Proceedings of the Eleventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2023)

Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection

Oana Inel[1], **Tim Draws**[2], **Lora Aroyo**[3]
[1] University of Zurich, Zurich, Switzerland
[2] Delft University of Technology, Delft, the Netherlands
[3] Google, NY, NY, USA
inel@ifi.uzh.ch, t.a.draws@tudelft.nl, l.m.aroyo@gmail.com

Abstract

The rapid entry of machine learning approaches in our daily activities and high-stakes domains demands transparency and scrutiny of their fairness and reliability. To help gauge machine learning models' robustness, research typically focuses on the massive datasets used for their deployment, *e.g.*, creating and maintaining documentation to understand their origin, process of development, and ethical considerations. However, data collection for AI is still typically a one-off practice, and oftentimes datasets collected for a certain purpose or application are reused for a different problem. Additionally, dataset annotations may not be representative over time, contain ambiguous or erroneous annotations, or be unable to generalize across domains. Recent research has shown these practices might lead to unfair, biased, or inaccurate outcomes. We argue that data collection for AI should be performed in a responsible manner where the quality of the data is thoroughly scrutinized and measured through a systematic set of appropriate metrics. In this paper, we propose a Responsible AI (RAI) methodology designed to guide the data collection with a set of metrics for an iterative in-depth analysis of the *factors influencing the quality and reliability* of the generated data. We propose a granular set of measurements to inform on the *internal reliability* of a dataset and its *external stability* over time. We validate our approach across nine existing datasets and annotation tasks and four input modalities. This approach impacts the *assessment of data robustness* used in real world AI applications, where diversity of users and content is eminent. Furthermore, it deals with fairness and accountability aspects in data collection by providing systematic and transparent quality analysis for data collections.

1 Introduction

As the use of machine learning (ML) and artificial intelligence (AI) becomes more ubiquitous in our daily activities, *e.g.*, to pick a restaurant for dinner (Burke 2002), as well as in high-stakes domains, *e.g.*, to select a job candidate (Li et al. 2021) or choose medical treatment for a patient (Shatte, Hutchinson, and Teague 2019), the need to scrutinize every aspect of AI systems is also increasing. This includes evaluating their training and testing data quality, as well as quantifying the level of fairness, transparency, accountability, and non-maleficence (Jobin, Ienca, and Vayena 2019) these systems have. Several actionable toolkits and checklists for both models and datasets have been proposed, such as Fairlearn (Bird et al. 2020), AI Fairness 360 (Bellamy et al. 2019), Aequitas (Saleiro et al. 2018), Model Cards (Mitchell et al. 2019), Datasheets for Datasets (Gebru et al. 2021), PAIR AI Explorables,[1] AI Test Kitchen.[2] Furthermore, this also led to an emerging data-centric research effort on how data quality can affect the robustness, reliability, and fairness of AI systems' performance in the real world (Mehrabi et al. 2021; Sambasivan et al. 2021; Kapania et al. 2020).

Traditionally, high-quality data for ML is collected from experts and inter-rater reliability (IRR) scores (*e.g.*, Cohen's κ (Cohen 1960), Fleiss' κ (Fleiss 1971), or Krippendorff's α (Krippendorff 2011)) measure their reliability. Employing experts, however, is often costly and time-consuming. Crowdsourcing is a widely used alternative to create ground truth datasets for ML applications. Due to the nature of crowdsourcing annotation studies (*i.e.*, raters who likely have limited or no domain expertise), a large body of research has primarily focused on data evaluation and aggregation techniques (Hovy et al. 2013; Dumitrache et al. 2018; Paun et al. 2018; Braylan and Lease 2020).

Under the assumption that each annotated input sample has only one correct interpretation (Nowak and Rüger 2010), crowdsourced annotations are typically aggregated using majority vote (MV) (Dumitrache et al. 2021). However, research has shown that data quality is complex and can be influenced by many *factors*, such as disagreement-prone or subjective tasks, ambiguous input samples, target annotations, and guidelines, diverse rater characteristics and perspectives (Welinder and Perona 2010; Aroyo and Welty 2014; Kairam and Heer 2016; Chang, Amershi, and Kamar 2017; Draws et al. 2022), ethical aspects and power structures in annotation processes (Miceli, Schuessler, and Yang 2020; Díaz et al. 2022), or cognitive biases (Eickhoff 2018; Santhanam, Karduni, and Shaikh 2020; Draws et al. 2022). In such cases, IRR scores may not always be able to capture the true annotations' reliability and MV could eliminate correct answers vetted by only a few raters. Additionally, IRR scores cannot be used to directly compare datasets, as they

[1] https://pair.withgoogle.com/explorables/
[2] https://blog.google/technology/ai/join-us-in-the-ai-test-kitchen/

only indicate raters' consistency rather than data quality.

These factors have led to several streams of research. First, the notion of ground truth currently adopts a perspectivist stance (Basile et al. 2021; Aroyo and Welty 2015) which highlights the need of diverse opinions and perspectives for a better knowledge representation compared to MV. Second, before runtime checklists have been proposed (Draws et al. 2021; Thomas et al. 2022) to help researchers consider cognitive biases and other human factors affecting their annotations. Third, attention is drawn to formulating dataset artifacts that describe the collection purpose, method, and raters (Bender and Friedman 2018; Ramírez et al. 2020; Gebru et al. 2021; Díaz et al. 2022). Fourth, existing datasets have been extensively judged, improved, and re-annotated based on empirical evidence suggesting that existing annotations are not representative anymore or contain ambiguous or erroneous annotations (Yun et al. 2021; Inel and Aroyo 2019; Aroyo and Welty 2015).

However, the research landscape is still lacking a unified framework that allows for cross-datasets comparisons and measurement of dataset stability for repeated data collections. Thus, or proposed approach complements existing research by proposing an *iterative metrics-based methodology* for thoroughly scrutinizing the *factors* influencing the intrinsic reliability of datasets and their stability over time and for various contexts or factors (Fig. 1).

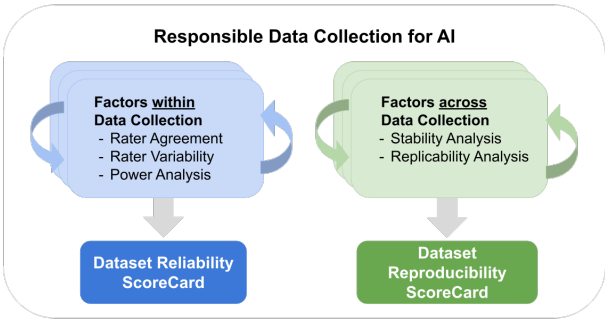


Figure 1: Methodology for measuring reliability and reproducibility of AI data collections.

Our proposed methodology enables a comprehensive analysis of data collections by applying reliability and reproducibility measurements in a systematic manner. The reliability metrics are applied on a single data collection and focus on understanding the raters. We then propose that data collection campaigns are repeated, either under similar or different conditions. This allows us to study in-depth the reproducibility of the datasets and their stability under various conditions or constraints, using a set of reproducibility metrics. In short, we propose a set of metrics that are applied (1) on a single repetition and (2) across repetitions to thoroughly evaluate the reliability and stability of annotated datasets. The overall methodology is designed to integrate responsible AI practices into data collection for AI. This allows data practitioners to follow our step-by-step guide to explore factors influencing reliability and quality, ensuring transparent and responsible data collection practices. We validate our methodology on nine existing data collections repeated at different time intervals with similar or different rater qualifications. The annotation tasks span different degrees of subjectivity, data modalities (text and videos), and data sources (Twitter, search results, product reviews, YouTube videos).

The following are the key contributions of this paper:[3]

1. a step-wise guide for practitioners consisting of a set of metrics to thoroughly investigate and explore factors that influence or impact the reliability of annotated datasets;
2. a validation and illustration of the proposed metrics-based iterative methodology for achieving transparency of the reliability of datasets and their stability over time on nine existing data collections; and
3. a discussion of implications and lessons learned for responsible data collection practices.

2 Related Work on Data Excellence

Data quality is an already established field of study (Wang and Strong 1996; Zaveri et al. 2016; Lukyanenko and Parsons 2015), spanning numerous domains and use cases, such as linked open data, user-generated data, to name a few. Data quality aspects are also addressed in several ISO standards (e.g., ISO 25012 (for Standardization 2008; Guerra-García et al. 2023), ISO 8000[4]). Research utilizing crowdsourced datasets has further broadened the community's views regarding factors influencing or affecting data quality. A large body of research has been focusing on acknowledging the impact of cognitive biases, such as confirmation bias or anchoring effect, on the crowdsourced data quality (Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019; Santhanam, Karduni, and Shaikh 2020; Draws et al. 2022). However, while various types of cognitive biases are known to impact data quality, in typical annotation studies, it is not a mainstream practice to account for raters' stances, opinions, or knowledge on various issues. Furthermore, despite several proposed mechanisms to mitigate biases (Eickhoff 2018; Hube, Fetahu, and Gadiraju 2019; Barbosa and Chen 2019), it is still unclear which mechanisms are suitable for a particular situation or which individual characteristics of the raters may lead to systematic biases (Draws et al. 2022). To further alleviate some of these issues, Draws et al. (2021) proposed a 12-item checklist for requesters to identify which cognitive biases might affect their data before the start of data collection. While this checklist offers a powerful tool for requesters, many times AI and ML practitioners reuse existing annotated datasets that might lack proper documentation or description of the annotation process, which makes the assessment difficult and leads to worrying outcomes when deployed in the real world (Paullada et al. 2021).

Unequal distribution of demographic characteristics among raters may subsequently lead to poor performance of ML models (Barbosa and Chen 2019). While investigating whether different cultural communities produce different gold standards and whether algorithms perform differently on gold standards from different cultural communi-

[3]Supplemental material and analysis at: https://github.com/oana-inel/ResponsibleAIDataCollection

[4]https://www.iso.org/standard/81745.html

ties, Sen et al. (2015) found that AMT-derived gold standards for knowledge-oriented tasks can not generalize across different communities and influence ML performance. In the context of image annotation, Dong et al. (2012) concluded that different cultures provide different tags, being highly influenced by their cognitive and emotional aspects. The same behavior was observed when performing news sentiment analysis (Balahur et al. 2010). Besides cultural differences, cognitive biases, and stereotypes, societal events or temporal aspects can also add variance in collected data (Aroyo and Welty 2015; Christoforou, Barlas, and Otterbacher 2021; Sen et al. 2015). Christoforou, Barlas, and Otterbacher (2021) showed empirical evidence that significant public health events might be reflected in the descriptive tags regarding a person's identity and body weight that raters use to annotate images of people.

In the remainder of this section, we review three streams of research regarding responsible data collection for AI: collection, assessment, and documentation and maintenance.

2.1 Data Collection

Data collection practices strongly affect the quality of crowdsourced data. This led to extensive explorations in evaluating raters' performance and identifying underperforming pools of raters (Ipeirotis, Provost, and Wang 2010; Bozzon et al. 2013; Soberón et al. 2013), improving the overall clarity of the annotation task design (Kittur, Chi, and Suh 2008; Gadiraju, Yang, and Bozzon 2017; Wu and Quinn 2017; Han et al. 2019), encouraging raters to reflect on their answers (Kutlu et al. 2020), or experimenting with several annotation designs to identify the most suitable one to capture the appropriate answers (Inel et al. 2018; Lau, Clark, and Lappin 2014; Roitero et al. 2018). More precisely, at the level of the task design, many studies experimented with annotation scales. For example, Roitero et al. (2018) showed that fine-grained annotation scales are more suitable and natural than coarse-grained annotation scales to capture web documents' relevance. Intrinsic motivation and incentives have also been shown to be beneficial in improving the quality of crowdsourced data (Ho et al. 2015; Kittur et al. 2013). In our research, we present extensive data analysis of nine existing data annotation tasks which vary in terms of collection criteria such as input data (tweets, product reviews, web documents, facial expression recordings, and news broadcasts) and annotation goal.

2.2 Data Assessment

To measure the reliability of crowdsourced annotations, research focused on quality control mechanisms (Daniel et al. 2018) and the definition of aggregation techniques (Hung et al. 2013; Li, Rubinstein, and Cohn 2019; Dumitrache et al. 2018; Hovy et al. 2013; Paun et al. 2018). Typically, current annotation campaigns rely on the use of multiple raters per annotated input and reporting of inter-rater reliability metrics (Park et al. 2012; Sigurdsson et al. 2016; Park, Shoemark, and Morency 2014), such as Cohen's κ (Cohen 1960), Fleiss' κ (Fleiss 1971), or Krippendorff's α (Krippendorff 2011). However, the choice of the IRR metric is less important than having a representative number of raters per input (Artstein and Poesio 2008). Furthermore, Popović and Belz (2022) found that raw counts are a more suitable input for computing and estimating inter-rater reliability compared to normalized counts or percentages in a machine translation use case. In a recent study, however, Braylan, Alonso, and Lease (2022) point out that Krippendorff's α relies on mean distances, which can lead to mistakenly discarding good data when dealing with subjective annotations and propose using more suitable distance functions depending on the task at hand. When dealing with subjective tasks, or tasks that could generate diverse opinions or perspectives and potentially multiple ground truths, achieving reliable results is thus even more challenging (Graham, Awad, and Smeaton 2018; Zahálka and Worring 2014; Basile et al. 2021). This could mean that different replications of such a task could give very different results.

To the best of our knowledge, only a few data collection experiments addressed repeatability (Blanco et al. 2011; Welty, Paritosh, and Aroyo 2019). Thus, our work proposes data collection repeatability as a responsible practice to measure data stability over time. We propose a set of metrics carefully chosen to scrutinize the human factors influencing various aspects of the data over time, thus fostering cross-comparison between datasets.

2.3 Data Documentation and Maintenance

By systematically reviewing 150 published papers dealing with classification tasks on Twitter data, Geiger et al. (2020) concluded that issues such as reliability, transparency, and accountability in data collection practices are not a mainstream approach in the ML community. A considerable amount of analyzed papers offer limited or no details regarding raters, their demographic information, compensation details, IRR scores of the collected datasets, annotation instructions, and overall setup of the annotation process. Following studies that also showed the extent to which potential biases are present in extensively used image datasets (Zhao et al. 2017; Hendricks et al. 2018; Otterbacher 2015), a lot of attention has been brought to data documentation and maintenance approaches, in many fields.

Inspired by medicine and psychology literature, Bender and Friedman (2018) proposed data statements for characterizing and understanding the raters of a natural language dataset, their potential biases, and how they might affect the deployment of ML models. Similarly, informed by the electronics industry where every component is thoroughly described in terms of characteristics, test results, or recommended usage, Gebru et al. (2018) proposed *datasheets* for datasets. Díaz et al. (2022) studied ethical considerations that affect the annotation of the dataset, such as, for instance raters' previous experience, and developed the CrowdWorkSheet framework to facilitate critical reflection and transparent documentation of dataset annotation decisions, processes, and outcomes. Pushkarna, Zaldivar, and Kjartansson (2022) proposed *data cards* to record key aspects of datasets and their life cycle (*i.e.*, explanations concerning the provenance, representation, usage, and fairness of ML datasets for all stakeholders), allowing for responsible AI development. Finally, (Wilkinson et al. 2016) proposed a set of guid-

ing principles to support researchers, industry practitioners, and funding and publishing agencies in scholarly data reuse; these guiding, measurable principles tackle four fundamental data principles — findability, accessibility, interoperability, and reusability (i.e., FAIR). In the crowdsourcing community, Ramírez et al. (2020, 2021) provide guidelines for requesters to improve dataset reporting and reproducibility.

While our work is not focusing on documenting the annotation process of human-annotated datasets, it complements existing approaches by proposing a set of reliability analysis metrics that foster responsible data documentation and adherence to proper data provenance and documentation guidelines for subsequent dataset alterations.

3 Reliability and Reproducibility Metrics for Responsible Data Collection

We introduce our proposed methodology for in-depth analysis of the *reliability* and *reproducibility* of data annotation studies. The proposed methodology brings together, in a systematic way, a set of measurements typically performed adhoc. However, the ability to observe their interaction allows data practitioners to provide a holistic picture of the data quality produced by these studies. Therefore, our proposed methodology provides a step-wise approach as a guide for practitioners to explore factors that influence or impact the reliability and quality of their collected data. While the reliability analysis focuses on the raters that participate in a data collection, the reproducibility analysis provides insights regarding the stability of the overall dataset. As observed in Figure 1, the chosen metrics provide input for a scorecard allowing for thorough and systematic evaluation and comparison of different data collection experiments.

Ultimately, the proposed reliability and reproducibility scorecards and analyses allow for more transparent and responsible data collection practices. This leads to the identification of factors that influence quality and reliability, the thorough measurement of dataset stability over time or in different conditions, and allows for datasets comparison.

3.1 Measuring Reliability of Annotations

We address the reliability of the crowdsourced annotations by looking at *raters agreement*, *raters variability*, and *power analysis* (i.e., determine the sufficient number of raters for each task). These analyses equip us with fundamental observations and findings for characterizing annotations' quality and reliability. It is important to note that depending on the nature and characteristics of the task (*i.e.*, difficulty, subjectivity, clarity), the assessment of the crowdsourced annotations' reliability is not always trivial and needs to be considered when generalizing the results across different tasks.

Rater agreement analysis: indicates the level of consistency among raters' annotations in an experiment. We compute the inter-rater reliability score (IRR) using Krippendorff's α (Krippendorff 2011) because is suitable for most annotation experiments, given that it can deal with multiple raters, various rating types (*i.e.*, categorical, ordinal, interval), and missing data (*i.e.*, not all units are annotated by all raters). Typically, α scores above 0.8 are considered to show strong or high agreement among raters, while values close to 0.6 are still considered acceptable (Landis and Koch 1977; Carletta 1996; Krippendorff 1980). Note that the IRR scores can be influenced by various characteristics of the task, which need to be taken into consideration in the overall analysis. For example, a low IRR reliability score (e.g., below or very close to 0.33) could indicate both high task subjectivity and unsuitable annotation guidelines or raters' qualifications, among others. Interpreting whether the agreement value is low, medium, or high, however, is often task-dependent and should be discussed on a case basis.

Rater variability analysis: gives insights regarding the variability in raters' answers distributions. We use two metrics to measure raters' precision for each individual annotated item in our datasets - we inspect the standard deviations in raters' annotations (Welty, Paritosh, and Aroyo 2019) when having binary or continuous value annotations and the index of qualitative evaluation (IQV) (Wilcox 1967) when having nominal or categorical values. IQV is a measure for assessing the variability of nominal variables with values between 0 (all raters' answers are in one category) and 1 (raters' answers are evenly distributed in each category). In our analysis, we consider IQV $\leq$ .33 as low variability, .33 $<$ IQV $>$ .66 as medium, and IQV $\geq$ .66 as high.

Power analysis: indicates whether the number of raters used in each annotation task is sufficient. To identify the optimal number of raters (*i.e.*, for which the variability in terms of IRR is not significant), we bootstrap the number of raters [3,4,5,..,n] per input item, where n is the maximum number of raters (Snow et al. 2008). For each number of raters, we perform 100 runs, where raters are randomly selected for each input item, and each time we compute the IRR. Then, we apply a chi-squared test for one standard deviation and test whether the standard deviation of the IRR scores for each number of raters is lower or equal to a threshold. In our experiments, we considered the threshold of .01 and for each number of raters, we test the following hypotheses: $H_0 : \sigma \leq .01$ and $H_a : \sigma > .01$, where σ refers to the standard deviation of the IRR scores. Given H_0, we conduct a right-tailed test, and we search for the lowest number of raters for which we fail to reject H_0, *i.e.*, p $<$ .05.

3.2 Measuring Reproducibility of Annotations

To investigate how rater populations influence the reliability of the annotation results, we propose *repeating* the annotations at different time intervals and in different settings, thus identifying the factors influencing their reliability. However, just by using the aforementioned reliability measures, we can not perform a proper comparison of the collected annotations. For instance, high IRR values in several repetitions indicate highly homogeneous raters' populations within each repetition, but it does not necessarily mean that the experiments are highly reproducible. For this, we perform two additional measurements: 1) stability analysis and 2) replicability analysis to understand how much variability the raters bring and how much we can generalize the results. For example, a high correlation between two task repetitions indicates that our results are stable, and the populations that

participated in the two repetitions are drawn from the same distribution.

Stability analysis: is the degree of association of the aggregated raters' scores across two annotation task repetitions. We measure the stability of the data collection with the correlation of the aggregated raters' annotations between pairwise repetitions of each task. The aggregated raters' annotation can be a mean value, majority vote, or any aggregation technique suitable for the task at hand. To compute the correlation, we use the Spearman's rank correlation (Xiao et al. 2016) for tasks with numerical values and Chi-square test of independence for tasks with categorical values.

Replicability similarity analysis: indicates the degree of agreement between two rater pools, making two data annotation tasks comparable. It consists of raters' agreement across repetitions of a particular task. To measure this, we use a metric called cross-replication reliability (xRR) (Wong, Paritosh, and Aroyo 2021). The xRR score between two repetitions is goes from 0 to the highest IRR score of the two repetitions. In a perfect replication, the xRR score is equal to the IRR score of each individual repetition (implying that the two repetitions also have equal IRR scores). This further means that similar IRR and xRR scores indicate both internal and external validity while much lower xRR values compared to IRR scores indicate low external validity.

4 Published Annotation Tasks and Datasets

We outline the experimental design to evaluate the reliability and reproducibility of nine published data annotation studies covering a wide range of content modalities and annotation tasks (Table 1). We first outline the data annotation studies (Section 4.1), and then describe the resulting datasets (Section 4.2). Thus, we use these nine data annotation studies as a two-fold objective: 1) to illustrate how practitioners should apply the methodology introduced in Section 4 to gain a better understanding of their data and 2) to validate the usefulness of the methodology to help practitioners explore factors that influence or impact data reliability.

4.1 Annotation Tasks

We first describe how the datasets (see Section 4.2) used in our experiments have been collected, *i.e.*, the task, the number of repetitions, and their settings (summarized in Table 1). All tasks and datasets have already been published.

Video Concepts Relevance (**VCR**; Inel and Aroyo 2022). The raters were asked to watch a video of 1-2 minutes and then select all relevant concepts for the content of the video from a list of machine-extracted concepts (an average of 11 concepts). Five different annotation experiments were run, each focusing on the identifications of different concept types, e.g *event* (**VCR_E**), *people* (**VCR_P**), *location* (**VCR_L**), and *organization* (**VCR_O**), and concepts of any type (**VCR_ALL**). The task was run on Amazon Mechanical Turk (AMT) with ten videos representing short English news broadcasts from YouTube annotated by 15 raters. Each task was repeated three times, at least three months apart, and each repetition used the same raters' qualifications, and raters were allowed to participate across repetitions.

Video Human Facial Expressions (**IRep**; Wong, Paritosh, and Aroyo 2021). The raters were given a video recording containing human facial expressions and were asked to select all facial expression labels (*i.e.*, emotions) that they perceived as being relevant from a predefined list of facial expression labels (Wong, Paritosh, and Aroyo 2021). The task was run on AMT. A total of 30 emotion labels (from the set defined by (Cowen and Keltner 2017)) were shown, together with the option "unsure" (raters were instructed to choose this option when it was not possible to determine the facial expressions expressed in the video recording). Each video recording was annotated by two raters. The task was repeated three times, each time with raters from a different pool, namely raters from Mexico City, Kuala Lumpur, Budapest, and internationals.

Product Reviews (**PR**; Qarout et al. 2019). The raters were given a product review and were asked to classify the issue described in the review into one of three possible classes (*i.e.*, "size aspects", "fit aspects", "no issue with size or fit"). The task was run on AMT, and each rater was required to annotate all 20 product reviews, which appeared in the same order for each rater, and each product review was annotated by at least 68 raters. The task was repeated five times at intervals of one week. The raters were not allowed to participate in more than one repetition.

Crisis Tweets (**CT**; Qarout et al. 2019). The raters were given a crisis-related Twitter message and were asked to categorize it into one of nine possible options (*i.e.*, "injured or dead people", "other useful information", "infrastructure and utilities damage", "not related or irrelevant", "sympathy and emotional support", "donation needs or offers or volunteering services", "missing, trapper or found people", "displaced people and evacuations", "caution and advice"). The task was run on AMT, and each rater was required to annotate all 20 tweets which appeared in the same order for all raters, and each tweet was annotated by at least 68 raters. The task was repeated five times at intervals of one week. Each rater was allowed to participate in just one repetition.

Words Similarity (**WS353**; Finkelstein et al. 2001; Welty, Paritosh, and Aroyo 2019). The raters were given a pair of words and were asked to rate the similarity of the two words on a scale from 0 to 10 (0 indicating the words are totally unrelated and 10 indicating the words are very closely related) (Finkelstein et al. 2001; Welty, Paritosh, and Aroyo 2019) (fractional scores such as .25, .5, and .75 are also possible). The task was first run by Finkelstein et al. (2001), and each pair of words was annotated either by 13 or 16 raters, and each rater annotated all pairs. The second time the task was run by (Welty, Paritosh, and Aroyo 2019) in 2019 (thus around 20 years apart), on AMT. In this repetition, each pair of words was annotated by 13 raters, and each rater was allowed to annotate as many pairs as they wanted.

4.2 Datasets

The tasks described in Section 4.1 resulted in nine annotated datasets covering different data modalities (text and videos of various lengths and duration) and sources (Twitter, product reviews, YouTube), as described in Table 2 and below.

Video Concept Relevance (VCR_E, VCR_P, VCR_L,

Task Type	Dataset Name	HITs	Annotation Template		Repetitions				
			Guidelines	Value	#	Distance	Pool	Platform	Repeat Raters
Video Concepts Relevance	VCR_ALL VCR_E VCR_P VCR_L VCR_O	88 19 23 22 9	Select all from list	Relevant concepts Relevant events Relevant people Relevant locations Relevant organiz.	3	>3 months	same	same	yes
Video Human Facial Emotions	IRep	1090	Select all from list	Facial expressions	4	-	same	same	no
Product Reviews	PR	20	Select one from list	Product issue	5	1 week	same	same	no
Crisis Tweets	CT	20	Select one from list	Crisis category	5	1 week	same	same	no
Words Similarity	WS353	353	Rate sim. of two words	From 0 to 10, increment of 0.25	3	20 years	diff	diff	no

Table 1: Overview of annotation tasks and their settings in terms of input data and annotation template.

Dataset	Input Modality	Content	Size
VCR_E	video	video - event pairs	208
VCR_P	video	video - people pairs	234
VCR_L	video	video - location pairs	223
VCR_O	video	video - organization pairs	59
VCR_ALL	video	video - concept pairs	969
IRep	video	human facial recordings	1065
PR	text	product reviews	20
CT	text	Twitter crisis messages	20
WS353	text	WordNet word pairs	353

Table 2: Overview of datasets used in our experiments.

VCR_O, VCR_ALL). Dataset of 208, 234, 223, 59, and respectively 969 video - concept pairs which have been annotated in terms of relevance in the data annotation tasks VCR_E, VCR_P, VCR_L, VCR_O, and, respectively VCR_ALL. The concepts were machine-extracted (video subtitles and video stream) from ten short English news broadcasts (*i.e.*, videos) published on YouTube, from a publicly available dataset (Inel, Tintarev, and Aroyo 2020; Jong et al. 2018; Inel and Aroyo 2022).

Video Human Facial Expressions (IRep). Dataset of 1090 video recordings of human facial recordings, part of the International Replication (IRep) dataset[5], published by Wong, Paritosh, and Aroyo (2021). Each video recording is annotated with emotions from 30 available emotions. The video recordings were generally very short, 5 seconds on average (a more extensive description of the recordings is found in (Cowen and Keltner 2017)).

Product Reviews (PR). Dataset of 20 English product reviews for fashion items (accompanied by a photo representative of the respective product), randomly selected from the dataset published by Chernushenko et al. (2018). Each product review is annotated with one of three possible issue classes, as described in Section 4.1.

Crisis Tweets (CT). Dataset of 20 English crisis-related Twitter messages (*e.g.*, earthquake, flood), randomly selected from the dataset published by Imran, Mitra, and Castillo (2016). Each tweet is annotated with one of nine possible crisis-related options, as described in Section 4.1.

WordSim (WS353)[6]. Dataset of 353 English word pairs (Finkelstein et al. 2001), used as benchmark for semantic similarity (Witten and Milne 2008) and word embeddings (Levy and Goldberg 2014; Bojanowski et al. 2017; Pennington, Socher, and Manning 2014). The word pairs were selected from WordNet, and include the 30 noun pairs from (Miller and Charles 1991). Each pair is annotated in terms of how similar the two words are on a 1 to 10 scale.

5 Results

In this section, we report on the results of the reliability analysis of the data collection studies described in Section 4.1, and in Table 1, and their repetitions. We apply our iterative metrics-based evaluation methodology to the nine datasets from these studies. In the analysis of the results, we denote each repetition as R_x, where x is the repetition index.

5.1 VCR: Annotation Tasks and Datasets

We first report on the reliability analysis of the VCR datasets (*i.e.*, VCR_ALL, VCR_E, VCR_P, VCR_L, VCR_O), as depicted in the first five rows in Table 3, columns R_1, R_2, and R_3. We observe that the datasets of all VCR annotation tasks and repetitions have mostly fair agreement and less often moderate agreement (R_1 & R_3 for VCR_ALL, R_1 for VCR_E, and R_1 & R_2 for VCR_P). The tasks VCR_O and VCR_L have, overall, the lowest inter-rater reliability. Similarly, the precision of the annotations in all tasks and repetitions is not substantial. In Table B1 in the Appendix, we show an overview of the variability of each repetition of

[5]https://github.com/google-research-datasets/replication-dataset

[6]https://aclweb.org/aclwiki/WordSimilarity-353_Test_Collection_(State_of_the_art)

the *VCR* annotation tasks. For each video-concept pair, we computed the standard deviation of their score. The majority of the experiments have a mean standard deviation (MSTD) of around 0.3, with the task *VCR_O* having a higher value of around 0.36. The standard deviation of deviations (STDD) is similar across tasks and repetitions, with the lowest value observed for the *VCR_O* task. These high values observed for MSTD and STDD show that this task for annotating relevant concepts in videos is subjective and raters consistently disagree. Concepts of type *organization* seem to generate the most disagreement among raters.

In our power analysis, we observe that all repetitions of each task tend to display similar variability in terms of IRR score. For each repetition of every task, according to the right-tailed Chi-squared test, we got very similar results in terms of the optimal number of raters needed to annotate a video-concept pair. According to the Chi-square test, the following number of raters is optimal (with minimal variation across repetitions): *VCR_ALL* - 6 raters, *VCR_E* - 12 raters, *VCR_P* - 11 raters, *VCR_L* - 11 raters, *VCR_O* - 14 raters. These, in addition to the high variability of IRR scores shown in Figure C1 in the Appendix, suggest that annotating the relevance of *organizations* in videos is a more difficult task that might require an even larger number of raters.

Although the IRR scores of the annotations gathered in all repetitions are rather low, the Spearman's rank correlation between the relevance score of the video-concept pairs (computed as the ratio of raters that picked the concept as relevant) in each pair of repetitions is high, above 0.85 for all tasks and repetitions, showing a statistically significant, strong positive correlation in Table 4. Furthermore, the pairwise xRR scores (see Figures E1, E2, E3, E4, E5) are very similar to the IRR scores of the repetitions. Thus, we observe that while the IRR scores are rather low, raters are similarly consistent in each repetition and across repetitions, showing that disagreement seems to be intrinsic to the task.

	R_1	R_2	R_3	R_4	R_5
VCR_ALL	0.43	0.40	0.44	-	-
VCR_E	0.44	0.37	0.41	-	-
VCR_P	0.41	0.40	0.39	-	-
VCR_L	0.30	0.38	0.34	-	-
VCR_O	0.25	0.30	0.30	-	-
IRep	0.25	0.23	0.50	0.13	-
PR	0.41	0.33	0.32	0.20	0.36
CT	0.59	0.70	0.68	0.72	0.65
WS353	0.59	0.57	0.50	-	-

Table 3: Krippendorff's α agreement for all datasets.

5.2 IRep: Annotation Task and Dataset

Wong, Paritosh, and Aroyo (2021) already provide an in-depth analysis of the IRR and xRR scores per emotion in three of the repetitions of the tasks. More precisely, they analyze the agreement among raters from three different regions, i.e., Mexico City (R_1), Kuala Lumpur (R_2), and Budapest (R_3). Their main conclusion is that raters seem to have more similar or divergent agreement values depending on their country of origin. In terms of individual emotions, they also observe that the most or least agreed-upon emotions are different for each country. Similarly, only a few emotions seem to have both internal and external validity, as the xRR scores between pairwise repetitions indicate.

In addition, in this paper, we also analyze R_4, a repetition of the task conducted with international raters, which could represent any possible region. We recall that in the IRep annotation task, the raters were able to select multiple expressions for each input video, which means that we deal with a multi-label annotation task. To compute rater agreement on this task, we used Cohen's κ implementation, which uses the MASI distance[7] (Passonneau 2006). In short, MASI is a distance metric used to compare two sets, in our case, two sets of annotated emotions. In Table 3, we observe that the repetition in which international raters are used, R_4, has the lowest agreement across all emotions. When inspecting agreement on individual emotions (see Table A3 in the Appendix), we observe that for almost all emotions, the IRR scores in R_4, the international repetition, have the lowest values.

Our stability analysis (see Table D1 in the Appendix) shows that emotion scores are poorly correlated across repetitions. We observe many weak correlations and only a few moderate correlations, statistically significant. Furthermore, the correlations with R_4 seem consistently lower. While the analysis performed by Wong, Paritosh, and Aroyo (2021) showed that for certain emotions such as "love" or "sadness" raters can have both high internal agreement and cross-replication agreement when comparing R_4 with the other three repetitions we can not draw such conclusions. Overall, both the internal agreement (see Table A3 in the Appendix) and the cross-replication agreement (see Figure E8 in the Appendix) indicate less consistency. More precisely, we can infer that disagreement seems to be intrinsic to the diversity of the raters and the way they interpret emotions.

5.3 PR: Annotation Task and Dataset

The inter-rater reliability scores computed on the PR datasets show, overall, fair agreement among raters. To better understand these agreement values, we also computed the IRR scores for each possible option that the raters could have chosen (see Table A1 in the Appendix), and we observed that the option "Fit & Aspect" is consistently generating lower agreement values, *i.e.*, in each repetition, compared to the other two options. Furthermore, we also observe a considerable difference of 0.21 in IRR scores between repetitions R_1 and R_4. As reported by Qarout et al. (2019), raters participating in R_4 had indeed lower accuracy compared to all other repetitions when compared against a ground truth, but the difference does not seem to be significant.

For the variability analysis, we computed for each unit in the dataset and, for each repetition, the index of qualitative variation (IQV). In Figure 2a, we observe that in all repetitions, the majority of the units annotated have high variability in terms of categories provided by raters, *i.e.*, it does not seem to be a predominant category that is chosen by the majority of the raters. This, in addition to the low IRR scores for

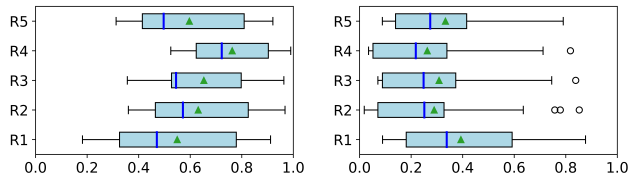
[7]NLTK: https://www.nltk.org/_modules/nltk/metrics/distance.html

	VCR_ALL	VCR_E	VCR_P	VCR_L	VCR_O
R1 & R2	ρ=0.90, p=0.0	ρ=0.90, p=5.05e-76	ρ=0.91, p=4.72e-89	ρ=0.91, p=9.24e-86	ρ=0.87, p=3.86e-19
R1 & R3	ρ=0.90, p=0.0	ρ=0.89, p=1.97e-72	ρ=0.90, p=4.25e-85	ρ=0.89, p=4.06e-77	ρ=0.86, p=1.05e-18
R2 & R3	ρ=0.90, p=0.0	ρ=0.90, p=2.60e-75	ρ=0.91, p=5.25e-85	ρ=0.91, p=4.50e-86	ρ=0.85, p=8.77e-18

Table 4: Spearman's ρ rank correlation of the relevance of each video-concept pair for each pair of VCR tasks repetitions.

the option "Fit & Aspect", could potentially indicate that the majority of the units annotated in this task are ambiguous or that the options they have to choose from are unclear or have overlapping meanings. The poor agreement and annotations stability across all units in this dataset is also confirmed by our power analysis, which indicates that a very high number of annotations is needed in each repetition to achieve stable results (see Figure C2 in the Appendix). In each repetition, the optimal number of raters is around 90, a number that is highly unlikely to be employed in such studies.

Further, we analyzed the stability of the experiments to understand to what extent the results of any two repetitions are similar. For this, we computed for each unit in each repetition the answer given by the majority of the raters (in case of ties, we selected the majority vote at random) and computed the Chi-square test of independence between every two repetitions. On the one hand, this analysis showed that the majority vote answers in any two repetitions are correlated and that there is no statistically significant difference among them (see Table D2 in the Appendix). On the other hand, the replicability analysis through the xRR measure showed similarly low values, just as the IRR scores. This result indicates that the agreement among raters is, while low, also consistent across repetitions.



(a) PR dataset (b) CT dataset.

Figure 2: IQV distribution for each unit and repetition (R_1 to R_5). The distribution is shown as a boxplot (median - blue line, mean - green triangle).

5.4 CT: Annotation Task and Dataset

The raters participating in all repetitions of the crisis tweets (CT) annotation task indicate moderate to substantial agreement, as observed in Table 3. Similarly, as for the PR datasets, we also computed the IRR scores for each possible option the raters were able to choose. These results are available in Table A2 in the Appendix. For this task, we observe that the majority of options generate moderate to substantial agreement, except for two possible answers, namely "missing, trapped, or found people" and "other useful information". However, when inspecting the data, we observe that these two options are rarely chosen by raters which might explain their very low agreement values (Brenner and Kliebsch 1996; Artstein and Poesio 2008).

For the variability analysis, we replicated the process described for the PR datasets. In contrast, however, we observe in Figure 2b that the index of qualitative evaluation for this dataset has more often values closer to 0, indicating that the annotated units have much lower variability in categories that the raters chose. More precisely, the annotated tweets seem to be easily annotated with a category that is often chosen by the large majority of the raters. R_1 of the dataset seems to have the lowest precision, which is consistent with the lower IRR score as well as with the lower overall accuracy presented by Qarout et al. (2019). In terms of power analysis, similarly to the PR annotation task and dataset, we observe that the mean IRR scores over 100 runs stabilize for a large number of raters (*i.e.*, 85-95 raters).

The stability analysis for this experiment shows that the majority vote answers are very similar across repetitions. More precisely, according to the Chi-square test of independence, we found no statistical difference between the majority vote answers of any two repetitions of the CT task (see Table D3 in the Appendix). In terms of cross-rater reliability, the xRR metric shows that the results for this dataset are consistent across repetitions (see Figure E7 in the Appendix). More precisely, we can infer that the high IRR values of these experiments generalize across different rater pools.

5.5 WS353: Annotation Task and Dataset

Among all the replicated studies we analyzed, the highest agreement scores are found for the word similarity datasets, $WS353$, namely 0.59, 0.57, and 0.50, as shown in Table 3. Such IRR values are typically acceptable in the context of natural language datasets. As reported by Welty, Paritosh, and Aroyo (2019), a thorough precision analysis indicated that while IRR scores have similar values across repetitions, there are certain word pairs for which the similarity score changed dramatically in the second and third repetition (*e.g.*, the pairs "Maradona"-"football" and "Arafat"-"peace" had higher similarity scores in the first repetition, and very low similarity scores in the second and third repetitions which were run almost 20 years from the first repetition). Our power analysis presented in Figure C4 in the Appendix indicates that around 12 raters could provide a reliable set of annotations in R_1, and even fewer raters in R_2 and R_3.

In terms of stability analysis, the Spearman's ρ correlation shows that all three repetitions are correlated with each other (statistically significant), and in particular R_2 and R_3, repetitions that were run on the same platform, with raters having similar characteristics (R_1 & R_2: ρ=0.87, p=9.93e-109; R_1 & R_2: ρ=0.84, p=4.61e-96; R_2 & R_3: ρ=0.95,

	Reliability			Reproducibility	
	Agreement	Variability	Power	Stability	Replicability Similarity
VCR_ALL	low	high	6 raters	high	high
VCR_E	low	high	12 raters	high	high
VCR_P	low	high	11 raters	high	high
VCR_L	low	high	11 raters	high	high
VCR_O	low	high	14 raters	high	high
IRep	low	-	-	low	low
PR	low to medium	high	90 raters	high	high
CT	medium to high	low	85-95 raters	high	high
WS353	medium	low to high	12 raters	medium to high	medium to high

Table 5: Scorecard summary of the reliability and reproducibility analysis of the nine experimental datasets.

p=2.88e-182). Similar results are observed in terms of cross-replication reliability, where the xRR values show higher agreement among raters that participated in the last two repetitions (*i.e.*, R_2 & R_3: xRR = 0.53), compared to raters that participated in the first repetition and the subsequent ones (*i.e.*, R_1 & R_2: xRR=0.49; R_1 & R_2: xRR=0.44).

6 Discussion

We discuss the results of our methodology for providing a coherent overview of data quality in terms of human factors influencing the reliability and reproducibility of a crowdsourced data collection. Our discussion is driven by the scorecards produced by our proposed methodology for in-depth analysis of the reliability and reproducibility of data annotation studies. The summary of our analysis is presented in Table 5. Furthermore, we provide lessons learned for responsible data collection practices, reflect on the limitations of our approach, and give directions for future work.

Factors influencing the quality of data collection. While we surveyed extensive literature in the area of crowdsourcing and human computation, we only found a handful of tasks and datasets that we could identify as repeated experiments. Furthermore, the nine annotation tasks and datasets we identified did not necessarily focus on identifying the factors that could influence the quality of data collection and neither on systematic measurement of their reliability and reproducibility. More precisely, current approaches typically use limited quality and reliability metrics, such as IRR scores or accuracy metrics against a gold standard to gauge data quality. Instead, our metrics provide a scorecard for comparing the reliability and reproducibility of each dataset and surfaces specific factors influencing results' quality.

In the **VCR** annotation tasks and datasets, we observed low IRR scores in all repetitions. However, the tasks and datasets have high stability, as the xRR analysis revealed similar cross-rater reliability and highly correlated relevance scores of the video concepts across repetitions. This indicates that raters are similarly consistent within each repetition and across repetitions and that the disagreement indicated by the low IRR scores is, in fact, intrinsic to the subjective nature of the task. One repetition of the **IRep** annotation task employed international raters. Overall, compared to the other repetitions (*i.e.*, region-specific), our analysis shows consistently low stability and cross-replication reliability. This indicates that not all rater pools are equally consistent across repetition and, more importantly, that the rater disagreement is correlated with the diverse background of the raters influencing the way they interpret emotions. More precisely, for similar tasks, our analysis indicates that *diverse raters should not be expected to produce a coherent view of the annotations* and we advise repeating the data collection by creating dedicated pools of raters with similar demographic characteristics and comparing their results. In the **PR** annotation task and dataset, we found that while stability can be achieved, the variability analysis and the power analysis indicated that even a very high number of raters (around 90) can exhibit high levels of consistent disagreement typically caused by the subjectivity of the task. In this case, we would advise optimizing the task design in order to decrease additional ambiguity in the annotation categories. The IRR analysis on the individual tweet categories on the **CT** task indicated that some categories may not be as clear as others or may only seldom be applicable. This indicates that careful attention should again be given to the design, instructions, and possible answer categories in the annotation task. Furthermore, the *high number of raters needed to obtain stable results* indicates that the task might benefit from a more thorough selection of raters and training sessions. Finally, the high variability for certain word pairs in the **WS353** task indicates that *data collection practices are affected by temporal and familiarity aspects*. This has serious implications for when data collections are reused, as certain annotations may become obsolete or change in interpretation over time.

Recommendations for responsible data collection In sum, applying our proposed methodology for responsible data collection does not pose any requirements on how data is structured or formatted. What we propose, does, however, affect the current practice and assumes a significant adaptation on the use of reliability and reproducibility metrics. The proposed methodology is centered around a set of systematic, iterative (i.e., repeated) pilots which allows to measure different characteristics of the data and task, as well as to capture raters characteristics and measure their potential biases. These aspects are captured with the proposed set of reliability and reproducibility metrics. Finally, we argue for systematic reporting on data collection provenance.

Systematic piloting: The proposed methodology for guiding data collection with a set of metrics for in-depth, iterative analysis of data reliability and replicability is primarily suitable as an investigative pilot of data annotation studies. Such early experimentation and thorough analysis of annotations would provide specific factors that could influence the data collection and could be ultimately mitigated for large-scale data collection.

Capture raters, task, and dataset characteristics: We argue that a responsible data collection practice should borrow guidelines for reporting human-centric studies from the fields of psychology, medicine, and even human-computer interaction, where human stances, opinions, and other meaningful characteristics are thoroughly recorded. While such a process would definitely increase the cost and time to gather the necessary data, it would also allow for more informed decisions on the proper process of collecting raters' annotations and possible future reuse.

Cognitive biases assessment: Recent research has demonstrated that raters' cognitive biases can strongly affect their annotations and reduce data quality (Hube, Fetahu, and Gadiraju 2019; Eickhoff 2018; Draws et al. 2022). To combat the influence of cognitive biases, Draws et al. (2021) introduced a checklist that can be used to identify and subsequently measure, mitigate, and document cognitive biases that may present an issue in the data collection tasks. We recommend using such a checklist between each iteration to surface possible cognitive biases that may affect annotations and allow for appropriate mitigation.

Provenance for data collection: To facilitate responsible reuse of datasets, data documentation, and maintenance approaches should thoroughly record its provenance, including quality scorecards. This would alleviate issues regarding the handling of data, reuse or modifications of annotation tasks, and platform selection. With proper provenance documentation, it is easier to identify factors that could influence data collections' quality. Such requirement becomes clear when data collection is influenced by temporal and regional aspects (see WS353 and IRep).

6.1 Limitations and Future Work

Diversity and scale of datasets. We experimented with nine datasets and annotation tasks with various goals, modalities, sizes, and overall setups. We repeated each task three to five times. While the overall number of units in some datasets was small (*e.g.*, $\sim$ 20 input units), the overall dataset size was much bigger as the number of raters providing annotations per item was significantly larger than in usual data collections. Our methodology is agnostic to the dataset size, and the significance of the results is not influenced by dataset size. In future work, we could extend the analysis to other data annotation tasks and input data modalities.

Scale and optimal number of repetitions. The repeatability experiments may not be scalable in terms of time and cost. However, our methodology provides optimization criteria that can mitigate this limitation in terms of input for the annotation tasks and the use of the bootstrap technique to optimize the number of raters needed for reliable results. As we have shown, iterative instances of *collect*, *measure*, *repeat* are suitable for adhering to responsible data collection practices. However, it is not trivial to decide on the number of repetitions necessary to determine that the collected data is reliable. This can be even more problematic for more subjective annotation tasks, which can be affected by raters' interpretations, opinions, perspectives, or familiarity with the items. Future work can address this limitation by investigating additional metrics for determining the suitable number of repetitions. Furthermore, future work could also focus on proposing a single suitable reproducibility score for repeated experiments similar to the one proposed by Belz, Popovic, and Mille (2022) for system reproducibility.

Raters' characteristics. The analyzed tasks included only limited information about the raters, besides some very general characteristics such as the platform on which they were recruited, country, or HIT approval rates. Furthermore, only one dataset out of the nine analyzed had a substantially different population of raters (i.e., from different countries). This aspect limits our analysis in terms of additional human factors that could influence rater agreement and the stability of the collected annotations. Future work could focus on replicating our analysis on more controlled data annotation experiments to study, for instance, the impact of age, gender, and other demographic information as additional reliability factors for responsible data collection. In future work, this iterative method of addressing responsible data collection should also investigate ways of properly maintaining and describing data provenance.

7 Conclusions

The continuous deployment of AI systems powered by crowdsourced data in real-world tasks has increased the attention of the research community to further scrutinize the quality and reliability of such datasets in diverse settings. In this paper, we propose a Responsible AI (RAI) methodology designed to guide the data collection with a set of metrics for an in-depth, iterative analysis of the *human factors influencing the quality and reliability* of the data they generate. The methodology consists of a set of metrics for a systematic analysis of data that brings transparency in how to interpret human disagreement and how to validate rater quality assuming diverse settings. We propose an iterative process to measure the reliability and stability of crowdsourced data from different perspectives. The repetition of experiments allows us to perform a comparative analysis across repetitions and measure changes both in the annotations and in the variance and consistency of raters. Due to the variety of quality metrics we employ, this research can have a strong impact on the way we measure data quality based on subjective human ratings. This further leads to increased diversity of AI systems and helps us deal with fairness and accountability aspects in data collection. By making the analysis process transparent through the set of metrics, we also deal with fairness and accountability aspects in data collection. We validated our methodology on nine existing annotation tasks and datasets. We found that our systematic set of metrics allows us to draw insights into the human and task-dependent factors that influence the quality of AI datasets.

Acknowledgements

We thank all reviewers for their valuable feedback that helped improve this submission. We also thank all authors who shared their collected datasets to help us validate our proposed methodology.

References

Aroyo, L.; and Welty, C. 2014. The Three Sides of CrowdTruth. *Journal of Human Computation*, 1: 31–34.

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.

Artstein, R.; and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4): 555–596.

Balahur, A.; Steinberger, R.; Kabadjov, M. A.; Zavarella, V.; der Goot, E. V.; Halkia, M.; Pouliquen, B.; and Belyaeva, J. 2010. Sentiment Analysis in the News. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Basile, V.; Cabitza, F.; Campagner, A.; and Fell, M. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4–1.

Belz, A.; Popovic, M.; and Mille, S. 2022. Quantified Reproducibility Assessment of NLP Results. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, 16–28. Association for Computational Linguistics.

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

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Blanco, R.; Halpin, H.; Herzig, D. M.; Mika, P.; Pound, J.; Thompson, H. S.; and Tran Duc, T. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 923–932.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Bozzon, A.; Brambilla, M.; Ceri, S.; and Mauri, A. 2013. Reactive crowdsourcing. In *Proceedings of the 22nd international conference on World Wide Web*, 153–164.

Braylan, A.; Alonso, O.; and Lease, M. 2022. Measuring Annotator Agreement Generally across Complex Structured, Multi-object, and Free-text Annotation Tasks. In *Proceedings of the ACM Web Conference 2022*, 1720–1730.

Braylan, A.; and Lease, M. 2020. Modeling and aggregation of complex annotations via annotation distances. In *Proceedings of The Web Conference 2020*, 1807–1818.

Brenner, H.; and Kliebsch, U. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, 199–202.

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4): 331–370.

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *arXiv preprint cmp-lg/9602004*.

Chang, J. C.; Amershi, S.; and Kamar, E. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17. New York, NY, USA: ACM.

Chernushenko, I.; Gers, F. A.; Löser, A.; and Checco, A. 2018. Crowd-Labeling Fashion Reviews with Quality Control. *CoRR*, abs/1805.09648.

Christoforou, E.; Barlas, P.; and Otterbacher, J. 2021. It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Cowen, A. S.; and Keltner, D. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38): E7900–E7909.

Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; and Allahbakhsh, M. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1): 1–40.

Díaz, M.; Kivlichan, I.; Rosen, R.; Baker, D.; Amironesei, R.; Prabhakaran, V.; and Denton, E. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2342–2351.

Dong, Z.; Shi, C.; Sen, S.; Terveen, L.; and Riedl, J. 2012. War versus inspirational in forrest gump: Cultural effects in tagging communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, 82–89.

Draws, T.; La Barbera, D.; Soprano, M.; Roitero, K.; Ceolin, D.; Checco, A.; and Mizzaro, S. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2114–2124.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 48–59.

Dumitrache, A.; Inel, O.; Aroyo, L.; Timmermans, B.; and Welty, C. 2018. CrowdTruth 2.0: Quality metrics for crowdsourcing with disagreement. *CoRR arXiv:1808.06080*.

Dumitrache, A.; Inel, O.; Timmermans, B.; Ortiz, C.; Sips, R.-J.; Aroyo, L.; and Welty, C. 2021. Empirical methodology for crowdsourcing ground truth. *Semantic Web*, 12(3): 403–421.

Eickhoff, C. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 162–170.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web*, 406–414.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

for Standardization, I. O. 2008. *ISO/IEC 25012: Software Engineering: Software Product Quality Requirements and Evaluation (SQuaRE): Data Quality Model*. ISO/IEC.

Gadiraju, U.; Yang, J.; and Bozzon, A. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT 2017*, 5–14. ACM.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H. M.; III, H. D.; and Crawford, K. 2018. Datasheets for Datasets. *CoRR*, abs/1803.09010.

Geiger, R. S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R.; and Huang, J. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336.

Graham, Y.; Awad, G.; and Smeaton, A. 2018. Evaluation of automatic video captioning using direct assessment. *PloS one*, 13(9): e0202789.

Guerra-García, C.; Nikiforova, A.; Jiménez, S.; Perez-Gonzalez, H. G.; Ramírez-Torres, M.; and Ontañon-García, L. 2023. ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design. *Data & Knowledge Engineering*, 145: 102152.

Han, L.; Roitero, K.; Gadiraju, U.; Sarasua, C.; Checco, A.; Maddalena, E.; and Demartini, G. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 321–329.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women Also Snowboard: Overcoming Bias in Captioning Models. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, volume 11207 of *Lecture Notes in Computer Science*, 793–811. Springer.

Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, 419–429.

Hovy, D.; Berg-Kirkpatrick, T.; Vaswani, A.; and Hovy, E. 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*, 1120–1130. The Association for Computational Linguistics.

Hube, C.; Fetahu, B.; and Gadiraju, U. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Hung, N. Q. V.; Tam, N. T.; Tran, L. N.; and Aberer, K. 2013. An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering*, 1–15. Springer.

Imran, M.; Mitra, P.; and Castillo, C. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016*.

Inel, O.; and Aroyo, L. 2019. Validation methodology for expert-annotated datasets: Event annotation case study. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Inel, O.; and Aroyo, L. 2022. Fine-tuning machine confidence with human relevance for video discovery. *Interactions*, 29(4): 78–82.

Inel, O.; Sauer, S.; Aroyo, L.; Bozzon, A.; and Venanzi, M. 2018. A Study of Narrative Creation by Means of Crowds and Niches. In *HCOMP (WIP&Demo)*.

Inel, O.; Tintarev, N.; and Aroyo, L. 2020. Eliciting User Preferences for Personalized Explanations for Video Summaries. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 98–106. ACM.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Jong, M. d.; Mavridis, P.; Aroyo, L.; Bozzon, A.; Vos, J. d.; Oomen, J.; Dimitrova, A.; and Badenoch, A. 2018. A Human in the Loop Approach to Capture Bias and Support Me-

dia Scientists in News Video Analysis. *Joint Proceedings SAD 2018 and CrowdBias 2018*, 2276: 32–40.

Kairam, S.; and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW 2016*, 1635–1646. ACM.

Kapania, S.; Sambasivan, N.; Olson, K.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. 2020. Data desiderata: Reliability and fidelity in high-stakes AI. In *1st Data Excellence Workshop at HCOMP*.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 453–456. ACM. ISBN 978-1-60558-011-1.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1301–1318.

Krippendorff, K. 1980. Validity in content analysis. *Computerstrategien Für Die Kommunikationsanalyse*.

Krippendorff, K. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC). University of Pennsylvania*.

Kutlu, M.; McDonnell, T.; Lease, M.; and Elsayed, T. 2020. Annotator rationales for labeling tasks in crowdsourcing. *Journal of Artificial Intelligence Research*, 69: 143–189.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Lau, J. H.; Clark, A.; and Lappin, S. 2014. Measuring gradience in speakers' grammaticality judgements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 821–826.

Levy, O.; and Goldberg, Y. 2014. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27.

Li, L.; Lassiter, T.; Oh, J.; and Lee, M. K. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176.

Li, Y.; Rubinstein, B.; and Cohn, T. 2019. Exploiting worker correlation for label aggregation in crowdsourcing. In *International Conference on Machine Learning*, 3886–3895.

Lukyanenko, R.; and Parsons, J. 2015. Information quality research challenge: adapting information quality principles to user-generated content. *Journal of Data and Information Quality (JDIQ)*, 6(1): 1–3.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Miceli, M.; Schuessler, M.; and Yang, T. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–25.

Miller, G. A.; and Charles, W. G. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1): 1–28.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Nowak, S.; and Rüger, S. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, 557–566.

Otterbacher, J. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1955–1964.

Park, S.; Mohammadi, G.; Artstein, R.; and Morency, L.-P. 2012. Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface. In *Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia*, 29–34. ACM.

Park, S.; Shoemark, P.; and Morency, L.-P. 2014. Toward crowdsourcing micro-level behavior annotations: the challenges of interface, training, and generalization. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, 37–46. ACM.

Passonneau, R. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 831–836. European Language Resources Association (ELRA).

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11): 100336.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6: 571–585.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Popović, M.; and Belz, A. 2022. On reporting scores and agreement for error annotation tasks. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, 306–315.

Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1776–1826. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Qarout, R.; Checco, A.; Demartini, G.; and Bontcheva, K. 2019. Platform-related factors in repeatability and reproducibility of crowdsourcing tasks. In *Proceedings of the*

AAAI Conference on Human Computation and Crowdsourcing, volume 7, 135–143.

Ramírez, J.; Baez, M.; Casati, F.; Cernuzzi, L.; and Benatallah, B. 2020. DREC: towards a Datasheet for Reporting Experiments in Crowdsourcing. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 377–382.

Ramírez, J.; Sayin, B.; Baez, M.; Casati, F.; Cernuzzi, L.; Benatallah, B.; and Demartini, G. 2021. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–34.

Roitero, K.; Maddalena, E.; Demartini, G.; and Mizzaro, S. 2018. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018*, 675–684. ACM.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Santhanam, S.; Karduni, A.; and Shaikh, S. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Sen, S.; Giesel, M. E.; Gold, R.; Hillmann, B.; Lesicko, M.; Naden, S.; Russell, J.; Wang, Z.; and Hecht, B. 2015. Turkers, scholars," arafat" and" peace" cultural communities and algorithmic gold standards. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*, 826–838.

Shatte, A. B.; Hutchinson, D. M.; and Teague, S. J. 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological medicine*, 49(9): 1426–1448.

Sigurdsson, G. A.; Russakovsky, O.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Much Ado About Time: Exhaustive Annotation of Temporal Data. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing*, 219–228.

Snow, R.; O'connor, B.; Jurafsky, D.; and Ng, A. Y. 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*, 254–263.

Soberón, G.; Aroyo, L.; Welty, C.; Inel, O.; Lin, H.; and Overmeen, M. 2013. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *CrowdSem 2013 Workshop*, volume 2.

Thomas, P.; Kazai, G.; White, R.; and Craswell, N. 2022. The Crowd is Made of People: Observations from Large-Scale Crowd Labelling. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 25–35.

Wang, R. Y.; and Strong, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4): 5–33.

Welinder, P.; and Perona, P. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, 25–32. IEEE.

Welty, C.; Paritosh, P.; and Aroyo, L. 2019. Metrology for AI: From Benchmarks to Instruments. *CoRR*, abs/1911.01875.

Wilcox, A. R. 1967. Indices of Qualitative Variation. Technical report, Oak Ridge National Lab., Tenn.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Witten, I. H.; and Milne, D. N. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*. AAAI press.

Wong, K.; Paritosh, P. K.; and Aroyo, L. 2021. Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 7053–7065. Association for Computational Linguistics.

Wu, M.-H.; and Quinn, A. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, 206–215. AAAI Press.

Xiao, C.; Ye, J.; Esteves, R. M.; and Rong, C. 2016. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14): 3866–3878.

Yun, S.; Oh, S. J.; Heo, B.; Han, D.; Choe, J.; and Chun, S. 2021. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2340–2350.

Zahálka, J.; and Worring, M. 2014. Towards interactive, intelligent, and integrated multimedia analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 3–12. IEEE.

Zaveri, A.; Rula, A.; Maurino, A.; Pietrobon, R.; Lehmann, J.; and Auer, S. 2016. Quality assessment for linked data: A survey. *Semantic Web*, 7(1): 63–93.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 2979–2989. Association for Computational Linguistics.