Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 44–61
Marseille, 24 June 2022

The Sensitivity of Annotator Bias to Task Definitions in Argument Mining

Terne Sasha Thorn Jakobsen[1], Maria Barrett[2], Anders Søgaard[3], David Dreyer Lassen[4]
[1,4]Copenhagen Center for Social Data Science, University of Copenhagen
[2]Department of Computer Science, IT University of Copenhagen
[3]Department 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)[1]. 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; van der Linden et al., 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 (Shnarch 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-

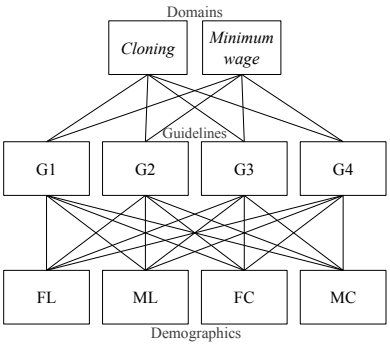


Figure 1: We re-annotate data in two *domains* across four annotation *guidelines* and four *demographics* (participant groups), as defined by binary gender (F/M) and political alignment (L/C) – to study the interaction of these three variables. We show that some guidelines promote cross-group differences and that this effect does not depend on task complexity.

[1]Lippi and Torroni (2016) identify three steps in a full argumentation mining pipeline: argumentative sentence detection, argument component boundary detection, and argument structure prediction. In this work, we focus on annotation schemes used for *argumentative sentence detection*.

tion of arguments that promote cross-group differences. All our annotations with demographic information will be publicly available along with IDs for corresponding sentences, but the sentences must be retrieved from Stab et al. (2018).[2]

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 *con*tra 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.

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

	Authors	Task focus	Guidelines	IAA
G1	Morante et al. (2020)	context-independent claim-like sentence detection	https://git.io/J1OKR	F-score $= 42.4$ (between token-level annotations)
G2	Levy et al. (2018)	context-dependent claim detection	See Figure 8, Appendix A	Cohen's $\kappa = 0.58$
G3	Stab et al. (2018)	context-dependent claim+premise detection	See Table 6, Appendix A	Cohen's $\kappa = 0.721$ for two expert annotators over 200 sents. For two non-experts $\kappa \approx 0.4$
G4	Shnarch et al. (2018)	context-dependent claim+premise detection	See Figure 9, Appendix A	Fleiss' $\kappa = 0.45$

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-*in*dependent 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 G(uideline)1-4. 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 4.1. 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 (Feinberg 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*.[3] 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

[3]Examples of the former can be found in G1, e.g., *if the text is [. . .] you should select Reject*, while G4 contains examples of the latter, e.g., *a candidate that [. . .] should not be accepted.*

	Guideline 1				Guideline 2				Guideline 3				Guideline 4				Total
	Lib		Cons		Lib		Cons		Lib		Cons		Lib		Cons		
	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	
n	65	66	61	62	66	62	62	61	65	66	62	64	61	64	63	63	1013
Avg sents	9.2	9.1	9.8	9.7	9.1	9.7	9.7	9.8	9.2	9.1	9.7	9.4	9.8	9.4	9.5	9.5	–

Table 2: The first row shows the distribution of the 1013 unique annotators of this study, and the second row shows the average number of sentences, out of 600, annotated by each individual in each annotator group.

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.[4]

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[5][6] with the relevant demographic backgrounds and a US nationality as pre-screening conditions, and they performed the annota-

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

[5] `https://www.prolific.co/`

[6] mTurk 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.

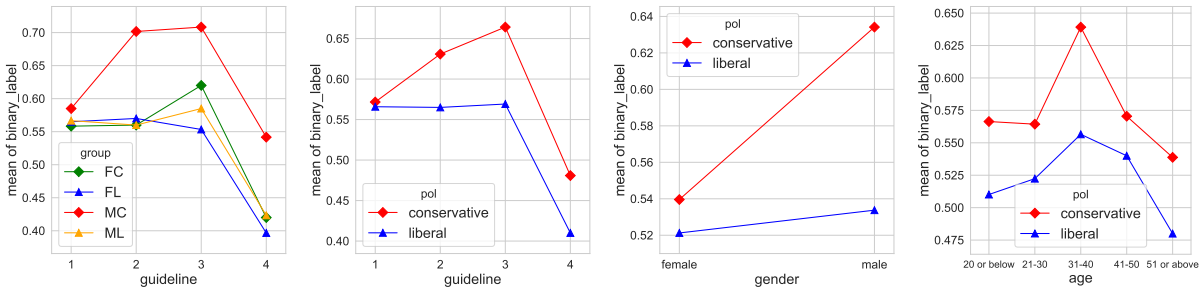


Figure 2: Interaction plots showing the interaction between variables (guideline, political alignment, gender and age) in terms of positive rate (the mean of binary labels). The plots furthermore illustrate the distribution of binary labels within demographic groups and guidelines.

tion task in a Qualtrics[7] 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

[7] https://www.qualtrics.com

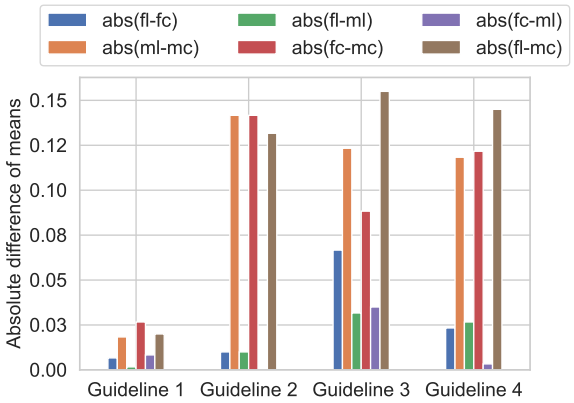


Figure 3: Absolute difference of positive rates of binarized annotations, i.e., the difference between annotator groups using the same guideline.

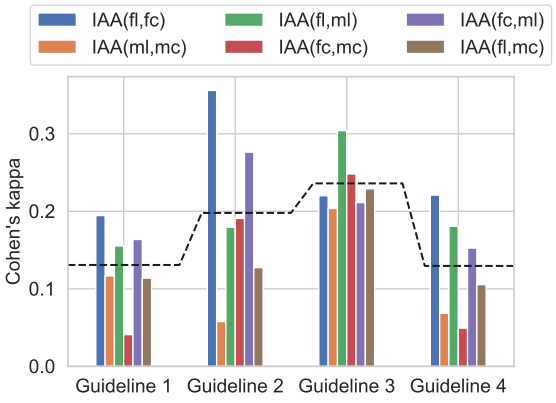


Figure 4: Agreement by Cohen's κ between the 600 (binarized) annotations from each group. The line indicates guideline means.

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 χ^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[8], and we found that there is *no mediation effect* of age.

	G1	G2	G3	G4
Political spectrum	ns	$\leq$0.01	$\leq$0.0001	$\leq$0.001
Gender	ns	$\leq$0.01	$\leq$0.01	$\leq$0.001
Age	ns	ns	$\leq$0.01	$\leq$0.0001

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

5.2. Agreement

We measure the inter-annotator agreement with Cohen's κ 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 κ-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 8-11) 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

[8]Performed with `statsmodels.stats.mediation.Mediation`.

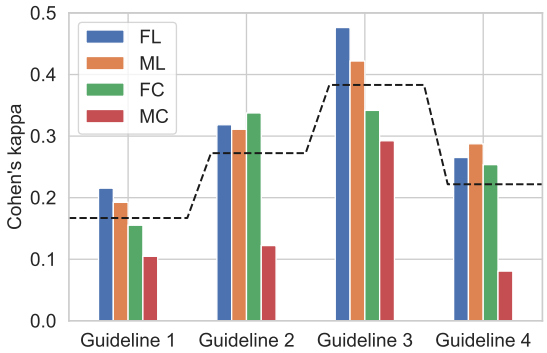


Figure 5: Agreement between the original annotations from the Stab et al. (2018) dataset and each set of our new annotations. Note that our κ-scores for G3 is higher than those reported for non-experts in Stab et al. (2018), see Table 1. This indicates that our annotation setup is generally of high quality and that low levels of agreement across groups reflect group differences rather than poor annotation conditions. We also compared our annotations to those gathered in a pilot study on mTurk, likewise finding the highest agreement with G3, with a κ-score of .34.

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 annotation 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 6 shows the results, but here we focus on the cross-group and cross-guideline differences. We, therefore, perform student's t-tests between the sets of F_1-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* F_1-scores differ significantly when comparing all guidelines except G1 and G3. The top half of Table 4 shows that *cross-guideline* F_1-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 7b). G1 also stands out with MT-DNN. (See scores of both models in Table 5.) However, χ^2-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.

		Mean diff.	p-value
FC	FL	0.02	ns
FC	MC	0.16	$\leq$0.001
FC	ML	0.08	ns
FL	MC	0.14	$\leq$0.001
FL	ML	0.06	ns
MC	ML	-0.08	ns
G1	G2	-0.11	$\leq$0.01
G1	G3	0.03	ns
G1	G4	-0.21	$\leq$0.001
G2	G3	0.14	$\leq$0.001
G2	G4	-0.09	$\leq$0.01
G3	G4	-0.24	$\leq$0.001

Table 4: We test the cross-topic performance of all pairs of annotations and perform pairwise, two-tailed student's t-test of F_1-scores, with Tukey's post hoc correction. The top half shows results from models evaluated on annotations from different guidelines (than train data), but by annotators with the same demographic attributes as train data and comparing these cross-guideline results to those of other demographic groups. The bottom half shows results from cross-group evaluations, evaluating models on annotations from a different demographic group (than train data) but using the same guideline as train data. All cross-group and cross-guideline scores are visualized in heatmaps in Figure 6.

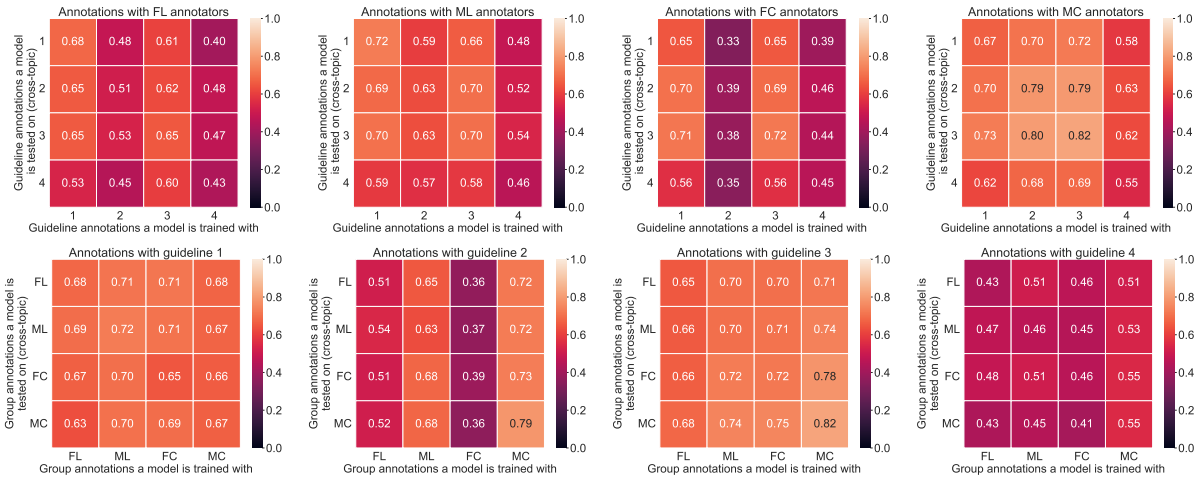
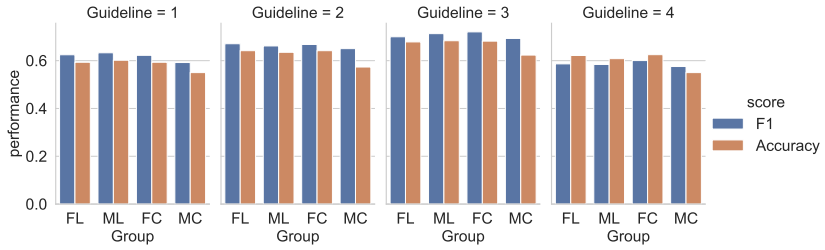
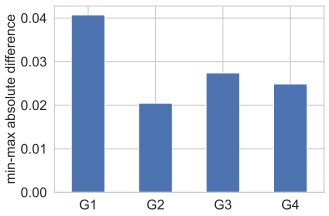


Figure 6: Cross-topic performance with binary F_1. **Top row**: evaluating models on annotations from different guidelines (than train data) but by annotators with the same demographic attributes as train data. Means from left to right: 0.55, 0.61, 0.53, 0.69. **Bottom row**: evaluating models on annotations from annotators with different demographic attributes (than train data) but from the same annotation guideline as train data. Means from left to right: 0.68, 0.57, 0.71, 0.48.



(a) Fine-tuned BERT. Baseline on original annotations F_1 = .8, accuracy = .802.



(b) BERT min-max absolute difference in F_1-scores.

Figure 7: These models are trained on all 8 topics of the dataset of Stab et al. (2018) and tested on our 300 sentences from the topics cloning and minimum wage, which we have re-annotated and removed from the training data. MT-DNN shows similar results, see Table 5.

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, 2016; Lawrence and Reed, 2019; Visser 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[9].

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-

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

	Guideline 1				Guideline 2				Guideline 3				Guideline 4			
	Lib		Cons		Lib		Cons		Lib		Cons		Lib		Cons	
	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂	♀	♂
BERT	.62	.63	.62	.59	.67	.66	.67	.65	.70	.71	.72	.69	.59	.58	.60	.58
MT-DNN	.62	.63	.60	.58	.67	.66	.66	.64	.70	.71	.69	.68	.60	.59	.60	.57

Table 5: F_1 scores of fined-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.

sis without the premise, likely due to the framing of the annotation task. Recently, Prabhakaran et al. (2021) investigated the impact of label aggregation (e.g. majority vote) on demographic biases, showing that aggregation under-represents, or ignores, a substantial number of annotators, and they encourage to release more information about annotators and transparency of selection biases. Davani et al. (2021) further tests the effectiveness of using individuals' annotations in a multi-task learning scheme and find it outperforms majority voting.

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.

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

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online, November. Association for Computational Linguistics.

Baron, J. and Jost, J. T. (2019). False equivalence: Are liberals and conservatives in the united states equally biased? *Perspectives on Psychological Science*, 14(2):292–303. PMID: 30836901.

Bentahar, J., Moulin, B., and Bélanger, M. (2010). A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33:211–259.

Bonica, A., Chilton, A. S., and Sen, M. (2015). The political ideologies of american lawyers. *Journal of Legal Analysis*, 8(2):277–335, 10.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT*.

Chen, W., Pacheco, D., Yang, K.-C., and Menczer, F. (2021). Neutral bots probe political bias on social media. *Nature Communications*, 12(5580).

Clarkson, J. J., Chambers, J. R., Hirt, E. R., Otto, A. S., Kardes, F. R., and Leone, C. (2015). The self-control consequences of political ideology. *Proceedings of the National Academy of Sciences*, 112(27):8250–8253.

Davani, A. M., D'iaz, M., and Prabhakaran, V. (2021). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *ArXiv*, abs/2110.05719.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In S. Bengio, et al., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Feinberg, M. and Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin*, 41(12):1665–1681. PMID: 26445854.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, mar.

Gampa, A., Wojcik, S. P., Motyl, M., Nosek, B. A., and Ditto, P. H. (2019). (ideo)logical reasoning: Ideology impairs sound reasoning. *Social Psychological and Personality Science*, 10(8):1075–1083.

Garrido-Muñoz, I., Montejo-Ráez, A., Martínez-Santiago, F., and Ureña-López, L. A. (2021). A survey on bias in deep nlp. *Applied Sciences*, 11:3184.

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November. Association for Computational Linguistics.

Ghassami, A., Khodadadian, S., and Kiyavash, N. (2018). Fairness in supervised learning: An information theoretic approach.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

Halpern, D. F. (2012). *Sex differences in cognitive abilities*. Psychology press, 4 edition.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, et al., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Huang, Y. and Wang, L. (2010). Sex differences in framing effects across task domain. *Personality and Individual Differences*, 48(5):649–653.

Jo, Y., Visser, J., Reed, C., and Hovy, E. (2020). Extracting implicitly asserted propositions in argumentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 24–38, Online, November. Association for Computational Linguistics.

Joseph, K., Friedland, L., Hobbs, W., Lazer, D., and Tsur, O. (2017). ConStance: Modeling annotation contexts to improve stance classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1124, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lakoff, G. (2006). *Moral Politics : How Liberals and Conservatives Think*. University of Chicago Press.

Lawrence, J. and Reed, C. (2019). Argument mining: A survey. *Computational Linguistics*, pages 765–818.

Levy, R., Bogin, B., Gretz, S., Aharonov, R., and Slonim, N. (2018). Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

Lindahl, A., Borin, L., and Rouces, J. (2019). Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence, Italy, August. Association for Computational Linguistics.

Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy, July. Association for Computational Linguistics.

Liu, X., Wang, Y., Ji, J., Cheng, H., Zhu, X., Awa, E., He, P., Chen, W., Poon, H., Cao, G., and Gao, J. (2020). The Microsoft toolkit of multi-task deep neural networks for natural language understanding. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics: System Demonstrations, pages 118–126, Online, July. Association for Computational Linguistics.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July.

Morante, R., van Son, C., Maks, I., and Vossen, P. (2020). Annotating perspectives on vaccination. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France, May. European Language Resources Association.

Musi, E., Ghosh, D., and Muresan, S. (2016). Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 82–93, Berlin, Germany, August. Association for Computational Linguistics.

Palau, R. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *ICAIL*.

Prabhakaran, V., Davani, A. M., and D'iaz, M. (2021). On releasing annotator-level labels and information in datasets. *ArXiv*, abs/2110.05699.

Preiss, D. D., Castillo, J. C., Flotts, P., and San Martín, E. (2013). Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences. *Learning and Individual Differences*, 28:193–203.

Rampersad, G. and Althiyabi, T. (2020). Fake news: Acceptance by demographics and culture on social media. *Journal of Information Technology & Politics*, 17(1):1–11.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June. Association for Computational Linguistics.

Salvi, C., Cristofori, I., Grafman, J., and Beeman, M. (2016). The politics of insight. *The Quarterly Journal of Experimental Psychology*, 69(6):1064–1072. PMID: 26810954.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July. Association for Computational Linguistics.

Shnarch, E., Alzate, C., Dankin, L., Gleize, M., Hou, Y., Choshen, L., Aharonov, R., and Slonim, N. (2018). Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605.

Shnarch, E., Choshen, L., Moshkowich, G., Aharonov, R., and Slonim, N. (2020). Unsupervised expressive rules provide explainability and assist human experts grasping new domains. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2678–2697, Online, November. Association for Computational Linguistics.

Sinnott-Armstrong, W. and Fogelin, R. (2014). *Cengage Advantage Books: Understanding Arguments: An Introduction to Informal Logic*. Cengage Learning.

Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2007). The affect heuristic. *European Journal of Operational Research*, 177(3):1333–1352.

Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium, October-November. Association for Computational Linguistics.

Sun, B., Mao, H., and Yin, C. (2020). Male and female users' differences in online technology community based on text mining. *Frontiers in Psychology*, 11:806.

Tsai, M.-J., Liang, J.-C., Hou, H.-T., and Tsai, C.-C. (2015). Males are not as active as females in online discussion: Gender differences in face-to-face and online discussion strategies. *Australasian Journal of Educational Technology*, 2015:263–277, 05.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

van der Linden, S., Panagopoulos, C., and Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, Culture & Society*, 42(3):460–470.

Vecchi, E. M., Falk, N., Jundi, I., and Lapesa, G. (2021). Towards argument mining for social good: A survey. In *ACL*.

Visser, J., Lawrence, J., Reed, C., Wagemans, J. H. M., and Walton, D. (2021). Annotating argument schemes. *Argumentation*, 35:101 – 139.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Welch, S. (1985). Are women more liberal than men in the U. S. congress? *Legislative Studies Quarterly*, 10(1):125–134.

Williamson, R. and Menon, A. (2019). Fairness risk measures. In Kamalika Chaudhuri et al., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, 09–15 Jun.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate

treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr.

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`.

(Stab et al., 2018)	*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) 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 most 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 no 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)

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.

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

topic	sentence	label1	label2	label3	label4
Cloning	God Bless you man.	NO CLAIM	Reject	Non-argument	Reject
Minimum wage	Regular increases allow workers' wages to keep pace with inflation.	CLAIM	Accept/Con	Supporting argument[1]	Accept
Minimum wage	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.	CLAIM	Accept/Con[2]	Opposing argument	Accept
Minimum wage	Proponents of minimum wages argue that giving workers more disposable income puts money back into the economy, which in turn creates jobs.	CLAIM	Accept/Pro	Supporting argument	Accept
Minimum wage	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.	CLAIM	Accept/Con[3]	Opposing argument	Accept
Minimum wage	Raising the wage will make it more expensive to hire younger and low-skill workers.	CLAIM	Accept/Pro	Opposing argument[4]	Accept

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: [1]MC disagreed and chose *Opposing argument*; [2]FL disagreed and chose *Accept/Pro*; [3]MC disagreed and chose *Accept/Pro*; [4]FC 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.

guideline	group	label
1	FL	CLAIM
	ML	CLAIM
	FC	CLAIM
	MC	CLAIM
2	FL	Reject
	ML	Reject
	FC	Accept / Con
	MC	Accept / Pro
3	FL	Non-argument
	ML	Non-argument
	FC	Non-argument
	MC	Supporting argument
4	FL	Reject
	ML	Reject
	FC	Reject
	MC	Accept

Table 8: *Lebowski-isms aside, among academics, the minimum wage debate really has become a war over arcane methodological differences.*

guideline	group	label
1	FL	NO CLAIM
	ML	CLAIM
	FC	NO CLAIM
	MC	CLAIM
2	FL	Reject
	ML	Reject
	FC	Accept / Pro
	MC	Accept / Pro
3	FL	Supporting argument
	ML	Non-argument
	FC	Non-argument
	MC	Supporting argument
4	FL	Reject
	ML	Accept
	FC	Reject
	MC	Accept

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.*

guideline	group	label
1	FL	CLAIM
	ML	CLAIM
	FC	NO CLAIM
	MC	CLAIM
2	FL	Accept / Pro
	ML	Reject
	FC	Accept / Pro
	MC	Accept / Pro
3	FL	Non-argument
	ML	Non-argument
	FC	Supporting argument
	MC	Supporting argument
4	FL	Reject
	ML	Reject
	FC	Reject
	MC	Reject

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

guideline	group	label
1	FL	CLAIM
	ML	NO CLAIM
	FC	CLAIM
	MC	NO CLAIM
2	FL	Accept / Con
	ML	Accept / Pro
	FC	Reject
	MC	Accept / Pro
3	FL	Supporting argument
	ML	Supporting argument
	FC	Non-argument
	MC	Opposing argument
4	FL	Reject
	ML	Reject
	FC	Reject
	MC	Reject

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.

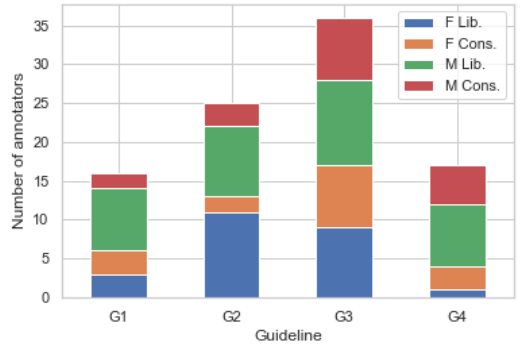


Figure 10: On the x-axis are the four guidelines and on the y-axis are the number of annotators who annotated following a given guideline. All 600 sentences were annotated once per guideline and demographic group. Annotator demographics are *not* balanced per guideline, and the total number of annotators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

[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.

	LIBERAL		CONSERVATIVE		μ
	♀	♂	♀	♂	
G1	0.650	0.517	0.690	0.363	0.555
G2	**0.805**	0.382	0.700	<u>0.342</u>	0.557
G3	0.733	0.487	0.683	0.653	0.639
G4	0.668	0.432	0.383	0.480	0.496
μ	0.714	0.454	0.638	0.460	–

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.

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.

[11]Meaning the model is comparable to simply fine-tuning bert-base.

	LIBERAL		CONSERVATIVE		μ
	♀	♂	♀	♂	
CLONING→MINIMUM WAGES					
G1	0.683	0.243	0.710	0.133	0.442
G2	0.950	0.217	0.753	0.073	0.498
G3	**0.963**	0.297	0.713	0.693	0.667
G4	0.670	<u>0.000</u>	0.133	0.217	0.255
μ	0.817	0.189	0.577	0.279	–
MINIMUM WAGES→CLONING					
G1	0.760	0.503	0.680	0.143	0.522
G2	0.783	0.183	0.543	0.137	0.412
G3	**0.977**	0.277	0.637	0.603	0.623
G4	0.603	<u>0.057</u>	0.127	0.233	0.255
μ	0.781	0.255	0.497	0.279	–

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.

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.

			LIBERAL		CONSERVATIVE	
			♀	♂	♀	♂
G1	LIB.	♀	1.000	–	–	
		♂	0.703	1.000	–	–
	CON.	♀	0.759	0.707	1.000	–
		♂	0.615	0.617	0.644	1.000
G2	LIB.	♀	1.000	–	–	–
		♂	0.612	1.000	–	–
	CON.	♀	0.855	0.647	1.000	–
		♂	0.570	0.624	0.608	1.000
G3	LIB.	♀	1.000	–	–	–
		♂	0.601	1.000	–	–
	CON.	♀	0.744	0.721	1.000	–
		♂	0.714	0.800	0.696	1.000
G4	LIB.	♀	1.000	–	–	–
		♂	0.639	1.000	–	–
	CON.	♀	0.577	0.634	1.000	–
		♂	0.665	0.687	0.651	1.000

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.

	LIBERAL		CONSERVATIVE		μ
	♀	♂	♀	♂	
CLONING→MINIMUM WAGES					
G1	0.833	0.498	0.850	0.426	0.651
G2	**0.871**	0.451	0.846	0.262	0.608
G3	0.833	0.579	0.818	0.785	0.754
G4	0.798	<u>0.000</u>	0.438	0.507	0.436
μ	0.832	0.382	0.738	0.495	–
MINIMUM WAGES→CLONING					
G1	0.862	0.656	0.825	0.413	0.689
G2	0.859	0.432	0.772	0.397	0.615
G3	0.846	0.449	0.797	0.736	0.707
G4	0.704	0.169	0.419	0.495	0.507
μ	0.818	0.427	0.703	0.510	–

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.