Argument from Old Man's View: Assessing Social Bias in Argumentation

Maximilian Spliethöver
Department of Computer Science
Paderborn University
Paderborn, Germany
mspl@mail.upb.de

Henning Wachsmuth
Department of Computer Science
Paderborn University
Paderborn, Germany
henningw@upb.de

Abstract

Social bias in language — towards genders, ethnicities, ages, and other social groups — poses a problem with ethical impact for many NLP applications. Recent research has shown that machine learning models trained on respective data may not only adopt, but even amplify the bias. So far, however, little attention has been paid to bias in computational argumentation. In this paper, we study the existence of social biases in large English debate portals. In particular, we train word embedding models on portal-specific corpora and systematically evaluate their bias using WEAT, an existing metric to measure bias in word embeddings. In a word co-occurrence analysis, we then investigate causes of bias. The results suggest that all tested debate corpora contain unbalanced and biased data, mostly in favor of male people with European-American names. Our empirical insights contribute towards an understanding of bias in argumentative data sources.

1 Introduction

Social bias can be understood as implicit or explicit prejudices against, as well as unequal treatment or discrimination of, certain social groups in society (Sweeney and Najafian, 2019; Papakyriakopoulos et al., 2020). A social group might be described by physical attributes of its members, such as sex and skin color, but also by more abstract categories, such as culture, heritage, gender identity, and religion. A typical, probably in itself biased, example of social bias is the old man's belief in classic gender stereotypes. In most cases, social bias is deemed negative and undesirable.

Recent research shows that bias towards social groups is also present in Machine Learning and Natural Language Processing (NLP) models (Chang et al., 2019), manifesting in the encoded states of a language model (Brown et al., 2020) or simply causing worse performance for underrepresented classes (Sun et al., 2019). Such bias has been studied for different NLP contexts, including coreference resolution (Rudinger et al., 2018), machine translation (Vanmassenhove et al., 2018), and the training of word embedding models (Bolukbasi et al., 2016). In contrast, Computational Argumentation (CA) has, to our knowledge, not seen any research in this direction so far. Given that major envisioned applications of CA include the enhancement of human debating (Lawrence et al., 2017) and the support of self-determined opinion formation (Wachsmuth et al., 2017), we argue that studying social bias is particularly critical for CA.

In general, social bias may affect diverse stages of CA: In argument acquisition, for example, researchers may introduce social bias unintentionally, for instance, by collecting arguments from web sources that are only popular in a certain part of the world. This is known as *sample bias* (Chang et al., 2019). In argument quality assessment, a machine learning model may develop a *prejudicial bias* and judge arguments made by a certain social group better, for instance, because it considers features inadequate for the task, such as the gender (Jones, 2019). And in argument generation, a model might produce arguments that have an *implicit bias* towards a certain social group, for instance, because the features chosen are based on prior experience of the researchers and may not properly represent the whole population (Fiske, 2004). As the examples indicate, social bias can, among other reasons, be caused by the source data and how it is being processed. As a starting point, this paper therefore focuses on social bias in the source data underlying CA methods. In particular, we ask the following questions:

Proceedings of the 7th Workshop on Argument Mining, pages 76–87
Barcelona, Spain (Online), December 13, 2020.

1. What, if any, types of social bias are present in existing argument sources and how do different sources compare to each other in this regard?
2. How much do certain groups of users contribute to the overall social bias of a source?
3. What kinds of linguistic utterances contribute towards certain types of social bias?

Applications such as those outlined above usually rely on web arguments for scaling reasons. To study the questions, we hence resort to five English debate portals that give access to arguments on versatile topics: *4forums.com*, *convinceme.net*, *createdebate.com*, *debate.org*, and *ChangeMyView*. All five have been deployed in CA corpora (Abbott et al., 2016; Durmus and Cardie, 2019; Al Khatib et al., 2020).

First, we analyze the general presence of social bias in each of the five debate portals. To this end, we train three custom word embedding models, one for each available corpus. Next, we evaluate the models for social bias using a widely used bias metric, called WEAT (Caliskan et al., 2017), and compare the results. We then inspect the debate.org portal more closely with regard to specific social groups. In particular, we group the texts based on the provided user information and apply the same evaluation. Lastly, to gain a better understanding of what makes some texts more biased than others, we explore their language by analyzing word co-occurrences with group identity words in the texts.

Our findings suggest that all three corpora are generally biased towards male (compared to female) and European-American (compared to African-American) people. This bias is not only reflected in the WEAT results, but also in unbalanced occurrences of identity words for certain social groups. More generally, we observe that the use of names as identity terms for social groups has unpredictable effects. With those insights, we contribute an initial understanding of social bias in sources of dialectical argumentative texts.

2 Related Work

In recent years, research on different types of bias in natural language has received a considerable amount of attention. Media bias is one prominent example (Fan et al., 2019), particular the political bias of news articles (Chen et al., 2020). In various sub-fields of NLP, studies on media bias are concerned with analyzing techniques utilized by media outlets when reporting news. These include the framing of an event by phrasing the report with positive or negative terms, and selective reporting by including or omitting facts depending on the tone and (political) stance a media outlet wants to convey (Chen et al., 2018; Hamborg, 2020; Lim et al., 2020). We do not target media bias here but *social* bias.

Social bias can emerge from pre-existing stereotypes towards any social group (Sweeney and Najafian, 2019), often leading to prejudices and discrimination. Stereotypes are understood as "beliefs about the characteristics of group members" (Fiske, 2004). They may be so powerful that they do not only limit the freedom of individuals, but also cause hate, exclusion, and — in the worst case — extermination (Fiske, 1993). Even if stereotypes, and with that social biases, are individually controllable, they persist to this day; prominent examples of social groups that have historically been subject to bias are ethnicities, genders, and age groups (Fiske, 1998). A major factor in carrying and reinforcing those biases is language (Sap et al., 2020). In spoken and written argumentation and debates, biased language can be present if, for example, one sides argues in self-interest (Zenker, 2011) to favor a certain social group or uses unbalanced arguments (Kienpointner and Kindt, 1997).

As CA methods are receiving more and more attention (Stede and Schneider, 2018), it is important to understand how existing social biases influence them. One possible source is the human-generated data on which automated systems are trained and evaluated (Chang et al., 2019). In CA, one of the main sources of data are online debate portals, both for research and for applications (Ajjour et al., 2019).

Prior work has evaluated different properties of dialogical argumentation on debate portals. For example, Durmus and Cardie (2019) evaluated the success rate of users in debates on debate.org based on prior experience, the users' social network in the portal, and linguistic features of their arguments. The authors find that information on a user is more informative in predicting the success compared to linguistic features. Al Khatib et al. (2020) retrieved all debates and posts from *Reddit's* discussion forum ChangeMyView to analyze the characteristics of debaters there. With this information, they were able to enhance existing approaches to predict the persuasiveness of an argument and a debater's resistance to be persuaded.

While the evaluation of debates notably misses work on social biases, other forms of bias have been studied. Stab and Gurevych (2016), for example, attempted to build a classifier that predicts the presence or absence of myside bias in monological texts. In contrast, this work offers an initial evaluation of social bias in dialogical argumentation by analyzing the posts of debate portals.

More generally, many approaches have been proposed to identify different types of social bias in word embedding models. In one of the first studies, Bolukbasi et al. (2016) showed that pre-trained models contain gender bias. They found that it is revealed when generating analogies for a given set of words. This suggests that the distance between word vectors can act as a proxy to identify biases held by a model. Building on that notion, methods to automatically quantify the bias in a pre-trained word embedding model were presented. Caliskan et al. (2017) introduce a metric named the *Word Embedding Association Test* (WEAT) that adapts the idea of the *Implicit Association Test* (Greenwald et al., 1998). Other methods include the *Mean Average Cosine Similarity (MAC)* test (Manzini et al., 2019), the *Relational Inner Product Association (RIPA)* test (Ethayarajh et al., 2019), the *Embedding Coherence Test (ECT)*, and the *Embedding Quality Test (EQT)* (Dev and Phillips, 2019). For a more detailed literature review on detecting and mitigating biases in word embedding models, see Sun et al. (2019). Our approach builds on the WEAT metric to evaluate social biases in embedding models generated from debate portal texts.

Probably closest to our work is the study of Rios et al. (2020). Building on a method developed by Garg et al. (2018), the authors analyzed scientific abstracts of biomedical studies published in a time span of 60 years to quantify gender bias in the field and to track changes over time. For this purpose, they generated separate word embedding models for each decade in their data and evaluated them using the WEAT metric. Using the RIPA test in addition, the authors identified the “most biased words” (Rios et al., 2020) of each model. While we will apply a similar method to detect social bias in textual data, our study differs in two main regards: First, instead of biomedical abstracts, we evaluate dialogical argumentative structures extracted from online debate portals and compare them to each other. Second, we additionally conduct a word co-occurrence analysis of the textual data as an attempt to get more insights into the WEAT results; something that is notably missing in previous studies.

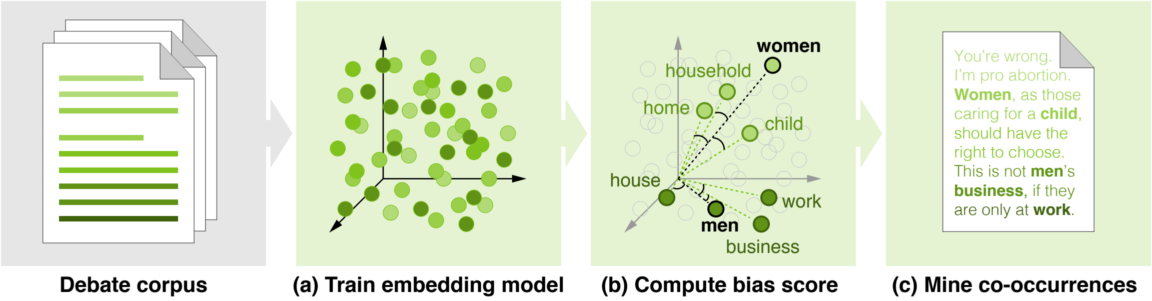
3 Data

To study social bias in argumentative language, we consider the dialogical argumentation found on online debate portals. As our analysis below is based on training word embedding models, which need sufficient data to find statistically meaningful co-occurrences, we resort to the three previously published corpora described below. The texts in these corpora have a similar argumentative structure, benefiting comparability:

IAC The *Internet Argument Corpus v2* (Abbott et al., 2016) contains around 16k debates from multiple debate portals, including *createdebate.com*, *convinceme.net*, and *4forums.com*. The debates tackle various topics that are led by users of the respective portals. As the number of arguments from the single portals are rather small compared to the other two corpora, we only consider this corpus as a whole.

CMV The *Webis-CMV-20* corpus (Al Khatib et al., 2020) covers posts and comments of roughly 65k debates from the internet platform *reddit.com*, specifically from its subreddit *ChangeMyView*. Each debate consists of multiple comments and arguments, including the opening post in which a user states an opinion and arguments on a given issue and prompts other users to provide opposing arguments. The most convincing counter arguments can then be awarded by the initiator of the debate.

debate.org The corpus of Durmus and Cardie (2019) is based on *debate.org*. It contains around 78k debates, each consisting of multiple arguments, written by over 45k users in total. In addition to the debates, the corpus has detailed self-provided user information, such as gender, ethnicity, and birth date. Adding information to one’s profile is voluntary and, thus, neither available for all users in the corpus nor fully reliable. Still, we will use the available information to analyze arguments of certain user groups.



Debate corpus (a) Train embedding model (b) Compute bias score (c) Mine co-occurrences

Figure 1: Overview of the methodology of our experiments: Given a debate corpus, (a) a custom embedding model is trained. (b) A bias score is computed using one of the WEAT tests. (c) Word co-occurrences that give hints about the bias are mined from the corpus (the shown example is made-up for emphasis).

4 Experiments

This section presents the experiments that we carried out on the given data in light of our three research questions, as well as the underlying methodology. We describe how we train embedding models, evaluate their bias, and analyze the word co-occurrences causing the observed bias. Figure 1 illustrates the process.[1]

4.1 Training of Word Embedding Models and Computation of Bias Scores

The Implicit Association Test (Greenwald et al., 1998) measures the response times of study participants for pairing concepts based on word lists and uses it as a proxy for bias (Caliskan et al., 2017). The test requires four lists of words: two target word lists, A and B, and two association word lists, X and Y. Target word lists implicitly describe a concept, such as a social group, while association word lists describe an association, such as being pleasant or unpleasant. The WEAT metric (Caliskan et al., 2017) aims to adapt this method to word embedding models. Under the assumption that word vectors with similar meaning are closer to each other, it computes the mean cosine distance between the four lists in a given model. The score represents the effect size of the difference in distances, which is formulated as follows:

$$\frac{\text{mean}_{x \in X}\, s(x, A, B) \;-\; \text{mean}_{y \in Y}\, s(y, A, B)}{\text{std\_dev}_{w \in X \cup Y}\, s(w, A, B)}$$

where $s(x, A, B)$ is the difference in cosine distances for each word in A to x and each word in B to x. The effect size then acts as a proxy to quantify bias.

Portal-specific Models To apply the WEAT metric to the texts of debate portals, we first extract all posts from the corpora and train one separate custom word embedding model on each corpus using the GloVe algorithm (Pennington et al., 2014). Given the custom models, we then evaluate social bias in them, focusing on three of the most common bias types: towards *ethnicity*, *gender*, and *age* (Fiske, 2004). These types are roughly represented by seven of the originally proposed WEAT tests, found in Table 1. As a notion of stability of the results, we additionally create five embedding models that are trained on random splits of the evaluated corpora and calculate the standard deviation of their WEAT scores.

Baseline Models To be able to assess the WEAT results obtained for the trained custom embedding models, we also evaluate two pre-trained models that have shown different levels of bias in previous work and use those in the sense of "baselines". They allow us to interpret the results in context. As our upper bias boundary, we choose the GloVe model pre-trained on CommonCrawl data (Pennington et al., 2014), as it has shown to comprise a high level of bias of different types (Caliskan et al., 2017; Sweeney and Najafian, 2019). Similarly, we use the pre-trained Numberbatch model 19.08 (Speer et al., 2017) as the lower boundary. Not only is this model claimed to be debiased in multiple ways (Speer, 2017), it has also been shown to be the least biased compared to other pre-trained models (Sweeney and Najafian, 2019).

[1] The code for reproducing the experiments can be found at `https://github.com/webis-de/argmining20-social-bias-argumentation`.

		Target words		Association words	
Type	Test	Compared concepts	Examples	Compared concepts	Examples
	WEAT-1	Flowers vs. insects	rose, spider	Pleasant vs. unpleasant	freedom, hatred
	WEAT-2	Instruments vs. weapons	guitar, gun	Pleasant vs. unpleasant	freedom, hatred
Ethnicity	WEAT-3	European- vs. African-Amer. names	Sara, Alonzo	Pleasant vs. unpleasant	freedom, hatred
	WEAT-4	European- vs. African-Amer. names	Brad, Darnell	Pleasant vs. unpleasant	freedom, hatred
	WEAT-5	European- vs. African-Amer. names	Brad, Darnell	Pleasant vs. unpleasant	joy, agony
Gender	WEAT-6	Male vs. female names	John, Amy	Career vs. family	executive, children
	WEAT-7	Math vs. arts	algebra, poetry	Male vs. female terms	man, woman
	WEAT-8	Science vs. arts	physics, symphony	Male vs. female terms	father, mother
	WEAT-9	Mental vs. physical disease	sad, cancer	Temporary vs. permanent	occasional, chronic
Age	WEAT-10	Young vs. old people's names	Tiffany, Bernice	Pleasant vs. unpleasant	joy, agony

Table 1: Overview of the 10 WEAT tests from Caliskan et al. (2017). As we focus on *ethnicity*, *gender*, and *age* type, we evaluate the given embedding models only on the tests printed in black. The given examples are drawn from the lists; equal words across tests indicate that the same list was used.

Group-specific Models As indicated in Section 3, the debate.org corpus comes with detailed meta-information about several users. To gain insights into the bias of different user groups, we therefore repeat the process outlined above for posts from self-identified "black" and "white" users,[2] female and male users, as well as users below the age of 23 and of 23+[3] in the debate.org dataset. While the age threshold may seem random, age data was sparse and, so, the boundaries were chosen to maximize balance, in order to allow for a rough evaluation of "younger" and "older" users. More or less, the three chosen pairs of user groups coincide with the evaluated social groups and, thus, may provide insightful comparisons.

4.2 Mining and Analysis of Word Co-occurrences

We further conduct a preliminary word co-occurrence analysis for each corpus with the aim to gain additional insights into the observed WEAT results. To achieve this, we first remove noise, such as stopwords, punctuation, and URLs from the corpora. Afterwards, we take as input all social group identity words of the WEAT tests, as exemplified in Table 1. For each list, we extract all words co-occurring with the words in the lists in a window size of 20 (ten words to the left and ten words to the right of the target word). Next, we count their total occurrences and manually evaluate the 100 most common words. As a by-product, we also retrieve the total number of occurrences of target words, which further contribute towards an understanding of the WEAT results.

As the WEAT evaluations only analyze the word embedding models, there is no notion of how often two target and association words are actually used in nearby context. Thus, to better understand such relations, we additionally evaluate the co-occurrences of words in the target and association lists on the sentence level. For each WEAT test, we first build all possible word pairs from the four lists in order to then filter the posts. If a post does not contain both words in a pair, it is discarded. The remaining posts are then split into sentences, allowing us to finally count the co-occurrences of each pair.

5 Results

Portal-specific and Baseline Models The WEAT results on the embedding models of the complete corpora in comparison to the baseline models can be seen in Table 2. Except for the WEAT-8 test, GloVe CommonCrawl yields the highest bias values (ranging from 1.0896 to 1.8734) and Numberbatch yields rather low bias values, respectively. Among the corpora, the Internet Argument Corpus v2 (IAC) shows

[2] In principle, we refrain from using those terms to refer to ethnicities, as they reduce groups to skin color, are thus stereotypes and not adequate to represent the groups (Bryc et al., 2015), promoting topological thinking (Jorde and Wooding, 2004). However, as they are present on debate.org, we use them here to refer to the respective user groups that self-identified as "black" or "white".

[3] The age of a user is the number of years from the specified birthday to 2017, the year of the data collection (Durmus and Cardie, 2019). In all our tests, we assumed those information to be true. As we identified some anomalies, e.g. more than 200 users are assigned an unlikely age of 118 years, we acknowledge that the results might not be particularly reliable.

		(a) Ethnicity			(b) Gender			(c) Age
Type	**Embedding Model**	**WEAT-3**	**WEAT-4**	**WEAT-5**	**WEAT-6**	**WEAT-7**	**WEAT-8**	**WEAT-10**
Pretrained	Numberbatch (debiased)	0.3203	*–0.0994*	0.5621	1.7527	*0.0153*	0.7429	0.8023
	GloVe CommonCrawl	**1.4367**	**1.5778**	**1.3803**	**1.8734**	**1.0896**	1.2780	**1.2527**
Custom	IAC	0.2933	0.3624	*0.1300*	*–0.3396*	0.4009	0.6265	*0.3208*
	CMV	*–0.0632*	0.4956	0.4175	1.3151	–0.3055	*0.4626*	0.7839
	debate.org: Full corpus	0.4125	0.5742	0.5811	1.2775	0.5833	**1.3053**	0.4018

Table 2: Bias values of the custom embedding models, trained on the three debate portal corpora, in comparison to the pretrained baseline models according to the WEAT metric. The higher the absolute value, the larger the bias. The highest value in each column is marked bold, the lowest italicized.

	(a) Ethnicity			(b) Gender			(c) Age
Embedding Model	**WEAT-3**	**WEAT-4**	**WEAT-5**	**WEAT-6**	**WEAT-7**	**WEAT-8**	**WEAT-10**
debate.org: ethnicity-black	0.2483	**1.1304**	**1.5238**	–0.2073	–0.0007	0.7710	n/a
debate.org: ethnicity-white	–0.1915	–0.4062	*0.2535*	0.5080	*0.0006*	**1.3182**	0.4592
debate.org: gender-female	0.3198	0.8689	1.3645	0.1251	**1.0460**	0.7991	0.9661
debate.org: gender-male	*0.0363*	*0.2730*	0.8751	**0.7665**	0.8377	1.3231	*0.2112*
debate.org: age-below-23	**–1.7250**	0.4406	–1.1198	–0.2220	0.5045	*0.5158*	n/a
debate.org: age-23-up	–0.9210	0.8369	0.5143	*0.0624*	0.1848	0.6525	**2.0051**

Table 3: Bias values for the group-specific embedding models of the debate.org corpus, according to WEAT. Higher absolute values mean larger bias, the highest in each column is marked bold, the lowest italicized. WEAT-10 has no result for *ethnicity-black* and *age-below-23* due to too many out-of-vocabulary tokens. The WEAT-10 value above 2 of the *age-23-up* corpus is probably caused by a floating point error.

the lowest values on average (e.g., 0.1300 for WEAT-5) and thus seems to be the least biased. In contrast, the debate.org corpus has higher values in almost all tests, with a noteworthy difference in the gender bias tests WEAT-6 and WEAT-8. In cases where it does not have the highest values (as for WEAT-6), it is surpassed by the CMV corpus. Compared to the WEAT scores of the baseline models, the results of all debate corpora are mostly closer to the debiased Numberbatch model than to GloVe CommonCrawl.

Regarding the general direction of the WEAT scores, we see in Table 2 that most of the observed effect sizes are positive, indicating a closer association of the first list of target words with the first list of association words (see Table 1 for the list ordering). This means that (a) European-American names are more associated with pleasant terms than African-American ones, (b) male names/terms are more closely associated with career, math, and science terms than female ones, and (c) young people's names are more associated with pleasant words that old people's names.

Group-specific Models A similar observation can be made for the debate.org sub-corpora in Table 3. Especially for the embedding models based on posts of female and black users, though, the standard deviation of the WEAT bias values is higher than for the whole corpus, suggesting less reliable results. For the respective user groups, the WEAT scores further seem to indicate a bias against the own social group. For example, the WEAT-3 and WEAT-4 test of the models for black and white users indicate closer associations to pleasant terms with the respective other social group (positive values for black, negative values for white). For female users, the analog notion applies to WEAT-6, WEAT-7 and WEAT-8. Another interesting observation can be made for the age groups: While the ethnicity WEAT tests indicate that older users are more biased towards European-American names, the exact opposite is true for younger users. In general, depending on the specific WEAT test, posts from all user groups seem to be biased in different regards. For example, while on average female users have the highest WEAT values, male users seem to be slightly more biased towards genders. Similarly, posts of younger users show the highest WEAT values in the ethnicity tests.

With some exceptions, the overall direction of the WEAT results in Tables 2 and 3 suggests that the three evaluated debate corpora are all biased towards men (compared to women) and the European-American

(a) Ethnicity	**WEAT-3**		**WEAT-4**		**WEAT-5**	
Corpus	**European**	**African**	**European**	**African**	**European**	**African**
IAC	1075:883	5:14	280:206	14:22	93:69	8:3
CMV	1503:1461	52:35	324:257	47:26	141:80	18:19
debate.org	1960:2009	23:26	530:429	45:73	180:108	32:19

(b) Gender	**WEAT-6**		**WEAT-7**		**WEAT-8**	
Corpus	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
IAC	313:712	35:74	1108:732	244:156	2648:682	278:133
CMV	865:1229	37:104	4586:3976	2006:1653	5840:4008	1396:1255
debate.org	456:844	31:101	3790:1957	572:438	5351:2011	446:405

(c) Age	**WEAT-10**	
Corpus	**Young**	**Old**
IAC	36:19	1:2
CMV	73:42	5:9
debate.org	41:21	3:3

Table 4: Absolute co-occurrences per WEAT evaluation between social group identity words and association words. The number left to the colon denotes the count of words from the first association list (e.g. *pleasant*), the number right to the colon the count for the second association list (e.g., *unpleasant*). Note that for WEAT-7 and WEAT-8, the numbers denote the counts of the respective target word lists.

ethnicity group (compared to the African-American group). That said, when building future CA systems, the IAC corpus is probably the best choice to reduce social biases in general. It is important to note, though, that it is also the smallest corpus of the three and does not include user information. The CMV corpus seems like a good compromise between the two, as it received lower WEAT values than the debate.org corpus and offers more data at the same time.

Word Co-occurrences The co-occurrence counts in Table 4 confirm the observed bias values in some cases. In the debate.org corpus, male identity words indeed occur more often with the math-related terms of the WEAT-7 test compared to female identity words. As a specific example, the word "he" co-occurs with math-related terms 1662 times, while the female counterpart "she" is only mentioned 178 times in a sentence with the same class of words. Similarly, in the CMV corpus, science-related terms of the WEAT-8 test appear 2006 times in the same sentence as the male identity word "his" while only sharing the same sentence 1396 times with all female identity words *combined*.

In most cases, however, words from both groups are only rarely mentioned in the same sentences, compared to the overall size of the corpora. For the tests that use names as social group identifiers, e.g. WEAT-6, this problem is even more noticeable, as indicated by the lower numbers presented in Table 4. In all three corpora, names generally co-occur less often with the association words in the same sentence compared to terms that describe a concept, as used, for example, in WEAT-7 and WEAT-8. This not only makes it harder to directly interpret the WEAT results. It also indicates that the cosine distance, on which the WEAT score is based, relies more on either a distant context, such as an entire post, or on common co-occurrences with potentially unrelated words. That said, the results presented above should be interpreted with care.

6 Discussion

To put the results into context, we will discuss two additional observations more closely in the following, namely the low number of occurrences of some identity words as well as the influence of using names for social group lexicons. We will then close this section with limitations observed during the analysis.

6.1 Low Occurrence of Identity Words

In some experiments, the number of occurrences of identity words should be considered when interpreting the results: In the WEAT-10 tests, the old people's names used for the social group of elderly people generally have a very low frequency. In the posts of black debate.org users, they sum up to only 18 occurrences in total. The same is true for users above the age of 22. Even though the overall number of occurrences seems sufficient to conduct the WEAT evaluation, this definitely calls the meaningfulness of the embedding model into question and, with that, the association tests for those groups.

Also, the occurrence *ratio* of identity words of two social groups being compared with WEAT deserves discussion. For all evaluated tests, at least twice as many identity words of one group are mentioned as for the other, in the highest case even 101 as many. The identity words of the European-American group utilized by WEAT-3, for example, occur on average 54.8 times as often as the African-American ones across the three debate corpora. While this ratio is not as high for the other tests, the general tendency is that European-American names are mentioned more often. The same is true for male names and terms, which are used more frequently than female ones, and young people's names that occur more often than old people's names across all corpora. Adding to this is the fact that the identity words are, compared to corpus sizes, only rarely used together in close context, say, in a sentence. The distance between two words in the vector space thus relies mostly on the co-occurrence with other, unrelated words.

On one hand, these observations make it less trivial to interpret WEAT results, as the different occurrences cause unequal probabilities for the two identity word lists to co-occur with the tested association words. On the other hand, they imply that the evaluated corpora are not very diverse and, so, provide imbalanced data with respect to the evaluated social groups. For debate.org, this imbalance can also be seen in the size of certain user groups, resulting in an unequal number of arguments from different groups.

Another potential confounding factor is the number of out-of-vocabulary (OOV) words. When the total number of identity words is low, many OOV words may make results even more unreliable. Among others, this is the case for the female user posts of the debate.org corpus. African-American names from the WEAT-3 test make up 41 of the 49 OOV words, leaving only 9 of the initial 50 names for the association test. A similar case is WEAT-10 for users above 22. From the eight old people's names, only one appears in the sub-corpus. An immediate effect of the missing words is, however, not visible from the results.

One practical consequence of the discussed unequal distribution is that automated systems, trained and evaluated on these corpora, may favor arguments of the majority group. The underlying discrimination against minority groups may lead to unfair behavior in all stages of CA, e.g., to a better ranking of arguments from majority groups in quality assessments or to generated arguments that reflect the opinion of the majority group mainly. These examples stress the need for more diverse debate corpora representing different social groups in a more balanced way. Creating such corpora will not be easy, though, since detailed user information is not often available on debate portals and other argument sources.

6.2 Using Names as Social Groups

A general issue underlying the results is that a list of names does not describe the concept of a social group to a level required by co-occurrence methods such as WEAT. For example, the token "palin" often co-occurs with the female list word "sarah" used in WEAT-6, referring to the politician *Sarah Palin*. Consequently, in the debate.org corpus, some other highly ranked terms relate to politics, such as "president", "conservative", and "supporter". While we did not observe this for all three corpora, it demonstrates that using names as identity words can cause public persons to somewhat act as representatives. Texts about them thus also influence the associations of the tests, independent of whether they are part of the group or their behavior is representative of it. This discrepancy leads to associations authors of evaluated texts have with the public persons rather than the social group and may ultimately influence the bias evaluation.

Another issues lies in the overall occurrence of names. Male and female *names*, for instance, appear less often in the debate texts than male and female *terms*. As already discussed above, a smaller number of occurrences of identity words may make the results more prone to distortions, due to small fluctuations in co-occurrences, and thus less reliable in general.

Together, these issues suggest that names might not be appropriate to represent a social group and to analyze bias for the evaluated corpora. Also, it seems questionable whether they are generally statistically representative of the target social groups. Partially, however, the lists of names also led to expected associations, as is the case for the WEAT-3, WEAT-4 and WEAT-5 test on the CMV corpus. The terms co-occurring with the African-American names suggest that they were at least to a small degree able to capture the associations, since some of the most co-occurring terms included "African" and "black".[4]

[4] We do not suggest that these terms adequately represent and describe the African-American social group, but simply state that the association between the social group and the terms generally holds.

6.3 Limitations

One limitation of our methodology is that it relies on the chosen embedding model to accurately model distances and associations between words. This makes it less applicable to smaller datasets, let alone single texts. Additionally, the influence of multiple factors on WEAT results is unclear. For example, choosing an alternative algorithm to generate an embedding model may yield different results, e.g., word2vec (Mikolov et al., 2013) or FastText (Bojanowski et al., 2017). Future work should explore more sophisticated methods to analyze social bias that do not depend on embedding models, as they are a point of uncertainty that might never be fully explainable due to the nature of generating the models.

Further, the results presented in this work are limited to the accuracy of WEAT. This dependence is problematic for three main reasons: First, it assumes that the calculation done by the test accurately models the associations between the lexicons. While Caliskan et al. (2017) show that they are able to reproduce results from previous psychology studies with humans, it remains open whether the notion generally applies. Second, it expects that the lexicons of the tests are representative of the social groups they ought to model. Especially with a list of names, however, this assumption can lead to unexpected results and might not hold, as shown above. The same is true for the co-occurrence analysis, which is also based on (and limited to) the social group lexicons of Caliskan et al. (2017). Lastly, the test assumes that all evaluated biases are quantifiable. While this allows for automation, this assumption is certainly questionable, among other reasons because every person might perceive bias differently.

Given those limitations, we would like to emphasize that the results presented in this work are meant as a first evaluation that needs to be further backed up and investigated with more tests in the future.

7 Conclusion

In this paper, we have analyzed social bias in three debate corpora commonly used in CA research. To this end, we have trained custom word embedding models and evaluated their associations for multiple social groups using WEAT. Further, we have analyzed co-occurrences of terms used to define the social groups. We have found all three corpora to show social bias, mostly towards women and African-American people. According to our evaluation, the smallest corpus, IAC (Abbott et al., 2016), carries the least social bias, whereas the debate.org corpus (Durmus and Cardie, 2019) shows mostly the highest. The CMV corpus (Al Khatib et al., 2020) seems like a middle ground, as it contains the most data and is less biased than the debate.org corpus. In all three corpora, we have found imbalances regarding the representation of the evaluated social groups.

Future work should investigate additional ways in which social bias may be present in CA methods, as the underlying data is not the only source. Other possible causes may be the features selected for a trained model or the way in which a model is applied to a real world problem. We believe that future CA corpora should be evaluated and mitigated for social bias. As data is underlying most CA methods, it is essential that it is as representative for as many social groups as possible in a balanced manner.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet Argument Corpus 2.0: An SQL schema for Dialogic Social Media and the Corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In *KI 2019: Advances in Artificial Intelligence - 42nd German Conference on AI, Kassel, Germany, September 23-26, 2019, Proceedings*, pages 48–59.

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Publisher: MIT Press.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165, Version: 3.

Katarzyna Bryc, Eric Y. Durand, J. Michael Macpherson, David Reich, and Joanna L. Mountain. 2015. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*, 96(1):37–53.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and Fairness in Natural Language Processing. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Learning to Flip the Bias of News Headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88, Tilburg University, The Netherlands. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the 4th Workshop on Natural Language Processing and Computational Social Science*. To appear.

Sunipa Dev and Jeff Phillips. 2019. Attenuating Bias in Word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 879–887.

Esin Durmus and Claire Cardie. 2019. A Corpus for Modeling User and Language Effects in Argumentation on Online Debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding Undesirable Word Embedding Associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. 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 6343–6349. Association for Computational Linguistics.

Susan T. Fiske. 1993. Controlling other people: The impact of power on stereotyping. *American Psychologist*, 48(6):621–628.

Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, volume 1-2, pages 357–411. McGraw-Hill, New York, NY, US, 4th edition.

Susan T. Fiske. 2004. *Social Beings: A Core Motives Approach to Social Psychology*. J. Wiley.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Felix Hamborg. 2020. Media Bias, the Social Sciences, and NLP: Automating Frame Analyses to Identify Bias by Word Choice and Labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 79–87, Online. Association for Computational Linguistics.

M. Tim Jones. 2019. Machine learning and bias. `https://developer.ibm.com/technologies/machine-learning/articles/machine-learning-and-bias/`. Last accessed: 2020-09-07.

Lynn B. Jorde and Stephen P. Wooding. 2004. Genetic variation, classification and 'race'. *Nature Genetics*, 36(11):S28–S33.

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.

John Lawrence, Mark Snaith, Barbara Konat, Katarzyna Budzynska, and Chris Reed. 2017. Debating Technology for Dialogical Argument: Sensemaking, Engagement, and Analytics. *ACM Transactions on Internet Technology*, 17(3):24:1–24:23.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 446–457, Barcelona, Spain. Association for Computing Machinery.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 1–13, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California USA.

Robyn Speer. 2017. ConceptNet Numberbatch 17.04: Better, less-stereotyped word vectors. `https://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/`. Last accessed: 2020-09-03.

Christian Stab and Iryna Gurevych. 2016. Recognizing the Absence of Opposing Arguments in Persuasive Essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118, Berlin, Germany. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

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

Chris Sweeney and Maryam Najafian. 2019. A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting Gender Right in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. Association for Computational Linguistics.

Frank Zenker. 2011. Experts and Bias: When is the Interest-Based Objection to Expert Argumentation Sound? *Argumentation*, 25(3):355.