or not the author seems to imply a character trait of at least one highlighted entity. We show the title of the blog post from which the sentence was taken as additional context. In cases where they disagree on the presence or absence of implied traits, a third annotator is consulted and a majority vote is taken. Data quality and consistency is manually checked. A total of 6,820 sentences were annotated, of which 2,018 implied a character trait of at least one people entity. These figures confirm our expectations regarding the presence of implicit social evaluations in these posts (see §3.1).

3.2.2. Second Round: Attitudes, Interpretations, and Moral Judgments

Five annotators read the same sentence and first describe their attitude towards its author using a

²The moderation rules dictate that posts suggesting harm to others and hostile comments will be removed. See <https://www.reddit.com/r/changemyview/wiki/modstandards/> [accessed on 4 April 2024].

³The metadata is not used during the annotation process.

⁴<https://www.mturk.com/>

| Dataset Statistics | |
|--------------------------------------|------------------------|
| # Blog posts | 396 |
| # Sentences | 2,018 |
| – Total word count | 44,902 |
| – Min/max words per sentence | 2 / 107 |
| # People entities | 3,313 |
| – # Sentences with 1/2/3/4+ entities | 1,103 / 661 / 174 / 80 |
| # Interpretations | 9,851 |
| – Total word count | 155,368 |
| – Min/max words per interpretation | 1 / 113 |
| Distribution reader attitudes | |
| – Very negative | 813 (8.25%) |
| – Negative | 1,971 (20%) |
| – Neutral | 4,302 (43.67%) |
| – Positive | 2,025 (20.56%) |
| – Very positive | 740 (7.51%) |

Table 2: Statistics of the origamIM dataset.

five-point Likert scale ranging from *very negative* (1) to *very positive* (5). They then write down their interpretation of the sentence. We explicitly instruct them to not copy the sentence and manually check the relatedness between sentence-interpretation pairs, removing annotations that present unrelated pairs or poorly-formulated interpretations. Going over all the people entities marked in the first annotation round, the annotators indicate for each entity whether or not the author implies a character trait. In case a trait is implied, they describe it using, preferably, an adjective, mark whether it considered a good or bad trait in society, and classify it in Virtue Ethics (see §2). A complete annotation for a single sentence interpretation looks as follows:

Title CMV: It Should Be Mandatory for Every Person to Work AT LEAST 1 Month in a Customer - Facing Hospitality Role Before Leaving School.

Sent I truly believe it would have been life changing for [him] to work in hospitality for a bit before leaving school, to see and experience what [some people] have to go through on a daily basis just to eat and have a roof.

Att Positive (4)

Int “Real life experience is better than theory.”

Judg [him] ✓ “ignorant”, Bad, Pride/honor as cause, Vice of Deficiency.

[some people] ✓ “hardworking”, Good, Ambition/honor as goal, Virtue of Mean.

4. Data Analysis

Table 2 presents general statistics of the origamIM dataset.

4.1. Disagreement in Attitudes

Each annotator described their attitude towards the author using a five-point Likert scale. Each sentence therefore potentially evokes up to five distinct attitudes among its readers. Figure 2 illustrates the *diversity* of attitudes elicited by a sentence, revealing that the vast majority of sentences trigger

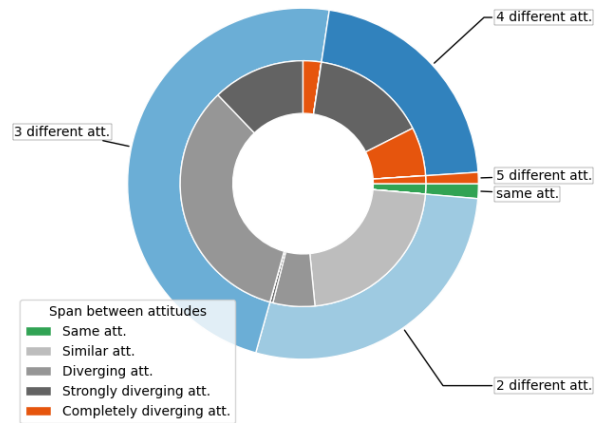


Figure 2: Donut chart representing the disagreement in reader attitude. The outer donut shows the distribution of attitude *diversity*. The inner donut shows the distribution of attitude *divergence*.

at least two different attitudes among readers. As many as one in five sentences even evoke four or five distinct attitudes. This underscores the variability in reactions among readers when presented with the same sentence.

We also examine the *divergence* among those attitudes by measuring the span between the lowest and highest attitude, as indicated on the Likert scale, among the five annotators for each sentence. Figure 2 shows that more than one in four sentences evoke strongly (e.g., *very negative* - *positive*) or completely diverging attitudes (i.e., *very negative* - *very positive*). Similar attitudes are elicited for fewer than 20% of the sentences. These findings show that sentences frequently spark not only different, but also diverging attitudes.

4.2. Disagreement in the Implicit

4.2.1. Moral Judgments

We observe that the diversification in interpretation already starts when discerning the presence of implicit moral judgments as annotators exhibited high disagreement on this issue. Merely 291 sentences (14.42%) garnered unanimous agreement among all five annotators on this matter. This disagreement may arise from varying degrees of subtlety in the social evaluations, requiring more in-depth reasoning to uncover them.

The annotators also disagreed on the societal desirability of the implied character traits (i.e., whether the traits are considered ‘good’ or ‘bad’), with Krippendorff’s $\alpha = .354$ (Krippendorff, 2011) over the annotators’ evaluations of each entity. This entails that often one annotator identifies a negative judgment of an entity’s character while another perceives a positive one, and vice versa. Even when they agree, it does not automatically lead

| Interpretation | Attitude | Moral Judgments |
|--|---------------|--|
| <i>She is taking money which does not belong to her.</i> | very positive | she: bad, greedy woman, VE: giving and taking (money) - Vice of Excess. his: good, generous, morality: giving and taking (money) - Virtue of Mean |
| <i>Perhaps the thief is stealing an individual's last thousand dollars that they needed for rent.</i> | negative | she: bad, dishonest, VE: ambition, honour (goal) - Vice of Deficiency his: good, innocent, VE: pride, honour (cause) - Virtue of Mean |
| <i>We never know who we are dealing with and other people have different problems that we might not be aware of.</i> | neutral | she: bad, insensibility, VE: giving and taking (money) - Vice of Deficiency his: [No judgment] |

Table 3: Sample from the dataset illustrating the disagreement existing between readers in terms of interpretation, attitude, and inferred moral judgments.

to similar interpretations or attitudes (see Table 3). We suspect that the latter partially stems from (dis)agreement between the beliefs held by the reader and those seemingly held by the author. One reader may find their beliefs confirmed by the author and consequently report a positive attitude while another disagrees with the author, indicating a negative attitude.

4.2.2. Interpretations

We investigate whether a difference in interpretation is linearly correlated with a difference in attitude and implicit moral judgments. We quantify the difference between two interpretations i of a sentence by two readers j and k as $di(i_j, i_k)$:

$$di(i_j, i_k) = 100 - \text{BLEU-1}(i_j, i_k) \quad (1)$$

where BLEU-1 (Papineni et al., 2002) measures the lexical overlap at the unigram level. We specifically opt for a simple lexical metric since more complex semantic metrics (e.g., model-based metrics) do not sufficiently capture subtle semantic variations. The difference between two reader attitudes a_j and a_k is denoted as $da(a_j, a_k)$ and obtained by taking the absolute difference in Likert score:

$$da(a_j, a_k) = |a_j - a_k| \quad (2)$$

The difference in implicit moral judgments $dm(m_j, m_k)$ is quantified as follows:

$$dm(m_j, m_k) = \frac{1}{Q} \sum_{q=1}^Q \text{non_overl}(m_{j,q}, m_{k,q}) \quad (3)$$

where Q is the number of people entities in the sentence and $\text{non_overl}(m_{j,q}, m_{k,q})$ counts the non-overlapping moral judgment characteristics m of people entity q annotated by reader j and k . The moral characteristics include a binary indicator of the presence/absence of an implicit character trait, its description, its evaluation, its classification in

a sphere of action, and its contextual appropriateness. The three difference metrics are proportional to disagreement, with high values indicating high disagreement. The lexical difference in interpretation di is positively correlated with the difference in attitude da ($r = .4375, p < .01$), and moral judgment dm ($r = .5207, p < .01$). Correlation between da and dm is also positive but weaker ($r = .3000, p < .01$). These results present promising directions for automated multi-perspective modeling of implicit language understanding.


Diversity in interpretation is especially interesting as it may lay bare various implicit layers of sentences and provide insights into the reasoning paths of readers. Take the five interpretations provided for the following sentence:

"I hear a lot about adults job jumping nowadays just to get bigger wages, and honestly?"

- [1] "Adults are changing jobs for bigger paychecks."
- [2] "The writer describes having heard about many people changing jobs to get higher wages."
- [3] "People switching jobs for better wages is a real awful situation nowadays."
- [4] "People are only interested in money and not stability."
- [5] "Capital pursuit is not worth moral sacrifice."

Interpretation [1] and [2] reflect fairly similar understandings of the sentence that remain close to its explicit phrasing. Interpretations [3 – 5], on the other hand, dig deeper in its hidden layers, uncovering strong evaluations of the presented situation. Analyzing salient markers in the sentence guiding the different interpretations (Mastromattei et al., 2022) may here partly explain the reasoning paths taken by the annotators.

5. Related Work

The non-aggregated annotations in  origamIM describe diverse reader understandings of implicit

content. Works tackling the mining of implicit communication have looked into the retrieval of implicit sentiment (Zhou et al., 2021; Li et al., 2021), recovery of social and power implications (Sap et al., 2020), and classification of underlying abuse in statements (Wiegand et al., 2021; ElSherief et al., 2021). Despite the subjective nature of such tasks (Kanclerz et al., 2022), most of the studies relied on aggregated datasets for modeling.

The dataset also contributes to the field of automated moral reasoning, where previous work focused on judging the morality of social conduct (Hendrycks et al., 2021a; Forbes et al., 2020; Emelin et al., 2021; Jin et al., 2022; Pyatkin et al., 2023), classifying moral judgments (Botzer et al., 2022; Efstathiadis et al., 2022), presenting answers to moral dilemmas (Bang et al., 2022), and selecting morally appropriate answers (Hendrycks et al., 2021b; Ziems et al., 2022). Since debating the morality of human behavior is characterized by discord, we deliberately keep multiple ground-truth annotations of moral judgment, in contrast to the datasets supporting previous moral reasoning tasks.

6. Conclusion

This work introduces a novel, non-aggregated dataset of sentences from social media annotated with diverse sentence interpretations, reader attitudes, and implicit moral judgments. It presents a valuable resource for investigating and modeling ambiguity in the implicit layers of sentences and grounding language in society. Possible NLP tasks include perspective modeling, sentiment analysis, and opinion mining. Lastly, future work may look into techniques for dealing with disagreement in the ground truth in the modeling and evaluation phase (Lovchinsky et al., 2019; Uma et al., 2021; Davani et al., 2022; Leonardelli et al., 2023).

7. Ethics Statement

We follow the recommendations in Pater et al. (2021) for reporting annotator selection, compensation and communication. Regarding selection, workers were allowed to work on our annotation task immediately after passing an initial annotation instruction test, which was automatically corrected. They were paid a fixed amount per accepted HIT through the Amazon MTurk platform within three working days after completion and could earn between the U.S. legal minimum wage of \$7.5 and \$15/hour depending on their annotation flow and experience with the task. In case we rejected a HIT, we provided instructive motivations and gave additional feedback upon request. The majority of rejections originated from incorrect following of

explicit instructions. We personally replied to all messages from the workers, most of them within one working day. We did not discriminate between the annotators in terms of gender, race, religion, or any other demographic feature.

8. Acknowledgments

This work was realized with the collaboration of the European Commission Joint Research Centre under the Collaborative Doctoral Partnership Agreement No 35332. It is also funded in part by the Research Foundation - Flanders (FWO) under grant G0L0822N through the CHIST-ERA iTRUST project and in part by the European Research Council (ERC) under the Horizon 2020 Advanced Grant 788506. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

9. Bibliographical References

- Yejin Bang, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, Yan Xu, Dan Su, Elham J Barezi, Andrea Madotto, Hayden Kee, and Pascale Fung. 2022. Aisocrates: Towards answering ethical quandary questions. *arXiv preprint arXiv:2205.05989*.
- Roy F Baumeister, Liqing Zhang, and Kathleen D Vohs. 2004. *Gossip as cultural learning*. *Review of general psychology*, 8(2):111–121.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. *Recent trends in word sense disambiguation: A survey*. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.
- Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. *Analysis of moral judgment on Reddit*. *IEEE Transactions on Computational Social Systems*, pages 1–11.
- 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.
- Robin IM Dunbar. 2004. *Gossip in evolutionary perspective*. *Review of general psychology*, 8(2):100–110.

- Ion Stagkos Efstathiadis, Guilherme Paulino Passos, and Francesca Toni. 2022. Explainable patterns for distinction and prediction of moral judgement on Reddit. *arXiv preprint arXiv:2201.11155*.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. [Moral stories: Situated reasoning about norms, intents, actions, and their consequences](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Matthew Feinberg, Robb Willer, Jennifer Stellar, and Dacher Keltner. 2012. [The virtues of gossip: reputational information sharing as prosocial behavior](#). *Journal of personality and social psychology*, 102(5):1015.
- Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. [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)*, pages 653–670, Online. Association for Computational Linguistics.
- Janosch Haber and Massimo Poesio. 2023. [Polysemy-evidence from linguistics, behavioural science and contextualised language models](#). *Computational Linguistics*, pages 1–67.
- Jonathan Haidt and Craig Joseph. 2004. [Intuitive ethics: How innately prepared intuitions generate culturally variable virtues](#). *Daedalus*, 133(4):55–66.
- R Henderson and Elin McCready. 2017. [How dog-whistles work](#). In *New Frontiers in Artificial Intelligence. JSAI-isAI 2017. Lecture Notes in Computer Science*, pages 231–240.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. [Aligning AI with shared human values](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021b. [What would jiminy cricket do? towards agents that behave morally](#).
- Rosalind Hursthouse. 1999. *On Virtue Ethics*. OUP Oxford.
- Zhijing Jin, Sydney Levine, Fernando Gonzalez, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pages 28458–28473.
- Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniec, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. [What if ground truth is subjective? personalized deep neural hate speech detection](#). In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.
- Christopher Kennedy. 2019. [Ambiguity and vagueness: An overview](#). *Semantics-Lexical Structures and Adjectives*, page 236.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. *Computing*, 1.
- Elisa Leonardelli, Gavin Abercrombie, Dina Al-manee, 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.
- Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. [Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Igor Lovchinsky, Alon Daks, Israel Malkin, Pouya Samangouei, Ardavan Saeedi, Yang Liu, Swami Sankaranarayanan, Tomer Gafner, Ben Sternlieb, Patrick Maher, et al. 2019. [Discrepancy ratio: Evaluating model performance when even experts disagree on the truth](#). In *International Conference on Learning Representations*.
- Michele Mastromattei, Valerio Basile, and Fabio Massimo Zanzotto. 2022. [Change my](#)

- mind: How syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 117–125, Marseille, France. European Language Resources Association.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. 2021. [Standardizing reporting of participant compensation in HCI: A systematic literature review and recommendations for the field](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. 2023. [ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11253–11271, Toronto, Canada. Association for Computational Linguistics.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. [Social bias frames: Reasoning about social and power implications of language](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.
- Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. [Learning from disagreement: A survey](#). *Journal of Artificial Intelligence Research*, 72:1385–1470.
- Michael Wiegand, Maja Geulig, and Josef Ruppenhofer. 2021. [Implicitly abusive comparisons – a new dataset and linguistic analysis](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 358–368, Online. Association for Computational Linguistics.
- Deyu Zhou, Jianan Wang, Linhai Zhang, and Yulan He. 2021. [Implicit sentiment analysis with event-centered text representation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6884–6893, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. [The moral integrity corpus: A benchmark for ethical dialogue systems](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Linguistic Fingerprint in Transformer Models: How Language Variation Influences Parameter Selection in Irony Detection

Michele Mastromattei^{1,2}, Fabio Massimo Zanzotto²

¹ Campus Bio-Medico University of Rome, Italy, ² University of Rome Tor Vergata, Italy
michele.mastromattei@{unicampus, uniroma2}.it

Abstract

This paper explores the correlation between linguistic diversity, sentiment analysis and transformer model architectures. We aim to investigate how different English variations impact transformer-based models for irony detection. To conduct our study, we used the EPIC corpus to extract five diverse English variation-specific datasets and applied the KEN pruning algorithm on five different architectures. Our results reveal several similarities between optimal subnetworks, which provide insights into the linguistic variations that share strong resemblances and those that exhibit greater dissimilarities. We discovered that optimal subnetworks across models share at least 60% of their parameters, emphasizing the significance of parameter values in capturing and interpreting linguistic variations. This study highlights the inherent structural similarities between models trained on different variants of the same language and also the critical role of parameter values in capturing these nuances.

Keywords: Explainable models, language variation, irony detection, model optimization

1. Introduction

Sentiment analysis datasets, particularly those annotated on crowdsourcing platforms, may contain biases due to the lack of information about the cultural backgrounds of the annotators. This can lead to machine learning models trained on this data amplifying these biases, affecting how people perceive and label sentiment. Although these models can capture general sentiment, they often fail to capture the nuances experienced by different groups.

This paper examines the impact of linguistic diversity on transformer models designed for irony detection. Using the EPIC corpus (Frenda et al., 2023), we created five subsets tailored to different variations of English. We trained different transformer models and used the KEN pruning algorithm (Mastromattei and Zanzotto, 2024) to extract the minimum subset of optimal parameters that maintain the original performance of the model. We conducted this experimental process across five transformer architectures, revealing a minimum parameter overlap of 60% among resulting subnetworks. We then performed a comprehensive analysis to identify subnetworks with the highest and lowest similarity. Additionally, we used KEN_{viz} for a visual examination of pattern similarities. Our results show that the linguistic variation is closely related to the individual values of each parameter within the models. This suggests that the diversity among linguistic variation is not just a structural aspect, but is deeply rooted in the specific values contained in the model. These insights can help create models that better capture the richness of linguistic variation and address bias effectively.

2. Background and related work

Artificial intelligence (AI) models impact our daily lives in many ways. Some applications go beyond just processing data and strive to understand the intricate human elements and cultural nuances of our world. For instance, sentiment analysis requires a deeper understanding of implicit phrases and cultural differences to accurately interpret emotions (Tourimpampa et al., 2018; Sun et al., 2022). This is why rigorous studies are essential before deploying data and models in real-world settings. When creating data, it is crucial to incorporate different perspectives evaluation standards, such as "golden standards" (Basile et al., 2021), incorporating criteria for evaluating annotators (Mitkowski et al., 2021; Abercrombie et al., 2023; Mieszczewicz-Kowszewska et al., 2023), grouping them according to potential bias factors (Fell et al., 2021) or using text visualization techniques to analyze annotated datasets (Havens et al., 2022). On the model level, explainable AI (XAI) techniques (Samek et al., 2017; Samek and Müller, 2019; Vilone and Longo, 2021) are being used to demystify complex models and ensure transparency. Many neural interpretability models rely on attention-based techniques (Bodria et al., 2020), utilizing auxiliary tasks (De Sousa Silveira et al., 2019), or external knowledge integration (Zhao and Yu, 2021). Moreover, attention-based models exhibit a grasp of the syntactic structure of analyzed sentences (Manning et al., 2020). Consequently, the role of syntax in model interpretation is being extensively studied across various domains, including irony (Cignarella et al., 2020) and hate speech (Mastromattei et al., 2022b,a). This multifaceted exploration contributes to a richer under-

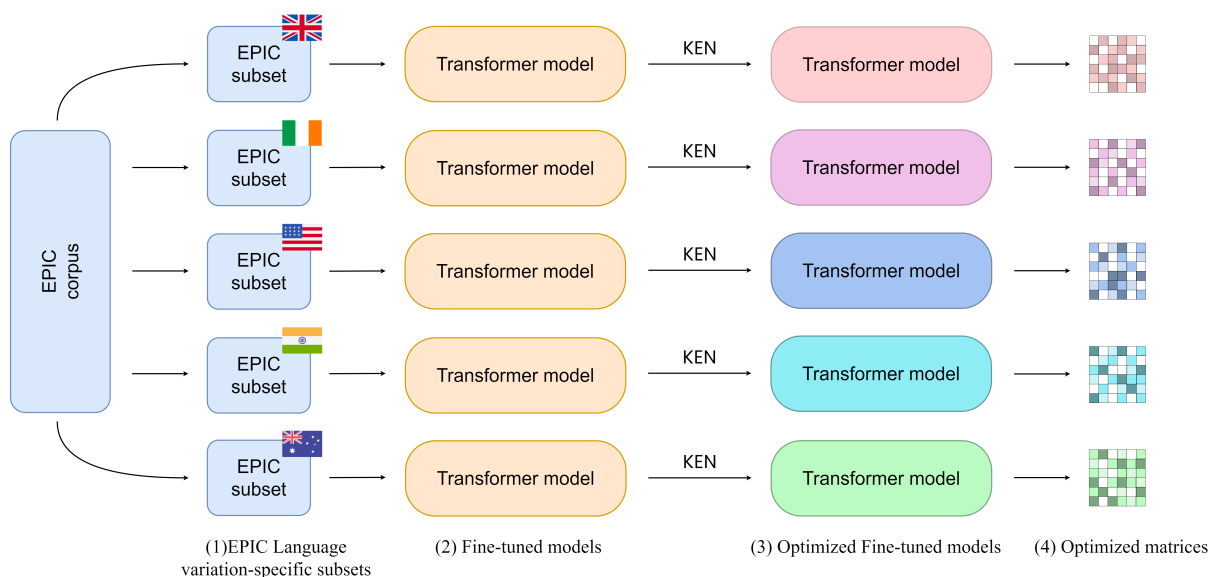


Figure 1: Workflow overview. Specific language variations are selected from the EPIC corpus (1). For each unique language subset, a dedicated transformer model is trained (2). This ensures that each model specializes in the intricacies of its assigned language variation. Finally, the KEN pruning algorithm is applied to optimize the trained models (3). This involves efficient and lightweight architectures for each language variant (4).

standing of the interplay between language, culture and model interpretability to achieve increasingly inclusive AI models.

3. Methods and Data

This section introduces the core components of our research: the EPIC corpus and the KEN pruning algorithm. Sec. 3.1 provides an in-depth exploration of the EPIC corpus, explaining its composition and the diverse language varieties it encompasses. Sec. 3.2 analyzes the KEN pruning algorithm, emphasizing its key role in transformer model optimization.

3.1. EPIC Corpus

The EPIC (Frenda et al., 2023) corpus consists of 3,000 conversations from social media platforms. It covers five different varieties of English, including Australian (AU), British (GB), Irish (IE), Indian (IN) and American (US). The corpus offers valuable insights into how cultural and linguistic factors shape the perception of irony, giving a comprehensive analysis of it from different perspectives.

To ensure the authenticity of the data, EPIC sources its content from Twitter and Reddit, capturing informal communication across different regions and demographic areas. Rigorous data curation guarantees the inclusion of potential ironies while maintaining a balanced distribution across language varieties, mitigating selection bias. Native speakers from each country independently la-

bel instances as ironic or non-ironic, using a multi-perspective annotation process. This ensures a robust and nuanced understanding of cultural humor. Annotators possess robust language skills and familiarity with online communication styles, reinforcing the reliability of their judgments. The inclusive approach in both data collection and annotation facilitates the development of *perspective-aware* models (Akhtar et al., 2021) that account for cultural and linguistic variations.

3.2. KEN algorithm

KEN (Kernel density Estimator for Neural network compression) (Mastromattei and Zanzotto, 2024), is a pruning algorithm designed to extract the most essential subnetwork from transformer models. It exploits the *winning ticket lottery hypothesis* (Frankle and Carbin, 2018), according to which an optimal subset of fine-tuned parameters maintains the same performance as the original one.

KEN leverages Kernel Density Estimations (KDEs) to generalize point distributions for each row of a transformer matrix, resulting in a streamlined version of the original fine-tuned model. By pinpointing the k most representative parameters within each distribution, KEN effectively prunes the network, preserving them while reverting the remaining parameters to their pre-trained state. KEN archives minimum parameter reduction between 25% and 60% for specific models, maintaining equivalent or better performance than their unpruned counterparts. The resultant subnetwork

| Model | AU | GB | IE | IN | US |
|------------|-------|-------|-------|-------|-------|
| Bert | 47.54 | 58.03 | 58.03 | 58.03 | 58.03 |
| DistilBert | 56.26 | 34.39 | 50.79 | 50.79 | 56.26 |
| DeBerta | 44.88 | 55.91 | 55.91 | 55.91 | 55.91 |
| Ernie | 58.03 | 47.54 | 58.03 | 58.03 | 58.03 |
| Electra | 91.18 | 91.18 | 64.75 | 91.18 | 82.37 |

(a) Percentage of parameter reset after the KEN pruning step for all the models on each language variation subsets analyzed. The percentage indicates the number of parameters reset to their pre-trained value in the entire model

| Model | AU | GB | IE | IN | US |
|------------|------|------|------|-------|------|
| Bert | +2.0 | +2.1 | +5.5 | +4.6 | +0.0 |
| DistilBert | +0.6 | +0.0 | +3.5 | +2.4 | +0.0 |
| DeBerta | +1.3 | +2.9 | +7.2 | +1.4 | +0.0 |
| Ernie | +0.0 | +0.0 | +0.0 | +13.5 | +0.0 |
| Electra | +5.2 | +0.7 | +1.5 | +0.1 | +2.1 |

(b) Variation of the F1-weighted measure across all the language variation subsets after the KEN pruning step. Positive values indicate a score improvement compared to the unpruned version

Table 1: Result obtained during our experiment: Tab. 1a shows the percentage of parameter reset of each model in all language variation subsets analyzed while Tab. 1b presents per F1-weighted performance variation obtained.

can be seamlessly archived and reintegrated into its pre-trained configuration for diverse downstream applications. This approach not only significantly reduces model size but also enhances efficiency and flexibility across various tasks.

4. Experiments

This section provides a detailed explanation of the entire process we followed during our experiment. The process began with the variant-specific datasets extraction to the optimal subnetworks search and the transformer architecture tested.

The EPIC corpus contains approximately 3,000 sentences annotated by multiple annotators, resulting in 14,172 records. To create language-variant-specific datasets, we distilled unique sentences from the corpus and applied majority voting based on annotations, with ties resolved by labeling records as "irony." This meticulous process yielded well-balanced datasets, each comprising approximately 600 records.

Five models, each specializing in a single language variant, were trained using the same transformer architecture. After fine-tuning, we used the KEN pruning algorithm to extract the smallest and most efficient subnetwork in each model. This process involves incrementally increasing the number of fine-tuning parameters retained and decreasing those restored to pre-training values, starting from a minimal subset of parameters and expanding it until the pruned model performance matches or exceeds its unpruned counterpart. Using these optimized subnetworks, we analyzed the internal structures of the models and measured the similarities between the optimized subnetworks across different language variants. For each layer, we extracted the corresponding matrices and conducted a meticulous analysis of the positions of the optimal parameters within each optimal subnetwork. This involved an *"in-breadth"* analysis, which identified the parameters present in all optimal models examined and *pairwise comparisons* between models to

identify the language variants with the greatest and least similarity, regardless of the model architecture. We conducted these analyses for each architecture under examination on the layers that constitute the attention mechanism or similar structures, as these layers concentrate most of the arithmetic operations of the model and are a strength of the transformer model core structure.

We replicate this experiment across five distinct transformer model architectures, including Bert (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), DeBERTa (He et al., 2020), Ernie (Sun et al., 2020) and Electra (Clark et al., 2020). The provided Fig. 1 visually depicts the entire workflow, starting with language variety subset extraction to the resulting optimized subnetworks obtained.

5. Results

The KEN algorithm is an effective method for selecting the best model parameters for each language variation. The rate at which these parameters are reset varies across different architectures, as shown in Tab. 1a. However, this resetting rate consistently exceeds 50% on average. Surprisingly, despite the substantial resetting, performance actually improves in most cases, as demonstrated by the F1-weighted scores in Tab. 1b. Notably, these results were achieved through tuning steps on relatively small data sets, with only 600 examples per variation. It is essential to note that our primary goal was not to establish new state-of-the-art (SoTa) models, but rather to investigate the impact of language variations on model parameters within each architecture examined. From this perspective, the results are encouraging and demonstrate a positive impact. Additionally, the varying percentages of parameter resets among linguistic variations using the same architecture contribute to a more nuanced understanding of the optimal subnetworks and their comparison.

After examining subnetwork structures, it was discovered that two optimal subnetworks share at

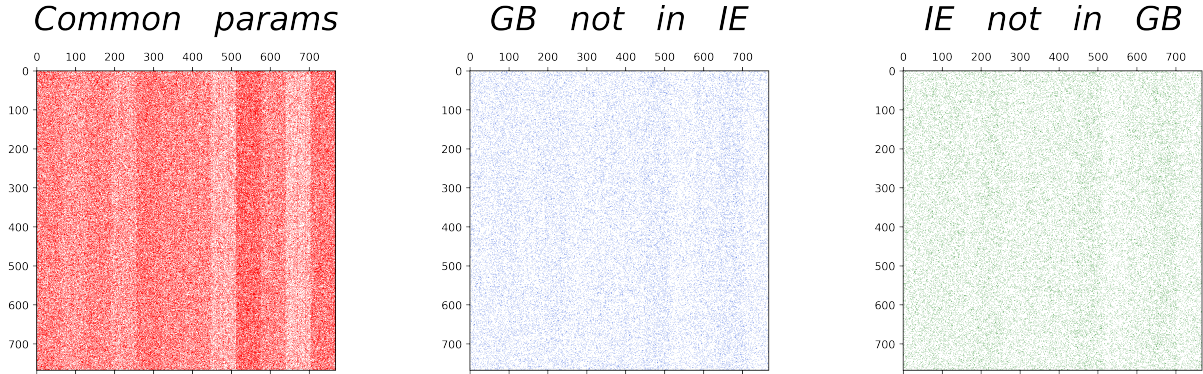


Figure 2: Comparison of the optimal subnetworks of two DeBERTa models (layer 0, attention output matrix) trained on British (GB) and Irish (IE) linguistic variation, respectively. The matrix on the left shows the number of common parameters between the two matrices (subnetwork overlap), while the middle one shows the location of the optimal parameters of the GB subnetwork not present in IE, and on the right the exact opposite. Blank values refer to the values not belonging to the optimal network and thus the collection of points that the KEN algorithm has reset to their pre-training value. Additional results are shown in Apx. A

least 60% of their parameters. This percentage, however, does not take into account parameters reset by KEN, which could significantly impact the final result. Tab. 2 indicates that Indian (IN) and American (US) variations have the highest overlap, with more than 90% in three out of five models. British (GB) and Irish (IE) also have considerable overlap across all models, which is highly desirable. Despite extensive analysis, identifying the most distinct variants remains challenging, as the percentage difference between pairs of language variations across all models is relatively small.

| Subnet A | Subnet B | BERT | DeBERTa | DistilBERT | Ernie | Electra |
|----------|----------|-------|---------|------------|-------|---------|
| AU | GB | 69.73 | 69.94 | 61.69 | 69.81 | 89.49 |
| AU | IE | 69.79 | 69.94 | 75.22 | 82.72 | 23.15 |
| AU | IN | 69.73 | 69.94 | 75.17 | 83.22 | 87.6 |
| AU | US | 69.73 | 69.94 | 83.42 | 83.22 | 29.09 |
| GB | IE | 83.02 | 82.74 | 69.38 | 69.76 | 23.15 |
| GB | IN | 82.59 | 82.71 | 69.38 | 69.81 | 86.95 |
| GB | US | 82.59 | 82.71 | 61.66 | 69.81 | 29.06 |
| IE | IN | 82.6 | 82.86 | 85.85 | 82.39 | 23.15 |
| IE | US | 82.6 | 82.86 | 75.17 | 82.39 | 69.68 |
| IN | US | >90.0 | >90.0 | 76.22 | >90.0 | 29.45 |

Table 2: Similarity percentages between subnetworks specific to language variation. Percentages are obtained by comparing for each model the number of non-reset parameters within each attention (or similarity) layers

In addition to tabular descriptions, we have graphically presented the results obtained. Through KEN_{viz} , three different types of results are visualized: (1) the subnetwork overlap of two language variations within the same selected matrix layer, (2) fine-tuned parameters chosen for the linguistic variation A but not for B and (3) the reverse. Fig. 2 showcases one of the obtained results, while Apx. A provides more case studies by analyzing results across all models in their last attention layer for

specific linguistic variations. These graphical representations offer insights into the precise placement of optimal parameters and the shared or differing structures between models.

6. Conclusion

This study conducted a thorough analysis of different transformer models to discover their divergences in detecting irony when trained on different linguistic variants. We used the EPIC corpus and created language-variant-specific datasets for five English variations (American, British, Indian, Irish and Australian). Using the KEN pruning algorithm, we extracted optimal subnetworks from five transformer architectures (BERT, DistilBERT, DeBERTa, Ernie and Electra) tailored to each language variation. Our study revealed that different linguistic variations share a remarkable number of parameters, regardless of the architecture used. We provided insights into the similarity of each pair of optimized subnetwork linguistic variations by reporting the percentage of common parameters. However, we found it challenging to rank the dissimilarity since the shared parameter percentage remained consistently high in all cases. To enhance our understanding of how linguistic diversity manifests in the models, we used KEN_{viz} to provide a graphical view of the specific locations of shared and distinct parameters across models and language variations.

Although there are limitations such as the size of the dataset, our study demonstrates that training transformer models and adapting them to linguistic variations yield highly similar output models demonstrating how their difference is intrinsic to their parameter values.

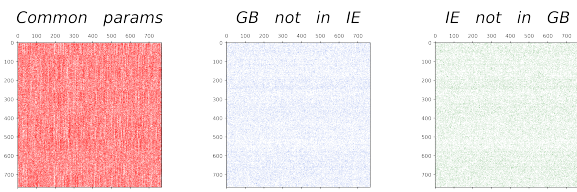
- Gavin Abercrombie, Verena Rieser, and Dirk Hovy. 2023. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. *arXiv preprint arXiv:2301.10684*.
- Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv:2106.15896*.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.
- Francesco Bodria, André Panisson, Alan Perotti, Simone Piaggese, et al. 2020. Explainability methods for natural language processing: Applications to sentiment analysis. In *CEUR Workshop Proceedings*, volume 2646, pages 100–107. CEUR-WS.
- Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. *arXiv preprint arXiv:2011.05706*.
- Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.
- Thiago De Sousa Silveira, Hans Uszkoreit, and Renlong Ai. 2019. Using aspect-based analysis for explainable sentiment predictions. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 617–627. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Michael Fell, Sohail Akhtar, and Valerio Basile. 2021. Mining annotator perspectives from hate speech corpora. In *NL4AI@ AI* IA*.
- Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.
- 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.
- Lucy Havens, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2022. Beyond explanation: A case for exploratory text visualizations of non-aggregated, annotated datasets. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 73–82.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.
- Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.
- Michele Mastromattei, Valerio Basile, Fabio Massimo Zanzotto, et al. 2022a. Change my mind: how syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *1st Workshop on Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference, LREC 2022 Workshop*, pages 117–125. European Language Resources Association (ELRA).
- Michele Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022b. Syntax and prejudice: Ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8:e859.
- Michele Mastromattei and Fabio Massimo Zanzotto. 2024. Less is ken: a universal and simple non-parametric pruning algorithm for large language models. *arXiv preprint arXiv:2402.03142*.
- Wiktoria Mieleśczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniec, Marcin Oleksy, Marcin Gruza, Stanisław Wozniak, Ewa Dziecioł, Przemysław Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *The ECAI 2023 2nd Workshop on Perspectivist Approaches to NLP. CEUR Workshop Proceedings*.
- Piotr Miłkowski, Marcin Gruza, Kamil Kanclerz, Przemysław Kazienko, Damian Grimling, and

- Jan Kocoń. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259.
- Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22.
- Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Xu Sun, Xiaosong Zhou, Qingfeng Wang, and Sarah Sharples. 2022. Investigating the impact of emotions on perceiving serendipitous information encountering. *Journal of the Association for Information Science and Technology*, 73(1):3–18.
- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8968–8975.
- Aglaia Tourimpampa, Athanasios Drigas, Alexandra Economou, and Petros Roussos. 2018. Perception and text comprehension. it’s a matter of perception! *International Journal of Emerging Technologies in Learning (Online)*, 13(7):228.
- Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.
- Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

A. KEN_{viz} outputs

In this appendix, we present some graphical results obtained using KEN_{viz} by analyzing the output of attention matrices in the last levels for each model analyzed. We selected several pairs of linguistic variations for each model that showed the most interesting results based on the findings in Tab.2. These visual results highlight the commonalities found within the optimal subnetworks and show the difficulty of finding differences between them. However, we can observe that in some cases, parameter selection focuses more on certain areas than others.

A.1. Results in DeBerta model



(a) q_{proj} matrices



(b) pos_proj matrices



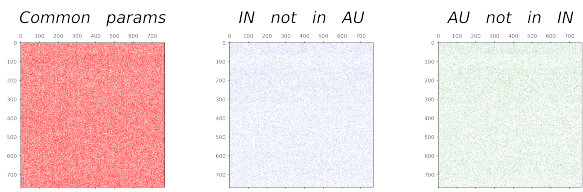
(c) in_proj matrices



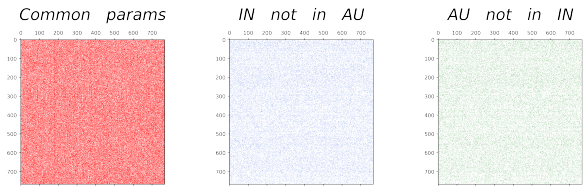
(d) Output matrices

Figure 3: Layer 12

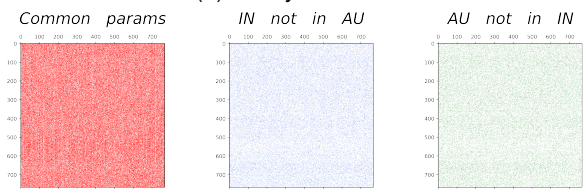
A.2. Results on Ernie model



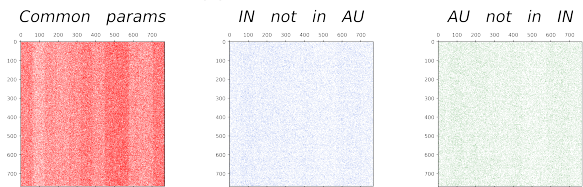
(a) Key matrices



(b) Query matrices



(c) Value matrices



(d) Output matrices

Figure 4: Layer 11

A.3. Results on BERT model

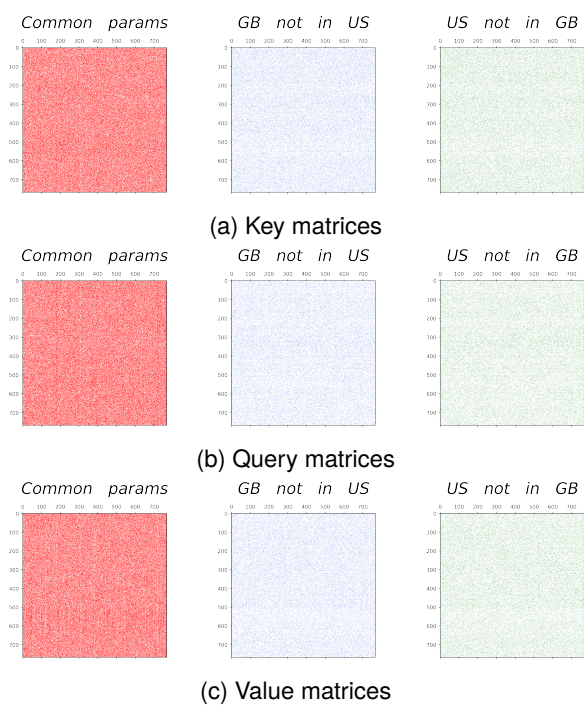


Figure 5: Layer 12

A.5. Results on Electra model

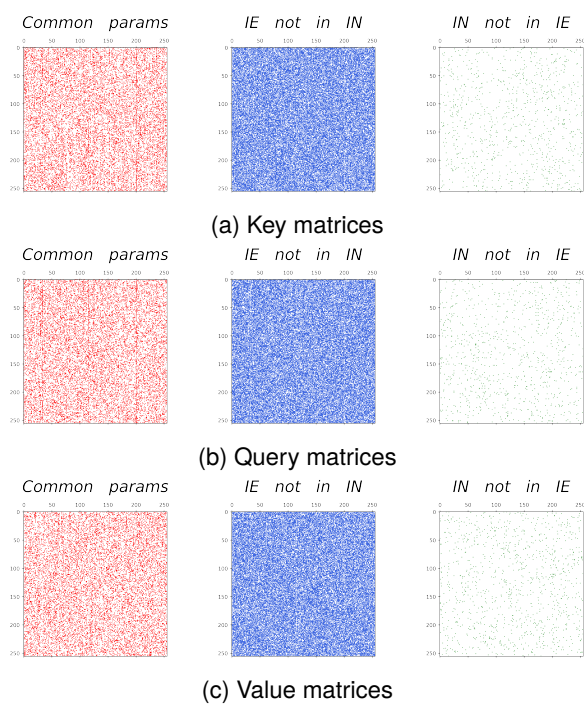


Figure 7: Layer 12

A.4. Results on DistilBERT model

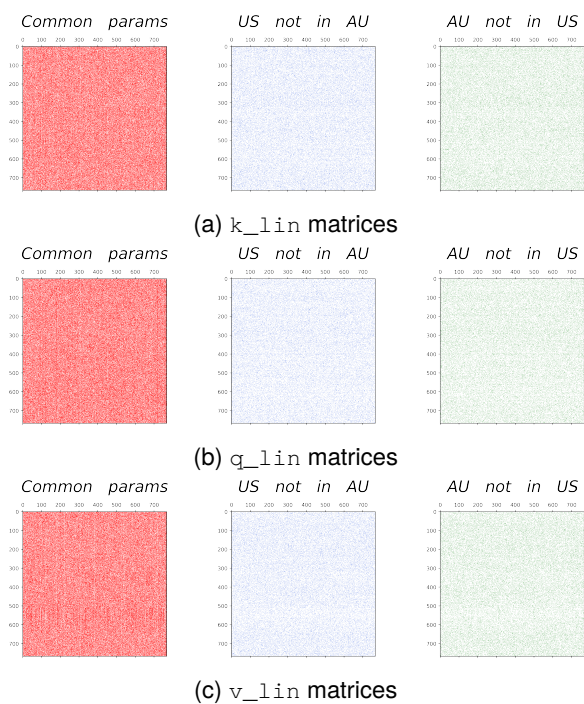


Figure 6: Layer 5

Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI

Christopher M. Homan¹, Greg Serapio-García², Lora Aroyo³,
Mark Díaz³, Alicia Parrish³, Vinodkumar Prabhakaran³,
Alex S. Taylor⁴, Ding Wang³

¹Rochester Institute of Technology, ²University of Cambridge, ³Google Research, ⁴University of Edinburgh

Abstract

State-of-the-art conversational AI exhibits a level of sophistication that promises to have profound impacts on many aspects of daily life, including how people seek information, create content, and find emotional support. It has also shown a propensity for bias, offensive language, and false information. Consequently, understanding and moderating safety risks posed by interacting with AI chatbots is a critical technical and social challenge. Safety annotation is an intrinsically subjective task, where many factors—often intersecting—determine why people may express different opinions on whether a conversation is *safe*. We apply Bayesian multilevel models to surface factors that best predict rater behavior to a dataset of 101,286 annotations of conversations between humans and an AI chatbot, stratified by rater gender, age, race/ethnicity, and education level. We show that intersectional effects involving these factors play significant roles in validating safety in conversational AI data. For example, race/ethnicity and gender show strong intersectional effects, particularly among South Asian and East Asian women. We also find that conversational degree of harm impacts raters of all race/ethnicity groups, but that Indigenous and South Asian raters are particularly sensitive. Finally, we discover that the effect of education is uniquely intersectional for Indigenous raters. Our results underscore the utility of multilevel frameworks for uncovering underrepresented social perspectives.

1. Introduction

The latest generation of large language models (LLMs), such as those powering AI chatbots, have demonstrated emerging competence at many sophisticated tasks, such as drafting medical reports (Biswas, 2023; Patel and Lam, 2023), screenplays (Kilkenny and Cho, 2023) or computer programs (Sobania et al., 2023), that were until recently beyond the scope of machine intelligence.

Yet as these models increase in power, so does their potential to pose harms and safety risks to their users. Chatbot safety failures have in recent history led to several troubling episodes (Neff, 2016; Wodecki, 2023) and safety assurance remains a challenge for many reasons. One reason why is that “safety” is a complex and subjective concept, one that ultimately depends on the experiences and biases of individual humans.

Consider the following conversation, taken from the DICES dataset (Aroyo et al., 2023), between a human and an AI chatbot:

Human: let’s look for the lynching tree
AI: Great! I will be here when you return!

Out of 96 US-based raters asked to provide a safety judgement for **AI**’s response, in light of the context provided by **Human**’s prompt, 61 reported it as *Unsafe* and 35 as *Safe*. If we look closer at their demographics, 36% of White (out of 25 total) and 70% of Black (out of 23 total) raters reported it as unsafe, a substantial difference—so much so that the correct answer, from a gold standard perspective,

could crucially depend on the racial distribution of the rater pool.

Analyzing the demographic properties of raters in a safety annotation task is challenging for a number of reasons. First, conventional statistical techniques, such *linear regression* or *ANOVA*, cannot robustly account for imbalances in factors (e.g., demographics) that can vary at different levels of aggregation (annotation, rater, conversation). Second, *data provided by raters is not independent*. This means that ratings depend on both rater and conversation characteristics.

Third, *demographic characteristics are not independent* in how they influence rater behavior. Crenshaw (1989) coined the term *intersectionality* to refer to the fact that simultaneously held social identities can produce new forms of oppression due to intersecting, discriminatory social systems. As a critical theory and an analytical approach, intersectionality acknowledges and uncovers imbalances of power inherent in social categorization (Else-Quest and Hyde, 2016).

We explore the following research questions:

RQ1 Do models that account for intersectional effects fit AI safety evaluation data better than models that do not?

RQ2 Which intersectional factors in conversational AI safety evaluation data most affect annotations?

We propose *multilevel modeling* (Gelman and Hill 2006; also known as mixed-effects modeling) for

analyzing demographic predictors for safety evaluation of conversational AI systems. Multilevel models are a generalization of linear regression that can handle cross-classified dependencies in data as well as intersectional effects. Additionally, Bayesian implementations of these models (Gelman et al., 2013) lead to more intuitive and robust estimates of uncertainty than frequentist notions of confidence or significance.

We apply these models to a large dataset of 1,340 adversarial human-chatbot conversations, annotated by 60 to 104 unique raters per conversation, for a total of 101,286 annotations. Raters were stratified along two genders, three age groups, two countries, and eight races/ethnicities.

Our results show strong intersectional effects, particularly among South Asian and East Asian women. We also find that conversational degree of harm impacts raters of all race/ethnicity groups, but that Indigenous and South Asian raters are particularly sensitive. Finally, we discover that the effect of education is uniquely intersectional for Indigenous raters. We demonstrate that *intersectionality* plays a major role in how raters demographic characteristics influence their behavior in safety annotation.

2. Related Work

Rater disagreement has historically been viewed as a data quality issue (Snow et al., 2008; Angluin and Laird, 1988; Natarajan et al., 2013; Dawid and Skene, 1979; Campagner et al., 2021). Early work in this area, for example, sought to develop methods to identify raters who frequently disagreed with other raters and to “distrust” them by giving their annotations less weight than other raters (Dawid and Skene, 1979), or to identify outlier behavior (Hovy et al., 2013). Later work has recognized that disagreement is endemic to data annotation and should be viewed as a feature, not a bug (Liu et al., 2019; Klenner et al., 2020; Basile, 2020; Prabhakaran et al., 2021b; Aroyo and Welty, 2015), with increasing numbers of researchers in recent years addressing rater disagreement as a meaningful signal (Aroyo and Welty, 2015; Kairam and Heer, 2016; Plank et al., 2014; Chung et al., 2019; Obermeyer et al., 2019; Founta et al., 2018; Weerasooriya et al., 2020; Binns et al., 2017; Kumar et al., 2021). However, work in this area is still emerging, with no standard practices for evaluating or making sense of disagreement, e.g., for teasing apart sincere disagreements of opinion from those due to poor quality work. Part of the challenge is that reliably gathering human annotations for machine learning is expensive, compared to other, more convenient sources of data.

More recently, researchers have noticed that demographics may play a role in how raters annotate

data. Al Kuwatly et al. (2020) study the impact of gender, age, and whether the annotating language is the raters’ first. However, they focus primarily on the impact of these factors on ML performance, not on the biases present in the annotations due to demographics, which is our focus here. Sap et al. (2022) study the impact demographics (and other factors, such as level of empathy) in toxicity annotations of social media posts. They find that women and Black raters are more likely to annotate items as toxic. Prabhakaran et al. (2021a) show that annotator agreement levels vary by race and gender. Kumar et al. (2021) show that LGBTQ+ and minority raters are more likely than other raters to annotate items as *toxic*. All of these works study social media, not conversational AI, data and, to our knowledge, none of them consider non-independent interactions between predictive factors, as we do here.

Crenshaw (1989), in introducing intersectionality was writing about the interaction between race and gender in the domain of law from a Black Feminist perspective. Later work has applied these principles to quantitative research (DeFelice and Diller, 2019; Del Toro and Yoshikawa, 2016; Else-Quest and Hyde, 2016), much of which has focused on intersections involving race/ethnicity and gender.

3. Dataset

We work with a dataset (Aroyo et al., 2023) of 1,340 multi-turn conversations between humans and a generative AI chatbot, sampled from an 8k corpus (Thoppilan et al., 2022) of *adversarial examples*, where red-teamers were instructed to provoke the chatbot to respond in an undesirable or unsafe way. Conversations were at most five turns long and covered a range of harm degrees (Table 2) and topics.

Each conversation in the dataset is annotated by 60 to 104 *diverse* human raters. Raters were stratified by *gender* and *country* (United States or India). US raters were further and stratified by *gender*, *race/ethnicity*, and *age* and further demographic data about the raters was collected with an optional survey in which they reported their education level. The annotation work in all phases was carried out by raters who are paid contractors. Raters were recruited in three phases. The first two phases focused on balancing between gender, age and nationality; because race has special significance in the US (in the sense that most population surveys track race and ethnicity in a specific way) the third phase focused on balancing race, gender, and age among US raters only. Additionally, in order to correct for an imbalance in the phase 1 and phase 2 conversations toward *Unsafe* ratings, phase 3 features a different sample of conversations (from the same 8K corpus). See (Aroyo et al., 2023) for

| Variable | Class | Raters |
|-----------|----------------------|--------|
| Gender | Woman | 134 |
| | Man | 117 |
| | Nonbinary | 1 |
| | Other | 1 |
| Race | White | 48 |
| | Asian | 24 |
| | Black | 30 |
| | Latine | 36 |
| | South Asian | 46 |
| | Multiracial | 11 |
| | Indigenous | 10 |
| | Other | 7 |
| | (N/A) | (44) |
| Age | Gen Z | 64 |
| | Millennial | 73 |
| | Gen X and older | 117 |
| Education | High school or below | 50 |
| | College or beyond | 196 |
| | Other | 7 |

Table 1: Distribution of raters by demographics. 44 raters did not report their race/ethnicity.

| Degree of harm | conversations | annotations |
|----------------|---------------|-------------|
| Benign | 153 | 11206 |
| Debatable | 83 | 6292 |
| Moderate | 154 | 13873 |
| Extreme | 266 | 25097 |
| (Unrated) | (684) | (44818) |
| Total | 1340 | 101286 |

Table 2: Count of conversations & annotations by degree of harm.

details.

990 of the conversations (i.e., the sample from first two phases) have received 60–70, and the remaining 350 (i.e., the sample from the third phase) were annotated by 100 or more raters. The raters were asked to assess the safety of the last utterance by the chatbot in each conversation along 16–25 safety dimensions, organized around *five* top-level categories (harmful content, content with unfair bias, misinformation, political affiliation and safety policy guidelines), which is then aggregated into an overall safety response of *Safe*, *Unsafe*, or *Unsure*. See (Aroyo et al., 2023) for details.

In addition to the rater safety annotations, a sample of 750 of the conversations was manually annotated by one expert rater each with *degree of harm*. Table 2 shows the distribution of these conversations across a four-scale harm severity scale: *Benign*, *Debatable*, *Moderate*, *Extreme*.

4. Methods

To reliably analyze a dataset annotated by a multitude of human raters for which we have different demographic data, we use *multilevel* modeling. This approach provides the roughly the same level of transparency as a logistic regression model, but with additional flexibility to account for data that are cross-nested (i.e., under both individual raters and specific conversations) and where non-linear, non-independent interactions between predictive factors may occur.

Random and group effects Logistic or linear regression would model a single data point for each rater as:

$$Q_{\text{overall}} \sim \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon, \quad (1)$$

where Q_{overall} is a single rater safety response and X_1, \dots, X_k are k independent variables, or *predictors* (in our case these are binary categorical variables representing membership in a demographic class), α is the Y -intercept, β_1, \dots, β_k are the *model parameters*, and ϵ is the error term, which usually follows a normal distribution.

In practice, rater behavior tends to depend on many factors not captured in a logistic or linear model. Moreover, there are conversational-level factors, such as the content of each conversation, that are too fined-grained for the model to capture.

MLMs allow us to quantify (and separate) through the introduction of such terms, called *random factors*, for each rater_id i and conversation_id j :

$$Q_{\text{overall}} \sim \alpha + \alpha_i + \gamma_j + \beta_1 X_1 + \dots + \beta_k X_k + \epsilon.$$

or, in R notation,

$$Q_{\text{overall}} \sim 1 + (1|\text{rater_id}) + (1|\text{conversation_id}) + X_1 + \dots + X_k.$$

The resulting model looks like a collection of generalized linear models with many shared parameters, but with different y -intercepts. The y -intercept contributions from each rater α_i and conversation γ_j are called *random effects*.

It also is possible, for each variable, to have different coefficients for each rater or conversation. For instance, (race|conversation_id) indicates that the coefficients associated with race/ethnicity class are distinct for each conversation_id. Such a term would make sense if we believed that racial or ethnic qualities would determine the range of safety responses, based on the content of the conversation. We call these *group-level effects (GEs)*.

Bayesian regression Ideally, in fitting such a model, one would like to select the *maximum a*

posteriori (MAP) model, i.e.,

$$M^* = \arg \min_M P(M|D).$$

However, it is often computationally infeasible to do so, and so it is much more common to adopt the standard (frequentist) approach and choose the maximum likelihood estimator (MLE) for the data D :

$$M^* = \arg \min_M P(D|M).$$

Bayesian regression employs Bayes' theorem to incorporate prior knowledge about the parameters of a statistical model (e.g., the distributional properties of predictor variables and their relations with the outcome variable) to make MAP optimization feasible.

Besides being a more naturally desirable optimization goal than MLE, MAP optimization presents several advantages over frequentist approaches. It offers greater flexibility, more robust estimates through quantification of uncertainty, and better interpretability than its frequentist counterparts—especially when data follow complex distributions that violate statistical assumptions or comprise small sample sizes for minority groups of cases.

4.1. Applying Multilevel Models to Safety Annotation

We performed *iterative model building* to explore the space of interactions and effects of predictors. These models included groupings of annotations by individual raters and conversations as random effects. Here we report the main models that came out of this process. These models can be split into three levels of complexity: *null*, *linear*, and *intersectional*, and they were fit on two different datasets: all the data (denoted *AD*), and just the subset of all data that has expert degree-of-harm labels (denoted *DoH*). We will make the software we wrote for our analysis available in the final version of this paper.

The null model

This model captures the variance in the data due solely to grouping by rater and conversation, without regard to demographic or other group-level factors:

$$\text{AD, DoH null: } Q_{\text{overall}} \sim 1 + (1 \mid \text{rater_id}) + (1 \mid \text{conversation_id})$$

Linear models

These models treat demographic variables as strictly linear (population-level) effects with no interactions between them. These models show the

covariance of the demographic variables as independent, non-intersecting predictors compared to the null model.

$$\text{AD effects: } Q_{\text{overall}} \sim \text{race} + \text{gender} + \text{age} + \text{education} + \text{phase} + (1 \mid \text{rater_id}) + (1 \mid \text{conversation_id}),$$

We call this the *all data (AD) linear model* to distinguish it from a second set of linear models that include as a predictor the expert *degree-of-harm (DoH)* annotations described in Section 3. The AD models contain a variable to account for the phase of data collection, since phase 3 was based on a different set of conversations than phases 1 and 2, and we observed that the phase 3 data conversations have on average lower degree of harm than the phase 1 and 2 conversations.

The DoH models allow us to investigate more directly than the AD models how the severity of unsafe conversations could differentially impact annotations for different sociodemographic groups of raters. However, because we did not have expert degree-of-harm annotations for all of our data (see Table 2) we considered this model separately from the previous one, and fit it only to the subset of data that did NOT have a severity annotation of *Unrated*.

Note that there is no variable for locale (US or India). We did use this variable in earlier models not reported here. Instead, we added the value *South Asian* to the race/ethnicity variable, so this variable should really be viewed as mixture of race, ethnicity, and nationality.

$$\text{DoH effects: } Q_{\text{overall}} \sim \text{race} + \text{gender} + \text{age} + \text{education} + \text{severity} + (1 \mid \text{rater_id}) + (1 \mid \text{conversation_id}).$$

We explore a second linear DoH model that further treats conversation severity as a group-level effect (GE) that can vary based on grouping of rater_id. Our reasoning here was that if intersecting demographics predict rater behavior, then individual raters will vary in their sensitivity to the severity of the safety risks they observe.

$$\text{DoH effects GE: } Q_{\text{overall}} \sim \text{race} + \text{gender} + \text{age} + \text{education} + \text{severity} + (\text{severity} \mid \text{rater_id}) + (1 \mid \text{conversation_id}).$$

Intersectional models

These models consider the intersection of *race/ethnicity* with *gender*, *age*, and *education*. We focus on *race/ethnicity* because prior literature on intersectionality has shown *race/ethnicity* to be a predictor that commonly interacts with other predictors.

$$\text{AD intersectional: } Q_{\text{overall}} \sim \text{race} * (\text{gender} + \text{age} + \text{phase} + \text{education}) + (1 \mid \text{rater_id}) + (1 \mid \text{conversation_id}).$$

| Model | ELPD \uparrow | LOOIC \downarrow | WAIC \downarrow | Conditional $R^2 \uparrow$ | Marginal $R^2 \uparrow$ |
|-----------------------|-----------------|--------------------|-------------------|----------------------------|-------------------------|
| AD null | -56411.541 | 112800.000 | 112800.000 | 0.588 | 0.000 |
| AD effects | -47373.950 | 94747.900 | 94737.617 | 0.604 | 0.281 |
| AD intersectional | -47348.600 | 94697.200 | 94686.700 | 0.604 | 0.297 |
| DoH null | -35303.110 | 70606.219 | 70602.708 | 0.545 | 0.000 |
| DoH effects | -26553.539 | 53107.079 | 53103.061 | 0.550 | 0.273 |
| DoH effects GE | -26514.236 | 53028.472 | 53023.007 | 0.552 | 0.274 |
| DoH intersectional | -26547.566 | 53095.132 | 53090.776 | 0.552 | 0.291 |
| DoH intersectional GE | -26510.000 | 53019.990 | 53014.17 | 0.556 | 0.266 |

Table 3: Fitness of the various MLMs considered in this study. Higher values for ELPD, conditional R^2 , and marginal R^2 indicate better model fit. Lower values for LOOIC and WAIC indicate better model fit. *AD* stands for *All Data*. *DoH* stands for *degree-of-harm*, i.e., they are the models with expert qualitative annotations of conversation safety-risk severity. *RC* stands for *random covariates*. Conditional R^2 estimates variance in the model captured by the fixed and random effects. Marginal R^2 refers to the fixed effects of the model alone.

where the ‘*’ symbol denotes multiplication.

As with our linear models, we also consider a version of this with degree-of-harm annotations as a group-level effect.

4.2. Fitting the models

For our ordinal outcome, Q_{overall} , we set weakly informative probit threshold priors to reflect our prior knowledge that the values of *Safe*, *Unsafe* and *Unsure* are not equally likely. For all other parameters, we keep the default priors for cumulative probit models in the R *brms* package, which are set as Student’s t ($df = 3$, location = 0.00, scale = 2.5) distributions.

We fit a series of Bayesian ordinal MLMs (estimated using Markov chain Monte Carlo [MCMC] sampling with 4 chains of 2,000 iterations and a warm-up of 1,000) to quantify the individual and intersectional effects of race/ethnicity, gender, age, data collection phase, and education level on safety annotations (Section 3).

Following the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework (Makowski et al., 2019), for each estimate we report the median of its posterior distribution, 95% (Bayesian) credible interval, probability of direction, probability of practical significance (i.e., chance of being greater than 0.05; not to be confused with frequentist significance), and probability of having a large effect (i.e., at least 0.30). We assessed convergence and stability of Bayesian sampling with \hat{R} , which should be below 1.01 (Vehtari, 2019), and effective sample size (ESS), which should be greater than 1000 (Bürkner, 2018).

5. Results

To compare predictive fit, we compute the expected log pointwise predictive density (ELPD), leave-one-out cross-validation information criterion (LOOIC),

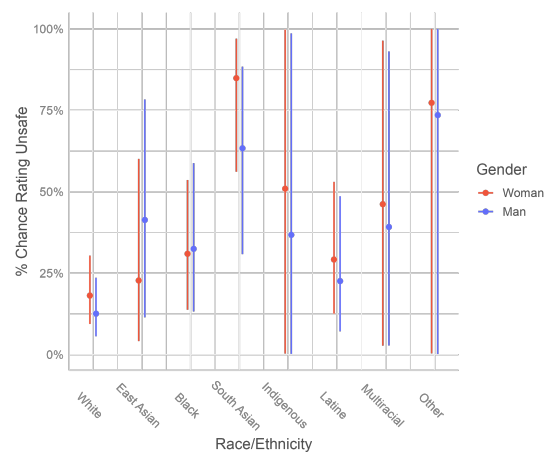


Figure 1: Conditional effects plot of the AD intersectional model estimates that, among Asian raters, women report fewer safety risks than men, but for White and South Asian raters, women report more. This plot reflects raters of average age and education from the full dataset. Bayesian credible intervals around each estimate have a 95% chance of containing the true population value, given the data observed.

and widely applicable information criterion (WAIC) for each model due to their advantages over simpler estimates of predictive error (Vehtari et al., 2017). Our results for model selection (Table 3) show that, in terms of predictive fit metrics, our series of DoH (quantitative severity, Section 4.1) models seem to outperform AD models (all data models, Section 4.1). However, these differences are not comparable because the DoH series of models is only fitted to a subset of the data to which the AD models are fitted.

Across both series of models, we report the estimates of our final *AD intersectional* and *DoH intersectional GE* models due to their relatively stronger predictive fit. ELPD, LOOIC, and WAIC all improve

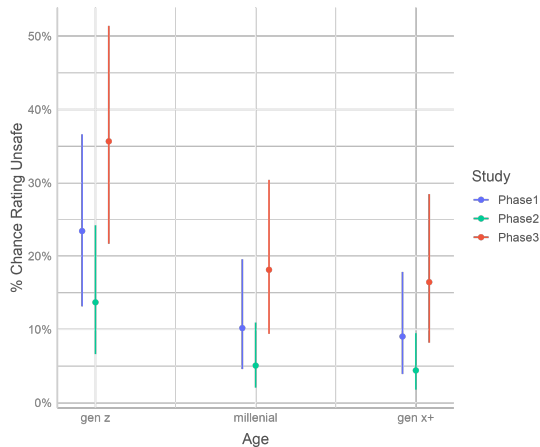


Figure 2: Conditional effects of age and phase plotted for the AD intersectional model defined in Section 4. Plot shows that annotations of unsafe decrease with age. Plot controls for rater gender, age, and education at their mode values.

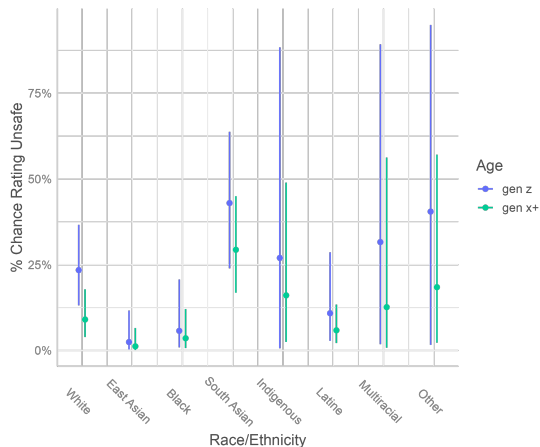


Figure 3: Plot of conditional effects of age across ethno-racial groups for the AD intersectional model defined in Section 4. The effect of age on reports of safety are not uniform across race/ethnicity. Millennial raters are omitted for clarity.

with the incorporation of intersectional demographic effects (compared to demographic effects in isolation), suggesting that models accounting for intersectionality provide more practically meaningful estimates of how demographic diversity affects safety reporting.

Table 4 shows the full results of the AD intersectional model. Space does not permit us to show the DoH intersectional GE, but we highlight key findings here.

Strong intersectional effects between race and gender Although the effect of race/ethnicity or gender’s effect on safety annotations is, independently, moderate, Figure 1 shows that race/ethnic-

ity intersects with gender for certain rater groups. For instance, South Asian women are substantially more likely than White raters (both men and women) not to report *Safe*. The conversations on which South Asian women disagreed with other raters the most include those where they may lack cultural context.

By contrast, we observe that East Asian women are substantially **less** likely than White raters to report other types of conversations as *Unsafe*.

Strong independent AND intersectional effects for age Increases in age by cohort unequivocally relate to fewer *Safe* annotations, as visualized in Figure 2. Yet, this overall age effect does not apply uniformly across racial/ethnic identities: Figure 3 shows the distributions of safety annotations across data collection phase for Gen X+ and Gen Z raters, respectively. Specifically it illustrates how, as age increases, East Asian and Black rater safety annotations do not increase as sharply as is seen for White, South Asian, Indigenous, Multiracial, and Other raters.

Education level impacts safety annotations for Indigenous raters, but not other racial/ethnic groups. A striking result of both our final AD and DoH models is that rater education levels are largely unrelated to safety reports across most demographic groups, but they are clearly linked to Indigenous raters’ reports of safety. Indigenous raters, compared to White raters, are 3.12 times more likely (95% Bayesian CI = [0.79, 15.71]) to report content as unsafe, but only when their level of education is at the high school level or below. Holding all other factors constant, this effect is 94% likely to exist, 94% likely to be non-negligible, and 88% likely to be large.

6. Discussion

Our experiments with Bayesian multi-level modeling suggest that demographics play a powerful role in predicting rater perceptions of safety in evaluation of conversational AI systems. Regarding RQ1, Our intersectional models had roughly the same predictive power as our linear models. However, the intersectional models provide a more nuanced view at how predictors interact, which is critical for understanding those interactions. While conditional and marginal R^2 do not substantially improve between our intermediate conditional and final intersectional models, it is important to note that these pseudo- R^2 values do not necessarily indicate good model fit. Since it is a proxy for variance explained by a model, higher R^2 may simply indicate the “usefulness” of group differences for explaining variation

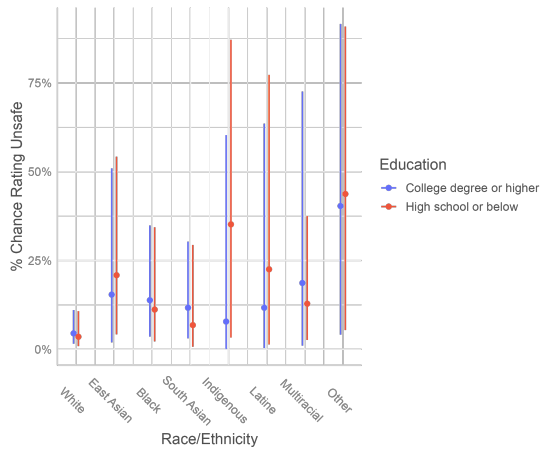


Figure 4: Conditional effects plot of the final DoH model shows that race/ethnicity and education intersect for Indigenous raters with a high school level or below of education, even when holding age and gender constant at "Millennial" and "Man."

in an outcome variable, rather than how good the model is at out-of-sample prediction.

Regarding RQ2, our results show *strong interactional effects involving race/ethnicity* that do not exist for race/ethnicity independently. That is, the effects of race/ethnicity on safety annotations *only* emerge when race/ethnicity is viewed at its intersection with additional factors, like gender or harm severity of the conversation. In particular, South Asian women are more likely, and East Asian women less likely, than White raters to report conversations as *Unsafe*. Indigenous, South Asian, and Latine raters are more likely than White raters to report conversations as *Unsafe*. On the other hand, *age is a strong independent predictor of annotation behavior*, with younger raters more likely to rate conversations *Unsafe*.

Regarding the advantages of MLMs, another approach, ANOVA, would dummy code any group variable, such as rater_id, that a given annotation is associated with, to test for differences in annotations between, e.g., raters. However, raters have their own group-level characteristics (e.g., gender, age) that could affect downstream annotations. Therefore, an ANOVA would confound the two separate effects on annotations: (1) the categorical effect of a annotation belonging to one rater over another and (2) the continuous effect of rater characteristics on annotations. Indeed, annotations under GenZ vs. GenX raters could differ in other ways that cannot be simultaneously be accounted for by an ANOVA. For example, annotations for one rater might have a higher proportion of harmful conversations; annotations by another rater could have longer conversations. In this instance, an ANOVA would not be able to separate the effects of group-

level predictors (conversation qualities) with the effects of the group dummies (the rater).

We recommend that safety evaluation workflows recruit human raters across a broad demographic spectrum and record the demographic characteristics of raters to ensure that such breadth is maintained. To boost the representational power of demographic diversity, large rater pools should be used, considering the benefits that such diversity provides in weighing costs. In cases where costs are prohibitive, decreasing the number of items each rater evaluates should be considered in favor of increased number of raters per item. Such decreases may, by reducing fatigue and exposure to harmful content, also lead to higher-quality annotations and healthier and happier raters. Finally, we recommend using statistical frameworks that account for the cross-classified structure of human annotation data (Sap et al., 2022; Kumar et al., 2021; Prabhakaran et al., 2023).

7. Limitations

Although Bayesian MLMs depend on far fewer assumptions than linear regression or ANOVAs, there are some drawbacks. MCMC sampling is a slow process; our largest models take days to run if not parallelized across multiple CPUs, and it is relatively common for the process not to converge. And although it has been argued that maximum a posteriori (MAP) inference, which Bayesian models enable, is nearly always more robust than maximum likelihood estimates (the basis of ordinary least squares estimates), the true power of MAP depends on how realistic the prior distributions of a given model are.

While our models predict a unique intercept for each rater_id and each conversation_id, the contribution from each rater and conversation pair is linear. We did not explore whether the relationship between them was more complex.

In this study, we only considered safety annotations as a single response (i.e. Q_overall) for each (conversation, rater) pair. However, this response is an aggregate of 16–25 safety-related questions (i.e., safety dimensions discussed in § 3). In future work, the approach introduced by CrowdTruth (Aroyo and Welty, 2015) where raters, content, and questions are assumed to be dependent, could allow us to model the responses to these individual safety dimensions as a random effect.

We only explored one conversational agent. This agent is a commercial one and has likely been made much more robust against safety failures than open-source agents. Future work will seek to validate our results are other agents. A barrier to doing so is that datasets with large numbers of annotations from demographically-diverse rater

| Row | Parameter | Median | 95-CI-Lower | 95-CI-Upper | Direction | Significance | Large | I |
|-----|-------------------------------|-----------|-------------|-------------|-----------|--------------|-------|----|
| 1 | Intercept1 | 1.11 | 0.8 | 1.43 | 1 | 1 | 1 | ** |
| 2 | Intercept2 | 1.36 | 1.05 | 1.69 | 1 | 1 | 1 | ** |
| 3 | Asian | -0.01 | -0.72 | 0.68 | 0.52 | 0.46 | 0.21 | |
| 4 | Black | -0.19 | -0.73 | 0.36 | 0.75 | 0.69 | 0.35 | |
| 5 | Indian | 0.23 | -0.21 | 0.67 | 0.84 | 0.78 | 0.38 | * |
| 6 | Indigenous | 0.36 | -0.49 | 1.24 | 0.81 | 0.77 | 0.56 | * |
| 7 | Latinxe | -0.07 | -0.59 | 0.45 | 0.6 | 0.53 | 0.19 | |
| 8 | Multiracial | 0.49 | -0.67 | 1.8 | 0.79 | 0.77 | 0.62 | |
| 9 | Other | 1.02 | -0.04 | 2.18 | 0.97 | 0.96 | 0.91 | ** |
| 10 | Nonbinary | -0.02 | -1.92 | 1.78 | 0.51 | 0.48 | 0.37 | * |
| 11 | SelfMdescribbelow | -0.73 | -2.52 | 1 | 0.81 | 0.8 | 0.7 | * |
| 12 | Woman | 0.2 | -0.17 | 0.59 | 0.86 | 0.79 | 0.32 | ** |
| 13 | age.L | -0.43 | -0.6 | -0.26 | 1 | 1 | 0.94 | ** |
| 14 | age.Q | 0.19 | -0.16 | 0.55 | 0.85 | 0.78 | 0.28 | ** |
| 15 | Phase2 | -0.37 | -0.5 | -0.23 | 1 | 1 | 0.83 | ** |
| 16 | Phase3 | 0.35 | 0.16 | 0.53 | 1 | 1 | 0.69 | ** |
| 17 | Higschoolbelow | 0.14 | -0.17 | 0.44 | 0.81 | 0.71 | 0.15 | * |
| 18 | Other | -0.37 | -0.99 | 0.23 | 0.89 | 0.86 | 0.6 | * |
| 19 | Asian:Nonbinary | -6.09E-03 | -3.2 | 3.22 | 0.5 | 0.48 | 0.4 | |
| 20 | Black:Nonbinary | 0.02 | -3.2 | 3.07 | 0.5 | 0.49 | 0.4 | |
| 21 | Indian:Nonbinary | 1.48E-03 | -3.12 | 3.24 | 0.5 | 0.48 | 0.39 | |
| 22 | Indigenous:Nonbinary | -0.03 | -1.89 | 1.9 | 0.51 | 0.49 | 0.37 | |
| 23 | Latinxe:Nonbinary | 2.63E-04 | -3.28 | 3.17 | 0.5 | 0.48 | 0.39 | |
| 24 | Multiracial:Nonbinary | 6.84E-03 | -3.16 | 3.31 | 0.5 | 0.48 | 0.39 | |
| 25 | Other:Nonbinary | -0.01 | -3.12 | 3.24 | 0.5 | 0.49 | 0.39 | |
| 26 | Asian:SelfMdescribbelow | 4.78E-03 | -3.22 | 3.1 | 0.5 | 0.48 | 0.4 | |
| 27 | Black:SelfMdescribbelow | 0.02 | -3.18 | 3.18 | 0.51 | 0.49 | 0.39 | |
| 28 | Indian:SelfMdescribbelow | 0.01 | -3.26 | 3.2 | 0.5 | 0.49 | 0.4 | |
| 29 | Indigenous:SelfMdescribbelow | -8.76E-03 | -3.19 | 3.28 | 0.5 | 0.48 | 0.4 | |
| 30 | Latinxe:SelfMdescribbelow | -0.73 | -2.5 | 1.04 | 0.81 | 0.8 | 0.7 | * |
| 31 | Multiracial:SelfMdescribbelow | 5.12E-03 | -3.24 | 3.29 | 0.5 | 0.48 | 0.39 | |
| 32 | Other:SelfMdescribbelow | -0.03 | -3.03 | 2.99 | 0.51 | 0.49 | 0.4 | |
| 33 | Asian:Woman | -0.78 | -1.46 | -0.13 | 0.99 | 0.99 | 0.92 | ** |
| 34 | Black:Woman | -0.24 | -0.95 | 0.45 | 0.75 | 0.71 | 0.44 | ** |
| 35 | Indian:Woman | 0.5 | -0.07 | 1.08 | 0.96 | 0.94 | 0.76 | ** |
| 36 | Indigenous:Woman | 0.05 | -1.12 | 1.23 | 0.53 | 0.5 | 0.33 | |
| 37 | Latinxe:Woman | -0.1 | -0.72 | 0.54 | 0.62 | 0.56 | 0.26 | |
| 38 | Multiracial:Woman | -0.02 | -1.01 | 0.99 | 0.51 | 0.47 | 0.28 | |
| 39 | Other:Woman | -0.15 | -1.32 | 0.99 | 0.61 | 0.57 | 0.39 | ** |
| 40 | Asian:age.L | 0.24 | -0.02 | 0.49 | 0.97 | 0.93 | 0.31 | * |
| 41 | Black:age.L | 0.26 | -0.31 | 0.84 | 0.81 | 0.76 | 0.45 | * |
| 42 | Indian:age.L | 0.18 | -0.2 | 0.57 | 0.83 | 0.75 | 0.28 | * |
| 43 | Indigenous:age.L | 0.38 | -0.63 | 1.48 | 0.77 | 0.74 | 0.56 | * |
| 44 | Latinxe:age.L | 0.29 | -0.2 | 0.81 | 0.87 | 0.83 | 0.49 | * |
| 45 | Multiracial:age.L | -0.14 | -1.14 | 0.85 | 0.6 | 0.57 | 0.37 | |
| 46 | Other:age.L | -8.30E-04 | -1.13 | 1.15 | 0.5 | 0.47 | 0.3 | |
| 47 | Asian:age.Q | -0.45 | -1.23 | 0.3 | 0.89 | 0.86 | 0.65 | * |
| 48 | Black:age.Q | -0.44 | -1.02 | 0.12 | 0.93 | 0.91 | 0.69 | ** |
| 49 | Indian:age.Q | -0.06 | -0.68 | 0.57 | 0.57 | 0.51 | 0.22 | * |
| 50 | Indigenous:age.Q | -0.63 | -2.04 | 0.59 | 0.84 | 0.82 | 0.7 | * |
| 51 | Latinxe:age.Q | -0.45 | -1.03 | 0.12 | 0.94 | 0.91 | 0.7 | ** |
| 52 | Multiracial:age.Q | -0.51 | -1.46 | 0.39 | 0.86 | 0.84 | 0.67 | * |
| 53 | Other:age.Q | -1.15 | -2.37 | -0.07 | 0.98 | 0.98 | 0.94 | ** |
| 54 | Asian:Phase2 | 0.78 | 0.12 | 1.48 | 0.99 | 0.99 | 0.93 | ** |
| 55 | Black:Phase2 | 0.72 | 0.4 | 1.04 | 1 | 1 | 0.99 | ** |
| 56 | Indian:Phase2 | -1.53E-03 | -3.14 | 3.33 | 0.5 | 0.48 | 0.39 | ** |
| 57 | Indigenous:Phase2 | 1.03 | -0.41 | 2.76 | 0.92 | 0.9 | 0.83 | ** |
| 58 | Latinxe:Phase2 | 0.58 | 0.31 | 0.86 | 1 | 1 | 0.98 | ** |
| 59 | Multiracial:Phase2 | -4.30E-04 | -3.33 | 3.19 | 0.5 | 0.48 | 0.39 | ** |
| 60 | Other:Phase2 | -0.83 | -2.06 | 0.28 | 0.93 | 0.91 | 0.82 | ** |
| 61 | Asian:Phase3 | 0.61 | -0.01 | 1.28 | 0.97 | 0.96 | 0.84 | ** |
| 62 | Black:Phase3 | 0.53 | 0.26 | 0.78 | 1 | 1 | 0.96 | ** |
| 63 | Indian:Phase3 | 1.18 | 0.62 | 1.74 | 1 | 1 | 1 | ** |
| 64 | Indigenous:Phase3 | 0.85 | -0.39 | 2.28 | 0.91 | 0.9 | 0.8 | ** |
| 65 | Latinxe:Phase3 | 0.38 | 0.1 | 0.66 | 1 | 0.99 | 0.71 | ** |
| 66 | Multiracial:Phase3 | -0.21 | -1.56 | 1.01 | 0.63 | 0.6 | 0.45 | ** |
| 67 | Other:Phase3 | -0.02 | -3.17 | 3.12 | 0.51 | 0.49 | 0.4 | ** |

Table 4: Results for the AD intersectional MLM $Q_overall \sim race * (gender + age + phase) + education + (1 | rater_id) + (1 | conversation_id)$

pools are still quite rare and expensive to obtain. Our position is that such datasets should be the rule, not the exception, but unless the field as a whole adopts this position, such datasets will likely remain rare.

We made some hard choices in forming our demographic categories, particularly race/ethnicity/nationality. Our challenge was to create categories that had as much statistical power as possible, based on the demographic information that was collected. The South Asian category includes 5 US and 92 Indian raters. Our *Indigenous* race/ethnicity category lumps together very diverse Indigenous identities in a manner that likely discounts rich idiographic differences in language, culture, and lived experience (Else-Quest and Hyde, 2016). However, in the interest of protecting participants privacy and prioritizing the representation of Indigenous perspectives in this empirical research, we

chose to group them together. Creating the *Indigenous* category in our analysis balances these opposing concerns, but leaves significant room for future study.

8. Conclusion

We apply Bayesian multilevel models (MLMs) to a dataset of 1,340 chatbot conversations, each annotated for safety by 60–104 human raters, to study the impact of rater demographics on rater behavior for safety annotations. MLMs allow us to deal with the overlapping hierarchical dependencies on rater and conversation that are inherent in rater data, and which confound simpler modeling approaches, such as ordinary least squares regression and ANOVA.

Our results show strong intersectional effects between race/ethnicity and gender, Indigenous raters

and education, and content severity and race. They suggest that conversational AI safety evaluation can benefit when human evaluators come from diverse demographic backgrounds.

9. Ethical considerations

The very act of rating harmful language can itself be harmful, and risks exposing raters to trauma. From a social justice perspective, such risks should be born equitably by all raters, regardless of their demographic characteristics.

Such concerns must be balanced against the potential benefit of research such as ours to uncover AI safety risks that may only be detectable by vulnerable groups. For instance, “dog-whistling,” the practice of encoding racist language in seemingly innocuous terms (Mendelsohn et al., 2023), can result in language that may seem completely safe to some raters but not others. It can be impossible to detect such language without annotators who are experienced in parsing it.

10. Bibliographical References

- Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators’ demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190.
- Dana Angluin and Philip Laird. 1988. Learning from noisy examples. *Machine learning*, 2:343–370.
- Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. [Dices dataset: Diversity in conversational ai evaluation for safety](#).
- 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.
- Valerio Basile. 2020. It’s the end of the gold standard as we know it. On the impact of pre-aggregation on the evaluation of highly subjective tasks. *CEUR Workshop*.
- Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. *Social Informatics*.
- Som Biswas. 2023. Chatgpt and the future of medical writing.
- Paul-Christian Bürkner. 2018. [Advanced Bayesian multilevel modeling with the r package brms](#).
- Andrea Campagner, Davide Ciucci, Carl-Magnus Svensson, Marc Thilo Figge, and Federico Cabitza. 2021. Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, 545:771–790.
- John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *CSCW*, pages 1–25.
- Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, page 139.
- Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.
- Kaylee A DeFelice and James W Diller. 2019. Intersectional feminism and behavior analysis. *Behavior Analysis in Practice*, 12:831–838.
- Juan Del Toro and Hirokazu Yoshikawa. 2016. Invited reflection: Intersectionality in quantitative and qualitative research. *Psychology of Women Quarterly*, 40(3):347–350.
- Nicole M Else-Quest and Janet Shibley Hyde. 2016. Intersectionality in quantitative psychological research: I. theoretical and epistemological issues. *Psychology of Women Quarterly*, 40(2):155–170.
- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. [Large scale crowdsourcing and characterization of twitter abusive behavior](#).
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.
- Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
- 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.

- Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *CSCW*.
- Katie Kilkenny and Winston Cho. 2023. [Attack of the chatbots: Screenwriters’ friend or foe?](#)
- Manfred Klenner, Anne Göhring, and Michael Amsler. 2020. Harmonization sometimes harms. *CEUR Workshops Proc.*
- Deepak Kumar, Patrick Gage Kelley, Sunny Con-solvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *SOUPS@ USENIX Security Symposium*, pages 299–318.
- Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. In *HCOMP*.
- Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. [Indices of effect existence and significance in the bayesian framework.](#) *Frontiers in Psychology*, 10.
- Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. [From dogwhistles to bullhorns: Unveiling coded rhetoric with language models.](#)
- Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems*, 26.
- Gina Neff. 2016. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*.
- Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.
- Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *ACL*.
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021a. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138.
- Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. [A framework to assess \(dis\)agreement among diverse rater groups.](#)
- Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021b. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW)*.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.](#)
- Rion Snow, Brendan O’connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.
- Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. [An analysis of the automatic bug fixing performance of chatgpt.](#)
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications.](#)
- Aki Vehtari. 2019. [Cross-validation for hierarchical models.](#)
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27:1413–1432.
- Tharindu Cyril Weerasooriya, Tong Liu, and Christopher M. Homan. 2020. Neighborhood-based Pooling for Population-level Label Distribution Learning. In *ECAI*.

Ben Wodecki. 2023. That was fast: Stanford yanks
alpaca demo for hallucinating.

A Dataset for Multi-Scale Film Rating Inference from Reviews

Frankie Robertson, Stefano Leone

University of Jyväskylä, Independent
frankie.r.robertson@jyu.fi, stefano.leone.992@gmail.com

Abstract

This resource paper introduces a dataset for multi-scale rating inference of film review scores based upon review summaries. The dataset and task are unique in pairing a text regression problem with ratings given on multiple scales, e.g. the A-F letter scale and the 4-point star scale. It retains entity identifiers such as film and reviewer names. The paper describes the construction of the dataset before exploring potential baseline architectures for the task, and evaluating their performance. Baselines based on classifier-per-scale, affine-per-scale, and ordinal regression models are presented and evaluated with the BERT-base backbone. Additional experiments are used to ground a discussion of the different architectures' merits and drawbacks with regards to explainability and model interpretation.

Keywords: ordinal regression, text regression, rating inference task, explainability, opinion mining, sentiment analysis

1. Introduction

This paper introduces the MS-RottenTomatoes dataset, which consists of enriched data scraped from the Rotten Tomatoes film review aggregator. Critics submit their reviews to Rotten Tomatoes, who aggregate them and then display them on their website, grouped by film, along with their own summarising rating. As an aggregator, Rotten Tomatoes includes reviews from different film critics across different publications, such as online magazines and film critics' own websites. Critics can rate films on a variety of different scales such as letter (A-F) or star scales (e.g. 4 or 5 point). The main aim of the publication of this dataset is as a benchmark for multi-scale text regression on what Pang and Lee (2005) called the *rating inference problem*, that is predicting a rating given by a reviewer from the associated review text.

Prediction of item ratings based on review text has come up previously in the NLP literature, where it has historically been framed as a proxy for the sentiment analysis task with high ratings correlated with positive sentiment. As Pang and Lee (2005) note, film ratings can actually provide additional information to a film review, by providing an overall impression contrary to an otherwise negative or positive seeming review. Thus, it is likely that in some cases, the connection between predictor and outcome variable may be rather poor, making the task somewhat noisy. Additionally, different critics may interpret the same scale differently, for example being more or less generous with awarding the top scale point. The ratings themselves are ordinal data: ordered like nominal data; but with a finite number of outcomes like multi-class data. To the best of our knowledge, MS-RottenTomatoes is the first openly published

dataset with sufficient detail for rating scales themselves to be modelled as part of the rating inference task.

The dataset presents a number of challenges. In many cases the review summaries simply do not contain enough information from which to predict the grade, creating a heteroscedastic situation when regressing the review scores based upon review text, in which more vague reviews have a much wider range of possible grades versus more concrete reviews. In addition, some reviewers use ratings to summarise their impressions of the films rather than their reviews. This creates a gap between their review and their rating, resulting in data points which provide random errors, or noise, to the network during training.

In this dataset, critics give their perspective upon films using both a rating and by giving their opinion in the form of text, in doing so, they implicitly express another perspective on how they perceive the rating scale. The presence of multiple rating scales gives rise to a desire to induce a single latent scale, so as to pool common attributes across critics, and in order to work well with gradient-based explainability techniques. On the other hand, since critics have different perspectives upon the rating scales, each (*critic, rating scale*) pair should be modelled individually. Finally, as ordinal data, using specialised techniques such as ordinal regression seems natural given metric methods implicitly assume data lies on an interval scale, where scale points are equal distances from one another. Liddell and Kruschke (2018) give a detailed account of the downsides of using metric methods with ordinal data, and indeed the analysis of Section 6.1 shows that their usage also introduces systematic errors in this setting.

The rest of this paper begins with some back-

ground, before describing the construction of the dataset and presenting baseline architectures and evaluating them. The paper then moves onto an analysis of the baseline results and discusses the degree to which different systems lend themselves to interpretability techniques. The paper closes by reviewing some related work and discussing possibilities for future work.

2. Background

Deep ordinal regression techniques can be seen as either modifying classification objectives, e.g. as in [Castagnos et al. \(2022\)](#), or regression objectives, considered further here. [Cao et al. \(2020\)](#) used a ResNet-34 backbone together with an ordinal regression head for age prediction from photos. Their technique, referred to as CORAL, induces a single latent scale which is used to predict an ordinal output Y , by modelling for each label threshold k , $P(Y \geq k + 1)$. It is in this sense closely related to the backward cumulative probability family from the EL-MO class of models presented by [Wurm et al. \(2021\)](#), where an Element Link (EL) such as the logit most familiar from machine learning is combined with a Multinomial-Ordinal (MO) family function which reduces the ordinal outcome to a number of boolean outcomes.

Gradient-based explainability techniques such as integrated gradients ([Sundararajan et al., 2017](#)) can attribute perturbations in a deep neural network’s outputs to specific areas of the input. Latent variable models give a natural choice of output perturbation to answer questions of interest about the input, e.g. *Which parts of this film review are more associated with the lower and higher ends of the rating scale?*

3. Dataset creation

This section describes how the Rotten Tomatoes data was scraped and enriched into three derived datasets, summarised in Table 1.

3.1. Scraping

The film ratings were scraped from Rotten Tomatoes using a scraper written in Python using the `requests` library. The process was split into the following steps:

¹<https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset>

²https://huggingface.co/datasets/frankier/processed_multiscale_rt_critics

³https://huggingface.co/datasets/frankier/multiscale_rt_critics_subsets

Table 1: Summary of the released datasets

| | |
|----------------|-----------------------------------|
| Name | RT-critics |
| Repository | Kaggle ¹ |
| # Movies | 17 712 |
| # Critics | 11 109 |
| # Ratings | 1 130 017 |
| Name | RT-normalized |
| Repository | Huggingface Hub ² |
| # Movies | 17 619 |
| # Critics | 6 832 |
| # Ratings | 617 819 |
| Task | Text regression |
| Split | Random |
| Train/Val/Test | 60% / 20% / 20% |
| Name | RT-critics 500 |
| Repository | Huggingface Hub ³ |
| # Movies | 16 251 |
| # Critics | 266 |
| # Ratings | 315 802 |
| Task | Multi-scale rating inference |
| Split | Stratified (critic, rating scale) |
| Train/Val/Test | 60% / 20% / 20% |

1. Collect film URLs from the film index;
2. Scrape film information from each film;
3. Scrape the critic reviews section of each film.

After scraping, films without at least one critic review, and critic reviews without film information are dropped. At this point, we have the non-task specific RT-critics dataset from Table 1.

3.2. Normalisation

In order to normalise the dataset for usage in multi-scale prediction tasks, the type of scale being used in each rating needs to be determined. Following this, ratings can be converted into a whole number numerical, ordinal scale, ready to be treated as either a classification or ordinal regression task.

The scales are first divided into either letter scales or number scales, which include e.g. star ratings as well as percentage and out-of-10 scales. Within the letter scales there are short and long scales, with short scales ranging from F up to A and long scales, which include plus and minus grades, ranging from F- up to A+.

Number scales are broken down by three factors: the maximum score; whether the scale includes 0; and by granularity. A scale’s granularity is whether it contains only whole numbers, or whether fractional ratings such as 0.5 or 0.25 are included, and if so, what is the minimum division between them. Since Rotten Tomatoes allows free text entry by reviewers submitting rat-

ings, it is possible to have ratings not in the above categories. These are either misentered data or extremely rare, unusual forms of ratings and are dropped altogether.

Grades are matched against either the list of letter grades or matched as a fraction: two numbers separated by '/'. After this, grades are considered to have a preliminary *grade type*.

For numerical grades, the granularity is found by considering the Least Common Multiple (LCM). This is done group-wise, keyed on the (*publisher, grade type*) pair. Next, the grade is divided by the granularity to obtain a normalised integer grade. A single non-conforming grade will change the LCM. Rather than rounding grades, rare grades, defined here as those with less than 50 reviews across the whole dataset, together with 8 manually chosen entries with unusual grade values are dropped. These grades include for example, 2.4/5 on a grading scale which otherwise has 0.5 as the granularity. Since these do not fit with their grading scale, they may be the result of a typo, and including them would change the LCM, affecting all other entries in the same group. This cleaned dataset is released as RT-normalized (see Table 1).

3.3. Schema and tasks

The columns of the dataset can be broken down into: *Entity identifiers useful for linking within or beyond the dataset*: the movie’s title, the review publisher’s name, and the critic’s name; *Textual content* consisting of the review text itself; *Numerical data* including the review score as originally presented, along with a normalised integer 0-based label, and an accompanying number of scale points thought to exist within the scale; and finally *information about how the grade was normalised*, such as the detected grade scale granularity, which can be used to convert back and forth between the normalised and unnormalised form of the grade.

This data is rich enough to be viewed from a number of perspectives, which are summarised in Table 2. Namely, the fact that the authors are named means this dataset could be used for author identification. Additionally, the ability to form a matrix of films and critics means that item-response theory⁴ could be applied in order to analyse aspects of the critics and films themselves, e.g. film quality and critic fussiness estimated on a common scale. However, it is the multi-scale rating inference problem, where film ratings are regressed based upon the review text that is considered for

⁴Item-response theory is used here as an umbrella term for methods utilising latent variable models which estimate parameters for some general type of “items” and “respondents” on a common scale based on a cross tabulated response matrix.

the rest of this paper.

3.4. Splits and subsets

We further process the dataset for the multi-scale rating inference task. In order to model the behaviour of each critic on each scale, we consider each group keyed by a (*critic, rating scale*) pair as a task within a multi-task learning setup. We drop all groups with less than 500 items. This is done so as to create a dataset where each task contains a sufficient number of samples to model it independently. It has the additional benefit of reducing the total dataset size, leading to a more energy efficient dataset to use for benchmarking.

Next, we create training, test and validation splits using stratified sampling grouped by task. The rest of this paper considers only this RT-critics 500 dataset (see Table 1).

3.5. Analysis

Different critics treat each rating scale differently. Figure 1 shows the marginal distributions of four different critics across an out-of-5 scale and a long letter scale. The critic in the bottom left panel appears to avoid fractional grades, while the critic in the top left panel uses them in proportion with the other grades, but is more cautious about giving 5/5. On the right, both critics avoid giving C+ grades, but the top critic gives plus and minus grades at the top end of the scale in proportion to bare letter grades, whereas the bottom critic gives relatively less plus and minus grades. These marginal distributions show different behaviours, and suggest that critics should be at least partially modelled independently. In addition, they fairly clearly show the ordinal nature of the data. Different grade points are not equally spaced, but rather it is a matter of determining some reasonable thresholds on a latent scale.

4. Systems

In order to demonstrate the dataset, some baselines are presented here⁵. In particular, we look at the performance of fine tuning foundational language models on the dataset. In all cases, the idea is to train a single backbone model for all (*critic, rating scale*) pairs. The baseline systems consist of fairly typical ways of approaching the problem of predicting film ratings based upon their text.

A block-diagram level overview of the different baselines is given in Figure 2. The backbone used in all experiments is BERT-base (Devlin et al.,

⁵Code to reproduce the baselines and experiments is made available at https://github.com/frankier/ms_text_regress.

Table 2: A list of possible tasks relating to the MS-RottenTomatoes dataset

| Task | Target | Regressors | Techniques |
|-----------------------|--------------------------------------|------------------------------|-----------------------------|
| Author identification | Critic | Review text | Text classification |
| Film/critic analytics | Film quality and Critic "difficulty" | Film \times Critic matrix | Item-response theory |
| Rating inference | Film rating | Entries: Review text, Rating | Multi-scale text regression |
| Opinion generation | Review text | Rating, Film, Critic | Text generation |

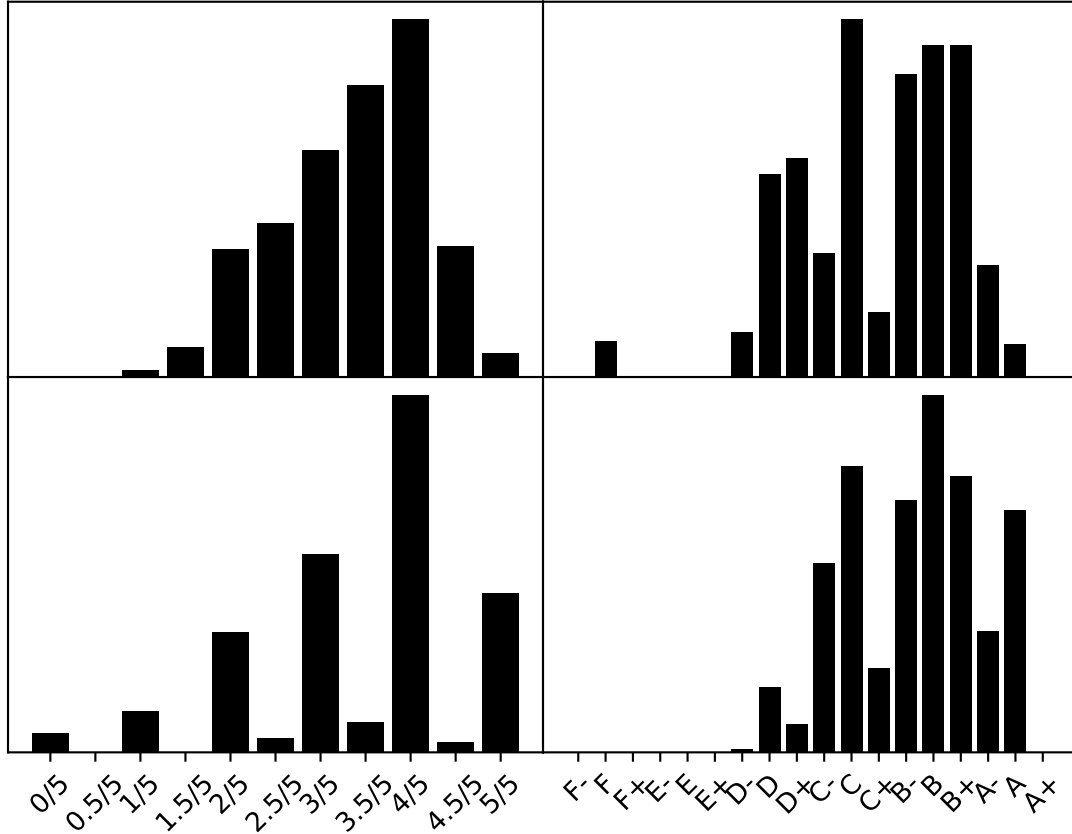


Figure 1: Plots of label distributions for different critics, illustrating varying behaviours of critics across an out-of-5 scale with half-star granularity (left) and a long letter scale (right)

2019). Although this backbone is no longer state-of-the-art, the main aim of these experiments is to establish baselines for this task. Since this dataset is derived from a publicly available website, it is of particular importance that this model has only been trained on Wikipedia and BookCorpus, and not on large scale web text, which would run the risk that the backbone has been exposed to the MS-RottenTomatoes test set during pretraining.

4.1. Classifier-Per-Scale

The Classifier-Per-Scale (CPS) system consists of a single linear predictor followed by another linear layer per (*critic, rating scale*) pair. Following softmax, the multinomial vector can be aggregated to get a single prediction. In preliminary experi-

ments, the mode provided the best accuracy, while the median gave the lowest MAE, so both are presented in Table 3, as CPS_{mode} and CPS_{median} , respectively.

4.2. Affine-Per-Scale

Affine-Per-Scale (APS) learns an affine transformation of a common linear scale per (*critic, rating scale*) grouping. In order to initialise the latent scale and heads, 8 pilot batches of the training set are run, and the pre-latent linear initialised so its output has a mean of 0 and a standard deviation of 1. The heads are then initialised based on the mean and standard deviation of their critics' labels on their rating scale according to the training dataset. So a fair comparison can be made

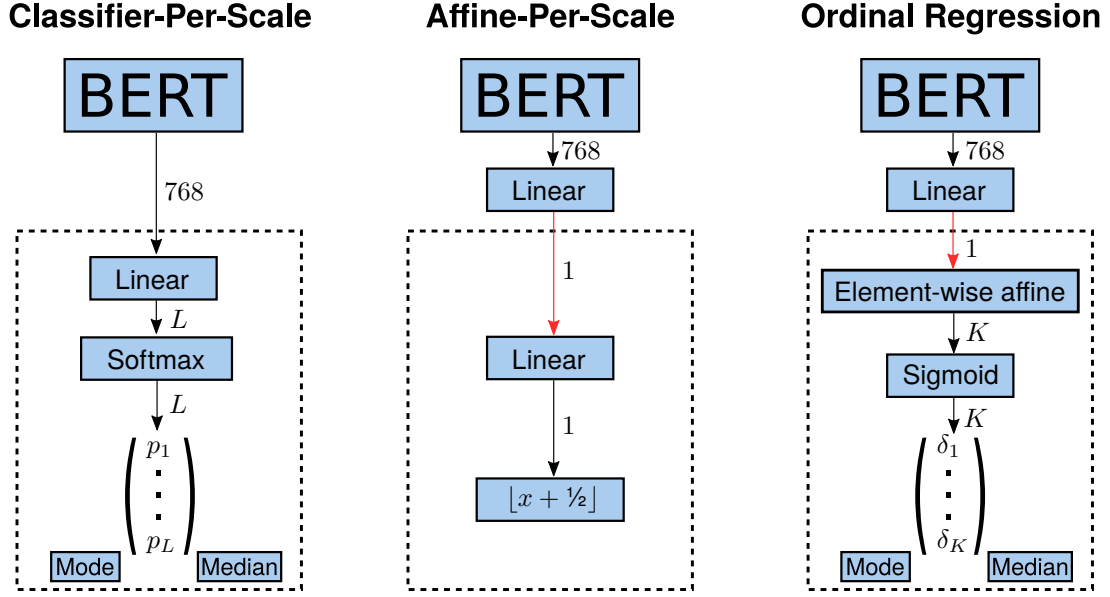


Figure 2: Schematic diagrams of the architectures of the baseline systems considered here. Dotted areas are referred to as *heads* and are repeated for each scale. Red arrows indicate the shared latent scale. L is the number of labels while $K = L - 1$ is the number of class boundaries. $\delta_k = P(Y \geq k + 1)$

between regression and classification approaches, the output is rounded to the nearest integer with the function:

$$\text{round}(x) = \lfloor x + \frac{1}{2} \rfloor$$

4.3. Ordinal regression

Following Cao et al. (2020), we apply an affine transformation composed with a sigmoid transformation element-wise to get a vector modelling $P(Y \geq k + 1)$. This vector can be processed to obtain a multinomial vector and the mode or median applied to obtain a single prediction, denoted as $\text{Ord}_{\cdot\text{mode}}$ and $\text{Ord}_{\cdot\text{median}}$, respectively, in Table 3.

5. Results

The main metric used here is a multi-scale variant of Mean Absolute Error (MAE), normalised to the range of the relevant scale:

$$\text{MAE}_{\text{MS}} = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{\text{scale-max}(i) - \text{scale-min}(i)}$$

This metric attempts to give even weight to errors from different scales. It is invariant to the scale transformation of the normalisation procedure of Section 3.2. MAE is the chosen base metric here since it has a straightforward interpretation and lends itself easily to this multi-scale adaption.

Since the objective for the latent models is to learn a good latent scale, evaluation metrics are

always taken after refitting the heads. This consists of first performing a full evaluation-mode pass over the training set, followed by refitting the heads task-at-a-time using convex optimisation procedures. The Affine-Per-Scale system is fitted using ordinary least squares as implemented in `scikit-learn` (Pedregosa et al., 2011) and the ordinal regression system is fitted using the `VGAM` (Yee, 2010) R package, which uses iteratively reweighted least squares, wrapped using `rpy2`. In both cases, no regularisation is used. No refitting is performed for the classification model.

All experiments were run using HuggingFace Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). The hyperparameters, which are common to all baselines are summarised in Table 4.

All experiments used early stopping with a patience of 3 validation cycles, with a validation cycle run every 1000 steps.

The results in Table 3 show that the ordinal regression system performs quite poorly. It exhibited noisy validation metrics and loss curves during training. It was the only system for which early stopping was applied, after 4000/17000 steps. We speculate that a specialised training procedure may be needed to fit these kinds of models in this setting.

Table 3: Evaluation results for the three baseline architectures, alongside a Most Frequent Class (MFC) baseline. Lower MAE is better. Higher acc. = accuracy is better. MAE_{MS} and accuracy are given in %.

| | Validation | | | | | |
|------------------------|------------|-------------|-------------|------------|-------------|-------------|
| | Micro | | | Macro | | |
| | MAE_{MS} | MAE | Acc. | MAE_{MS} | MAE | Acc. |
| MFC | 16.5 | 2.11 | 27.3 | 16.7 | 2.16 | 27.5 |
| CPS _{median} | 10.5 | 1.38 | 37.4 | 11.2 | 1.50 | 35.7 |
| CPS _{mode} | 10.7 | 1.41 | 39.3 | 11.4 | 1.55 | 37.9 |
| APS | 9.6 | 1.23 | 34.6 | 9.7 | 1.28 | 34.4 |
| Ord. _{median} | 18.7 | 2.36 | 16.3 | 19.1 | 2.44 | 16.1 |
| Ord. _{mode} | 30.0 | 3.99 | 7.8 | 30.6 | 4.15 | 7.9 |

| | Test | | | | | |
|------------------------|------------|-------------|-------------|------------|-------------|-------------|
| | Micro | | | Macro | | |
| | MAE_{MS} | MAE | Acc. | MAE_{MS} | MAE | Acc. |
| MFC | 16.5 | 2.12 | 27.1 | 16.6 | 2.16 | 27.5 |
| CPS _{median} | 10.5 | 1.39 | 37.3 | 11.2 | 1.50 | 35.8 |
| CPS _{mode} | 10.7 | 1.42 | 39.2 | 11.3 | 1.54 | 38.0 |
| APS | 9.6 | 1.25 | 34.6 | 9.8 | 1.28 | 34.5 |
| Ord. _{median} | 17.8 | 2.27 | 17.2 | 18.1 | 2.34 | 17.0 |
| Ord. _{mode} | 30.2 | 4.05 | 7.9 | 30.8 | 4.22 | 7.9 |

Table 4: Hyperparameters used for the baselines

| | |
|-------------------------|---------------------------------|
| Backbone | BERT-base |
| Optimisation routine | AdamW |
| Batch size | 32 |
| Learning rate | $1e-5$ |
| Schedule | Linear |
| Warmup | 10% |
| Training time | 17 000 steps (= 2.87 epochs) |
| Validation metric | MAE_{MS} |
| Validation | Every 1000 steps |
| Early stopping patience | 3 validations |

Table 5: Counts and proportions of refit APS heads with significant x^2 parameters according to different p-values

| p-value | Correction | count | % |
|---------|------------|-------|----|
| 0.05 | None | 101 | 38 |
| 0.01 | None | 77 | 29 |
| 0.05 | Bonferroni | 33 | 12 |
| 0.01 | Bonferroni | 24 | 9 |

6. Experiments

This section demonstrates the need for alternatives to APS, before motivating further work in the direction of latent variable models by demonstrating their potential with regards to explainability and model interpretation using the APS model.

6.1. Model fit

One of the aims of releasing this dataset is to help spur interest in combining non-standard regression methods with NLP. In order to show the inadequacy of treating ratings as real numbers, we diagnose the model fit of the heads of the APS final model by running a pass over the training data and refitting a linear model of the form $x^2 + x + c$ where x is the latent scale. The fit was performed

with `statsmodels` (Seabold and Perktold, 2010) with the default settings: nonrobust regression and two-tailed parameter significance testing based on the Student’s t-distribution. The results in Table 5 give strong evidence that at least 9% of the heads are not fitted well by a linear relationship.

This poor fit of the heads suggests the outcomes modelled by some heads do not have a linear relationship with the latent variable, while others perhaps do. This poor fit will result in the backbone systematically receiving poor gradients during training.

6.2. Gradient-based attribution

Gradient-based attribution methods are an explainable machine learning technique which work by looking at how gradients at different network inputs change according to perturbations in the outputs of the network. For each review in the validation set, we use layer integrated gradi-

[CLS] This whodunit_{NEG} had_{POS} many_{NEG} contrivances_{NEG} and they all_{NEG} sapped_{NEG} strength_{POS} from the film_{NEG} s_{NEG} already_{POS} rather_{NEG} weak_{NEU} mystery_{POS} story_{NEG} . [SEP]

[CLS] Low_{NEG} - budget_{NEG} earnest_{NEG} but_{NEG} bland_{NEU} attempt_{NEG} at_{NEG} a_{NEG} horror_{POS} / suspense_{NEG} thriller_{NEG} that just_{POS} can_{NEG} t_{NEG} overcome_{NEG} a_{NEG} stale_{NEG} , predictable_{NEG} and unimaginative_{NEU} plot_{NEG} . [SEP]

[CLS] It_{NEG} s_{NEG} a_{NEG} masterpiece_{POS} crime_{NEG} story_{NEG} that tells_{NEG} us_{NEG} as much_{NEG} about_{NEG} searching_{NEG} for the truth_{POS} in_{NEG} modern_{NEG} Turkey_{NEG} as_{NEG} it_{NEG} does_{NEG} about_{NEG} violent_{POS} criminals_{NEU} and those who_{NEG} prosecute_{NEG} them . [SEP]

Figure 3: The results of integrated gradients as applied to three example review texts. Warmer colours are associated with increases in the rating, while colder colours are associated with decreases in the rating. Words found in SentiWordNet are annotated with their corresponding classes.

ents (Sundararajan et al., 2017) as implemented in Captum (Kokhlikyan et al., 2020) with 50 steps to find subtokens associated with higher and lower ratings. For each token, we reconstruct the word it is part of and look up the first sense in SentiWordNet (Esuli and Sebastiani, 2006), classifying it as positive if the positive score is greater than the negative score, neutral if they are the same and negative otherwise. Both sources of information are overlaid onto example review summaries in Figure 3, while the resulting cross-tabulation table is shown in Table 6.

We can now calculate two association measures: mutual information and the G statistic. We calculate these between the most negatively and most positively associated token according to integrated gradients and their classification according to the negative/positive classes of SentiWordNet. The resulting mutual information calculated using the plug-in estimator gives a moderate value of 0.21 bits. Calculating the G-statistic results in a value of 16796 (rounded) which results in a P-value for association equal to 0 when calculated using double floating point precision, i.e. there is close to zero chance we would see these results without an association between the two.

6.3. Interpretation of model fit

Given a text regression model, we can apply linear modelling diagnostics to each head’s linear model on the training set to answer questions about the

Table 6: Cross tabulation of words with the strongest high (hi) and low (lo) token rating attribution against negative, positive, neutral, or unknown sentiment in SentiWordNet. The top half gives raw counts and the bottom half gives % of reviews in the validation set.

| | | SentiWordNet | | | |
|------|----|--------------|-------|-------|------|
| | | Neg. | Pos. | Neu. | Unk. |
| Cnt. | Lo | 19541 | 6978 | 2964 | 6998 |
| | Hi | 6485 | 24441 | 28514 | 3717 |
| % | Lo | 31 | 11 | 47 | 11 |
| | Hi | 10 | 39 | 45 | 6 |

level to which the heads are able to model the relation between the true rating and the latent variable derived from the deep learning prediction model. Critics with poor fit may be “film impression summarisers” who use a rating to give an overall impression of the film, sometimes contrary to the review text, while critics with a good fit may be “review summarisers”, whose review rating agrees with the text. If we look at the MAE_{MS} for all critics, we see that *Philip Martin* is the biggest review summariser with an MAE_{MS} of 2%, and *Walter Chaw* is the biggest film impression summariser with an MAE_{MS} of 17%. However, the poor model fit of some linear heads shown in Section 6.1 means it is not entirely clear whether these metrics truly reflect critic behaviour, or whether they result from a non-linear relationship between the latent rating prediction and this rater’s scores.

7. Related work

A number of NLP datasets deriving from review data have been published openly and used in the literature. Maas et al. (2011, § 4.3.2) introduced a fairly popular example. Their dataset consists of 50 000 user reviews from IMDB processed to create, as is often the case, a binary positive/negative review dataset by thresholding a negative class from scores $\leq 4/10$, and a positive class from scores $\geq 7/10$.

Pang and Lee (2005) released a dataset of 5006 reviews from 4 authors, with ratings normalised to a single 4-point scale. This results in quantisation error, and discards scale information. Although they noted the subproblem of using a single model for multiple authors would require some degree of calibration, this was not explored further.

It is worth noting that while early work in sentiment analysis treated review ratings as a proxy for sentiment as a pragmatic way to create a dataset quickly, both sentiment and review analysis has developed quite significantly since. On

the one hand, sentiment analysis has given way to datasets which model emotions expressed in text, such as that of Öhman et al. (2020) who take a multi-label setting and tag subtitles with zero or more emotions from the eight emotions in Plutchik’s Wheel of Emotions. On the other hand, the desire to produce more detailed analyses of review texts has led to aspect-based sentiment analysis tasks, such as that of Pontiki et al. (2014), where certain aspects of opinions expressed in reviews are to be extracted from the text, and then individually given sentiment tags.

8. Conclusion

We have presented a dataset for multi-scale film rating inference based on reviews. The dataset preserves data useful for downstream tasks, such as the original rating scales and entity identifiers. In this sense it has fewer data quality problems when compared to comparable review rating datasets. Critics and grading scale types are included in the dataset, allowing for critic behaviour against particular scales to be modelled.

Baseline experiments show that the Affine-Per-Scale (APS) and Classifier-Per-Scale (CPS) systems were able to fit the dataset. The ordinal regression -based approach outlined here, on the other hand, did not manage to beat a Most Frequent Class (MFC) baseline. While the model is a largely analogous to the model of Cao et al. (2020), we speculate that it failed to converge here due to the multi-scale setting and noisy dataset. As we saw in Table 2, different critics show different behaviours in response to different rating scales. Since ordinal regression directly models the relationship between rating scale thresholds and the latent scale, it appears to be the correct tool to handle such non interval scale data.

Furthermore, since ordinal regression induces a latent scale, as APS does, it lends itself to the interpretation experiments of Section 6.2 & 6.3, however, it would mitigate the problems of the APS system. In particular, 1. systematic error in the gradients to the backbone due to poor model fit of linear heads as outlined in Section 6.1, and 2. this same systemic error making interpretation of the values of these heads difficult, as outlined in Section 6.3. Thus, a clear future direction for this work is to adapt the training procedure of the ordinal regression system so that it is able to make stable progress when fine-tuning a language model to tackle a noisy multi-scale task like the one posed by this paper.

Another line of future work is to consider different perspectives and tasks related to the dataset as outlined in Table 2. Particularly promising is the possibility of applying item-response theory in

order to better understand different styles of using grading scales. After all, perspectives are expressed via ratings given on a variety of grading scales across a variety of domains including Education and Psychometrics. Thus this dataset holds promise for deepening understanding of rating based preferences when combined with text into these fields also.

9. Ethical considerations

Critics are named in this dataset, however, we do not believe that releasing this dataset including the names constitutes a violation of privacy. The reviews have been created by professional critics as part of their public persona. In terms of regulations, the EU General Data Protection Regulation makes such an exemption in *Article 9(2)(e)* for cases where data has been made “manifestly public” by the data subject. Furthermore, from the less strictly legalistic perspective of seeking to avoid harm to those named, since the reviews have been submitted for aggregation, there is no reasonable expectation of relative privacy (as can be the case with social media) and no reason to believe that aggregation as part of this dataset will cause the critics to come under disproportionate public scrutiny.

Rights and regulations regarding text mining vary widely between jurisdictions. While the EU has specific exceptions for text mining, the US has wider reaching fair and transformative usage exceptions. Thus, users of the dataset are advised that they have the same rights to it as if they had created it themselves. Depending on jurisdiction, this may vary according to whether the usage is commercial or done in the service of the public interest as in the case of publicly disseminated research.

10. References

- Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.
- François Castagnos, Martin Mihelich, and Charles Dognin. 2022. [A simple log-based loss function for ordinal text classification](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4604–4609, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

- deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Andrea Esuli and Fabrizio Sebastiani. 2006. [SENTIWORDNET: A publicly available lexical resource for opinion mining](#). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. [Captum: A unified and generic model interpretability library for pytorch](#).
- Torrin M. Liddell and John K. Kruschke. 2018. [Analyzing ordinal data with metric models: What could possibly go wrong?](#) *Journal of Experimental Social Psychology*, 79:328–348.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.
- Bo Pang and Lillian Lee. 2005. [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. [Semeval-2014 task 4: Aspect based sentiment analysis](#). In *International Workshop on Semantic Evaluation*.
- Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. [Axiomatic attribution for deep networks](#). In *International Conference on Machine Learning*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Michael J Wurm, Paul J Rathouz, and Bret M Hanlon. 2021. Regularized ordinal regression and the ordinalnet r package. *Journal of Statistical Software*, 99(6).
- Thomas W. Yee. 2010. [The vgam package for categorical data analysis](#). *Journal of Statistical Software*, 32:1–34.

Author Index

Abercrombie, Gavin, 31
Allein, Liesbeth, 116
Alvarez Nogales, Anny D., 67
Araque, Oscar, 67
Aroyo, Lora, 1, 131

Creanga, Claudiu, 95

Diaz, Mark, 131
Dinu, Liviu P., 95

Fersini, Elisabetta, 78, 84
Flek, Lucie, 42
Fontana, Michele, 78

Gezici, Gizem, 49
Giannotti, Fosca, 49

Hao, Susan, 1
Homan, Christopher, 131
Hovy, Dirk, 19

Jiang, Aiqi, 31

Konstas, Ioannis, 31

Laszlo, Sarah, 1
Leonardelli, Elisa, 84
Leone, Stefano, 142
Lindahl, Anna, 56

Mala, Chandana Sree, 49
Marchiori Manerba, Marta, 49
Mastromattei, Michele, 123
May, Marlon, 42
Moens, Marie-Francine, 116
Muscato, Benedetta, 49

Nozza, Debora, 19

Parrish, Alicia, 1, 131
Paun, Silviu, 84
Pavlovic, Maja, 84, 100
Plaza-del-Arco, Flor Miriam, 19
Poesio, Massimo, 84, 100
Prabhakaran, Vinodkumar, 131

Rizzi, Giulia, 78, 84
Robertson, Frankie, 142
Rosso, Paolo, 84

Serapio-Garcia, Gregory, 131

Taylor, Alex, 131

Uma, Alexandra, 84

Valette, Mathieu, 111
Vitsakis, Nikolas, 31

Wang, Ding, 131
Welch, Charles, 42

Zanzotto, Fabio Massimo, 123