ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Beck, Jacob

Article — Published Version

Quality aspects of annotated data

AStA Wirtschafts- und Sozialstatistisches Archiv

Provided in Cooperation with:
Springer Nature

Suggested Citation: Beck, Jacob (2023) : Quality aspects of annotated data, AStA Wirtschafts- und Sozialstatistisches Archiv, ISSN 1863-8163, Springer, Berlin, Heidelberg, Vol. 17, Iss. 3, pp. 331-353, https://doi.org/10.1007/s11943-023-00332-y

This Version is available at:
https://hdl.handle.net/10419/313157

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

AStA Wirtschafts- und Sozialstatistisches Archiv (2023) 17:331–353
https://doi.org/10.1007/s11943-023-00332-y

ORIGINALVERÖFFENTLICHUNG

Quality aspects of annotated data

A research synthesis

Jacob Beck

Received: 25 May 2023 / Accepted: 31 October 2023 / Published online: 27 November 2023
© The Author(s) 2023

Abstract The quality of Machine Learning (ML) applications is commonly assessed by quantifying how well an algorithm fits its respective training data. Yet, a perfect model that learns from and reproduces erroneous data will always be flawed in its real-world application. Hence, a comprehensive assessment of ML quality must include an additional data perspective, especially for models trained on human-annotated data. For the collection of human-annotated training data, best practices often do not exist and leave researchers to make arbitrary decisions when collecting annotations. Decisions about the selection of annotators or label options may affect training data quality and model performance.

In this paper, I will outline and summarize previous research and approaches to the collection of annotated training data. I look at data annotation and its quality confounders from two perspectives: the set of *annotators* and the *strategy* of data collection. The paper will highlight the various implementations of text and image annotation collection and stress the importance of careful task construction. I conclude by illustrating the consequences for future research and applications of data annotation. The paper is intended give readers a starting point on annotated data quality research and stress the necessity of thoughtful consideration of the annotation collection process to researchers and practitioners.

Keywords Data quality · Data annotation · Training data · Human annotation · Research synthesis

✉ Jacob Beck
Ludwig-Maximilians-University Munich, Munich, Germany
E-Mail: jacob.beck@stat.uni-muenchen.de

1 Introduction

Typically, Machine Learning (ML) applications are evaluated by assessing how well an algorithm models a hold-out portion of its respective training data. However, learning from and reproducing erroneous data inevitably limits the value of a model's practical implementation. Therefore, a complete view of ML quality must go beyond the sole assessment of performance and include a perspective on the data and how it was sourced. However, best practices for the collection of training data are not taught in data science programs and often do not exist. This lack of guidance leaves researchers to make arbitrary decisions when collecting training data, such as the selection of annotators or the (number of) label options. These decisions impact the accuracy of the collected training data and, ultimately, model performance. Researchers find little guidance in the literature about how best to collect training data for ML models.

In this paper, I will outline previous research and approaches to the collection of annotated training data. The article highlights studies from diverse academic disciplines where parameters that affect annotated data quality are identified, estimated, discussed, or accounted for. As a result, the paper serves as a starting point for annotated data quality research. It stresses the necessity of thoughtful consideration of the annotation collection process to researchers and practitioners.

Initiated by a study on task structure and annotator effects in hate speech annotation (Beck et al. 2022), we have established the structural similarity between annotation tasks and surveys (i.e., the provision of a stimulus and fixed response/annotation options). We learned how research on data annotation could potentially benefit from past research from survey methodology. Theoretical concepts and previous findings from survey methodology will serve as an additional angle in this study in order to understand and detect mechanisms and challenges for data annotation.

I will look at data annotation and its concomitant quality confounders from two perspectives: the set of *annotators* and the *strategy* of data collection. The first section will feature studies that tackle the connection between the composition and behavioral patterns of annotators and the resulting dataset. In the second section, I will outline different strategies for constructing, realizing, and evaluating an annotated data collection process. This section will highlight the many decisions that come along with data collection. For most cases and decisions, best practices have not been developed (yet), and those that do exist are highly task- and data-specific. Rather than generating new empirical results, the paper will outline and summarize the various implementations of previous annotation collections and stress the importance of careful task construction. Both overarching sections are again divided into subsections that each cover specific aspects of data annotation. The paper concludes with potential routes for future research and applications of data annotation.

2 Annotators

A common way to obtain training data for ML models is through human data annotation (Tab. 1). Annotators have different backgrounds, i.e., they can be re-

Table 1 Overview table for cited studies that feature empirical use of annotations

Application Domain/ Data Type	Text	Image
Hateful/Offensive/ Toxic/Abusive Language	Binns et al. 2017; Founta et al. 2018; Sap et al. 2019, 2022; Xia et al. 2020; Davidson and Bhattacharya 2020; Al Kuwatly et al. 2020; Larimore et al. 2021; Excell and Moubayed 2021; Arhin et al. 2021; Beck et al. 2022; Davani et al. 2023; Huang et al. 2023	/
Other Linguistics	Nédellec et al. 2006; Dandapat et al. 2009; Fort and Sagot 2010; Ho et al. 2015; Guillaume et al. 2016; Fort 2016; Fort et al. 2018; Geva et al. 2019; Chen et al. 2020; Biester et al. 2022; Pyatkin et al. 2023; Yu et al. 2023	/
(Bio-)Medicine	Figueroa et al. 2012; Richter and Khoshgoftaar 2020	Rogstadius et al. 2011; Chandler and Kapelner 2013
Other	Sheng et al. 2008; Maaz et al. 2009; Shaw et al. 2011; Eickhoff 2018; Hube et al. 2019; Kutlu et al. 2020; Thorn Jakobsen et al. 2022; Ding et al. 2022; Kuzman et al. 2023; Pangakis et al. 2023	Goh and Lee 2011; Mekler et al. 2013; Khetan et al. 2018; Chen and Joo 2021; Zhao et al. 2021

searchers, domain experts, company employees, student assistants or crowdworkers. The distribution of crowdworking[1] labor has increasingly been organized through crowdworking platforms such as Amazon Mechanical Turk[2] or Prolific (Cefkin et al. 2014; Belletti et al. 2021). These platforms serve as quick and efficient tools to split data annotation into microtasks and retrieve the required annotations through human crowdworkers. In a usual setting, researchers or companies act as task requesters who set up and provide the pool of crowdworkers with microtasks. Crowdworkers then select themselves to work on specific tasks in exchange for a payment usually defined in advance by the task requester.

The following section will begin with a general segment on different annotator profiles. I will then look at annotators from two perspectives: annotator characteristics and annotator behavior. Within each subsection, I will highlight certain concepts and parameters that may affect annotated data quality.

2.1 Annotator profile

Before addressing annotators' specific characteristics and behavioral patterns in general, it is crucial to acknowledge the existence of various annotator profiles. These profiles encompass different roles, such as crowdworkers, student assistants, researchers, and domain experts, each with its own set of advantages and disadvantages. Notably, certain profiles may show tendencies toward specific biases discussed throughout this paper.

[1] Also referred to as crowdsourcing.

[2] Also referred to as Amazon MTurk or MTurk.

In the following, I will elaborate on two profiles that hold particular significance in the context of data annotation for ML applications: crowdworkers and domain experts.

Since crowdworkers appear to be a very frequently requested annotator profile, most of the existing research examining phenomena around annotation logically revolves around crowdsourced annotations. Generally, crowdsourced annotations can be retrieved at a high velocity and with comparably little cost and effort. Wang et al. (2013) provide a good overview of the various realizations of employing crowdworkers ranging from "Games with a purpose" to "Wisdom of the crowd" implementations. However, many factors, such as the (precarious) work standards, the incentive structure and the commitment of annotators, raise doubts about the inherent data quality of crowdsourced annotations. In addition, the susceptibility to unwanted bot annotations threaten data quality and replicability. To address these concerns, crowdworking platforms constantly aim to improve data quality and increasingly provide relevant information such as annotator demographics or metadata (e.g., response times). Moreover, some promising results have been drawn regarding bias mitigation and data quality improvement methods for crowdsourced annotations (Zhang et al. 2017).

In contrast to a large, unknown crowd of annotators, a significant portion of annotation tasks are conducted by domain experts. This profile of annotators usually consists of professional domain experts that either are recruited for the annotation (e.g., doctors for skin cancer classification tasks or in-house experts at a company) or annotate data for their own ML application (e.g., researchers). Here, the relation between the annotator and the application can play an important role in the expert's identification and resulting motivation for the task. While domain experts supposedly contribute higher-quality annotations due to their expertise, insufficient availability and high cost can become problematic factors. Furthermore, the assumption of gold-standard annotations can lead to overreliance and "bias-blind spots".

Generally, it is important to note that there is no one-size-fits-all "ideal" annotator profile. The choice of profile depends on task-specific requirements, resource availability, and contextual constraints.

2.2 Annotator characteristics

Annotators are mostly a very selective sample of individuals as they either self-select into data annotation (e.g., on crowdsourcing platforms) or are assigned to the annotation task (e.g., doctors obliged to label medical documents). In the past, the annotator composition of a crowdsourcing platform like Amazon MTurk was found to be more balanced with respect to socio-demographic characteristics than other "convenience samples", such as groups of self-selected college students, but clearly less balanced towards the national distribution than high-quality internet panels or probability panels (Berinsky et al. 2012). In web surveys representation matters a great deal, as the goal of most surveys is to draw inference about the entire population. However, in contrast to (web) surveys, sets of annotated data are not collected for this purpose. Consequently, the pool of data annotators does not necessarily have to be a random population sample. Even though this practice for

Table 2 Cited studies by confounding characteristics and data type

Annotation Confounder/Data Type	Text	Image
Gender	Binns et al. 2017; Al Kuwatly et al. 2020; Excell and Moubayed 2021; Beck et al. 2022; Biester et al. 2022; Sap et al. 2022; Ding et al. 2022	Chen and Joo 2021; Zhao et al. 2021
Race	Sap et al. 2019; Larimore et al. 2021; Arhin et al. 2021	/
Other demographics	Al Kuwatly et al. 2020; Beck et al. 2022	/
Cognitive Bias	Dandapat et al. 2009*; Shaw et al. 2011; Eickhoff 2018; Hube et al. 2019; Davani et al. 2023	Chandler and Kapelner 2013

* Expert annotator

inferential conclusions does not strictly apply to annotated data, annotations may still differ by annotator characteristics and potentially distort the resulting set of annotations in an unwanted manner. Previous studies have observed and analyzed a variety of annotator characteristics (Tab. 2). Their impact on the annotated dataset will be outlined in this section.

2.2.1 First language

The effect of the annotator's first language appears to be an important demographic feature for judging the outcome of a language annotation task. Generally, the first language is measured as a proxy variable for (English) language proficiency. If tasks are designed featuring any understanding of the English language, the language proficiency level should logically be an important determinant of an annotator's aptitude for this task. However, many annotation platforms and tasks do not restrict the set of annotators to a certain degree of language proficiency. Crowdworkers who complete tasks in English reside around the globe and do not have to meet formal language prerequisites. In 2009 36% of the Amazon MTurk workforce resided in India (Ross et al. 2010). Since English is the first language to only 0.02% of India's population (Census of India 2011), it is likely, that most of the Indian MTurkers are not first language English speakers. In an even stronger example of how nonexistent eligibility criteria with respect to language proficiency can affect the sample of annotators in a text annotation task, 48% of the individuals who labeled English-language tweets for hate speech were Venezuelan residents (Founta et al. 2018). While learning English as a second/foreign language does not automatically come along with insufficient language proficiency, it appears reasonable that, especially for complex, multilayered tasks like hate speech detection, a very high degree of English understanding is needed to create a high quality set of annotated data. Slang, irony, and sarcasm are important linguistic concepts that annotators should be able to grasp. In an experimental study in the realm of hate speech on Twitter, Beck et al. (2022) find that non-native English speakers labeled significantly fewer tweets as hateful compared to their native English speaker counterparts. Only US residents were eligible to complete the annotation task in their sample.

Further empirical results that detect first evidence for differences in data quality can be found in Al Kuwatly et al. (2020). Here, annotators examined whether comments on Wikipedia contained a form of personal attack. Subsequently, sets of annotations were split between first language English speaking annotators and their non-native counterparts. The resulting models trained on data annotated by native English speakers showed to be significantly more sensitive.

2.2.2 Expertise

Expertise or the individual's qualification for an annotation task plays an important role for the annotation behavior and the resulting data. While task qualification should naturally be a continuous variable, researchers and developers mainly differentiate between laypersons[3] (such as crowdworkers or student assistants) and domain experts (e.g., radiologists for annotating X-ray images). Annotator expertise has to be tackled from the perspective of data quality and concomitant resource efficiency. In most cases, crowdsourced labels are much easier and cheaper to obtain. However, we make the assumption that an expert's label inherits higher quality and that certain tasks cannot reasonably be completed by laypersons (like X-ray image annotation). While task-specific and difficult to quantify, one would optimally strive to understand how large the loss of annotation quality between an expert and a layperson is in order to calculate which approach serves the resulting model better. Additionally, this decision is influenced by both the necessity and the presence of experts for a given task. It might be unclear whether domain experts exist (e.g., hate speech detection) or whether the domain expertise is required for the annotation task (e.g., a biologist classifying images of cats and dogs).

In some (industry) applications, the annotation process is conducted by in-house experts. A report by Muller et al. (2021) describes how expert annotation is done at IBM Research and illustrates concepts to increase efficiency and data quality. Most importantly, the group of expert annotators is always led by one responsible person and annotation tools range from manual coding in a spreadsheet to specifically programmed annotation software.

Maaz et al. (2009) thoroughly examine how expert annotators perform compared to laypersons in an occupation-coding task. Overall, they find very comparable agreement scores within experts, within laypersons, and between experts and laypersons, which suggests that the added value of an expert annotation is rather small in this context. Despite these findings, independent coding by two coders, followed by expert adjudication, is often considered best practice for high-quality occupation coding (Biemer and Caspar 1994).

2.2.3 Race/Ethnicity

Fewer studies have analyzed differences in annotated data (quality) with respect to the annotator's race/ethnicity. Generally, past results have been mixed as Arhin et al.

[3] Layperson in terms of a specific, certified qualification for one task. Certainly, crowdworkers can be experts at annotation tasks as well.

(2021) find black annotators to more frequently deviate from the majority label in a toxic text classification task. In addition, Larimore et al. (2021) observe significant differences in annotations between White and non-White annotators when asked to assess the racial sentiment of tweets. On the contrary, race/ethnicity did not have a significant impact on hate speech annotation of tweets (Beck et al. 2022).

Independent of annotator race/ethnicity, Sap et al. (2019) observe higher predicted toxicity of statements in African American English (AAE) compared to non-AAE statements. In a subsequent experiment, instructing annotators to consider the dialect of and race/ethnicity of a statement's creator led to fewer toxic annotations. Racial bias can be found in various data sets ranging from hate speech detection (identified using topic modeling; Davidson and Bhattacharya 2020) to image captioning (Zhao et al. 2021). Increased awareness of racial bias in annotation and training data is key to preventing models from picking up racist patterns and corroborate them when being deployed. When detected, a posteriori bias mitigation methods can ideally be applied (Xia et al. 2020).

2.2.4 Gender

Looking at annotator gender, previous results have been mixed. Al Kuwatly et al. (2020) do not find significant differences in model sensitivity and specificity for models trained on male and female annotators' data, respectively. In line with this, sets of annotated data did not meaningfully differ by gender across four different Natural Language Processing (NLP) tasks in a study by Biester et al. (2022). Binns et al. (2017) observe small differences in the resulting training data when grouping the data by annotator gender, however, when training two separate models, these turned out to be very similar. Utilizing a previously annotated corpus of Wikipedia comments, they trained models to predict the toxicity of a statement. On the contrary, clear gender differences were found in toxicity annotation (Excell and Moubayed 2021), facial recognition tasks (Chen and Joo 2021),offensive language and racism annotation (Sap et al. 2022) and sentiment analysis across four different annotation modalities (Ding et al. 2022). With respect to toxicity/hate speech, Cowan and Khatchadourian (2003) observe that women generally take a more negative stance towards the harm of hate speech and, on average, value freedom of speech as less important than men do.

2.2.5 Education

Even though education (as a proxy variable for skill and task qualification) could be an important confounding variable when looking at the quality of annotated data, its impact has only been sparsely assessed by previous research. While Beck et al. (2022) find no effect of educational attainment, Al Kuwatly et al. (2020) observe models trained on data from annotators with lower education attainment to inherit a higher sensitivity.

2.2.6 *Age*

Depending on the annotation task at hand, looking at the annotator age seems to be important in order to detect possible differences in annotation patterns. If models differ by annotator age and the sample of annotators deviates from the national/global distribution with respect to age, assessing the distribution of annotations by age appears to be worthwhile to prevent unwanted distortions. Along these lines, Al Kuwatly et al. (2020) detect significant differences in both sensitivity and specificity between models trained on annotations that were grouped by age. In their assessment of Amazon MTurk respondent characteristics, Berinsky et al. (2012) report the age distribution of MTurkers to be significantly younger than other convenience samples and the national distribution.

2.2.7 *Political orientation*

The annotator's political orientation can serve as an observable covariate that explains (some part of) beliefs and values held by an individual, which could be especially important in subjective annotation tasks.

Argument annotations in two political contexts (cloning and minimum wage) were significantly different by political leaning of annotators, measured as self-reported categorization of conservative or liberal. These differences in annotations transformed downstream into algorithmic bias (Thorn Jakobsen et al. 2022). In addition, conservative annotators annotated AAE as toxic more frequently while at the same time flagging fewer instances of racist language as toxic (Sap et al. 2022). In a more abstract fashion, Davani et al. (2023) observe a correlation between stereotypes held by annotators and their hate speech annotation behavior and subsequent errors of the resulting hate speech classifier.

In 2012, a sample of Amazon MTurkers was on average more democratic-leaning and more liberal compared to a sample of individuals that adequately depicts the national distribution (Berinsky et al. 2012). While this distribution might have shifted in the past decade, it seems to be possible that crowdworkers comprise a skewed sample regarding political orientation, which might affect certain types of (more subjective) annotation tasks and the consequent models.

In general, it remains unclear when and how annotator characteristics affect the data generation process. However, collection and careful monitoring of potential differences by certain characteristics appears to be an important step. The relevance of (demographic) characteristics can differ largely by annotation task. Furthermore, the question about the required covariate (e.g., socio-demographic) distribution of annotators remains unsolved and likewise task-specific. While it seems reasonable that, e.g., genome sequences are labeled by expert biologists and not by a probability sample, models that are supposed to inherit societal beliefs and values can be distorted by heavily biased annotator samples.

2.3 Annotator behavior

Similar to surveys, annotation tasks provide the individual with a stimulus (survey question or annotation item) and fixed response options (response or label options). Due to this similarity, certain well-studied conscious or subconscious cognitive processes that influence how respondents answer a survey question might as well be present in annotation tasks. For data annotation, some of these cognitive processes, whose theoretical background often stems from social psychology, have previously been examined, such as anchoring (disproportionate focus on one piece of information) or confirmation bias (tendency to perceive information in a way that confirms previously held beliefs) (Eickhoff 2018; Hube et al. 2019). Others still lack succinct research in the realm of data annotation, e.g., speeding (annotating at an unreasonable velocity) or straightlining (repeatedly selecting the same label option irrespective of the annotation item presented) (Zhang and Conrad 2014; Schonlau and Toepoel 2015). The following subsections will illustrate three behavioral concepts that might affect data annotation behavior and have received some attention in past research.

2.3.1 Motivation

It is important to understand what motivates individuals to participate in an annotation task to help task requesters designing annotation tasks in line with the annotators' motivations and, if applicable, make use of additional motivating factors. Motivations can range from a merely monetary incentive to intrinsic interest in the resulting model (e.g., when annotating data for one's own research). Systematically studying interactions and conversations in the largest Amazon MTurk forum, "Turker Nation", Martin et al. (2014) observe that monetary motivations seem to be by far the most important motivating factor among crowdworkers. Even though the enjoyability of a task appeared to have an impact on a task's popularity (i.e., a slightly worse paid task was accepted if it was reported to be enjoyable), the monetary aspect was essential to the interacting crowdworkers. In addition to a more positive perception of enjoyable tasks, Chandler and Kapelner (2013) find annotators to be more active in a task if the task is framed to be somewhat meaningful. To create a meaningful frame, some annotators were told that their work would be used for medical research. No context related to the task was given to another part of the annotators, and some were additionally informed that their work would be discarded after the annotation process. They observed the perception of meaningfulness to increase participation rates, annotation quantity and data quality (Chandler and Kapelner 2013).

While the importance of monetary motivations is a natural characteristic of crowdworking, a construct that is specifically advertised as an easy way to earn money, task requesters should keep in mind that annotators do not necessarily have an interest in creating high quality data or well-performing models. Survey methodologists have developed theories and practical approaches on how to collect and evaluate survey participation reasons that could be applicable and beneficial for annotation tasks (Singer 2011; Keusch 2015; Haensch et al. 2022). However, crowdworkers may

misreport their motivations for engaging in annotation work due to social desirability bias (Antin and Shaw 2012). The response behavior when reporting motivations is likely to be affected by the power asymmetries between crowdworker and task requester that arise from crowdworking being a crucial source of income for many crowdworkers (Martin et al. 2014; Miceli et al. 2022).

2.3.2 Dishonesty

Dishonest behavior or misreporting can occur in surveys and in annotation tasks, driven by the individual's motivations and incentives. Within the setting of a (crowdsourced) annotation task, one can imagine multiple reasons for dishonest behavior. Incorrect information can be submitted to meet the eligibility criteria for an annotation task, or to mitigate the task burden, a phenomenon called "motivated misreporting" (Kreuter et al. 2011; Tourangeau et al. 2012; Eckman et al. 2014). If these behaviors do not occur at random, the resulting training data is prone to bias, and data quality is threatened by dishonest annotator behavior. Determining elements of annotation tasks that encourage misreporting can help researchers and task requesters to construct annotation tasks such that this behavior is prevented.

Some studies have examined the presence of dishonest annotation behavior. Suri et al. (2011) find evidence that annotators were willing to provide wrong answers for better payment. The level of fraudulent behavior decreased when annotators sensed being detected. When looking at misreporting individual characteristics in order to be admitted to an annotation task, Chandler and Paolacci (2017) report a clear tendency toward incorrect answers. The annotator sample for a task where being parent to an autistic child was a prerequisite showed approximately double the share of (reported) parents of an autistic child compared to the same task where a child with autism was not a mandatory prerequisite. In line with that, a similar experiment in the same study shows that at certain payment levels, annotators are willing to report a different gender for study eligibility (Chandler and Paolacci 2017).

2.3.3 Networking among annotators

When trying to understand annotating behavior and the self-selecting process of annotators, researchers and task requesters need to take networking and information exchange between annotators into account. Annotators use online forums to exchange annotation strategies and information with others (Martin et al. 2014). Forum users exchanged ways to earn money easier and faster, as well as which tasks were more enjoyable. The community generally condemned fraudulent behavior or cheating but not the use of loopholes within tasks or the exploitation of tasks with low payment, e.g., through reduced effort in the annotation process. Furthermore, annotators shared intelligence about task requesters they considered good and bad requesters (Martin et al. 2014). Even though not everyone is obliged to be active in an online forum and its users are a selective sample, the paper shows that the assumption of independence between observations (i.e., annotation responses) may not be valid. In addition, it seems possible that a quality assessment of the requesters will be shared among annotators.

Table 3 Cited studies by strategical dimension and data type

Strategical Dimension/Data Type	Text	Image
Gamification	Guillaume et al. 2016; Fort et al. 2018; Chen et al. 2020	Goh and Lee 2011; Mekler et al. 2013
Other task design	Nédellec et al. 2006*; Maaz et al. 2009*; Fort and Sagot 2010*; Kutlu et al. 2020; Thorn Jakobsen et al. 2022; Pyatkin et al. 2023	–
Resource allocation strategy	Ho et al. 2015	Rogstadius et al. 2011; Khetan et al. 2018

(* Expert annotator)

3 Data collection strategy

Virtually every decision with respect to annotated data collection strategy may have implications for the resulting set of annotations and the subsequently trained models. These decisions are located in the entire data collection process and range from considerations about required sample sizes to task design and data evaluation approaches. Building on the insights on annotator characteristics and behavior, the following section will illustrate four broad areas of strategic decisions around data annotation.

3.1 Task design

Designing annotation tasks in different ways can lead annotators towards different annotation patterns (Pyatkin et al. 2023). Hence, when designing the annotation task, many decisions have to be made that affect the resulting data and its quality. While these decisions might appear of minor importance, they are usually not based on empirical results but rather seem to be arbitrary choices. Without understanding how certain design features of annotation tasks affect the annotation behavior, these arbitrary choices can lead to a distorted training dataset. This section showcases a variety of task design options and potential effects on data quality (Tab. 3).

3.1.1 Label options

Which and how many label options are provided is not always straightforward or suggested by the data or model. The level of annotation detail ranges between a binary and a continuous annotation scale. As shown in Maaz et al. (2009), the number of label options can be varied depending on the desired degree of aggregation achieved by the annotation. Kutlu et al. (2020) discuss response scales in annotation tasks. They denote a clear tradeoff between the information gained and the burden imposed on the annotators, where increasing label options generally increases both. One potential tweak to the scale of label options is the addition of a label that provides annotators with the possibility to express their uncertainty, such as a "don't know" option. While the provision of a "don't know" option prevents forcing annotators into unwanted labels, it might encourage annotators to not thoroughly think about an

annotation decision and merely label "don't know" in case of a slight doubt. However, its value in annotation tasks appears to be unclear. Beck et al. (2022) report an insignificantly small (around 2%) share of "don't know" labels selected. Similarly, the set of label options can potentially be extended by a residual category (e.g., "not elsewhere classified") in order to prevent unwanted wrong label assignments. It is worth noting that the discussed studies address language annotation tasks, and it remains unclear to what degree these findings apply to other types of data, such as images.

3.1.2 Rationale

Kutlu et al. (2020) experiment with asking annotators to provide the rationale behind every annotation judgment made. In general, the authors argue that requesting rationales improves the quality of the resulting data and yields additional information. However, they observe that experienced Amazon MTurk crowdworkers (that completed 20 or more tasks) were more likely not to take the additional time to provide the rationale, as it was not mandatory (Kutlu et al. 2020). While this does not appear to be a feasible option for a full-scale annotation process, asking annotators for their judgment rationale might help requesters in an earlier stage. A more practicable application could be asking for rationales in a potential "pre-test" setting of an annotation task. Similar to conducting cognitive interviews (Beatty and Willis 2007), where respondents are asked to express their full thought process, in preparation of experiments or surveys, a smaller number of annotations with an extensive rationale could be collected prior to the main data collection in order to detect potentially unwanted behavioral patterns.

3.1.3 Guidelines

Another component of an annotation task that can potentially bias or anchor the subsequent annotation process is the initial annotation guidelines or tutorials. In her book "Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects" Karën Fort provides a theoretical framework on how annotation guidelines are an important, yet often resource-intensive, part of the design process (Fort 2016). When constructing an annotation task tutorial, requesters have to make important decisions such as the number and the selection of examples and balance the degree of leeway that is given to the annotators. Empirically, Nédellec et al. (2006) observe an improvement in annotated data quality through the provision of guidelines. Thorn Jakobsen et al. (2022) conclude that annotations and annotator bias are impacted by the instructing guidelines.

3.1.4 Order

As suggested by the theory of contrast and assimilation, previously perceived pieces of information impact the perception of the information at hand (Bless and Schwarz 2010). Beck et al. (2022) experimentally examine the presence of order effects and find a tweet to be labeled as hateful less frequently when preceded by a more

hateful tweet, compared to the hatefulness annotation of the same tweet preceded by a less hateful tweet. In line with the theory by Bless and Schwarz (2010), this result provides first evidence for the presence of a contrast effect. The contrast effect implies that an item is seen as more dissimilar to the previously annotated item(s), which results in the individual's judgment significantly depending on the already perceived body of data (Bless and Schwarz 2010). Thus, items to be annotated should be arranged in random order. While random ordering is seen as a best practice for many applications, it could potentially be problematic for Active Learning (AL). AL describes an ML approach in which the model predicts the annotation of which item would currently provide the model with the greatest benefit in terms of model performance (Settles 2009). This purposeful ordering (by the model) could foster unwanted order effects. However, since AL generally holds the potential for multiple benefits, such as the reduction of annotation costs, the expected bias introduced by non-random ordering needs to be weighed against the anticipated AL benefits (Zhang et al. 2022). In addition, the ordering of multiple different tasks needs to be assessed empirically. If two annotations are to be made for one item (e.g., the brightness and the resolution of an image), a task design decision with respect to order needs to be made. Both annotations could be retrieved in one screen, brightness annotations could be followed by resolution annotations, or each image could first be annotated regarding brightness, immediately followed by that image's resolution annotation.

3.1.5 Gamification

Transforming an annotation task into a (somewhat) enjoyable game could theoretically have some positive implications for the annotation (Goh and Lee 2011; Mekler et al. 2013; Chen et al. 2020). First, in line with Martin's observation, enjoyable tasks facilitate the recruitment process as crowdworkers are more likely to accept these kinds of tasks (Martin et al. 2014). More importantly, Fort, who makes an effort in gamification in linguistic annotation tasks, observes promising results in terms of annotation quantity and quality (Fort 2016). Collecting the annotations in a gamified setting increased the annotation output per person and provided some evidence for higher data quality (Guillaume et al. 2016; Fort 2016; Fort et al. 2018). Here, the challenge factor between players (e.g., high scores or leaderboards) could play a motivating role. Nevertheless, the author stresses that setting up an annotation game is costly in terms of financial and time resources (Fort 2016). Therefore, it seems like there are only a few settings where gamification is a feasible and reasonable approach to annotation, such as repeated data collection phases. A possible solution to overcome the high costs could be an annotation platform, where specific tasks can be embedded in a gamified platform design.

3.1.6 Pre-Annotation

Another design choice that has shown some promising results with respect to reduced annotation time (and therefore cost) and increased quality is the pre-annotation of items by an algorithm or an individual. Here, annotators don't perceive the annotation items in an unlabeled fashion but with the pre-annotation, which they are supposed

to confirm or reject. Fort and Sagot (2010) state that pre-annotation led to faster annotation but also confirmation bias, meaning that annotators were excessively less likely to deviate from the pre-annotated label. These findings are in line with previous research that sees the potential for data quality but raises the threat of confirmation bias (Dandapat et al. 2009).

3.2 Data composition

Two relevant questions concerning the desired data composition of annotations that have been tackled in past research will be portrayed in the following section.

3.2.1 Train-test split

Geva et al. (2019) examine the train-test split from a data annotation perspective. In their study, where annotators were asked to create new text examples in order to train an NLP model, they conclude that test and train set annotators should be separated. Put differently, annotators for the test and train data should be distinct groups of individuals. The most important reason for the strict segregation of test and train data annotators is the prevention of one (or very few) annotators creating large shares of both train and test data, resulting in overfitted models (on that particular annotator's data) (Geva et al. 2019).

Furthermore, in a study that evaluated models with extremely small datasets and costly annotations (here: autism classification and neuroimaging), decreased model accuracy was observed counterintuitively with increasing sample size (Vabalas et al. 2019). Upon further evaluation, the authors conclude that training data for models trained on extremely small datasets was not split into test and train sets in order to make the most use of every (sparse) annotation. However, this produced largely overfitted models that achieved high accuracy scores but did not generalize well outside the training data. This again stresses that conducting a train-test-split between annotators should be considered best practice, even with sparse training data. When performing strict annotator segregation, differences in the distribution of observed annotator characteristics between train and test data annotators should be assessed to detect unwanted imbalances.

3.2.2 Annotations per item

In addition to concerns about the train-test split, the design of an annotation task requires a decision about the number of annotations collected per item (e.g., per image or phrase). Precisely, an assumption needs to be made whether an additional annotation for an item outweighs the benefit of an annotation of a new item. Here, multiple parameters need to be taken into account, such as the costs per annotation of a (new) item, task complexity, annotation quality, and the desired model outcome.

In a theoretical paper from 2008, Sheng et al. tackle the tradeoff problem between an additional annotation and an additional example. They stress that the decision is highly dependent on the annotation quality and the cost of collecting an additional annotation in relation to adding a new example. If annotator quality is high, collecting

one label per item appears to be the most efficient strategy. On the contrary, low annotator quality suggests collecting multiple annotations per item (Sheng et al. 2008).

A more recent study by Khetan et al. (2018) confirms the importance of the cost structure. They observe that when adding a new example is cheap and annotations are costly, it appears to be more efficient to collect an annotation for a new item than an additional annotation for an already labeled item. This especially holds true if models reach a quality threshold, where they argue that the addition of new annotated examples is most important to increase the model's quality.

3.3 Monetary incentives

In many cases, annotators receive payments for completing the task. How the monetary incentive is structured is likely to influence annotation and, ultimately, quality of data and model. This section tackles the issue of annotator payment from two angles: the general wage level and the more complex design of flexible payment schemes.

3.3.1 Payment level

In the process of annotation task design, an appropriate wage level needs to be derived. Given a financial budget, higher wages lead to fewer total annotations to be collected. However, insufficient wages are as well likely to come with negative consequences that might offset the benefit of more generated annotations. Inappropriately paid tasks will have difficulties to crowdsource annotators, especially when competing with other tasks on the crowdsourcing market. Furthermore, even if annotators are willing to complete an underpaid task, Martin et al. (2014) observed the general notion among annotators that exploiting (e.g., speeding through) poorly paid tasks is less unacceptable than properly paid ones. Especially if crowdworkers have an approximate desired hourly wage in mind, underpaid tasks should be more likely to be sped through. Contrary to this theoretical argumentation, multiple studies conclude that higher wages do increase the quantity of work done (i.e., they facilitate recruitment) but not the quality of annotations (Rogstadius et al. 2011; Buhrmester et al. 2011; Litman et al. 2015; Vaughan 2018).

3.3.2 Payment flexibility

However, how the annotation behavior eventually turns out is greatly affected by the existence and structure of incentives. Here, the first and most obvious decision is between a fixed payment per task and payment per time. In general, none of the options is strictly better, as fixed payments incentivize speeding and unthoughtful annotation, whereas payment per time incentivizes taking needless amounts of time per task while not necessarily guaranteeing higher quality. Similar to a multitude of other paid tasks (e.g., responding to surveys or the entire service sector), different strategies need to be assessed and validated. A more fine-grained approach to annotation incentives is the implementation of performance-based bonus payments. The

idea to provide additional payments for high-quality annotation is intended to improve the annotation behavior. However, previous findings have been mixed. While some studies observe improved data quality through performance-based payments (Ho et al. 2015), others could not confirm the existence of this relationship (Shaw et al. 2011; Lou et al. 2013). This could have multiple reasons, such as an insufficient incentive in relation to the required additional effort or merely that annotators already completed the task to their best knowledge.

While these incentives theoretically sound like a promising tool to improve data quality, the estimation of annotator performance raises another problem. Without gold standard data at hand (which is often generated in the annotation process), performance can only be measured with imperfect indicators such as response time or agreement score with the majority label. If parameters were known to perfectly measure annotator quality, the annotation would be obsolete. Ultimately, reducing the annotator's leeway through incentives (or extreme guidelines) increases the degree to which the resulting dataset is a function of the task requester. Even though task requesters often know what the intended outcome is supposed to be, this must not necessarily be the case for every annotation task.

3.4 Data requirements

During the task design process, an educated decision regarding the required sample size needs to be made for both the sample size of annotators and the number of total annotations. These theoretical considerations should be done by initially putting constraints like budget or annotator availability (e.g., domain experts) aside. Then, in a second step, these constraints need to be taken into account and a final strategy for the targeted sample size can be made. However, as I will outline in this section, requirements are not necessarily fixed but can be adjusted in a flexible manner. In general, a thoughtful a priori estimation of the required number of observations (annotations), such as power calculations, appears to be important within a scientific and data-driven approach to model training. Merely collecting data until the model seems to have reached a sufficient performance or until the money is spent seems like a suboptimal strategy.

3.4.1 Required sample size

An approach to flexibly adjust the annotations collected could be predicting the required sample size during data collection. This can be done by parallelizing data collection and model training process and modeling the estimated performance curve (e.g., based on MAE or RMSE). Mukherjee et al. (2003) provide theoretical groundwork on how performance curves can be used to estimate the benefit of a data point in classifier models. Based on the observed trajectory of the performance curve, the added value of an additional annotated data point can be predicted and weighed against the costs under the assumption of constant annotation quality. According to Figueroa et al. (2012), the performance curve generally takes on the shape of the "inverse power law" and modeling the learning curve is essential for finding the optimal sample size. They describe the common process for annotated data collection

as "an initial number of samples in an ad hoc fashion to annotate data and train a model" (Figueroa et al. 2012, p. 9). The number of annotations is then steadily increased if the target model performance has not been reached. The authors argue that this strategy is "based on the vague but generally correct belief that performance will improve with a large sample size" (Figueroa et al. 2012, p. 9). Even though an additional data point is unlikely to decrease model performance, it still needs to be weighed against its costs. Therefore, the authors strongly argue for modeling efforts and stress that the final strategy also depends on the required model performance and annotation costs. Here, Active Learning could serve as an effective data collection framework to estimate the information gained by an annotation and, with that, minimize the required sample size.

This adaptive approach to sample size is in clear contrast to data collection processes with other applications, such as surveys or experiments, where the sample size is mostly derived a priori (e.g., through power calculations).

3.4.2 Required positive instances

A slightly different approach to estimating the desired sample size is focusing on the required positive instances in an annotated dataset for binary classification (e.g., positive instances of breast cancer on mammography results). A study by Richter and Khoshgoftaar (2020) aims at modeling learning curves depending on the number of positive instances in very large training datasets that inherit very low positive rates (e.g., melanoma or other rare medical incidents). They analyze four datasets with more than 1 million observations. In three of four cases, high levels of model performance could be achieved with less than 2500 positive instances. The findings underline that the number of positive instances is potentially a better explanatory variable to model performance than the total sample size and call for the inspection of learning curves to make an informed judgment on sample requirements. Multiclass classification tasks add another layer of complexity to estimating sample size requirements from the number of instances per class.

3.4.3 Required number of annotators

In addition to the previously discussed (demographic) distribution of annotators, an insufficient number of annotators is threatening the quality of the resulting dataset and model quality, as it potentially provides single annotators with excessive leverage. Put differently, by requesting an additional annotation and thereby consulting another individual, the quality of the data can be greatly enhanced, similar to seeking opinions from multiple doctors to make a diagnosis. Annotator constraints such as availability, costs, and quality should be weighed against the associated benefits and assist task requesters in estimating a target number of total annotators. This can lead to the development of best practices for annotation in certain domains, such as the utilization of independent double coding followed by expert adjudication for occupation coding (Biemer and Caspar 1994). Geva et al. (2019) nicely underline the importance of considering the number of annotators by showing that, in many cases, a small number of annotators is responsible for a very large proportion of the

annotations. An example they give is the Multi-Genre Natural Language Inference (MNLI) dataset, where an eighth of the annotators was responsible for around 90% of the total annotations. Since annotations are nested within annotators (similar to survey interview responses nested within interviewers), allowing these large shares of annotations per individual provides excessive leverage to single annotators and renders the training data more prone to bias. They find empirical support for this assumption, as adding an annotator identifier as a model feature increased model performance across three of four examined datasets. In addition, the clear individual component of annotations became obvious when models trained to predict annotators based on their annotations performed quite well in their study. Ultimately, in a setup where the annotators created new examples (to be annotated), a single-annotator trained model generalized worse to the test data of other annotators (Geva et al. 2019). Overall, this shows that very small numbers of annotators or large shares of annotations per individual can come along unwanted consequences. While, especially with difficult or domain-specific tasks, the potential annotator pool is often small, the variance explained by the annotator should at least be evaluated. Adding more annotators decreases the individual's impact on the model and may reduce the risk of a biased training dataset.

4 Conclusion

In this paper, I have taken a data perspective aiming at outlining which features and decisions within the data annotation realm can affect data and ML model quality. The paper was divided into two main sections: First, I highlighted some potentially biasing features and mechanisms on the annotator side, e.g., demographic characteristics like first language or behavioral concepts like misreporting. The second part contained a variety of strategic data collection decisions that can or have to be made and their empirical examinations. This study demonstrated how broad and complex dealing with annotated data can be. In addition, it showed how mechanisms and decisions on both annotator and strategy level can affect data quality and sketched potential roads to account for that.

5 Future work

5.1 Implications for task requesters

Since biasing mechanisms and distorting task design seem to be very task-specific, it may be difficult to develop overarching theories or best practice guidelines. However, even if best practices are not available, task requesters should still aim at prioritizing which kinds of errors should mostly be avoided and design annotation tasks accordingly. The combination of the present data, annotators and resources could give a clear indication of how to make certain choices, such as deciding between a fixed payment and an hourly wage for a task. Furthermore, if similar annotation tasks are deployed repeatedly, it might make sense to experiment with differently

designed tasks and evaluate the quality of the respective annotations. Practitioners in the area of data annotation need to find a middle way between seeing errors everywhere and expecting the model to account for everything. Not all biases can be eradicated in the training process, and not every possibility of biased training data renders annotated data unusable.

5.2 Implications for future research

Generally, a large share of the featured research has not investigated the research questions in real-world, experimental settings. Many studies used simulated annotations, synthetic data, or evaluated design effects a posteriori rather than in a planned experimental setup. Most of the theoretical studies are predestined for experimental validation, such as the paper by Sheng et al. (2008) on collecting additional versus new annotations. It could be worth attempting to transfer more theoretical constructs from survey methodology into data annotation in order to examine the presence of similar mechanisms. Concepts like speeding, straightlining, or acquiescence (the tendency to agree with the interviewer/task) have been studied around surveys and could assist in determining threats to annotated data quality. To improve annotation task design, it seems worthwhile to examine whether annotation behavior is affected by increasing expertise and/or fatigue throughout a task. In addition, future research could experiment with flexible annotation designs taking the uncertainty into account. For example, annotation tasks could always collect two labels per item, an additional one if annotators disagreed, and no further if they agreed. Moving forward, it seems important whether the heterogeneity between annotations should be eradicated (e.g., by very strict guidelines) or valued and implemented in the model or data collection process. Other relevant data types, such as audio or video data have not been featured in this paper. However, especially annotation of audio data is taking up a prominent spot in the field and a growing body of studies has been published (e.g., Wang et al. 2019; Cartwright et al. 2019; Meyer et al. 2020).

Ultimately, the rapidly developing large language models (LLMs) like ChatGPT might turn out as an extension to the "data annotation toolbox". First studies on LLM data annotation have found promising results for text genre identification (Kuzman et al. 2023), hate speech detection (Huang et al. 2023), LLM-assisted grammar analysis (Yu et al. 2023) and various text annotation tasks (Gilardi et al. 2023). Under which circumstances these models can be used to annotate data or assist a human annotator remains to be determined. Pangakis et al. (2023) conclude that the benefit and quality of LLM annotation is highly dependent on annotation task and data.

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest J. Beck declares that he/she has no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the

material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Al Kuwatly H, Wich M, Groh G (2020) Identifying and measuring annotator bias based on annotators' demographic characteristics. In: Association for Computational Linguistics (ed) Proceedings of the fourth workshop on online abuse and harms, pp 184–190

Antin J, Shaw A (2012) Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India. In: Proceedings of the SIGCHI Conference on human factors in computing systems, pp 2925–2934

Arhin K, Baldini I, Wei D et al (2021) Ground-truth, whose truth?—examining the challenges with annotating toxic text datasets

Beatty PC, Willis GB (2007) Research synthesis: the practice of cognitive interviewing. Public Opin Q 71:287–311. https://doi.org/10.1093/poq/nfm006

Beck J, Eckman S, Chew R, Kreuter F (2022) Improving labeling through social science insights: results and research agenda. In: Chen JYC, Fragomeni G, Degen H, Ntoa S (eds) HCI international 2022—late breaking papers: interacting with eXtended reality and artificial intelligence. Springer Nature Switzerland, Cham, pp 245–261

Belletti C, Erdsiek D, Laitenberger U, Tubaro P (2021) Crowdworking in France and Germany. Report. Leibniz-Zentrum für Europäische Wirtschaftsforschung (ZEW)

Berinsky AJ, Huber GA, Lenz GS (2012) Evaluating online labor markets for experimental research: amazon.com's mechanical Turk. Polit anal 20:351–368. https://doi.org/10.1093/pan/mpr057

Biemer P, Caspar R (1994) Continuous quality improvement for survey operations: some general principles and applications. J Off Stat 10:307

Biester L, Sharma V, Kazemi A et al (2022) Analyzing the effects of annotator gender across NLP tasks. In: Proceedings of the 1st workshop on perspectivist approaches to NLP@ LREC2022, pp 10–19

Binns R, Veale M, Van Kleek M, Shadbolt N (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: Social Informatics: 9th International Conference, SocInfo 2017, Oxford, September 13–15, pp 405–415 (Proceedings, Part II 9)

Bless H, Schwarz N (2010) Chapter 6—mental construal and the emergence of assimilation and contrast effects: the inclusion/exclusion model. In: Advances in experimental social psychology. Academic Press, pp 319–373

Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6:3–5. https://doi.org/10.1177/1745691610393980

Cartwright M, Dove G, Méndez Méndez A, Bello J, Nov O (2019) Crowdsourcing multi-label audio annotation tasks with citizen scientists. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–11

Cefkin M, Anya O, Dill S et al (2014) Back to the future of organizational work: crowdsourcing and digital work marketplaces. In: Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing. Association for Computing Machinery, New York, pp 313–316

Chandler D, Kapelner A (2013) Breaking monotony with meaning: motivation in crowdsourcing markets. J Econ Behav Organ 90:123–133

Chandler JJ, Paolacci G (2017) Lie for a dime: when most prescreening responses are honest but most study participants are impostors—jesse J. Chandler, Gabriele Paolacci, 2017. https://journals.sagepub.com/doi/abs/10.1177/1948550617698203. Accessed 2 Nov 2022

Chen Y, Joo J (2021) Understanding and mitigating annotation bias in facial expression recognition. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, pp 14960–14971

Chen C-M, Li M-C, Chen T-C (2020) A web-based collaborative reading annotation system with gamification mechanisms to improve reading performance. Comput Educ 144:103697. https://doi.org/10.1016/j.compedu.2019.103697

Cowan G, Khatchadourian D (2003) Empathy, ways of knowing, and interdependence as mediators of gender differences in attitudes toward hate speech and freedom of speech. Psychol Women Q 27:300–308. https://doi.org/10.1111/1471-6402.00110

Dandapat S, Biswas P, Choudhury M, Bali K (2009) Complex linguistic annotation—no easy way out! A case from Bangla and Hindi POS labeling tasks. In: Proceedings of the third linguistic annotation workshop (LAW III), pp 10–18

Davani AM, Atari M, Kennedy B, Dehghani M (2023) Hate speech classifiers learn normative social stereotypes. Trans Assoc Comput Linguist 11:300–319. https://doi.org/10.1162/tacl_a_00550

Davidson T, Bhattacharya D (2020) Examining racial bias in an online abuse corpus with structural topic modeling. arXiv preprint arXiv:2005.13041

Ding Y, You J, Machulla T-K et al (2022) Impact of annotator demographics on sentiment dataset labeling. Proc Acm Hum Comput Interact 6:1–22. https://doi.org/10.1145/3555632

Eckman S, Kreuter F, Kirchner A et al (2014) Assessing the mechanisms of misreporting to filter questions in surveys. Public Opin Q 78:721–733. https://doi.org/10.1093/poq/nfu030

Eickhoff C (2018) Cognitive biases in crowdsourcing. In: Proceedings of the eleventh ACM international conference on web search and data mining. Association for computing machinery New York, pp 162–170

Excell E, Moubayed NA (2021) Towards equal gender representation in the annotations of toxic language detection. arXiv preprint arXiv:2106.02183

Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH (2012) Predicting sample size required for classification performance. BMC Med Inform Decis Mak 12:8. https://doi.org/10.1186/1472-6947-12-8

Fort K (2016) Collaborative annotation for reliable natural language processing: technical and sociological aspects. John Wiley & Sons

Fort K, Sagot B (2010) Influence of pre-annotation on POS-tagged corpus development. In: The fourth ACL linguistic annotation workshop Uppsala, pp 56–63

Fort K, Guillaume B, Constant M et al (2018) "Fingers in the nose": evaluating speakers' identification of multi-word expressions using a slightly Gamified Crowdsourcing platform. In: Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018), pp 207–213

Founta A, Djouvas C, Chatzakou D et al (2018) Large scale crowdsourcing and characterization of twitter abusive behavior. In: Proc Int AAAI Conf Web Soc Media https://doi.org/10.1609/icwsm.v12i1.14991

Geva M, Goldberg Y, Berant J (2019) Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. arXiv preprint arXiv:1908.07898

Gilardi F, Alizadeh M, Kubli M (2023) ChatGPT outperforms crowd-workers for text-annotation tasks. Proc Natl Acad Sci 120:e2305016120. https://doi.org/10.1073/pnas.2305016120

Goh DH, Lee CS (2011) Perceptions, quality and motivational needs in image tagging human computation games. J Inf Sci 37:515–531. https://doi.org/10.1177/0165551511417786

Guillaume B, Fort K, Lefèbvre N (2016) Crowdsourcing complex language resources: playing to annotate dependency syntax. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 3041–3052

Haensch A-C, Weiß B, Steins P et al (2022) The semi-automatic classification of an open-ended question on panel survey motivation and its application in attrition analysis. Front Big Data 5:880554. https://doi.org/10.3389/fdata.2022.880554

Ho C-J, Slivkins A, Suri S, Vaughan JW (2015) Incentivizing high quality crowdwork. In: Proceedings of the 24th international conference on world wide web International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. In, pp 419–429

Huang F, Kwak H, An J (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, pp 294–297

Hube C, Fetahu B, 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. Association for computing machinery New York, pp 1–12

Keusch F (2015) Why do people participate in Web surveys? Applying survey participation theory to Internet survey data collection. Manag Rev Q 65:183–216. https://doi.org/10.1007/s11301-014-0111-y

Khetan A, Lipton ZC, Anandkumar A (2018) Learning from noisy singly-labeled data. arXiv preprint arXiv:1712.04577

Kreuter F, McCulloch S, Presser S, Tourangeau R (2011) The effects of asking filter questions in Interleafed versus grouped format. Sociol Methods Res 40:88–104. https://doi.org/10.1177/0049124110392342

Kutlu M, McDonnell T, Elsayed T, Lease M (2020) Annotator rationales for labeling tasks in crowdsourcing. J Artif Intell Res 69:143–189. https://doi.org/10.1613/jair.1.12012

Kuzman T, Mozetič I, Ljubešić N (2023) ChatGPT: beginning of an end of manual linguistic data annotation? Use case of automatic genre identification. arXiv, abs/2303.03953
Larimore S, Kennedy I, Haskett B, Arseniev-Koehler A (2021) Reconsidering annotator disagreement about racist language: noise or signal? In: Association for Computational Linguistics (ed) Proceedings of the ninth international workshop on natural language processing for social media, pp 81–90
Litman L, Robinson J, Rosenzweig C (2015) The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. Behav Res Methods 47:519–528. https://doi.org/10.3758/s13428-014-0483-x
Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. In: Association for Computing Machinery (ed) Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining New York, pp 623–631
Maaz K, Trautwein U, Gresch C et al (2009) Intercoder-Reliabilität bei der Berufscodierung nach der ISCO-88 und Validität des sozioökonomischen Status. Z Erzieh 12:281–301. https://doi.org/10.1007/s11618-009-0068-0
Martin D, Hanrahan BV, O'Neill J, Gupta N (2014) Being a turker. In: Association for Computing Machinery (ed) Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing New York, pp 224–235
Mekler ED, Brühlmann F, Opwis K, Tuch AN (2013) Disassembling gamification: the effects of points and meaning on user motivation and performance. In: Association for Computing Machinery (ed) CHI '13 extended abstracts on human factors in computing systems New York, pp 1137–1142
Meyer J, Rauchenstein L, Eisenberg J, Howell N (2020) Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In: Proceedings of the twelfth language resources and evaluation conference, pp 6462–6468
Miceli M, Posada J, Yang T (2022) Studying up machine learning data: why talk about bias when we mean power? Proc Acm Hum Comput Interact 6:34:1–34:14. https://doi.org/10.1145/3492853
Mukherjee S, Tamayo P, Rogers S et al (2003) Estimating dataset size requirements for classifying DNA microarray data. J Comput Biol 10:119–142. https://doi.org/10.1089/106652703321825928
Muller M, Wolf CT, Andres J et al (2021) Designing ground truth and the social life of labels. In: Proceedings of the 2021 CHI conference on human factors in computing systems. ACM, Yokohama, pp 1–16
Nédellec C, Bessieres P, Bossy RR et al (2006) Annotation guidelines for machine learning-based named entity recognition in microbiology. In: Proceeding of the ACL workshop on data and text mining for integrative biology, pp 40–54
Pangakis N, Wolken S, Fasching N (2023) Automated annotation with generative AI requires validation. arXiv preprint arXiv:2306.00176
Pyatkin V, Yung F, Scholman MCJ et al (2023) Design choices for crowdsourcing implicit discourse relations: revealing the biases introduced by task design. arXiv preprint arXiv:2304.00815
Richter AN, Khoshgoftaar TM (2020) Sample size determination for biomedical big data with limited labels. Netw Model Anal Health Inform Bioinform 9:12. https://doi.org/10.1007/s13721-020-0218-0
Rogstadius J, Kostakos V, Kittur A et al (2011) An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. Proc Int AAAI Conf Web Soc Media 5:321–328. https://doi.org/10.1609/icwsm.v5i1.14105
Ross J, Irani L, Silberman MS et al (2010) Who are the crowdworkers? shifting demographics in mechanical turk. In: CHI '10 extended abstracts on human factors in computing systems. Association for computing machinery New York, pp 2863–2872
Sap M, Card D, Gabriel S et al (2019) The risk of racial bias in hate speech detection. In: Association for Computational Linguistics (ed) Proceedings of the 57th annual meeting of the association for computational linguistics Florence, pp 1668–1678
Sap M, Swayamdipta S, Vianna L et al (2022) Annotators with attitudes: how annotator beliefs and identities bias toxic language detection. arXiv preprint arXiv:2111.07997
Schonlau M, Toepoel V (2015) Straightlining in Web survey panels over time. Surv Res Methods 9:125–137. https://doi.org/10.18148/srm/2015.v9i2.6128
Settles B (2009) Active learning literature survey. Computer sciences technical report 1648. University of Wisconsion-Madison
Shaw AD, Horton JJ, Chen DL (2011) Designing incentives for inexpert human raters. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. Association for computing machinery New York, pp 275–284

Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD 08. ACM Press, Las Vegas, p 614
Singer E (2011) Toward a benefit-cost theory of survey participation: evidence, further tests, and implications. J Official Stat 27(2):379–392
Suri S, Goldstein DG, Mason WA (2011) Honesty in an online labor market. Hum Comput 11(11):61–66
Thorn Jakobsen TS, Barrett M, Søgaard A, Lassen D (2022) The sensitivity of annotator bias to task definitions in argument mining. In: European Language Resources Association (ed) Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022 Marseille, pp 44–61
Tourangeau R, Kreuter F, Eckman S (2012) Motivated underreporting in screening interviews. Public Opin Q 76:453–469. https://doi.org/10.1093/poq/nfs033
Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14:e224365. https://doi.org/10.1371/journal.pone.0224365
Vaughan JW (2018) Making better use of the crowd: how crowdsourcing can advance machine learning research. J Mach Learn Res 18(1):7026–7071
Wang A, Hoang CD, Kan M-Y (2013) Perspectives on crowdsourcing annotations for natural language processing. Lang Resour Eval 47:9–31. https://doi.org/10.1007/s10579-012-9176-1
Wang Y, Mendez A, Cartwright M, Bello J (2019) Active learning for efficient audio annotation and classification with a large amount of unlabeled data. In: ICASSP 2019–2019 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, pp 880–884
Xia M, Field A, Tsvetkov Y (2020) Demoting racial bias in hate speech detection. arXiv preprint arXiv:2005.12246
Yu D, Li L, Su H, Fuoli M (2023) Using LLM-assisted annotation for corpus linguistics. arXiv preprint arXiv:2305.08339
Zhang C, Conrad F (2014) Speeding in Web Surveys: the tendency to answer very fast and its association with straightlining. Surv Res Methods 8:127–135. https://doi.org/10.18148/srm/2014.v8i2.5453
Zhang J, Sheng V, Li Q (2017) Consensus algorithms for biased labeling in crowdsourcing. Information Sciences 382–383:254–273. https://doi.org/10.1016/j.ins.2016.12.026
Zhang Z, Strubell E, Hovy E (2022) A survey of active learning for natural language processing. arXiv preprint arXiv:2210.10109
Zhao D, Wang A, Russakovsky O (2021) Understanding and evaluating racial biases in image captioning. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, pp 14810–14820

Online References

Census of India (2011) https://censusindia.gov.in/census.website/. Accessed 30 Mar 2023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.