The Sensitivity of Annotator Bias to Task Definitions in Argument Mining

Jakobsen, Terne Sasha Thorn; Barrett, Maria; Søgaard, Anders; Lassen, David Dreyer

Published in:
Proceedings of the 16th Linguistic Annotation Workshop, LAW 2022 - held in conjunction with the Language Resources and Evaluation Conference, LREC 2022 Workshop

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NC

Citation for published version (APA):
The Sensitivity of Annotator Bias to Task Definitions in Argument Mining

Terne Sasha Thorn Jakobsen¹, Maria Barrett², Anders Søgaard³, David Dreyer Lassen⁴
¹,⁴Copenhagen Center for Social Data Science, University of Copenhagen
²Department of Computer Science, IT University of Copenhagen
³Department of Computer Science, University of Copenhagen
{terne.thorn, david.dreyer.lassen}@sodas.ku.dk, mbarrett@itu.dk, soegaard@di.ku.dk

Abstract

NLP models are dependent on the data they are trained on, including how this data is annotated. NLP research increasingly examines the social biases of models, but often in the light of their training data and specific social biases that can be identified in the text itself. In this paper, we present an annotation experiment that is the first to examine the extent to which social bias is sensitive to how data is annotated. We do so by collecting annotations of arguments in the same documents following four different guidelines and from four different demographic annotator backgrounds. We show that annotations exhibit widely different levels of group disparity depending on which guidelines annotators follow. The differences are not explained by task complexity, but rather by characteristics of these demographic groups, as previously identified by sociological studies. We release a dataset that is small in the number of instances but large in the number of annotations with demographic information, and our results encourage an increased awareness of annotator bias.

Keywords: Annotation, bias, argument mining

1. Introduction

Argument mining is one of the most important and popular tasks at the intersection of natural language processing and the social sciences. Still, it suffers from “a lack of a standardized methodology for annotation” (Lawrence and Reed, 2019). Approaches to argument mining are diverse, i.e. there are various definitions of what constitutes an argument, how to assess its quality (Vecchi et al., 2021), how to model arguments, the granularity of both the input and the target, and hence how arguments are annotated for training (Lippi and Torroni, 2016). Simultaneously, what constitutes an argument may be sensitive to social biases among annotators. Such social biases have already been documented for related tasks such as fake news identification (Rampersad and Althiyabi, 2020) and stance detection (Joseph et al., 2017). One way in which annotation guidelines differ is how much evidence they require for something to be an argument, from guidelines that essentially equate claims with arguments (Morante et al., 2020) to guidelines in which evidence is a necessary component of an argument (Shmarc et al., 2020). In addition to fairness, annotation guidelines must be applicable across topics or domains (Stab et al., 2018). This paper compares how annotators from different demographic backgrounds interpret annotation guidelines of varying complexity and to what extent they subsequently agree on how to annotate for arguments. To this end, we crowd-source an argument annotation task in conjunction with demographic attributes, as visualized in Figure 1, creating a dataset of sentences with multiple annotations balanced across four argument annotation guidelines, gender, and political alignment. We show that the agreement cross-group is much lower than the agreement reported in previous work, suggesting social group differences in how guidelines are interpreted. We further demonstrate clear differences in how much group annotations vary when annotating with different guidelines, and we demonstrate the annotator bias effect on model performance, observing significant differences in performance across some groups and guidelines. We stress that bias – not disagreement – is what has to be mitigated. We need to recruit a diverse set of annotators if we are interested in a defini-
2. Task Definitions in Argument Mining

2.1. What is an Argument?

An argument consists of propositions, which are statements that are either true or false. Such statements are also commonly known as claims. An argument needs to have at least two claims, one being the conclusion, also sometimes referred to as the major claim, and at least one reason backing up the conclusion, often called the premise. Arguments are used to justify or explain claims, and argumentation is usually connected to the task of convincing or persuading others, but that need not be the purpose of any argument (Sinnott-Armstrong and Fogelin, 2014). According to Palau and Moens (2009), there are several definitions of an argument, but the (minimal) definition given above – namely that an argument is formed by premises and a conclusion made up of propositions – is common to all. The definition given here deals with explicit arguments. However, implicit arguments can be inferred from less than two propositions (i.e. only one proposition from where both the conclusion and premise can be inferred) and from sentences that are not propositions (e.g. questions and imperatives). Such implicit arguments are naturally more complex (and ambiguous) and, therefore, rarely touched in argument mining (Jo et al., 2020).

2.2. Task Definitions

NLP papers are not always explicit about what they mean by claim. Sometimes claim means conclusion, while at other times it seems to indicate either the premise or both the conclusion and premises (as both parts are formally claims/propositions). The lack of explicitness can make it difficult to compare data and systems. This section describes the definitions used in four argument mining papers and their respective guidelines that we will explore further in this study. The four papers have been chosen based on the availability of annotation guidelines, the extent to which they have been cited, and, most importantly, on the goals of the annotations being very similar, although formulated in different ways. In the following, we will underline how their definitions fit with the definition given above and each other.

Morante et al. (2020) use the term claim to refer to the conclusion and the term premise for the rest of the argument. They use the term “claim-like” to describe sentences that are either claims or premises which resemble claims and focus the annotation task on finding such claim-like sentences. They furthermore define claims as opinionated statements wrt some topic, but do not require annotators to distinguish between supporting or opposing claims.

Levy et al. (2018) define the term claim as “the assertion the argument aims to prove”. Hence, they similarly use this term to describe the conclusion. They do not mention the argument’s premises, but they use a simple annotation guideline that focuses on finding statements that clearly support or contest a given topic. In their guideline, they put forward a rule of thumb for correctly identifying such statements: “If it is natural to say ‘I (don’t) think that <topic>, because <marked statement>’, then you should probably select ‘Accept’. Otherwise, you should probably select ‘Reject’”. For this rule of thumb, the example topic is “We should ban the sale of violent video games to minors”. The example seems to contradict the earlier definition of a claim because the topic itself is a proposition (claim) that functions as a conclusion. In contrast, the statement functions as the premise of the argument. However, they work with claims under the definition of “context-dependent claims”, which explains the seeming contradiction. They define context-dependent claims as “a general, concise statement that directly supports or contests the given Topic” and require annotators to distinguish whether the claim is pro or contra a topic.

Stab et al. (2018) likewise use a context-dependent approach. Still, while Levy et al. (2018) use topics that resemble the conclusions of arguments, Stab et al. (2018) use more general topics such as “minimum wage”, that does not reflect a conclusion in itself. Unlike both Morante et al. (2020) and Levy et al. (2018) who use the word claim as the subject of interest, Stab et al. (2018) do explicitly use the word argument. They also use an additional explicit requirement in their definition of an argument: it must provide evidence or reasoning that can be used to support or contest the topic (which essentially says that there should be a claim or premise backing up another claim or conclusion). Like Levy et al. (2018), they require annotators to distinguish between supporting and opposing arguments.

Shnarch et al. (2018) use the term claim as meaning the conclusion and define the premise as a type of evidence. They work specifically with what they call evidence sentences and try to detect sentences that contain evidence that can be used to clearly support or contest a given topic. The topics are the same conclusion-like topics as Levy et al. (2018). Although detecting evidence might sound like a different task, it very much resembles the approach of Stab et al. (2018) who say that a sentence should not be accepted if it only contains a claim – some evidence must back up the claim. Since Stab et al. (2018) also accepts reasoning as sufficient backing of a claim, Shnarch et al. (2018) are a bit more strict concerning this requirement.

Annotations, annotation guidelines and code is available on www.github.com/terne/Annotator-Bias-in-Argmin

Stab et al. (2018)
We call G1 context-dependent because the topic in connection to the sentence is an integral part of the argument and evaluating stance. and inter-annotator agreement (IAA) are those reported in the respective papers. We describe G2-4 as context-dependent because the topic in connection to the sentence is an integral part of the argument and evaluating stance.

Table 1: Overview of annotation guidelines used in our experiments. Descriptions of the unmodified guidelines and inter-annotator agreement (IAA) are those reported in the respective papers. We describe G2-4 as context-dependent because the topic in connection to the sentence is an integral part of the argument and evaluating stance. We call G1 context-independent because, even though the topic is provided, it does not ask annotators to take the topic nor stance towards it into account for recognizing a claim.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Task focus</th>
<th>Guidelines</th>
<th>IAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1 Morante et al. (2020)</td>
<td>context-independent claim-like sentence detection</td>
<td><a href="https://git.io/J1OKR">https://git.io/J1OKR</a></td>
<td>F-score = 42.4 (between token-level annotations)</td>
</tr>
<tr>
<td>G2 Levy et al. (2018)</td>
<td>context-dependent claim detection</td>
<td>See Figure 8 Appendix A</td>
<td>Cohen’s $\kappa = 0.58$</td>
</tr>
<tr>
<td>G3 Stab et al. (2018)</td>
<td>context-dependent claim+premise detection</td>
<td>See Table 6 Appendix A</td>
<td>Cohen’s $\kappa = 0.721$ for two expert annotators over 200 sents. For two non-experts $\kappa \approx 0.4$</td>
</tr>
<tr>
<td>G4 Shnarch et al. (2018)</td>
<td>context-dependent claim+premise detection</td>
<td>See Figure 8 Appendix A</td>
<td>Fleiss’ $\kappa = 0.45$</td>
</tr>
</tbody>
</table>

Table 1: Overview of annotation guidelines used in our experiments. Descriptions of the unmodified guidelines and inter-annotator agreement (IAA) are those reported in the respective papers. We describe G2-4 as context-dependent because the topic in connection to the sentence is an integral part of the argument and evaluating stance. We call G1 context-independent because, even though the topic is provided, it does not ask annotators to take the topic nor stance towards it into account for recognizing a claim.

2.3. Complexity

In Table 1, we give an overview of the four studies just described and directions to their guidelines. We enumerate them and refer to their guidelines as G1-4. The order reflects the level of requirements that must be fulfilled before a sentence can be marked as a claim/argument – which we may also refer to as complexity – with G4 requiring most. While G3 and G4 require backing (premises) for claims, G2 and G1 only require claims to be present and opinionated. Before using these annotation guidelines for re-annotating data, we make some important modifications which we explain in section 4.2. Most importantly, the exact role of the context-dependency is modified such that all guidelines may work with non-conclusive topics. In Table 1, we show the agreement between annotators in the original studies, further indicating the complexity of the respective tasks.

3. Bias

In this paper, we study bias in the annotations of arguments in online debates. The ability to mine arguments for and against positions in online debates is critical in monitoring public sentiment and combating misinformation. Often such debates are controversial, associated with high engagement, and susceptible to bias. We define bias as an inclination or prejudice for or against something, e.g., groups, individuals, concepts and behaviors. The term social bias can be used in two senses: an individual’s bias which is explained by the (social) group the individual belongs to, and bias against (social) groups. The latter is typically the focus of bias studies in NLP (as in e.g., Sap et al. (2019); Rudinger et al. (2018); see also Garrido-Muñoz et al. (2021) for more bias definitions).

Men and women are known to exhibit different behavior in online communities (Sun et al., 2020), with men being more active than women (Tsai et al., 2015). There is some evidence of gender differences in both the formulation of and reasoning about arguments (Preiss et al., 2013), and overwhelming evidence of gender differences in perception and attention in general (Halpern, 2012). Similar differences in online debate behavior have been found for conservatives and liberals (Fenigberg and Willer, 2015; Chen et al., 2021), as well as differences in how arguments are perceived (Lakoff, 2006; Gampa et al., 2019). Based on this, we hypothesize that the subjective nature of the task, as well as these observations, lead to demographic differences in how arguments are annotated. Being unaware of such differences may lead to biased models. Of course, the extent to which argument annotation is subjective and susceptible to bias depends on how arguments are defined in the task definitions or annotation guidelines. Different definitions may be more or less sensitive to disparate interpretations. We expect that political alignment is likely to produce biased annotations in the annotation of arguments, partially because of what is known as the affect heuristic (Slovic et al., 2007). The affect heuristic can be described as a cognitive shortcut whereby a decision is made based on an emotional response, such as evaluating the quality of an argument based on your attitude towards it and will be predominant when the task involves a high degree of uncertainty (ambiguity).

Disparate interpretations may also result from framing effects (Tversky and Kahneman, 1981). Something that could potentially affect annotators in different ways is the degree to which a task is defined by what you should do versus what you should not do. Investigating such framing effects in detail is outside the scope of this paper and would require meticulous experiments with subtle changes in the languages. Some studies show gender differences in framing effects (Huang and Wang, 2010). Finally, Clarkson et al. (2015) found that conservatives exhibit greater self-control relative...
to liberals due to their enhanced endorsement of free will. This potentially makes conservatives more prone to confirmation bias (Baron and Jost, 2019), more reluctant to follow complex guidelines, and more reluctant to change (Salvi et al., 2016). This may partly explain our observation below that (male) conservatives disagree the most with other groups.

**Bias and fairness** Our study of bias in annotations is closely related to the concept of fairness because annotator biases could skew the representation of certain phenomena in data, which would, in turn, result in unfair treatment for some users. E.g. while an image gender classification system may struggle with classifying dark-skinned females (Buolamwini and Gebru, 2018) due to lack of representation in the data, a text classifier could struggle with potential arguments that would be treated systematically different by annotators with different backgrounds if people of both or all backgrounds are not represented among the annotators. In argument mining, this could lead to discrimination against certain ways of formulating an argument and against arguments expressing certain political viewpoints. What it actually means for a system to be fair is purely value-based, and some notions of fairness can be completely contradictory (Friedler et al., 2021). Hence, what attributes are important when investigating annotator bias depends on which aspects we value as important to be fair towards, and our beliefs about how to successfully be fair, and hence it is crucial that researchers and developers are explicit about the values their work embodies. In this study, we operationalize fairness as demographic parity wrt protected attributes that are sensitive to bias in the context of argumentation.

### 4. Experiments

#### 4.1. Modifications of guidelines

To be able to compare annotations resulting from different guidelines, some modifications of the guidelines were necessary: Firstly, G1 was changed from token-level (marking spans of claims in documents) to sentence-level annotation, and an extra task of identifying claim source was omitted. Secondly, the topics used in G2 and G4 are different from those in G3 (as described in section 2.2). The data we are using in this study is from Stab et al. (2018) (G3), where topics are short and without stance, and therefore we changed the wording of the topics in G2 and G4, such that they could work with the topics "cloning" and "minimum wage". Furthermore, in G2, we changed the wording of a rule-of-thumb and removed the underlining of claims/statements in example sentences. Thirdly, the guideline of Stab et al. (2018) is not public. Therefore we constructed a guideline based on the description in their paper and sent it to the authors who confirmed the similarity.

#### 4.2. Data collection

From the corpus created by Stab et al. (2018) for cross-topic argument mining, we re-annotated 600 sentences. The source is web documents and a wide range of text types within eight controversial topics. Of the 600 sentences we extracted from their corpus, half is from the cloning topic half from the minimum wage topic, i.e. two distant topics; one from the medical domain and one from the political domain. Each sentence was annotated following G1–4 and, within each guideline, by individuals with different demographic backgrounds.

**Demographics** We defined demographic backgrounds by gender identifications (female or male) and political alignments (liberal or conservative). Binary genders were chosen due to the lower frequency of non-binary individuals and the need for having balanced sets of annotators in this study – but when asked about their gender, respondents could choose “other”. The political alignments chosen are well suited for the dataset, which seems to consist of instances mostly discussing topics from a US perspective. Only annotators with a US nationality were invited to participate in the study. It is standard to study liberals and conservatives as opposing ideologies in a US political scene, where the large majority of the population identifies as either liberal or conservative, though with a larger part conservative.

**Process** Importantly, a meticulous process was used to balance the number of annotators and the number of sentences each annotator was given, to ensure reliable statistical tests of differences: Firstly, annotators were recruited through Prolific with the relevant demographic backgrounds and a US nationality as pre-screening conditions, and they performed the annotation.

---

4According to a recent Gallup poll [https://tinyurl.com/45nadh6z](https://tinyurl.com/45nadh6z)

4mTurk does not enable balanced recruitment across participant groups. We include an mTurk replication of our study *without balanced groups*, which served as a pilot study, in Appendix C for interested readers.
tion task in a Qualtrics survey. Annotators who passed the pre-screening were directed to the Qualtrics survey designated to annotators with their background, and here they were firstly met with a few questions on their background to confirm the pre-screening conditions and to get further information that could be confounding factors: age, ethnicity, and education. Survey question formulations followed standards from European Social Survey and US Census. Secondly, when an annotator had passed the pre-screening conditions and the confirmation of these, one of the four guidelines was presented, at random, to the annotator, followed by a set of 10 random sentences. The randomization in Qualtrics made sure each element (guideline and sentences) was presented evenly. However, when annotators left the survey without finishing, a count of the presented items would still be added and, therefore, some manual checks and new recruiting had to be done to make sure all sentences where annotated with each guideline and by an annotator of each demographic background.

**End-result** Table 2 shows that the number of annotators, and the number of sentences each annotator received, were balanced across groups and guidelines. In our final dataset, the individuals representing different demographic backgrounds are composed of between 61-66 annotators within each guideline, giving a total of 1013 annotators used in this study, as there are 4(guidelines)×4(backgrounds) set of annotations. With this process, each sentence was re-annotated a total of 16 times (and by 16 individuals). To be able to compare the annotations across both guidelines and demographics, we binarized all non-binary annotations before later model training and analysis, such that 1 equals a claim/accept/supporting argument/opposing argument, and 0 equals no claim/reject/no argument.

### 4.3. Models

We fine-tuned BERT-base on one topic and evaluated on the other using each of the 16 sets of re-annotated sentences. We used a batch size of 5, learning rate of 5e-5 and fine-tuned each model over 5 epochs and 10 random seeds (of which we took the majority label). The models were fine-tuned and tested with binarized labels.

We then fine-tuned another BERT-base and a model for multi-task learning on the entire corpus of Stab et al. (2018), the source of the re-annotated sentences, but those 600 sentences were removed from the training and validation set of the corpus before fine-tuning, leaving approx. 17,000 sentences, herein approx. 3,500 sentences from the cloning and minimum wage topics. We used Huggingface’s BertForSequenceClassification for the single-task setup, and for multi-task learning, we used Microsoft’s MT-DNN (Liu et al., 2019; Liu et al., 2020) with a pre-trained BERT-base as the main (shared) layer and eight classification heads, i.e. for each topic. Using 5 epochs, a batch size of 8, cross-entropy loss for MT-DNN, and otherwise default hyperparameters, we trained and tested each model over 10 random seeds and collected the majority predictions for analysis.

### 5. Analysis

#### 5.1. Demographic (dis)parity

We analyze the interaction between the positive rate of binarized annotations and four variables of interest: the guideline and three demographic attributes of the annotator: gender, political alignment, and age. Expectantly, positive rates differ between guidelines: the guideline containing most requirements for detecting a claim (G4) also exhibits the lowest positive rates. This holds for all annotators, but there are notable gaps between the positive rates of female/male and liberal/conservative annotations with G2–4: males and conservatives – and especially male conservatives – annotate more sentences as claims or arguments than other annotators. The following will explore the differences across demographic groups of the annotators. We analyze the per guideline difference in positive rates between all groups: female liberal (FL), male liberal (ML), female conservative (FC) and male conservative (MC), shown in Figure 3. The differences vary greatly between groups, and most importantly, they vary in

Figure 2: Interaction plots showing the interaction between variables (guideline, political alignment, gender and age) in terms of positive rate (the mean of binary labels). The plots furthermore illustrate the distribution of binary labels within demographic groups and guidelines.
a meaningful way; we observe minor differences between groups that are, from a social science empirical perspective, also more similar: female conservatives are more similar to male liberals than to male conservatives and female liberals; all groups are distant from male conservatives; male conservatives are in particular distant from female liberals. Table 3 summarizes where significant differences were found using a $\chi^2$-test. G2–4 exhibit significant differences across political spectrum and gender, and annotations with G3 and G4 also show significant differences across ages. Only G1 exhibits no significant proportional differences in labels across these three attributes. The positive rate is higher for middle-aged (31–40) annotators, and this is a bit more pronounced for conservatives. See Figure 2. Since the group of male conservative annotators are on average older than the other groups, it is reasonable to question whether age may be a mediator for the relationship between this group and its higher fraction of positive annotations. We performed a mediation analysis\footnote{Performed with statsmodels.stats.mediation.\texttt{Mediation}.} and we found that there is no mediation effect of age.

<table>
<thead>
<tr>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Political spectrum</td>
<td>ns</td>
<td>$\leq 0.01$</td>
<td>$\leq 0.0001$</td>
</tr>
<tr>
<td>Gender</td>
<td>ns</td>
<td>$\leq 0.01$</td>
<td>$&lt; 0.01$</td>
</tr>
<tr>
<td>Age</td>
<td>ns</td>
<td>ns</td>
<td>$&lt; 0.01$</td>
</tr>
</tbody>
</table>

Table 3: $p$-values from $\chi^2$-tests of differences of label frequencies given different backgrounds across the four guidelines. $\chi^2$-tests were made over contingency tables of non-binarised labels.

### 5.2. Agreement

We measure the inter-annotator agreement with Cohen’s $\kappa$ between each set of annotations from each guideline, and for all guidelines, we find the highest agreement within genders and political alignments (Figure 4). The lowest agreements are found between male conservatives and all other groups, even female conservatives. This aligns with findings in social science that female conservatives are more liberal than male conservatives (Welch, 1985; Bonica et al., 2015). We note that when measuring the agreement between females–males and liberal–conservatives (both at approx. 0.2 highest $\kappa$-score), i.e. of higher-level groups, there is a lot of information loss, including insight to considerable disagreements between female and male conservatives. We emphasize that more fine-grained knowledge of background (including more attributes) expose such hidden patterns. We also see, in Figure 4 that the agreement varies depending on guidelines. G3, based on Stab et al. (2018), has low differences in agreement. Counterintuitively, the guideline exhibiting the lowest difference in label distributions (and positive rates), i.e. G1, also shows low agreement. We include examples of sentences that were easiest to agree on (Table 7) and more difficult to agree on (Table 8) in Appendix B. In general, it seems easier to agree on sentences that clearly state a thought outcome (e.g. of raising the minimum wage). Agreeing on the stance of the argument is of course more difficult than agreeing on whether it is an argument at all. More difficult sentences to agree on seem to include factual statements, and statements with unclear stance relations, but also statements with a clear political narrative such as, “And, of course, you can also expect to hear conservatives shout back that the idea is a job killer.”

We compare our annotations to the original from Stab et al. (2018) in Figure 5. For three out of four guidelines, annotations by liberals match the original annotations best. The min-max difference in agreement is fairly equal across G2–3, with a difference of 0.2. Even though Figure 4 show that G3 has the most stable cross-group agreement, when we compare them to the original annotations, there is a clear hierarchy in the
agreements, indicating that the original annotators were likely liberal and also mostly female. The higher mean Cohen’s kappa scores may also be explained by using female, liberal annotators, as they agree most with other groups, as we saw in Figure 4.

5.3. Algorithmic bias

We have shown that annotator bias exists in the annotations of arguments. We now investigate the consequence of guideline differences and annotator bias on model performance. As described in §4.3, we firstly trained and tested models, cross-topic, on each combination of the 16 sets of annotations. Figure 5 shows the results, but here we focus on the cross-group and cross-guideline differences. We, therefore, perform student’s t-test between the sets of F1-scores (i.e. between each map in fig. 6). Models trained on data annotated using different guidelines produce significantly different cross-group performances. The bottom half of Table 4 shows that cross-group F1-scores differ significantly when comparing all guidelines except G1 and G3. The top half of Table 4 shows that cross-guideline F1-scores are significantly different when comparing the scores of models trained by annotations by male conservatives to models trained on both annotations by female conservatives as well as by female liberals. This aligns with the findings above, that male conservatives disagree more with other groups.

We then fine-tuned BERT and MT-DNN on the entire original dataset. From Figure 5, we infer that annotations from male conservatives are most likely underrepresented in the dataset of Stab et al. (2018). In effect, the large models systematically perform worse when evaluated on this group’s annotations. With BERT, we see that the min-max difference between groups is more pronounced when data is annotated using G1 and G3 (Figure 7D). G1 also stands out with MT-DNN. (See scores of both models in Table 5.) However, χ²-tests with proportions of correct and incorrect predictions of MT-DNN tell us that group differences within each guideline are only significant when including MC. I.e. differences in performance between FL, ML and FC are not significant given the same guideline. Differences between guidelines for each group are significant at the 95% significance level for all except MC.

Based on the above analysis, it seems that differences in annotator bias, depending on task definitions, cannot be simply explained by differences in guideline complexity. If this was the case, we would expect that more complex tasks, given by G3 and G4, contain more instances of ambiguity where intuition will play are larger role in the annotations. Vice versa, we would expect less intuition-lead annotations with G1 and G2. This may hold true when comparing positive rates, but when comparing agreement and model performance, differences seem to derive from annotator characteristics, with especially one demographic group standing out.

<table>
<thead>
<tr>
<th></th>
<th>Mean diff.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>0.02</td>
<td>ns</td>
</tr>
<tr>
<td>FC</td>
<td>0.16</td>
<td>≤0.001</td>
</tr>
<tr>
<td>MC</td>
<td>0.08</td>
<td>ns</td>
</tr>
<tr>
<td>FL</td>
<td>0.14</td>
<td>≤0.001</td>
</tr>
<tr>
<td>FL</td>
<td>0.06</td>
<td>ns</td>
</tr>
<tr>
<td>MC</td>
<td>-0.08</td>
<td>ns</td>
</tr>
<tr>
<td>G1</td>
<td>G2</td>
<td>-0.11</td>
</tr>
<tr>
<td>G1</td>
<td>G3</td>
<td>0.03</td>
</tr>
<tr>
<td>G1</td>
<td>G4</td>
<td>-0.21</td>
</tr>
<tr>
<td>G2</td>
<td>G3</td>
<td>0.14</td>
</tr>
<tr>
<td>G2</td>
<td>G4</td>
<td>-0.09</td>
</tr>
<tr>
<td>G3</td>
<td>G4</td>
<td>-0.24</td>
</tr>
</tbody>
</table>
6. Related Work

6.1. Evaluating argument annotation schemes

Argument annotation schemes (and specifically argument schemes that define the annotation of relations between argumentative discourse units) have been theoretically compared and evaluated extensively (Bentahar et al., 2010; Lippi and Torroni, 2018; Lawrence and Reed, 2019; Yisser et al., 2021), and to a lesser degree practically or directly, by annotating the same data with different guidelines (Habernal et al., 2014). Most related to ours, wrt practically comparing annotations deriving from different annotation guidelines, is the work of Lindahl et al. (2019) who investigate annotations of argument schemes, following the schemes by Walton et al. (2008). Here, an argument — consisting of a conclusion and a set of premises — is given an additional label reflecting the type (scheme) of the argument, such as argument from analogy, practical reasoning, or argument from consequences. They find low inter-annotator agreement in both the selected schemes and the selected conclusion and premises and observe that annotators may recognize and annotate argument conclusions, premises and types very differently, even when having expert (linguistic) knowledge.

6.2. Annotator bias

Geva et al. (2019) show that conditioning on annotator ID leads to better performance in question answering and natural language inference (NLI). Al Kuwatly et al. (2020) investigate annotator bias in hate speech classification, focusing on the role of gender, first language, age and education on annotators’ ability to identify personal attacks and on model performance and find all variables except gender to affect the annotation of hate speech. A different approach is taken by Gururangan et al. (2018) who investigate what they call annotation artifacts in NLI datasets, and they find that simple classifiers perform well when only observing the hypothe-

The challenges in identifying argument schemes and ways of improving schemes and annotation guidelines have also previously been identified by Musi et al. (2016).
6.3. Fairness

The paper contributes to the fairness literature by pointing out how group-level biases may have a severe influence on our gold standards. In our point-of-view, models should be insensitive to protected attributes such as gender and political leaning. How fairness is defined varies, with some seeing fairness as (approximately) equal positive class rates (or equal odds) (Hardt et al., 2016; Ghassami et al., 2018), and others are seeing fairness as (approximately) equal risk (Donini et al., 2018) or equal error (Zafar et al., 2017). Our study has been focused on fairness defined by demographic parity. See Williamson and Menon (2019) and Mehrabi et al. (2021) for surveys of fairness definitions.

7. Conclusion

We have shown that annotator bias is sensitive to task definitions. By re-annotating data from two domains of online debate, using four guidelines and four groups of annotators with distinctly different demographic backgrounds known to affect argumentation (political leaning and gender), we find significant differences in demographic disparity, agreement and algorithmic bias depending on both the guideline and the background of the annotators. Differences in group disparity are not explained by task complexity; instead they seem to be driven by social characteristics from the differences in demographic backgrounds.

Acknowledgments

Many thanks to Anna Rogers and Carsten Eriksen for their insightful comments.

Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN.

Table 5: $F_1$ scores of fine-tuned BERT and the multi-task learning model MT-DNN. MT-DNN is trained with the 8 topics as separate tasks, and predictions are made with the classification heads for the two topics of interest. BERT results are visualized in Figure 7.

<table>
<thead>
<tr>
<th>Guideline 1</th>
<th>Guideline 2</th>
<th>Guideline 3</th>
<th>Guideline 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lib</td>
<td>CONS</td>
<td>Lib</td>
<td>CONS</td>
</tr>
<tr>
<td>BERT</td>
<td>62</td>
<td>66</td>
<td>67</td>
</tr>
<tr>
<td>MT-DNN</td>
<td>62</td>
<td>63</td>
<td>62</td>
</tr>
</tbody>
</table>

Impact Statement

**Broader impact**  Our work shows the importance of recruiting a balanced set of annotators and considering the impact of guideline biases across different demographics. We hope this work will contribute to pushing for a more fair dataset and model development.

**Informed consent**  Annotators were informed of the overall aim of the study, to study demographics and natural language understanding, and they consented to the sharing and use of their responses for research purposes.

**Sensitive and personal information**  Responses were anonymous and voluntary. We did not ask for any information that could be reasonably linked to any individual. We present experiments with annotators that are grouped by their gender and political leaning. Annotators were also asked about their level of education and ethnicity, but since we did not balance based on these attributes, we did not include further analysis based on them. Most annotators identified as white (75%) and were college-educated (86%), which is important to keep in mind for the interpretation of our results.

**Remuneration**  Annotators were paid an average of $10.7 hourly wage.

**Intended use**  The collected annotations and demographic information will be publicly available for research purposes.

**Institutional approval**  The study is exempt from IRB approval at the authors’ institutions because it deals with anonymous responses.

8. Bibliographical References


Appendix A: Annotation guidelines

We present the guidelines used for annotating the referenced corpora either as screenshots of the actual guidelines, when these are provided by the authors or as extracts from the articles, describing the annotation rules and process. Our slightly modified guidelines are available on [www.github.com/terne/Annotator-Bias-in-Argmin](http://www.github.com/terne/Annotator-Bias-in-Argmin).

We define an argument as a span of text expressing evidence or reasoning that can be used to either support or oppose a given topic. An argument need not be “direct” or self-contained – it may presuppose some common or domain knowledge or the application of commonsense reasoning – but it must be unambiguous in its orientation to the topic. (...) unlike (other) models, which are typically used to represent (potentially deep or complex) argument structures at the discourse level, ours is a flat model that considers arguments in isolation from their surrounding context. A great advantage of this approach is that it allows annotators to classify text spans without reading large amounts of context and without considering relations to other topics or arguments. (...) Annotators classified the sentences using a browser-based interface that presents a set of instructions, a topic, a list of sentences, and a multiple-choice form for specifying whether each sentence is a supporting argument, an opposing argument, or not an argument with respect to the topic.

Table 6: Extracts from [Stab et al. (2018)](http://www.github.com/terne/Annotator-Bias-in-Argmin) describing the rules and process of annotation.

---

**Assessing the value of potential claims**

In this task you are given a topic and possibly-related statements, each marked within a particular sentence.

For each candidate, you should select "Accept", if you think that the marked statement can be used "as is" during discourse, to directly support or contest the given topic. Otherwise, you should select "Reject".

If you selected "Accept", you should further indicate whether the marked text supports the topic ("Pro") or contests it ("Con").

Note that if the marked text is non-coherent, hence cannot be used "as is" during a discussion about the topic, you should select "Reject".

Similarly, if the marked text supports/contests a different topic, even if it is somewhat related to the examined topic, you should typically select "Reject".

As a rule of thumb, if it is natural to say "I don’t think that <topic>, because <marked statement>; then you should probably select "Accept". Otherwise, you should probably select "Reject".

Finally, if you are unfamiliar with the examined topic, please briefly read about it in a relevant data source like Wikipedia.

**Examples for the topic "We should ban the sale of violent video games to minors"**

1. "The researchers found that adolescents that play violent video games are more at risk for violent behavior (but without statistical significance)." – Accept / Pro.
2. "Previous reports suggested that kids playing Doom are not at a greater risk for violent behavior." – Accept / Con.
3. "The researchers found that adolescents that play violent video games are at risk for violent behavior." – Reject. Due to the prefix "found that", the marked text is not coherent and cannot be used "as is" while discussing the topic.
4. "While violent video games are often associated with aggressive behavior, recent studies are starting to suggest otherwise." – Reject. Due to the prefix "While", the marked text is not coherent and cannot be used "as is" while discussing the topic.
5. "Many people believe that some TV shows increase youth violence." – Reject. The marked text is not directly supporting/contesting the topic.

Figure 8: Annotation guidelines of [Levy et al. (2018)](http://www.github.com/terne/Annotator-Bias-in-Argmin)
1. General instruction

In this task you are given a topic and evidence candidates for the topic. Consider each candidate independently. For each candidate please select Accept if and only if it satisfies ALL the following criteria:

1. The candidate *clearly supports* or *clearly contests* the given topic. A candidate that merely provides neutral information related to the topic should not be accepted.

2. The candidate represents a *coherent, stand-alone* statement, that one can articulate (nearly) “as is” while discussing the topic, with no need to change/remove/add more than two words.

3. The candidate represents valuable evidence to *convince one* to support or contest the topic. Namely, it is not merely a belief or merely a claim, rather it provides an indication whether a belief or a claim is true.

Note, if you are unfamiliar with the topic, please briefly read about it in a relevant data source like [Wikipedia](https://en.wikipedia.org).

Figure 9: Annotation guidelines of Shnarch et al. (2018). Besides the general instructions shown here, the guideline also includes some examples.
### Appendix B: Annotation examples

<table>
<thead>
<tr>
<th>topic</th>
<th>sentence</th>
<th>label1</th>
<th>label2</th>
<th>label3</th>
<th>label4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloning</td>
<td>God Bless you man.</td>
<td>NO CLAIM</td>
<td>Reject</td>
<td>Non-argument</td>
<td>Reject</td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Regular increases allow workers’ wages to keep pace with inflation.</td>
<td>CLAIM</td>
<td>Accept/Con</td>
<td>Supporting argument</td>
<td>Accept</td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Scarda says that the downside to a $15 minimum wage is that some minimum wage earners will lose their jobs or have their hours cut.</td>
<td>CLAIM</td>
<td>Accept/Con</td>
<td>Opposing argument</td>
<td>Accept</td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Proponents of minimum wages argue that giving workers more disposable income puts money back into the economy, which in turn creates jobs.</td>
<td>CLAIM</td>
<td>Accept/Pro</td>
<td>Supporting argument</td>
<td>Accept</td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Despite the inevitable negative outcomes that will surely result from a $15 minimum wage – we’ve already seen negative effects in Seattle’s restaurant industry – politicians and unions seem intent on engaging in an activity that could be described as an “economic death wish.”</td>
<td>CLAIM</td>
<td>Accept/Con</td>
<td>Opposing argument</td>
<td>Accept</td>
</tr>
<tr>
<td>Minimum wage</td>
<td>Raising the wage will make it more expensive to hire younger and low-skilled workers.</td>
<td>CLAIM</td>
<td>Accept/Pro</td>
<td>Opposing argument</td>
<td>Accept</td>
</tr>
</tbody>
</table>

Table 7: Examples of sentences that were easy to annotate with all guidelines, based on all annotators agreeing on whether the sentence contained a claim/argument or not. Numbering signifies instances with one disagreement wrt stance: ¹MC disagreed and chose Opposing argument; ²FL disagreed and chose Accept/Pro; ³MC disagreed and chose Accept/Pro; ⁴FC disagreed and chose Supporting argument. Agreeing on the stance of the argument is more difficult than agreeing on whether it is an argument at all.
Table 8: Lebowski-isms aside, among academics, the minimum wage debate really has become a war over arcane methodological differences.

Table 9: In cloning, the nucleus of an ordinary cell, such as skin or muscle, is placed in an egg from which the nucleus has been removed.

Table 10: The White House proposed to increase minimum wages to $10.10.

Table 11: And, of course, you can also expect to hear conservatives shout back that the idea is a job killer.
Appendix C: Mechanical Turk pilot study

In this appendix we describe the method and results of a pilot study on Amazon Mechanical Turk (mTurk), for the interested reader. In this pilot study, we learned that mTurk does not, at the time of writing, facilitate complex data collection and experiments with options to balance across attributes (demographics and guideline), randomize presented items and present them evenly among participants. When collecting annotations in a standard fashion, i.e. with none on the balancing and randomization methods, the resulting distribution of annotators is very unbalanced and there are large differences in how many items (HITs) each annotator choose to work on. This pilot motivated us to use the platforms Prolific and Qualtrics[10] for our data collection for the main study.

Data collection

We designed an MTurk survey in which annotators could self-report demographic information and express interest in a text annotation task. Based on this survey, we recruited annotators that were then presented with different annotation guidelines (the same as in the main study) and asked to annotate texts for arguments according to these guidelines across the two different domains, cloning and minimum wage.

Figure 10: On the x-axis are the four guidelines and on the y-axis are the number of annotators who annotated following a given guideline. All 600 sentences were annotated once per guideline and demographic group. Annotator demographics are not balanced per guideline, and the total number of annotators also varies across guidelines.

Figure 10 shows the number of annotators involved with annotating the 600 sentences within each guideline and demographic group. The varying number of annotators across these dimensions reflect that in some groups, more individuals were involved in annotating the 600 sentences; hence they annotated fewer sentences each, while in other groups, only a few (as little as one individual with Guideline 4 with the Female and Liberal background) participated, and hence annotated more sentences each. Annotations with Guideline 3 is the most balanced wrt. the number of annotators with backgrounds who participated. Annotators could annotate using another guideline if at least one day passed from their last annotation task using another guideline. Furthermore, they were given instructions saying it was essential that they only considered the new instructions given in the new guideline and followed these closely.

Model training

We trained a model on one topic and tested it on the other using each of the 16 sets of re-annotated sentences. We used Microsoft’s MT-DNN ([Liu et al., 2019] [Liu et al., 2020]) with a pre-trained bert-base as the main (shared) layer but trained the model with the single classification task. [11] Using 5 epochs, a batch size of 5, cross-entropy loss, and otherwise default hyperparameters, we trained and tested each model over 10 random seeds and collected the majority predictions for analysis. Table 13 show the positive rate of all predictions and Table 15 show F1 scores between the predictions and the matching guideline-group annotations.

Results

We briefly outline some of the main results from the pilot. Due to attributes not being balanced, we caution against too much interpretation of the results.

Female liberals and male conservatives disagree the most The agreement between two different groups can be calculated from our data as pairwise F1 scores and can be seen in Table 14. The agreement is generally highest within genders and political leanings. The macro-averaged agreement across the four guidelines is 0.734 between female conservatives and female liberals, but only 0.641 between male conservatives and female liberals. The agreement is 0.677 between female conservatives and male liberals.

<table>
<thead>
<tr>
<th></th>
<th>Liberal</th>
<th>Conservative</th>
<th>μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>0.650</td>
<td>0.517</td>
<td>0.363</td>
</tr>
<tr>
<td>G2</td>
<td><strong>0.805</strong></td>
<td>0.382</td>
<td><strong>0.342</strong></td>
</tr>
<tr>
<td>G3</td>
<td>0.733</td>
<td>0.487</td>
<td>0.653</td>
</tr>
<tr>
<td>G4</td>
<td>0.668</td>
<td>0.432</td>
<td>0.480</td>
</tr>
<tr>
<td>μ</td>
<td>0.714</td>
<td>0.454</td>
<td>0.638</td>
</tr>
</tbody>
</table>

Table 12: Positive rate, i.e., the fraction of sentences labeled as claims or arguments across guidelines (G1–4) and demographics, averaged over both topics. The highest value is boldfaced, lowest is underlined.

---

[10] We note that Qualtrics is a fairly costly platform and we therefore see the development of open-source JavaScripts for controlled data collection as a direction for future research which many could benefit from.

Cross-group argument mining is hard  From Table[14] we immediately see that cross-group argument mining is hard. This follows directly from the low agreement rates. We also see clear performance drops when evaluating our models across different groups. Training a model on one domain with annotations from liberal females following Guideline 1, for example, lead to an F1 score of 0.86 on the other domain (on average, across both directions), when the test data is also annotated by liberal females; for the other three groups, F1 scores drop to 0.85, 0.76, and 0.66. Similar results are observed across the other group combinations.

Table 13: Positive rate of cross-topic predictions of fine-tuned argument mining models. To understand how to read the table, take this example: the first value, 0.683, is the mean of the predictions over the minimum wage sentences by a model trained with the cloning sentences that were annotated by liberal females using Guideline 1. Highest value is boldfaced, lowest is underlined.

Table 14: Agreement between groups within guidelines calculated with F1 for the positive class. These align well with the reported inter-annotator agreement scores in the literature; see Table[1] Average agreement for Guideline 1-4 is .67, .65, .71 and .64, respectively.

Table 15: Cross-topic F1 score of fine-tuned argument mining models across different guidelines. F1-scores are for the positive class between predictions and annotations of same guideline-group combination, e.g. cross-topic predictions over the minimum wage sentences from a model trained on cloning sentences annotated by liberal females using guideline 1 are compared to the annotations for the minimum wage sentences by liberal females. Highest value is boldfaced, lowest is underlined.