The Sensitivity of Annotator Bias to Task Definitions

Anonymous ACL submission

Abstract

NLP models are biased by the data they are trained on, including how it is annotated, but NLP research increasingly examines the *social biases* of models, often in the light of their training data. This paper is first to examine to what extent 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 groups, as previously identified by sociological studies.

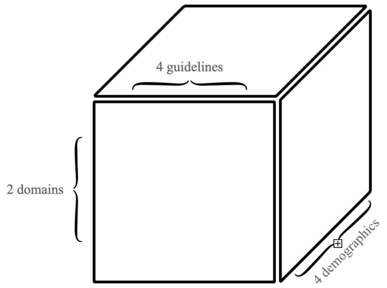


Figure 1: We re-annotate data in two *domains* across four annotation *guidelines* and four *demographics* (participant groups), as defined by political alignment and gender – to study the interaction of these three variables. We show that some guidelines promote cross-group differences and that this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

1 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).[1] Simultaneously, what constitutes an argument may be sensitive to social biases. 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). In this paper, we compare how different demographics interpret different guidelines and how they subsequently agree on how to annotate for arguments.

Contributions We use Prolific[2] to crowd-source an argument annotation task in conjunction with a demographic survey, creating a dataset of sentences with multiple annotations, balanced across four guidelines, gender, and political alignment. We show that cross-group Cohen’s kappa is significantly lower than the values reported in previous work, suggesting social differences in how guidelines are interpreted. We see clear differences in how much groups vary when annotating with different guidelines: Cross-group differences are particularly pronounced when comparing *male conservatives* with other groups, except in the case of

[1] Lippi and Torroni (2016) provide a survey of the life of argument(ation) mining in its first, approximately, ten years. They also clearly outline how diverse the approaches to argument mining are, i.e. there are various definitions of what constitutes an argument, how to model arguments, the granularity of both the input and the target, and hence how arguments are annotated for training. They 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*.

[2] mTurk does not enable balanced recruitment across participant groups. We include an mTurk replication of our study *without balanced groups* in the appendix for interested readers. Tendencies are similar, but with less support for minority participant groups.

the annotation guidelines presented in Stab et al. (2018), which is also the guideline exhibiting the highest inter-annotator agreement scores in general. We stress that bias – not disagreement – is what has to be mitigated. If we are interested in a definition of arguments that promotes cross-group differences, we need to recruit a diverse set of annotators (to avoid downstream bias), while for other task definitions, this is less important. Annotations and demographic survey responses will be made publicly available along with IDs for corresponding sentences that are from the dataset of Stab et al. (2018).

2 Task Definitions in Argument Mining

What is an Argument? An argument is made up 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).

Task Definitions NLP papers are not always explicit about what they mean by *claim*. Sometimes *claim* means conclusion, while 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 comparing data and systems tricky at times. 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, with the reasoning that: "Since premises are frequently claim-like statements and express the stance of the author, we do not exclude them from the annotation task." Morante et al. (2020) therefore 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 contraction. They define context-dependent claims as "a general, concise statement that directly supports or contests the given Topic". Therefore, they are in practice not working with claims in the form of conclusions, but instead, they are working with any claim/proposition/premise directly linked to the topic/conclusion. They 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)

No.	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 6	Cohen's $\kappa = 0.58$
G3	Stab et al. (2018)	context-dependent claim+premise detection	See Table 4	Cohen's $\kappa = 0.721$ for two expert annotators over 200 sentences, for two non-experts $\kappa \approx 0.4$
G4	Shnarch et al. (2018)	context-dependent claim+premise detection	See Figure 7	Fleiss' $\kappa = 0.45$

Table 1: Overview of annotation guidelines used in our experiments. Descriptions are of the unmodified guidelines and inter-annotator agreement (IAA) are those reported in the respective papers. G2-4 are in the appendix. 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.

use topics that resemble the conclusions of arguments, Stab et al. use more general topics such as "minimum wage", that does not reflect a conclusion in itself. Hence, in principle, the (rest of the) conclusion should be present in the sentence itself since the sentence should, in principle, contain a complete argument. Unlike both Morante et al. and Levy et al. who use the word *claim* as the subject of interest, Stab et al. 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., 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 support or clearly 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. also accepts *reasoning* as sufficient backing of a claim, Shnarch et al. are a bit more strict concerning this requirement.

Complexity In Table 1, we give an overview of the four studies and directions to their guidelines. We enumerate them and refer to their guidelines as G(uideline)1–4. We try to make the numbering reflect 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 the fourth guideline requiring most. While G3 and G4 require backing (premises) for claims, G2 and G1 only require claims to be present and opinionated. Ranking G1 and G2 is difficult; G1 is longer and has more examples than the others, but G2 requires annotators to distinguish between pro and claims con a given topic. Before using these annotation guidelines for re-annotating data, we make some important modifications which we explain in section 3.2. 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, indicating the complexity of the respective tasks.

3 Bias

In this paper, we study bias in the 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 social bias. Men and women, for example, are known to exhibit different behavior in such debates (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 writing 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. Of course, the extent to which argument annotation is subjective and susceptible to demographic bias depends on how arguments are defined in the task definitions or annotation guidelines. Different definitions will be more or less sensitive to disparate interpretations. One reason we think political alignment is likely to surface as a bias is 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 to liberals due to their enhanced endorsement of free will. This potentially makes conservatives more prone to confirmation bias (Baron and Jost, 2019) and more reluctant to follow complex guidelines and more reluctant to change (Salvi et al., 2016). This may explain our observation below that (male) conservatives disagree the most with other groups.

4 Experiments

Modifications of guidelines To be able to compare annotations resulting from different guidelines, some modifications of the guidelines were necessary: Firstly, the Morante et al. (2020) guideline (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 Levy et al. (2018) (G2) and Shnarch et al. (2018) (G4) are different from those in Stab et al. (2018) (as described earlier). The data we are using in this study is from Stab et al. (2018) (see the next section for a description of the data), 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". Specifically, in G2 and G4, the example topic (which is the same for both guidelines) "We should ban the sale of violent video games to minors" was changed to "Banning the sale of violent video games to minors". Furthermore, in G2, we changed the wording of a rule-of-thumb from *If it is natural to say "I (don't) think that <topic>, because <marked statement>", then...* to instead being *If it is natural to say "I (don't) think that <topic> is good, because <statement>", then...* and in that guideline, we also removed the underlining of claims/statements in the example sentences. Thirdly, Stab et al. (2018) have not published their guideline, and therefore we constructed a guideline based on the description in their paper and sent it to the authors who confirmed the similarity.

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 four individuals with different demographic backgrounds. We defined demographic backgrounds by gender identifications (female or male) and political alignments (liberal or conservative). This means that each sentence was re-annotated a total of 16 times. Annotators were recruited through Prolific with the relevant demographic backgrounds and a US nationality as pre-screening conditions, and they performed the annotation task in a Qualtrics survey. Annotators who passed the pre-screening were also given a survey

[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*.

on their background to confirm the pre-screening conditions and to get further information that could serve as confounding factors: age, ethnicity, and education. Survey question formulations followed well-tested standards from European Social Survey and US Census. The number of annotators, and the number of sentences each annotator received, were *balanced across groups and guidelines* (see table 5 in the appendix for more information). Lastly, while G1 and G4 asks annotators to select a binary label (1: claim or no claim, 4: Accept or reject), G2 and G3 asks for one of three possible labels (2: Accept_pro, Accept_con or Reject, 3: supporting argument, opposing argument or no argument). Therefore, to be able to compare the annotations across both guidelines and demographics, we binarized all non-binary annotations before the model training and analysis, such that 1 equals a claim/accept/supporting argument/opposing argument, and 0 equals no claim/reject/no argument.

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 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, 2020) with a pre-trained BERT-base as the main (shared) layer and eight classification heads, i.e. for 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.[4]

5 Analysis

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. Expectedly, 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 2. The differences vary greatly between groups, and most importantly, they vary in 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 2 summarizes where significant differences were found using 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 8 in the appendix. 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, therefore, performed a mediation analysis, 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 2: p-values from χ^2 tests of differences of label frequencies given different backgrounds across the four guidelines.

[4] Scripts for training and testing with MT-DNN and BERT, as well as all model outputs, are available on www.github.com/...

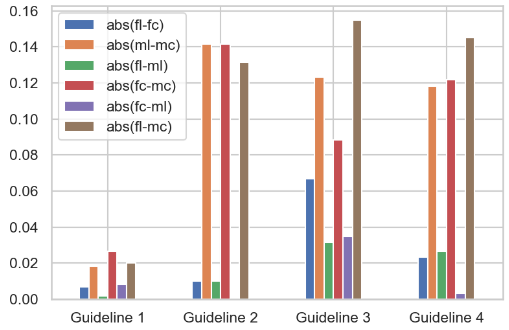


Figure 2: Absolute difference in positive rates of binarized annotations.

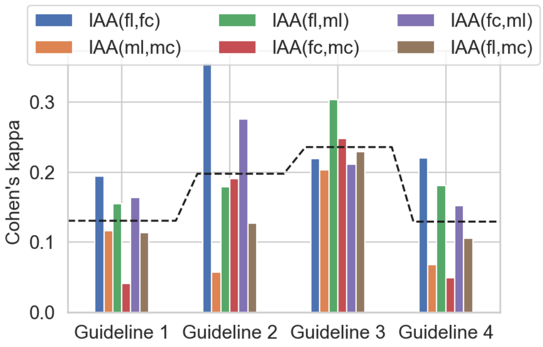


Figure 3: Agreement by Cohen's κ between the 600 (binarized) annotations from each group. The line indicates guideline means.

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 3). 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 3, 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 the appendix.

We compare our annotations to the original from Stab et al. (2018) in 4. 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 3 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 are agreeing most with other groups, as we saw in Figure 3.

Algorithmic bias We have so far 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, we firstly trained and tested models, cross-topic, on each combination of the 16 sets of annotations. Figure 9 and 10 show the results, but here we focus on the cross-group and cross-guideline differences. Models trained on data annotated using different guidelines produce significantly different cross-group performances. The bottom half of Table 3 shows that *cross-group* F_1-scores differ significantly when comparing all guidelines except G1 and G3. The top half of Table 3 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, as described in §4, fine-tuned BERT and MT-DNN on the entire original dataset. From Figure 4, 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 5b). G1 also stands out with MT-DNN. However, χ^2 tests with proportions of correct and incorrect predic-

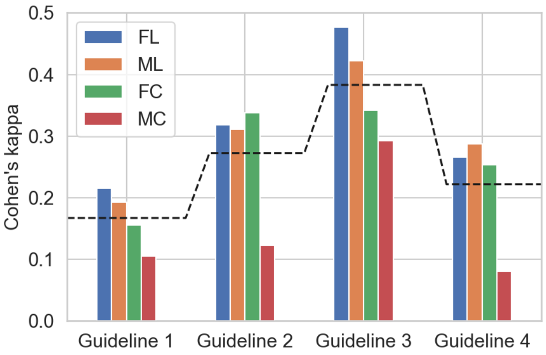


Figure 4: Agreement between the original annotations from the (Stab et al., 2018) dataset and each set of our new annotations. Note that our κ-scores for G3 (Stab et al., 2018) is higher than those reported for non-experts in Stab et al. (2018). This indicates that our annotation setup is generally of 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 findng the highest agreement with G3, with a κ-score of .34.

		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

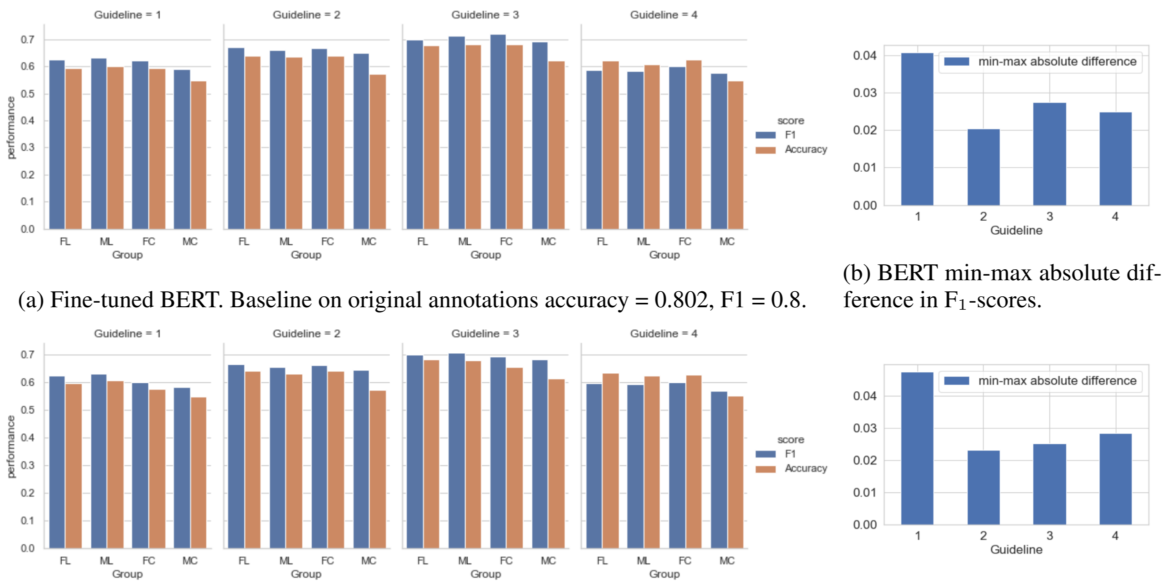
Table 3: We test the cross-topic performance of all pairs of annotations and perform pairwise, two-tailed student's t-test of F_1-scores, with Tukey's post hoc correction. The top half shows results from models evaluated on annotations from different guidelines (than train data), but by annotators with 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 can be found in the appendix in Figure 9 and 10.

tions of MT-DNN tells 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, which again stresses the bias towards this group. Differences between guidelines for each group are significant at the 95% significance level for all *except* MC. This aligns with findings from social science, described in §3, that conservatives may be more reluctant to change, and we do not see the same pattern with female conservatives.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 were 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, and 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.

6 Related Work

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, in terms of 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 constructed, 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[5].

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



(a) Fine-tuned BERT. Baseline on original annotations accuracy = 0.802, F1 = 0.8.

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

(c) MT-DNN. Baseline on original annotations: accuracy = 0.823, F1 = 0.819.

(d) MT-DNN min-max absolute difference in F_1-scores.

Figure 5: These models are trained on all 8 topics 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 is trained with the 8 topics as separate tasks, and predictions are made with the classification heads for the two topics of interest.

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 hypothesis 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.

Fairness The paper contributes to the fairness literature by pointing out how group-level biases may have a severe influence on our gold standards. Fair NLP models should be insensitive to protected attributes such as gender and political leaning. How exactly 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 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.

Ethics Statement

We present experiments with annotators that are grouped by their gender and political leaning. Annotators were also asked about the level of education and ethnicity, but since we did not balance based on these attributes, we did not include further analysis based on these attributes. We note that most annotators identified as white and were college-educated, which is important to keep in mind for the interpretation of our results. The annotators provided demographic information voluntarily and consented to the sharing of this information for research purposes. Annotators were paid an average of $10.7 hourly wage, exceeding the US federal minimum wage ($7.25). Our work shows the importance of recruiting a balanced set of annotators and considering the impact of guideline biases across different demographics. We encourage others to do further analyses using our data. We hope this work will contribute to pushing for a more fair dataset and model development.

References

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

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

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

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

Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. 2021. Neutral bots probe political bias on social media. *Nature Communications*, 12(5580).

Joshua J. Clarkson, John R. Chambers, Edward R. Hirt, Ashley S. Otto, Frank R. Kardes, and Christopher Leone. 2015. The self-control consequences of political ideology. *Proceedings of the National Academy of Sciences*, 112(27):8250–8253.

Aida Mostafazadeh Davani, Mark D'iaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *ArXiv*, abs/2110.05719.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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

Anup Gampa, Sean P. Wojcik, Matt Motyl, Brian A. Nosek, and Peter H. Ditto. 2019. (ideo)logical reasoning: Ideology impairs sound reasoning. *Social Psychological and Personality Science*, 10(8):1075–1083.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 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. Association for Computational Linguistics.

AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. 2018. Fairness in supervised learning: An information theoretic approach.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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. Association for Computational Linguistics.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In *ArgNLP*.

Diane F Halpern. 2012. *Sex differences in cognitive abilities*, 4 edition. Psychology press.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. 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. Association for Computational Linguistics.

Kenneth Joseph, Lisa Friedland, William Hobbs, David Lazer, and Oren Tsur. 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. Association for Computational Linguistics.

George Lakoff. 2006. *Moral Politics : How Liberals and Conservatives Think*. University of Chicago Press.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, pages 765–818.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081.

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

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

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 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. Association for Computational Linguistics.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The Microsoft toolkit of multitask 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. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France. European Language Resources Association.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 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. Association for Computational Linguistics.

R. Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *ICAIL*.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark D'iaz. 2021. On releasing annotator-level labels and information in datasets. *ArXiv*, abs/2110.05699.

David D. Preiss, Juan Carlos Castillo, Paulina Flotts, and Ernesto San Martín. 2013. Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences. *Learning and Individual Differences*, 28:193–203.

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

Carola Salvi, Irene Cristofori, Jordan Grafman, and Mark Beeman. 2016. The politics of insight. *The Quarterly Journal of Experimental Psychology*, 69(6):1064–1072. PMID: 26810954.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 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.

Eyal Shnarch, Leshem Choshen, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 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. Association for Computational Linguistics.

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

Paul Slovic, Melissa L. Finucane, Ellen Peters, and Donald G. MacGregor. 2007. The affect heuristic. *European Journal of Operational Research*, 177(3):1333–1352.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 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. Association for Computational Linguistics.

Bing Sun, Hongying Mao, and Chengshun Yin. 2020. Male and female users' differences in online technology community based on text mining. *Frontiers in Psychology*, 11:806.

Meng-Jung Tsai, Jyh-Chong Liang, Huei-Tse Hou, and Chin-Chung Tsai. 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.

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

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

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

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

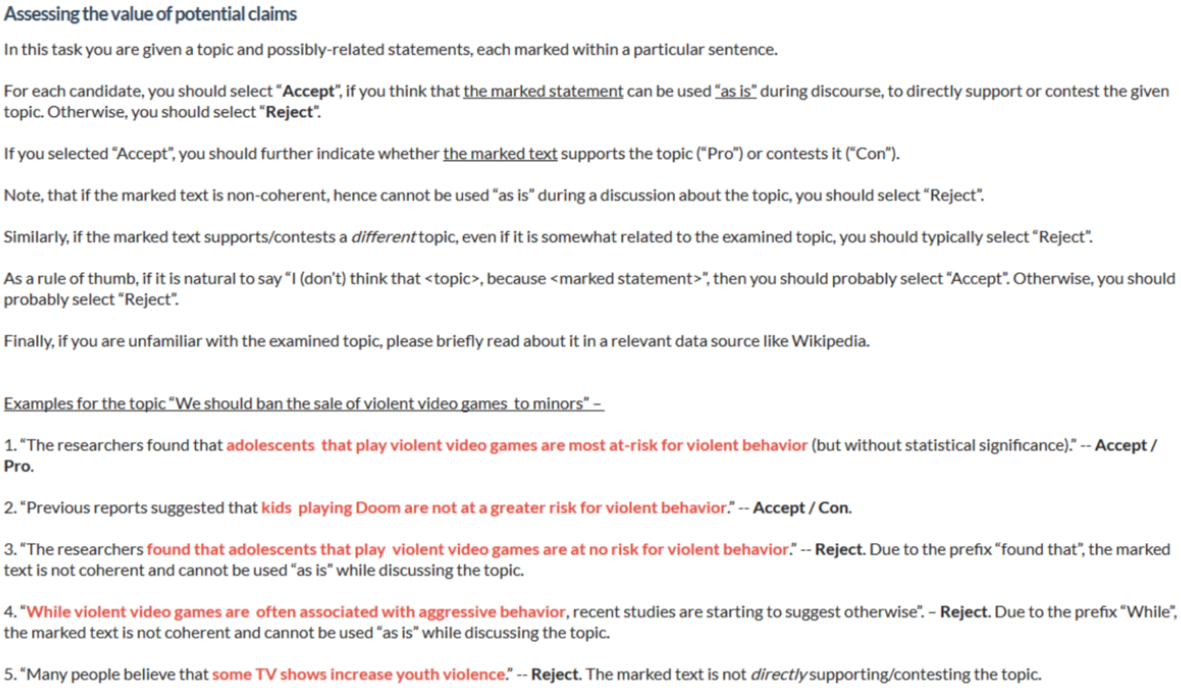
Susan Welch. 1985. Are women more liberal than men in the U. S. congress? *Legislative Studies Quarterly*, 10(1):125–134.

Robert Williamson and Aditya Menon. 2019. Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*.

A Appendix

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. The slightly modified guidelines are available on www.github.com/....



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.**

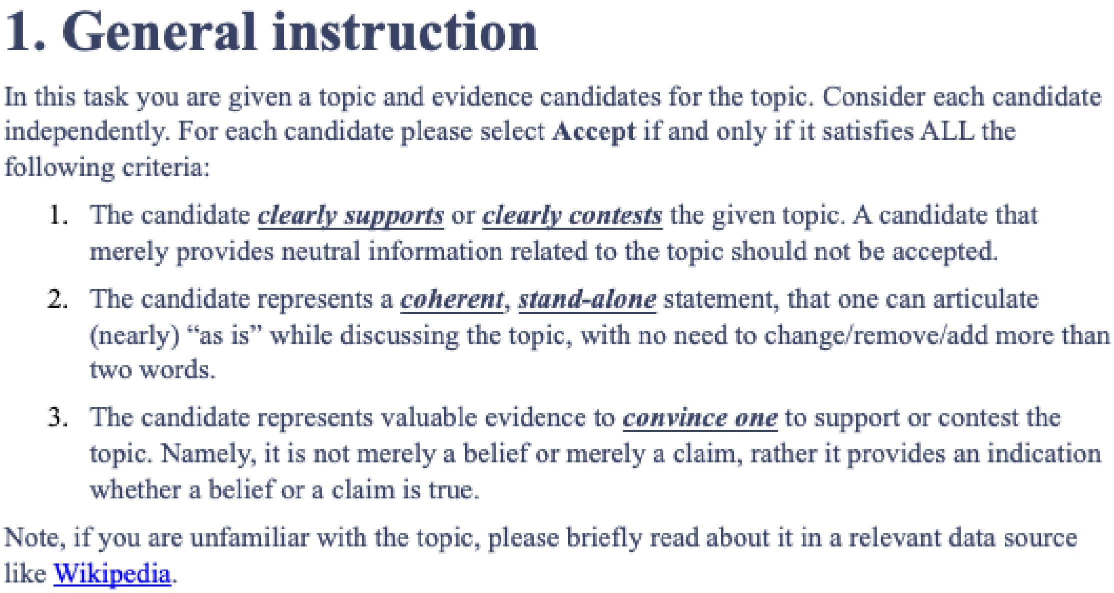
2. "Previous reports suggested that kids playing Doom are not at a greater risk for violent behavior." -- **Accept / Con.**

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 6: 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 7: Annotation guidelines of Shnarch et al. (2018). Besides the general instructions shown here, the guideline also includes some examples.

(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 4: Extracts from Stab et al. (2018) describing the rules and process of annotation.

			No. annotators	Avg. sent
G1	Lib.	♀	65	9.2
		♂	66	9.1
	Con.	♀	61	9.8
		♂	62	9.7
G2	Lib.	♀	66	9.1
		♂	62	9.7
	Con.	♀	62	9.7
		♂	61	9.8
G3	Lib.	♀	65	9.2
		♂	66	9.1
	Con.	♀	62	9.7
		♂	64	9.4
G4	Lib.	♀	61	9.8
		♂	64	9.4
	Con.	♀	63	9.5
		♂	63	9.5
			1013	9.5

Table 5: The table shows the number of annotators and the average number of sentences annotated by each annotator, given a guideline and a demographic background. The number of annotators and sentences are balanced. Sentences were randomized.

label	Non-argument	Opposing argument	Supporting argument
pol			
conservative	403	372	425
liberal	517	360	323

Table 6: Example of a contingency table for G3, with label proportions given the political alignment attribute. Contingency tables and χ^2-tests were made for each guideline and attribute of interest (political alignment, gender and age).

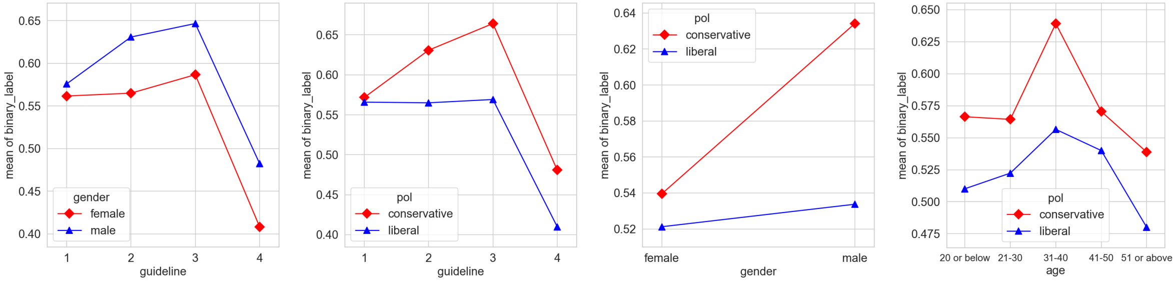


Figure 8: Interactions between variables (guideline, political alignment, gender and age) in terms of positive rate (the mean of binary labels).

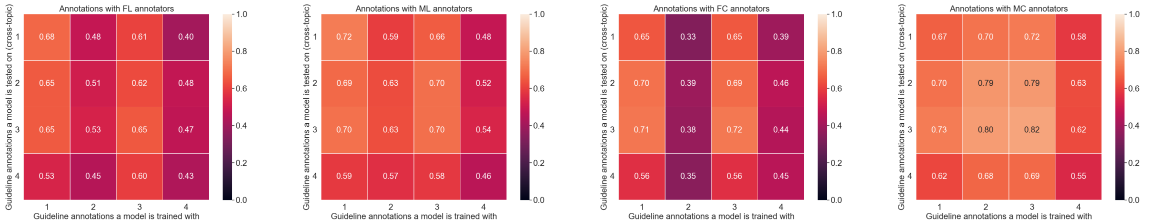


Figure 9: Cross-topic performance with binary F_1 – evaluating models on annotations from different guidelines (than train data) but by annotators with the same demographic attributes as train data. Means from left to right: 0.55, 0.61, 0.53, 0.69.

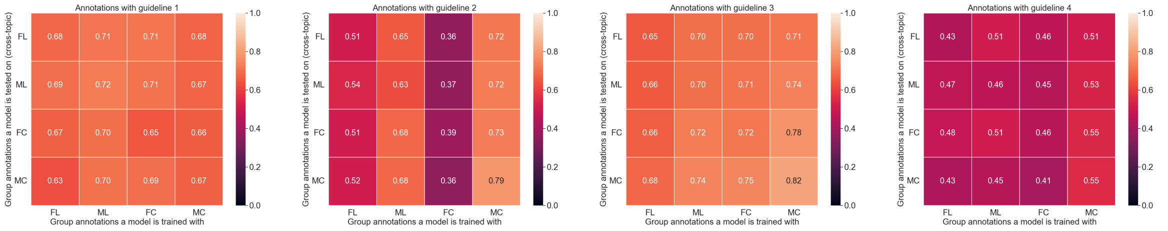


Figure 10: Cross-topic performance with binary F_1 – evaluating models on annotations from annotators with different demographic attributes (than train data) but from the same annotation guideline as train data. Means from left to right: 0.68, 0.57, 0.71, 0.48.

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.

		label
guideline	group	
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.*

		label
guideline	group	
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.*

		label
guideline	group	
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.*

		label
guideline	group	
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.*