University of Groningen

Differential Bias

Palomino, Alonso; Potthast, Martin; Al-Khatib, Khalid; Stein, Benno

Published in:
Findings of the Association for Computational Linguistics

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

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

Publication date:
2022

Differential Bias: On the Perceptibility of Stance Imbalance in Argumentation

Alonso Palomino[1]   **Martin Potthast**[1]   **Khalid Al-Khatib**[2]   **Benno Stein**[3]

[1] Leipzig University `<first>.<last>@uni-leipzig.de`
[2] University of Groningen `khalid.alkhatib@rug.nl`
[3] Bauhaus-Universität Weimar `benno.stein@uni-weimar.de`

Abstract

Most research on natural language processing treats bias as an absolute concept: Based on a (probably complex) algorithmic analysis, a sentence, an article, or a text is classified as biased or not. Given the fact that for humans the question of whether a text is biased can be difficult to answer or is answered contradictory, we ask whether an "absolute bias classification" is a promising goal at all. We see the problem not in the complexity of interpreting language phenomena but in the diversity of sociocultural backgrounds of the readers, which cannot be handled uniformly: To decide whether a text has crossed the proverbial line between non-biased and biased is subjective. By asking "Is text X more [less, equally] biased than text Y?" we propose to analyze a simpler problem, which, by its construction, is rather independent of standpoints, views, or sociocultural aspects. In such a model, bias becomes a preference relation that induces a partial ordering from least biased to most biased texts without requiring a decision on where to draw the line. A prerequisite for this kind of bias model is the ability of humans to perceive relative bias differences in the first place. In our research, we selected a specific type of bias in argumentation, the stance bias, and designed a crowdsourcing study showing that differences in stance bias are perceptible when (light) support is provided through training or visual aid.

1 Introduction

Bias is a multifaceted phenomenon that can be wittingly or unwittingly introduced into language. Walton (1999) traces the term's history with argumentation and concludes that bias implies one-sidedness. Van Laar (2007) distinguishes two forms of biased argumentation, namely the exclusion of arguments of the pro or con position ("stance bias") and the exclusion of arguments discussing a certain aspect (frame) that is relevant to an issue. Figure 1 illustrates the spectrum of stance
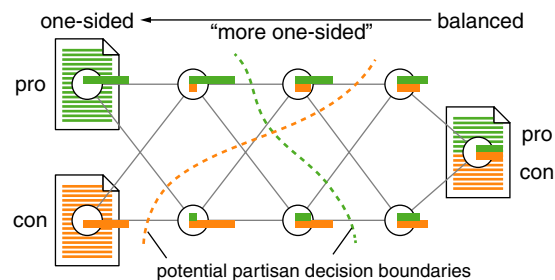

Figure 1: Stance bias induces a partial ordering of argumentations, the extremes being purely one-sided (left) or perfectly balanced (right). Where supporters of one side or the other draw the line between biased and non-biased depends on their degree of conviction.

balance from a one-sided (left) to a balanced (right) situation. In between, an argument may be perceived as one-sided (biased) despite the inclusion of arguments from the other side if no sufficient balance is maintained. Nevertheless, a cleverly chosen stance imbalance can also serve as a rhetorical device for persuasion (Walton, 1999). Regardless of the actual balance of their arguments, both sides may accuse the other of bias due to one-sidedness.

The diversity of sociocultural backgrounds, environmental conditioning, or educational attainment makes it difficult to treat bias as an absolute concept for binary classification, except for extreme cases. Granted, when focusing on the relation between texts regarding their bias, one takes a step back both in terms of problem difficulty and predication. But, trying to model the "differential nature of bias" (differential bias) is a valid strategy to eliminate individual, subjective factors. By developing a measure of argumentative balance with respect to the stance (stance balance) that induces a partial ordering of argumentative texts, a gradual preference relation as shown in Figure 1 is established. As a consequence, the evaluation of one-sidedness cannot be decided for a single text but is constrained to insights from answers to the question: "Is text X more [less, equally] one-sided than text Y?"

411

Using stance as an example, we investigate for the first time the extent to which differential bias is perceptible to human annotators—an important prerequisite for practical applications of debate technologies. Based on a model differential stance bias (Section 3), 720 human preference judgments are collected in a carefully designed crowdsourcing study (Section 4). Our analysis of the judgments shows that extreme imbalance is perceptible, whereas more subtle imbalance is not unless (mild) training is provided (Section 5).[1] This result is important for annotating language bias in general, for argument search engines, and for developing curricula for argument analysis (Section 6).

2 Related Work

After a brief review of research on bias in natural language processing (NLP), we survey bias and one-sidedness in argumentation, and the application of pairwise judgments in corpus construction.

2.1 Bias in Natural Language Processing

Blodgett et al. (2020) survey 146 papers that study various forms of bias, finding that "quantitative techniques for measuring or mitigating 'bias' are [often] poorly matched to their motivations." Sheng et al. (2021) survey 90 papers on societal biases in language generation that tackle gender, profession, race, religion, and sexuality, among which gender bias stands out as the most frequently studied form of bias. The literature review of 61 papers by Sun et al. (2019) focuses explicitly on recognizing and mitigating gender bias in NLP, concluding that the subfield still lacks a shared understanding, standardization, as well as evaluations that demonstrate the generalizability of current techniques. Shah et al. (2020) study bias formally and focus on how and where it is introduced into an NLP pipeline (e.g., semantic bias in embeddings, label and selection bias in data sources). Their survey of 93 papers overviews suggested countermeasures. Bender and Friedman (2018) proposes the use of "data statements" as a means to raise awareness of ethical issues among authors, which the ACL board has meanwhile taken up, and Hovy and Spruit (2016) outlines the ethical implications and impacts NLP systems have on society.

Although (social) bias has attracted much attention, to the best of our knowledge, only Spliethöver and Wachsmuth (2020) explicitly study social bi-

ases in argumentation, showing that current argument corpora are biased "in favor of male people with European-American names." However, two types of non-social biases have been studied more in-depth because of their relevance for argumentation: cognitive biases and one-sidedness bias.

2.2 Cognitive Biases in Argumentation

Huang et al. (2012) show that exposing decision makers—who exert confirmation bias via selective reading habits—to counterarguments will improve their decision outcomes. Wright et al. (2017) propose a visual "argument mapping" aid for intelligence analysts, which organizes arguments and counterarguments to better manage cognitive biases such as confirmation bias, anchoring, framing effects, and neglect of probability. These biases have recently also been studied by Kiesel et al. (2021), who discuss the challenges and opportunities of developing argumentative conversational search engines. For educators who teach argument evaluation, Diana et al. (2019, 2020) develop a measure that predicts if student assessments of argument strength are affected by confirmation bias; the measure is based on the alignment of individuals' values with values in political arguments. (Amorim et al., 2018) find that confirmation bias impacts peer-evaluation of student essays, which then propagate into automatic essay scoring systems.

2.3 One-Sidedness Bias in Argumentation

The following studies are of particular relevance to our contributions because of their focus on perceptions of bias in argumentation. Walton (1999) and Van Laar (2007) argue that bias in argumentation implies one-sidedness because a biased argumentation typically fails to be balanced, favoring one side of a topic, aspect, or stance over others. One-sidedness is a phenomenon that influences the perception of information seekers: Schlosser (2011) and Chen (2016) investigate the relation between reviewers' expertise, their bias, and product type when evaluating the utility of online product feedback. The authors observe that customers typically consider one-sided reviews more helpful than two-sided reviews because, in one-sided reviews, users usually consider aspects such as the expertise of the reviewer to be more significant in their purchase decision. Wolfe et al. (2013) analyze the process of students in understanding one-sided arguments: In two experiments that focus on measuring read-

---

[1]Code and data: https://github.com/webis-de/AACL-22

ing times and analyzing how students summarize neutral texts, the authors find that the processing of one-sided arguments is based on a "belief bias." Kienpointner and Kindt (1997) study one-sidedness of argumentation to understand the antagonistic climate of political debates about political assylum. They examine which aspects are included or excluded (global bias) and what strategies are used to address certain aspects (local bias).

Operationalizations of one-sidedness in argumentation have been contributed by Potthast et al. (2018) and Kiesel et al. (2019), who study writing-style-based approaches in order to detect hyper-partisan news; they organized a shared task that received more than 40 submissions. Stab and Gurevych (2016) choose a semantics-based approach to predict the presence or absence of opposing arguments in an argumentative text. Both approaches target only the extreme cases of one-sidedness (i.e., only pro or only con arguments; see Figure 1 on the far left). More generally, Küçük (2021) formulates the problem (without its operationalization) of predicting the ratio of pro, con, and neutral stances in argumentative corpora. Our work complements theirs by first answering the question whether stance bias can be perceived at all.

2.4 Relative Judgments in Argumentation

Kienpointner and Kindt (1997) point out that argument bias assessment requires judgment in relation to the argumentative context because of their complex properties. Lacking an objective reference, another piece of argumentation can be used instead. Two recent studies follow this approach: Habernal and Gurevych (2016) propose the task of predicting the convincingness of arguments, where given a pair of arguments with the same stance on an issue, the more convincing one must be chosen. From many such comparisons a global ranking can be statistically derived. Gienapp et al. (2020) apply the same approach to annotate argument quality while minimizing the number of comparisons required.

Beyond argumentation, (Howcroft et al., 2020) employ a relative judgment of summaries to rank them based on quality criteria, such as fluency and readability. Simpson et al. (2019) utilize pairwise comparisons to infer labels for humorous and metaphoric texts, and Gooding et al. (2019) do so to annotate words. Cattelan (2012) survey such analyses beyond NLP and computer science.

Text X on "Tidal Energy"	Text Y on "Tidal Energy"
Anchors of tidal energy systems can damage ecosystems. They might damage salt marshes, estuaries, and near-shore reefs, or alter the natural processes that maintain ocean and coastal ecosystems, such as the movement of sand, silt, animals, and larvae.	Many forms of energy have potential environmental impacts. The focus should be on designing regulations that minimize or eliminate these impacts. Regulations can minimize the environmental impact of tidal energy.
Some fishing communities worry that their fishing grounds could be disturbed. Toxic hydraulic fluids might leak. Tidal energy could disrupt local fishing industries.	Bays and estuaries are always naturally "flushed" or cleansed and replenished by tide-waters flowing in and out. To the extent that tidal energy impairs the natural flows of these tides, it impairs this natural "flushing" mechanism. This can alter and even destroy an ecosystem.
While there may still be some concerns regarding the safety and local environmental impact of nuclear energy, nuclear energy releases 0-emissions and so is an important part of the fight against global warming. It should be pushed forward; tidal energy should not "replace" it.	Tidal energy can slow the movement of water in a bay or estuary, which reduces the amount of kinetic energy and causes the body of water to freeze-over more often or for longer periods of time. This has consequences for marine eco-systems. Tidal energy can lead to undesirable prolonged winter icing.
There is only one major tidal generating station in operation. This is a 240 megawatt station at the mouth of the La Rance river estuary on the northern coast of France (a large coal or nuclear power plant generates about 1,000 MW of electricity). Tidal energy generates very little energy comparatively.	Offshore turbines do not alter the flow of tides as much as barrages can, so they have a smaller environmental impact. Offshore turbines don't hamper he flow of tides.

Figure 2: Which of the two argumentations (X vs. Y, left vs. right) is more biased? Answer is given below.

3 Measuring Differential Bias Perception

To measure the perceptibility of differential stance bias in argumentation in a controlled environment, we develop a basic experiment as shown in Figure 2. For two argumentations, X and Y, on the same topic, the central question to participants is: "Which of the two argumentations is more biased?" With a corpus of topic-labeled arguments at hand, the construction of pairs of argumentations for a given topic is straightforward. First, the number of pro arguments for X and Y is selected, each ranging from zero to four arguments. A corresponding number of pro arguments are randomly selected and randomly assigned to the available slots. Then, the remaining slots are randomly filled with random con arguments on the same topic. The two argumentations are shown side by side to be read one after the other and then compared. No 1:1 correspondence is to be assumed between arguments at the i-th position, $i = 1, \ldots, 4$, and therefore arguments are not aligned across argumentations.

Alternative experimental setups were explored as part of a pilot study to find the simplest ("atomic") setup to investigate our experiment variables of interest. In the following, an overview these variables is given (D = dependent, I = independent, C = controlled, U = uncontrolled):
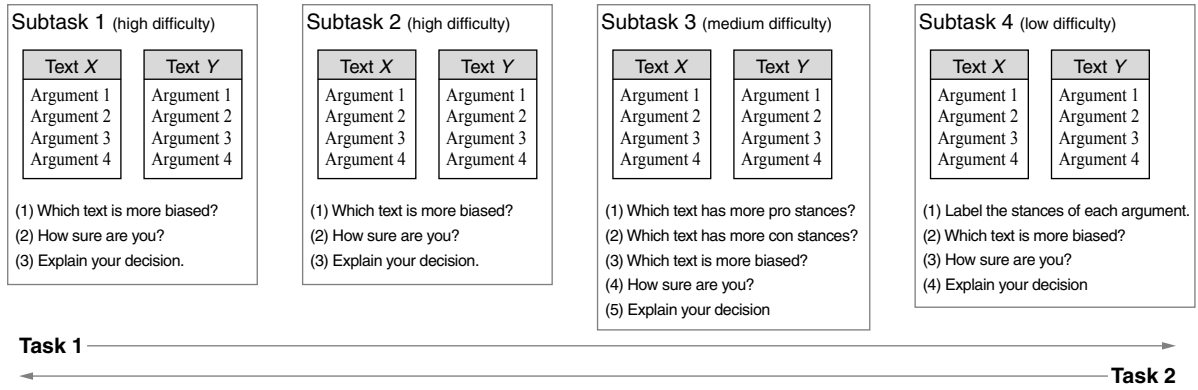
Figure 3: Two tasks organize the experiments in decreasing and increasing order of difficulty, respectively.

D perceptibility of differential stance bias

I_1 differential stance bias: 3 levels
I_2 difficulty: low, medium, high
I_3 participant training: autodidact vs. trained
I_4 participant expertise: stance labeling accuracy
I_5 participant confidence: 3-level self-assessment

C_1 argumentation length: 4 arguments each
C_2 argument length: 2-3 sentences each
C_3 argument frames: all arguments from 1 frame
C_4 order effects I: random order of arguments
C_5 order effects II: no textual coherence
C_6 opinion diversity: 9 participants per topic

U_1 stance perceptibility: e.g., explicit vs. implicit
U_2 other language biases: e.g., at the lexical level

Key to measuring D, the perceptibility of differential stance bias, is a model for independent variable I_1: Let X denote an argumentation on a given topic that argues pro or con a proposition, where $X = X^+ \cup X^-$ is the union of a set of pro arguments X^+ and con arguments X^-, so that $X^+ \cap X^- = \emptyset$. The stance balance δ between pro and con arguments in X is measured as the absolute size difference between X^+ and X^-:

$$\delta(X) = \left| |X^+| - |X^-| \right|.$$

The closer δ is to 0, the more balanced the two sides are in X. Given a second argumentation Y, $|X| = |Y|$, the differential stance bias can then be quantified as the absolute difference between the stance balance scores of X and Y:

$$\Delta(X,Y) = |\delta(X) - \delta(Y)|.$$

The closer Δ is to 0, the less strong the differential bias is between X and Y. Given the designated argumentation length of $|X| = |Y| = 4$, the image of both δ and Δ is $\{0, 2, 4\}$.

Consider the "Tidal Energy" example in Figure 2: Argumentation X is an extreme case with only con arguments, argumentation Y is the balanced case with two pro (first, fourth) and two con arguments (second, third). This yields $\delta(X) = 4$ and $\delta(Y) = 0$ and thus $\Delta(X,Y) = 4$, the maximum possible differential bias in this setting. It follows that X is more biased than Y. Suppose Y consists of only pro arguments instead ($\delta(Y) = 4$): Although X and Y argue exclusively on either the pro or the con side, both argumentations are equally imbalanced, so that $\Delta(X,Y) = 0$. The intermediate case of $\Delta(X,Y) = 2$ results when Y includes three pro and a single con argument or vice versa.

Regarding independent variable I_2 (see Figure 3), high difficulty means no indication is given about argument stance or stance balance before asking which of the two argumentations is more biased. Medium difficulty means we first ask which argumentation contains more pro arguments and which contains more con arguments before asking to select the more biased one—a subtle hint at stance balance. Low difficulty means the stance of each argument has to be labeled first as pro, neutral, con, or unknown, coloring each argument according to the chosen label. In this way, stance is strongly emphasized when answering the question which argument is more biased.

Experiments are presented in either decreasing (Task 1) or increasing (Task 2) order of difficulty (I_3; see Figure 3). Task 2 is also preceded by an explanation of the importance of stance balance in relation to bias, as well as an admission test for argument stance labeling that must be successfully completed. Note the different learning experiences participants have as they move through the two tasks. Two groups of participants are distinguished (I_4): those who succeed in stance label-

414

ing and those who get it wrong more than once. Also, participants are asked to self-assess their confidence (I_5).

The variables C_1-C_6 are tightly controlled by setting a specific value or by introducing diversity through randomization. Text lengths (C_1 and C_2) are kept as short as possible, and each argument has a similar length to the others, so that confounding by length differences is avoided. C_3 specifically avoids mixing up stance bias with Van Laar's second type of one-sidedness in argumentation with respect to topic frames. C_5 is due to the requirement to generate argumentative texts rather than reusing existing ones, although the texts are generally not extremely incoherent either. Variables beyond our control include the difficulty of distinguishing pro and con stances in the argument corpus (some arguments are more subtly pro or con than others; U_1), and (social) language biases that may be due to word choice (U_2). Nevertheless, we manually checked for hate speech and removed corresponding arguments, as well as analyzed for biased terminology using dictionaries.

4 Crowdsourcing Study

Our crowdsourcing study implements the assessment of the perceptibility of differential stance bias on Amazon Mechanical Turk (MTurk). We give an overview of the task design and its instantiation using an argument corpus.

4.1 Argument Corpus

Arguments from the Webis-Argument-Framing-19 corpus (Ajjour et al., 2019) were used to populate the tasks. It includes 12,000 arguments on 465 topics sourced from debatepedia.org. Each argument is labeled with its topic, pro or con stance, and frame. Table 1a shows three examples. Topics with fewer than 8 arguments of comparable length available with the same frame label were excluded (C_2, C_3). From the remaining topics, 40 argument pairs were randomly generated (see Figure 2). Table 1b shows the distribution of argument pairs across the six differential bias cases implied by our model (I_1).

To exercise "limited control" over possible linguistic biases (U_2), we use the Linguistic Inquiry and Word Count (LIWC) 2022 (Tausczik and Pennebaker, 2010),[2] a dictionary-based tool for identifying psycholinguistic, thematic, and tonal properties of language. All selected arguments score low on tone and swearing. Nevertheless, we manually reviewed each argument for quality and to ensure that they do not contain offensive language.

4.2 Study Execution

We created 20 "Human Intelligence Tasks" (HITs) on MTurk, 10 for each of the two tasks, and distributed the 40 argument pairs to them as subtasks (see Figure 3). A total of 571 workers participated, each of whom was allowed to participate only once. Repeated participation would have undermined the learning experience of a task and thus the results achieved. One task took 45 minutes and was paid at 6 USD or 8 USD per hour.

The tasks are implemented in the form of a dialog, one subtask per step. For quality control (see below) JavaScript code was inserted to monitor the workers. As in Figure 2, a pair of arguments was presented side by side in columns as titled boxes with borders. The dialog could only be worked through forward, using JavaScript to ensure that workers complete a subtask first before moving on. Moving back was prevented retrospective revisions, which is particularly relevant for Task 1, where subtasks are arranged in decreasing order of difficulty (I_2). Since nine workers (C_6) worked independently on the same subtask, the arguments in each argumentation X and Y were randomly shuffled for each worker (C_4).

The following measures were taken for quality control: (1) work times: exclusion of workers who took less than 3 minutes to read instructions or less than 8 minutes to complete a subtask; (2) MACE score (Hovy et al., 2013): exclusion of workers whose score is below 80%; (3) approval rate: exclusion of workers whose MTurk approval rate was below 90%; and (4) language proficiency: exclusion of workers who submitted many grammatical or spelling errors or random text in the mandatory comment fields for decision rationales in each subtask. For Task 2, an admission test had to be passe. Finally, we ensured that each task contained subtasks with ($\Delta(X, Y) = 0$) and without ($\Delta(X, Y) > 0$) differential stance bias. In this way, workers who are able to distinguish the two types can be distinguished from those who cannot.

---

[2]https://www.liwc.app/

(a)

Labels		Argument
Topic:	Abortion	No individual has rights over another individual. Therefore, a fetus cannot be said to have an inviolable right to a woman's body and sustenance from that body. A woman can, therefore, decide to deprive the fetus of the usage of her body (abortion). A fetus cannot have a right to a woman's body to sustain its life.
Stance:	Pro	
Frame:	Fetus rights	
Topic:	Death penalty	The U.N. does not support the death penalty. In all the courts we have set up (U.N. officials) have not included death penalty. United Nations oppose the death penalty.
Stance:	Con	
Frame:	Internat. law	
Topic:	Capitalism vs. Socialism	This type of helpful framework neglects appeals for human rights and any other framework of deontology, morality, ethics, etc. Capitalism can embrace the utilitarian framework while not precluding any form of decision calculus in policymaking to protect human rights. Socialist leadership cannot protect human rights effectively.
Stance:	Pro	
Frame:	Rights	

(b)

Differential Bias		Experiments	
$\Delta(X,Y)$	$\delta(X){:}\delta(Y)$	HITs	Workers
0	0:0	6	54
0	2:2	7	63
0	4:4	6	54
2	2:0 / 0:2	7	63
2	4:2 / 2:4	7	63
4	4:0 / 0:4	7	63

(c)

Bias	Precision		Accuracy
$\delta(X){:}\delta(Y)$	Pro	Con	
0:0	0.80	0.90	0.84
2:2	0.80	0.66	0.77
4:4	0.61	0.40	0.62
4:2 / 2:4	0.44	0.35	0.44
4:0 / 0:4	0.80	0.62	0.76
2:0 / 0:2	0.52	0.60	0.55
Avg.	0.66	0.59	0.66

Table 1: (a) Examples of arguments from the top three most frequent topics in the Webis-Argument-Framing-19 corpus. (b) The stance distributions and the number of instances considered for the crowdsourcing study. (c) The crowd worker precision and accuracy for labeling pro/con arguments.

5 Results and Analysis

This section reports workers' perception accuracy for differential stance bias with respect to the independent variables discussed in Section 3.

5.1 Perceptibility of Differential Stance Bias

Table 2a shows workers' overall perception accuracy of differential stance bias (I_1). Argument pairs without differential stance bias $\Delta(X,Y) = 0$) are less well perceived as such, with an average of 0.42 in Task 1 and 0.57 in Task 2, than argument pairs with differential stance bias ($\Delta(X,Y) > 0$; average of 0.55 in Task 1 and 0.72 in Task 2). The highest accuracy is obtained for the 2:0 / 0:2 distribution (0.56 and 0.76 in Tasks 1 and 2) and the lowest for the 0:0 distribution (0.46 and 0.51 in Tasks 1 and 2). The higher a differential stance bias, the easier it is for workers to perceive it.

Furthermore, we find that workers who are successful at labeling the stance of individual arguments (I_4; part of subtasks of low difficulty) are also more successful at perceiving differential stance bias. Table 1c shows the workers' precision and accuracy for stance labeling. Only the pro and con arguments are displayed. Workers who could not decide (label "unknown") or label an argument as neutral are excluded from further analysis, since all arguments are either pro or con. An average

accuracy of 0.62 is achieved for stance labeling, ranging from 0.44 to 0.84 depending on the differential bias distribution. Table 2d shows that workers with high competence in stance labeling (accuracy > 0.7) achieve a perception accuracy for differential stance bias of 0.59 in Task 1 and 0.67 in Task 2, but workers with low competence achieve only 0.31 in Task 1 and 0.42 in Task 2.

An examination of whether workers' educational attainment (collected by questionnaire) affects perception accuracy is negative. The highest degrees attained are high school (20%), bachelor (60%), and master or higher (20%), and the Pearson correlations range from -0.08 to 0.0 for Task 1 and 0.01 to 0.0 for Task 2.

5.2 Dependence on Difficulty and Training

Table 2c shows perception accuracy as a function of difficulty (I_2) and difficulty order (I_3). In both tasks, perception accuracy increases the lower the difficulty of the subtask. The accuracy at high difficulty is 0.38 for Task 1 and 0.58 for Task 2. Accuracy at medium difficulty is 0.55 for Task 1 (0.17 improvement over high difficulty) and 0.67 for Task 2 (0.09 decrease below high difficulty). At low difficulty, workers achieve the highest accuracy of 0.64 in Task 1 and 0.74 in Task 2. Although an improvement in perception accuracy can be observed between high and low difficulty, workers in

(a)

Differential Bias		Accuracy per Subtask and Difficulty							
$\Delta(X,Y)$	$\delta(X):\delta(Y)$	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4
0	0:0	0.27	0.33	0.58	0.66	0.60	0.55	0.42	0.50
0	2:2	0.33	0.28	0.56	0.57	0.62	0.67	0.56	0.54
0	4:4	0.11	0.36	0.44	0.64	0.70	0.64	0.52	0.50
2	2:0 / 0:2	0.41	0.60	0.55	0.68	0.88	0.86	0.62	0.70
2	4:2 / 2:4	0.44	0.44	0.55	0.65	0.77	0.69	0.66	0.59
4	4:0 / 0:4	0.45	0.55	0.68	0.70	0.88	0.66	0.75	0.66

(b)
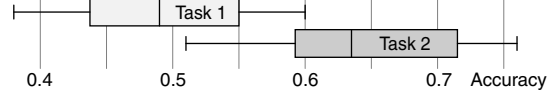
(c)

Average Accuracy per Task and Subtask Difficulty

Sec.	Type	Acc.	Δ Acc.	Sec.	Type	Acc.	Δ Acc.
1.1 / 1.2	1	0.38	0.00	2.1	3	0.74	0.00
1.3	2	0.55	0.17	2.2	2	0.67	−0.07
1.4	3	0.64	0.09	2.3 / 2.4	1	0.58	−0.09

(d)

Differential Bias		Accuracy per Worker Competence Level															
		Task 1								Task 2							
		Low				High				Low				High			
$\Delta(X,Y)$	$\delta(X):\delta(Y)$	1.1	1.2	1.3	1.4	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	2.1	2.2	2.3	2.4
0	0:0	0.11	0.00	0.18	0.28	0.37	0.55	0.66	0.53	0.55	0.44	0.33	0.24	0.61	0.63	0.55	0.66
0	2:2	0.14	0.33	0.42	0.11	0.44	0.22	0.52	0.55	0.43	0.25	0.22	0.12	0.70	0.65	0.50	0.55
0	4:4	0.00	0.26	0.33	0.24	0.68	0.44	0.66	0.48	0.50	0.33	0.48	0.44	0.62	0.55	0.66	0.58
2	2:0 / 0:2	0.20	0.44	0.55	0.35	0.72	0.66	0.63	0.68	0.66	0.60	0.50	0.54	0.80	0.75	0.58	0.72
2	4:2 / 2:4	0.48	0.45	0.42	0.28	0.55	0.77	0.60	0.66	0.52	0.57	0.55	0.33	1.0	0.88	0.70	0.66
4	4:0 / 0:4	0.55	0.36	0.48	0.60	0.66	0.62	0.74	0.76	0.61	0.46	0.56	0.20	0.75	0.68	0.68	0.64
Average		0.25	0.31	0.40	0.31	0.57	0.54	0.63	0.61	0.54	0.44	0.40	0.29	0.74	0.69	0.61	0.63

Table 2: Perception accuracy of differential stance bias (a) dependent on subtask, and subtask difficulty (high: 1.1, 1.2, 2.1, and 2.2; medium: 1.3 and 2.3; low: 1.4 and 2.4), (b) dependent on task, (c) dependent on task and subtask difficulty, and (d) dependent on worker competence at argument stance labeling.

Task 1 seem to be confused by the change from high to medium difficulty. However, the relative improvement in accuracy between high and low difficulty is the same for both tasks.

The perception accuracy differs in absolute terms between Task 1 (high to low difficulty) and Task 2 (training, then low to high difficulty). Table 2b shows that workers achieve an accuracy of 0.38 to 0.60 with an average of 0.52 in Task 1 and an accuracy of 0.51 to 0.76 with an average of 0.64 in Task 2. This accuracy difference (about 0.15) suggests that there is a positive effect of training workers to perform more difficult subtasks, rather than expecting them to detect differential stance bias off hand. The difference is significant according to Welch's t-test, with a test statistic of 4.28 and a p value of 0.001.

5.3 Worker Confidence and Feedback

Participants had to self-assess their confidence in each subtask as "very sure," "reasonably sure," and "uncertain" (I_5). When workers gave either of the first two self-assessments, they have higher perception accuracies, albeit dependent on of the presence of differential stance bias. For Task 1, an accuracy of 0.58 is achieved in cases with differential stance bias, and an accuracy of 0.69 for Task 2. The lowest accuracies are observed when workers self-assess their confidence as "not sure", where the accuracy falls to 0.42 in Task 1 and 0.65 in Task 2.

For a qualitative assessment of the workers' abilities, we asked them to justify their decision as to which of the two argumentations is more biased. In reviewing the collected justifications, we found that some workers cited other types of bias in addition to stance bias. Table 3 shows selected justifications of the workers on subtasks with different differential bias distributions, depending on whether the workers chose the more biased argumentation correctly, or incorrectly. While one worker argues about the emotional tone of the argument chain on a subtask with differential bias distribution of 4:4, another worker addresses the factuality of the arguments on a subtask with a distribution of 4:0/0:4.

Differential Bias		Decision Justification		
$\Delta(X,Y)$	$\delta(x):\delta(y)$	Topic	Correct	Incorrect
0	0:0	Globalization	I'm not completely sure about the second paragraph in text X but I think overall both texts have 2 con and 2 pro paragraphs: X: pro-con-con-pro Y: con-pro-con-pro	It is a complex issue and at least one argument on both texts could be constructed as in favor or against globalization, depending on the reader's initial stance.
0	2:2	Adult Incest	Both texts have pros and cons regarding incest.	Text Y features three pro arguments, whereas text X has an equal share of pro and con arguments.
0	4:4	Animal Testing	Both X and Y have an even amount of pro and con arguments. The first argument in Y is in a slight gray area, but it seems to be against the spirit of animal testing overall so. I counted it as a con.	X favors the side of animal life being sacred more and a more emotional argument chain
2	2:0 / 0:2	Banning Cell Phones in Cars	Y posts three points on the good cellphones in cars can do and though it does present the cons they are vastly outweighed in the argument	Because X is more pro arguments
2	4:2 / 2:4	Tidal Energy	Text Y is more one-sided as it explains more cons of tidal energy than pros of it, whereas text X has an equal number of pros and cons.	Y is talking about megawatts and energy and anchors and cables while X speaks in a Reader's Digest form that most are able to understand
4	4:0 / 0:4	Atheism	X only makes arguments against atheism and tries to justify acts of religion. Y gives quite logical arguments and doesn't seem to be very one sided.	X seems more to add their own opinion, while Y explains more of the facts.

Table 3: Examples of worker decision justifications dependent on differential bias.

Some workers base their decisions on the perceived degree of subjectivity or objectivity of their arguments. Others argue about the constructiveness, persuasiveness, rationality, and verifiability of the arguments. Comparing the justifications in Task 1 with those in Task 2, many workers in Task 1 stick to their initial explanations used in the first sub-tasks. In contrast, workers in Task 2 more often realize that the study is about stance balance.

6 Conclusions and Future Work

How one-sided can an argumentation be without this one-sidedness going unnoticed? In political debates, parties often accuse each other of one-sidedness. An advanced debating skill hence is to subtly shift the balance of an argumentation so that it goes unnoticed, especially by potential voters. But political debate is perhaps only one of the most extreme cases in which skewed argumentation can be found, as Kienpointner and Kindt (1997) point out in their in-depth analysis of letters to the editor regarding the topic of political asylum. We believe that a perfectly balanced argument is the exception rather than the rule because, after all, the goal is to convince the audience of one's own opinion. For this reason, audiences must try to be alert to (subtle) manipulation opportunities, and here formal training to detect differences in stance bias seems a promising future direction. The results of our study indicate that differences in stance bias are perceptible, and even more so when (as in our crowdsourcing setting) workers are trained with a visual aid in the form of green/red signals that emphasize the stance of the arguments. In an online setting where arguments are read at leisure, a tool that highlights the distribution of stances could help readers make more informed decisions, not to mention formal training in the context of a debate.

Our study further demonstrates, to our knowledge for the first time, the viability of assessing bias via relative judgments rather than absolute ones. In this regard, we focus on stance bias for its ease of operationalization. It is likely that many other types of bias can be similarly addressed in future work, although not all of them can be operationalized as easily as in our study. In any case, the obvious ben-

efits of reducing the subjective bias of annotators certainly justify further investigation and the development of a theory in this direction. This should, of course, include research on the interactions of biases, since not all biases can be easily controlled. For instance, some workers based their judgments not only on differences in stance, but also pointed to other language properties that, in their opinion, made one argumentation seem more one-sided than another. Although we took great care to exclude examples of language bias, subtle biases may still have been present in our sample. A straightforward and fascinating application of our differential bias model is to extend it to the study of frame bias.

Ethical Statement

Our study involved human annotators recruited via the Amazon Mechanical Turk (MTurk) crowdsourcing platform. In accordance with fair compensation guidelines for crowd workers, we ensured an appropriate hourly rate. As is common with Mturk, some worker accounts attempted to submit fake data. We reviewed all cases that met our study's exclusion criteria regarding rejection of submitted work. As part of the quality control process, in addition to open rejection, we also introduced "internal rejection," which resulted in acceptance of the submitted work with full payment but exclusion from our analysis. We decided to play it safe and discard internally in ambiguous cases. All workers were informed about the study and agreed to the quality controls and compensation conditions before work began.

One limitation of our study is that it focuses on differences between stance biases. We are aware that several factors play a role in making an argumentation biased such as the frame of an argumentation, logical connections between arguments, (lack of) cohesion and strength of arguments, or social biases resulting from loaded language. However, it remains an open problem to define the semantics of these biases. Of course, laboratory experiments are always limited in this or similar ways, and our experiment is no different. We do not foresee any unethical uses of our results or its underlying tools, but hope that it will contribute to advancing the discourse on bias, and help to depart from from absolute bias claims.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling Frames in Argumentation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP 2019)*, pages 2922–2932. ACL.

Evelin Amorim, Márcia Cançado, and Adriano Veloso. 2018. Automated Essay Scoring in the Presence of Biased Ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 229–237. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Trans. Assoc. Comput. Linguistics*, 6:587–604.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics.

Manuela Cattelan. 2012. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433.

Ming-Yi Chen. 2016. Can Two-sided Messages Increase the Helpfulness of Online Reviews? *Online Information Review*, 40(3):316–332.

Nicholas Diana, John C. Stamper, and Ken Koedinger. 2019. Predicting Bias in the Evaluation of Unlabeled Political Arguments. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 1640–1646. cognitivesciencesociety.org.

Nicholas Diana, John C. Stamper, and Ken Koedinger. 2020. Towards Value-Adaptive Instruction: A Data-Driven Method for Addressing Bias in Argument Evaluation Tasks. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–11. ACM.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient Pairwise Annotation of Argument Quality. In *58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 5772–5781. Association for Computational Linguistics.

Sian Gooding, Ekaterina Kochmar, Advait Sarkar, and Alan Blackwell. 2019. Comparative judgments are more consistent than binary classification for labelling word complexity. In *Proceedings of the 13th*

Linguistic Annotation Workshop, pages 208–214, Florence, Italy. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H. Hovy. 2013. Learning Whom to Trust with MACE. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1120–1130. The Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Hsieh-Hong Huang, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku. 2012. Understanding the Role of Computer-mediated Counter-argument in Countering Confirmation Bias. *Decision Support Systems*, 53(3):438–447.

Manfred Kienpointner and Walther Kindt. 1997. On the problem of bias in political argumentation: An investigation into discussions about political asylum in germany and austria. *Journal of pragmatics*, 27(5):555–585.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 Task 4: Hyperpartisan News Detection. In *13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 829–839. Association for Computational Linguistics.

Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *3rd Conference on Conversational User Interfaces (CUI 2021)*, New York. ACM.

Dilek Küçük. 2021. Stance Quantification: Definition of the Problem.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. In *56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 231–240. Association for Computational Linguistics.

Ann E. Schlosser. 2011. Can Including Pros and Cons Increase the Helpfulness and Persuasiveness of Online Reviews? The Interactive Effects of Ratings and Arguments. *Journal of Consumer Psychology*, 21(3):226–239.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5248–5264. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4275–4293. Association for Computational Linguistics.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from Old Man's View: Assessing Social Bias in Argumentation. In *7th Workshop on Argument Mining*, pages 76–87. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the Absence of Opposing Arguments in Persuasive Essays. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1630–1640. Association for Computational Linguistics.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Jan Albert Van Laar. 2007. One-Sided Arguments. *Synthese*, 154(2):307–327.

Douglas Walton. 1999. *One-sided Arguments: A Dialectical Analysis of Bias*. SUNY Press.

Michael B. Wolfe, Shawna M. Tanner, and Andrew R. Taylor. 2013. Processing and Representation of Arguments in One-Sided Texts About Disputed Topics. *Discourse Processes*, 50(7):457–497.

William Wright, David Sheffield, and Stephanie Santosa. 2017. Argument Mapper: Countering Cognitive Biases in Analysis with Critical (Visual) Thinking. In *21st International Conference Information Visualisation, IV 2017, London, United Kingdom, July 11-14, 2017*, pages 250–255. IEEE Computer Society.