Investigating Bias in Computing Science Literature

Using Natural Language Processing to Investigate Bias in Abstracts of Published Computing Science Papers and the Role the Diversity of a Team of Co-authors plays in it

Marie-Sophie Simon
Radboud University, Nijmegen

ABSTRACT

This research analyses abstracts of papers published at top computing science conferences in 2017 and the bias they may exhibit. With computing science continuing to have a very low proportion of female researchers, earlier research has hypothesised that low participation among female researchers is partly due to a self-reinforcing cycle. Our research aims at investigating this hypothesis by correlating bias in published abstracts of computing science to diversity of a team of co-authors in terms of nationality and gender representation. We use the Dbias python package for classification and identification of biased words and logistic regression to investigate factors influencing the bias. Our results show, that an overwhelmingly large part of the published abstracts of computing science in 2017 did exhibit some form of bias. The recognized biased words, contained numerous typical computing science terminologies, indicating a need for future work to investigate the degree to which such terminology is biased and which effects this may have on the sociology within the field. We further found that there was a statistically significant effect of the diversity of a team of authors in terms of nationalities and gender distribution reducing bias in an abstract. This continues to stress the importance of diverse teams to pave the path for a more inclusive future.

1 INTRODUCTION

With only about 15-30% of female researchers, the gender gap within Computing Science is a well-known problem and a popular topic of research [6, 7, 14]. Furthermore, there is scientific evidence indicating that gender stereotypes and biases perpetuate their view on individuals [4] as well as the participation of women in academic fields [10]. Due to observations such as these, it is generally accepted that increasing diversity among computing science researchers would be beneficial for both society as well as science. It is unclear, however, why computing science continues to be a field dominated by men and what interventions would possibly help to increase participation among female researchers. One hypothesis attributes part of the reason of low female participation to the chicken-and-egg nature of the problem, where it is believed that part of the factors that decrease participation among female researchers, is the low participation of female researchers [12]. This research aims to investigate part of this hypothesis and the earlier observation that bias continues to perpetuate the low number of female researchers in computing science. To this end, this research analyses bias in abstracts of computing science published at top conferences in the year 2017. As an indicator to test the hypothesis that a less diverse environment creates a less inclusive atmosphere, we use logistic regression to correlate the bias that is exhibited to two factors of diversity within a team of co-authors, that is nationalities and genders represented in the group.

Copyright held by the owner/author(s).
TxMM'23-'24, Radboud University, Nijmegen, Netherlands

2 RELATED WORK

2.1 Gender representation in CS

The fact that there is a gender gap in the field of computing science is well known and a lot of research has investigated the specific gender ratio or possible reasons for the gap. In the following, we will look at previous research to argue *why* it is important to investigate gender representation and bias within the field in more detail.

According to research, more females enrolling in tertiary scientific education correlates with a decrease in explicit and implicit national gender-science stereotypes across nations [10]. This highlights the importance of gender representation in science for overall gender equity within a country. Additionally, issues of gender representation can be self-exacerbating, as, for example, [12] discussed in their article. A common theme in their suggested reasons for the gender gap in computer systems research is the "the chicken-and-egg nature of the problem" where factors reducing women's participation also amplify the effect of their reduced participation[1].

We see from the short investigation above, that the way we present gender and the way genders are represented influence the way they will be perceived and thus contribute to the degree of equitable participation of genders. Because of this, it is relevant to investigate gender bias, specifically in fields such as computing science, where the participation among females remains low. Finding gender bias within the abstracts of these publications would indicate another relevant factor in the quest for a fair representation in the field.

2.2 Gender bias in NLP

Bias and specifically gender bias are not new topics within NLP research. And yet, research analysing how gender bias is investigated within the field of NLP, concluded that oftentimes, the underlying definition of gender bias is lacking [3, 13]. Among others, one of the lacking factors continues to be the representation of gender as a binary variable [13] and often the lack of consideration with regards to intersectionality [3]. While especially more recently published papers do acknowledge this as a limitation [3], the complexity and fluidity of the concept are not yet being integrated into this research [3, 13].

[1] For example, because there are few women to cooperate with, in computer systems research, women will be less likely to stay in fields because there are no women to collaborate within their field.

As our research relies on earlier work for the dataset and to detect bias, it will be out of scope for us to address the full complexity of these issues. We do, however, need to consider this as a limitation of our research, which we will discuss in subsection 5.1 and consider the possible ethical repercussions of this, which we do in subsection 3.4. Furthermore, we will need to be aware of possible underlying assumptions that may have been made by the bias detector we will be using (see subsection 2.3).

2.3 Detecting bias in texts – Dbias

To be able to detect gender bias in scientific texts, we will use the research result of [11], the Dbias package in Python. The purpose of this study was to develop a fair pipeline to train models by recognising when textual data is biased, to identify and remove the bias, such that models using that data, do not reinforce the earlier existing biases. For our research, we will mainly use the first two steps of the pipeline, that is the first step which detects bias (the classifier of Dbias) and the second step, which recognises the biased entities, which will be used as a way to verify and interpret our outcomes.

As Dbias was developed aiming for a **fair** pipeline, it detects, next to gender bias, also age, racial and ethnicity, disability and mental health bias. Religion, education and political ideology are also biases that Dbias tries to mitigate [11]. Using a broader conceptualisation of bias, we hope to correct for some of the possible concerns brought up in subsection 2.2, such as fluidity of gender and intersectionality. However, neither gender nor any of the other attributes are elaborately conceptualised in the research by [11], therefore this is mere speculation and we may need to investigate this via the recognised entities of Dbias.

In contrast to [11], who concentrated and fine-tuned their model on news media, our focus is on scientific articles. As obtaining a labelled dataset to fine-tune a new model is unfeasible within the scope of this research, we rely on the idea that bias is a transferable concept independent of the domain. We again will be using the entities recognized by the model to investigate the implications this may have for our results.

3 APPROACH

To answer our research questions, our investigation consists of two steps. First, we classify a dataset of abstracts published at computing science conferences in 2017 to be either biased or non-biased and retrieve the words that indicated this bias. In our second step, we investigate a possible correlation between biased texts and aspects of diversity in the team of co-authors, that is, the distribution of male and female-identified co-authors and the variety of nationalities within a team.

3.1 Dataset

To analyse bias that may be present in publications of computing science conferences, we used the data repository created and published by Frachtenberg [5]. This repository contains some of the top computing science conferences that took place in 2017, each of the articles published through the respective conference, the abstracts and authors of each of the publications and data about each of the authors such as their identified gender and nationality. Our final dataset containing papers and the information about the diversity of a team contained 2391 different papers from the original 2439 papers over 56 out of the 75 conferences.

We decided to only use the abstracts of each publication to recognize possible bias instead of the full text as we are using bias to analyse the way computing science as a field is perceived. As it is very common to only read abstracts of papers, this is an accurate proxy for the way someone would perceive computing science literature. If we were more interested in the implicit biases that may live within research teams of the computing science field, analysing the full texts would be more appropriate.

3.2 Variables

To answer our research question our dependent variable is a binary label of whether a text is either "Biased" or "Non-Biased". Using the classifier function of Dbias [11] over each of the abstracts of our papers returns a label and a score which we add to our dataset. Additionally, we use the recognizer function on the texts that were found to be biased which gives the words that are recognized as output, adding those to our dataset as well.

Our independent variables are aspects of the diversity of a team of coauthors in terms of genders and nationalities represented. The diversity of nationalities is represented by the number of unique countries, which we thus encode as a simple fraction of the uniquely different nationalities over all the authors in the team where we knew the nationality. This will result in values between 0 and 1, where the closer it is to 0, the more authors there are with the same nationality and a score of 1 would mean maximum diversity, i.e. every author has a different nationality. By using only the cases where the nationality of a coauthor was known, we avoid underestimating the diversity of teams due to unknowns.

The diversity of gender within a team is a slightly more difficult feature to represent in just one number. If we just took the percentage of one of the genders, we do not get a linear factor for diversity (as both 0% of one gender, as well as 100% of one gender, are non-diverse). The same reasoning holds if we were to use the ratio of one gender over another gender. Therefore, we decided to summarise the diversity of genders in a function according to:

$$f(x) = \begin{cases} \frac{1}{50}x, & \text{if } x \leq 50 \\ -\frac{1}{50}x + 2, & \text{if } x > 50 \end{cases}$$

Where x is the percentage of coauthors that are male-identified, of all authors whose gender is known. As we know that x is always between 0 and 100, the function will always output a value between 0 and 1, where 0 is no diversity at all (either only men or no men at all) and 1 is the highest possible diversity, so equal numbers of both genders. Importantly, this function would return the same value if we were to use the percentage of female co-authors.

3.3 Methods used for Analysis

As already said in subsection 2.3 we use the Dbias package to detect bias and biased words in the published abstracts of computing science papers [11]. Using the results of this model, we will investigate the words that were recognised to be biased by first lemmatizing them, to get a more accurate measurement of frequencies for our following analysis.

We use a logistic regression to analyse any possible correlation between bias, the number of co-authors in a team and our two aspects of diversity. To improve our analysis, we had to balance our dataset using the random oversampling algorithm from the imbalanced-learn repository [9]. By using oversampling instead of undersampling we avoid losing information. Additionally, we use our model to interpret the coefficients and not for prediction, making overfitting less relevant, thus making the simple random algorithm more appropriate than SMOTE or ADASYN. To overcome the fluctuations in a random oversampling algorithm, we run the model 100 times and calculate the average over all the parameters.

3.4 Ethical Considerations

As we are using published texts of authors in a different way than they intended, we need to consider the ethics of our approach. First, we want to stress that any bias found is not ascribed to any individual or believed to be a consequence of wrongdoing, but that we hope to use these results to give insight into possible obstacles towards a more inclusive future within computing science research and society. As the goal for a more inclusive field a value endorsed by both society and research, we argue that despite the abstracts not being intended to be analysed for bias, investigating ways to create a more inclusive environment in computing science is in the interest of the researchers and therefore ethically responsible.

Possible ethical concerns about our research arise from how we analyzed diversity, particularly our use of gender as a binary variable despite gender now being viewed as a spectrum and fluid concept and subject to change [3, 13]. By not addressing these shortcomings in our research, we are continuing to perpetuate these outdated views which may be harmful to both research and society. Unfortunately, to address shortcomings such as the non-binary nature of gender, we will need to create a diverse and inclusive environment first, making this, in our opinion, another chicken-and-egg problem. We hope, therefore, that despite this research continuing to use gender as a binary variable, it still paves the way for a more inclusive environment which will eventually overcome such shortcomings. Thus, we believe this research is still ethical, even though it perpetuates old stereotypes and views.

4 RESULTS AND ANALYSIS

4.1 Bias in Publications

The classifier function of the Dbias module [11] is able to classify 2366 of the 2391 abstracts that we provide as input. Inspecting the labels, we see that 2228 of the abstracts are classified to be biased, which means about 93.2% of abstracts in our dataset were recognised to contain some bias. For further interpretation of these results, we investigate the recognized words that contained a bias. Unfortunately, this function only recognises words in 811 of the cases of biased abstracts. While it is a bit unclear why this only returns words for so few cases due to a lack of documentation and description of the functions of Dbias, we assume the fact that the model has not been trained on scientific literature plays a role. After lemmatizing the words, we calculated the frequencies to inspect the most common words indicating bias in our dataset. In Figure 1 we can see the ten most common words that were recognized to exhibit bias.

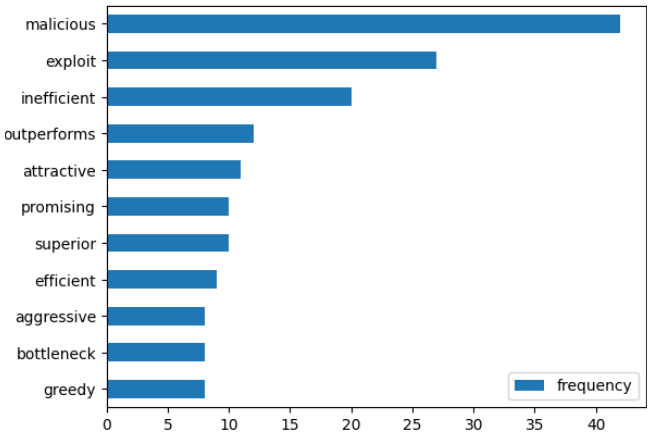

Figure 1: The frequencies of the top 10 most common biased words

As we can see, some of these words are very commonly used in computing science literature. For example, "malicious" and "exploit" are very common terminologies used in cyber-security or "efficient" and "inefficient" which are commonly used in the context of algorithm design or evaluation. The fact that these words are commonly used within computing science, raises the question of whether the words are truly indicating bias or whether they have a different meaning in the context of computing science publications compared to news media (which is what the Dbias model was trained on). We will discuss this in more detail in the section 5. To get further insight into the words that were recognized to contain bias, we also created a word cloud, giving us a more general overview of the recognized words.

Figure 2: The word cloud created from the recognized entities

4.2 Bias and Diversity in a Co-Author Team

Our second step was to investigate whether there is a correlation between a published abstract being biased and aspects of diversity within a team of co-authors. The results of our logistic regression indicate that both the diversity of gender within a coauthor group as well as the diversity of nationalities have a statistically significant impact on whether or not an abstract of a computing science publication is biased. The number of co-authors in a team did not

have a significant effect which is why we removed it as a feature in our final analysis. Diversity of gender within a co-author group has a coefficient of -0.33 (p-value 0.005) and the variety in nationalities within a group influences the bias with a coefficient of -0.49 (p-value 0.000). Since we coded the biased class as 1 and the non-biased class as 0, a negative coefficient means here that the feature leads to less bias. This means, according to our model, having a more diverse background of nationalities within a team of co-authors has a slightly stronger effect towards less biased text outcomes than an even distribution of identified genders.

5 DISCUSSION AND OUTLOOK

Our results show, that a large part of the abstracts from published computing science papers in 2017 indicate some bias. Further, we showed that a more diverse team of co-authors, that is, more even participation among male and female researchers and more unique nationalities within a team, decrease the likelihood for their abstract to contain bias, where nationalities have a slightly stronger effect. These findings corroborate the hypothesis, that factors reducing the participation of minorities may amplify the effect of their reduced participation (i.e. a more homogenous group is more likely to exhibit a bias which in turn makes it less attractive for new minorities to join this group).

What is important to consider are some of the words that were identified to be exhibiting bias in our dataset. As we have seen in Figure 1, most of the most frequently used biased words are common terminology within the field. This raises the question, of whether the fact that the bias detector was fine-tuned on news data resulted in wrongful detections. For example, one might argue that it is impossible to talk about cyber security without words such as "malicious" or "exploit" (e.g. an attacker with malicious intentions exploits a software vulnerability).

One possible answer to this question, could come from insights of the attribution theory of psychology. This theory describes the way individuals tend to attribute causes to another person's behaviour, without having true insight into their reasoning [2]. In this field, researchers have identified a "fundamental attribution error", which refers to the tendency of people to attribute other people's behaviour to their internal and personal characteristics, not taking into account possible external and situational circumstances. An example from [2] could be that individuals attribute homelessness to the character (and their flaws) of that person, ignoring possible circumstances that could have led to a person's homelessness. This could apply similarly to the way science attributes terms such as maliciousness and exploits to fields such as cyber security (e.g. someone hacking into software may not be malicious but there may be external reason that drove them to do this).

The example above is meant to illustrate that dismissing the recognized biased words as common computing science terminology risks overlooking inherent bias within our field. While this may not hold equally for every word, we do encourage using these results for future research, possibly in an interdisciplinary setting between computing scientists and linguists or social scientists to explore the implied bias that may be perpetuated by common terminology within our field.

Our results additionally show a correlation between the diversity of the team of co-authors and whether or not the text is classified as biased. This corroborates the hypothesis that because of the non-diversity within computing science, the field presents itself to the outside (through its publications) as a non-inclusive environment, possibly reducing the participation of minority groups. While this stresses the importance of diverse research teams, it is important to acknowledge that this is unfeasible in the current state of the field as there are simply not enough participants of some minorities (e.g. female researchers). As a temporary solution, conferences could consider using tools, such as the Dbias pipeline used in this research, to identify bias within research papers and even to de-bias texts [11]. As some conferences have already started with diversity efforts in their organisational strategies, this would be an easy additional step. If conferences would be interested in implementing such language processing on a larger scale, it could be of relevance to calculate their energy usage to be able to consider trade-offs between different sustainability goals of energy conservation and equal inclusion of everyone.

5.1 Limitations

There are a few limitations in our research approach that we could not overcome at this time and are therefore important to consider. First, our dataset represents gender as a binary variable that has been identified through an outsider. This goes against the new insights that gender is not only a spectrum but also a fluid concept that can change over time [3, 13]. As using gender as a binary concept goes partly against the aim of this research as it is not inclusive towards everyone we discuss the ethical considerations about this in subsection 3.4.

Next to the way we conceptualised the gender variable, it is also important to address the other independent variables. For one, we used only two aspects of diversity within our analysis, while there are of course endlessly many to choose from. Due to the scope of this research and dataset, it was not feasible to include more measures of diversity within this research. Future work, however, could analyse bias within computing science abstracts with the topics of articles or conferences or make more distinctions between nationalities based on cultural or language backgrounds.

Another limitation is the fact that we used the Dbias model to recognize bias within publications of computing science, while that model was fine-tuned on news articles. While it is reasonable to argue that bias is a concept that transfers over different domains, there are some problems that we need to consider. For example, if a text discusses a bias that people exhibit against women who are perceived to be attractive [8], then the word "attractive" occurs in a very different context than when it is a news article announcing the most attractive women of a year [1]. These nuances will not be taken into account in our research and therefore require us to take our results lightly.

It is furthermore important to note that we now handled a very general conceptualisation of bias within our texts. The Dbias function classifies texts into a binary category that is either biased or non-biased, without specifying which bias is present. While we hoped that this would lead to a less narrow view of bias, possibly

including intersectionality in a more natural way than by, for example, counting male and female pronouns, it was out of this scope to validate this in our research.

5.2 Future Work

Despite overcoming the limitations within our research as mentioned above, future work will also be needed to place our research in a larger context and continue to address the non-diverse environment within computing science research.

For example, our results indicate that increased diversity leads to less biased texts. At the same time, however, we know that it is currently unfeasible to create only diverse teams within computing science. This raises the question of whether it is possible to train awareness of diversity in such a way, that the effect is similar. Future research could aim to investigate whether researchers who received diversity, equity and inclusion awareness training, exhibit less bias in their research.

Additionally, it could be interesting to investigate the way research teams are being formed and whether it is possible to increase awareness about the consequences the composition of a team can have. It is important here, to not force diversity by all means as this could backlash in the way diversity is perceived or could risk minorities being seen as a means for diversity without being taken into account as an end of themselves.

6 CONCLUSION

In this research, we set out to investigate the question of whether abstracts of published computing science papers are exhibiting some form of bias. By analysing the correlation between the two aspects of diversity, gender and nationality within a team, we addressed part of the chicken-and-egg problem that formulates the complication that a non-diverse environment becomes less attractive for people outside of the majority group.

Our results show, that a large part of the published abstracts of computing science in 2017 did exhibit some form of bias. When we investigated the words that were recognized to indicate bias, we were confronted with some typical computing science terminology. Future work should investigate the degree to which such terminology is biased and which effects this may have on the sociology within the field.

We further found that there was a statistically significant effect of the diversity in nationalities and an even gender distribution within a team of co-authors on the biased outcome of an abstract. This continues to stress the importance of diverse teams to pave the path for a more inclusive future.

REFERENCES

[1] 2023. Top most beautiful women in the world 2023 - Daily News. https://dailynews.co.tz/top-most-beautiful-women-in-the-world-2023/

[2] Brooks J Baumgartner, Lisa M Bauer, and Khanh Van T Bui. 2012. Reactions to Homelessness: Social, Cultural, and Psychological Sources of Discrimination. *Psi Chi Journal of Psychological Research* 17, 1 (2012).

[3] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2083–2102.

[4] Naomi Ellemers. 2018. Gender stereotypes. *Annual review of psychology* 69 (2018), 275–298.

[5] Eitan Frachtenberg. 2021. Systems conferences analysis dataset. https://doi.org/10.5281/zenodo.5590575 Version 2021-10-21.

[6] Eitan Frachtenberg and Rhody D Kaner. 2022. Underrepresentation of women in computer systems research. *Plos one* 17, 4 (2022), e0266439.

[7] Eitan Frachtenberg and Noah Koster. 2020. A survey of accepted authors in computer systems conferences. *PeerJ Computer Science* 6 (2020), e299.

[8] Stefanie K Johnson and Elsa Chan. 2019. Can looks deceive you? Attractive decoys mitigate beauty is beastly bias against women. *Archives of Scientific Psychology* 7, 1 (2019), 60.

[9] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365

[10] David I Miller, Alice H Eagly, and Marcia C Linn. 2015. Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology* 107, 3 (2015), 631.

[11] Shaina Raza, Deepak John Reji, and Chen Ding. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics* (2022), 1–21.

[12] Alexis Richter, Josh Yamamoto, and Eitan Frachtenberg. 2023. Why Are There So Few Women in Computer Systems Research? *Computer* 56, 2 (Feb. 2023), 101–105. https://doi.org/10.1109/MC.2022.3219633 Conference Name: Computer.

[13] Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. http://arxiv.org/abs/2112.14168 arXiv:2112.14168 [cs].

[14] Josh Yamamoto and Eitan Frachtenberg. 2022. Gender differences in collaboration patterns in computer science. *Publications* 10, 1 (2022), 10.

A WORK REPORT

Most of the details of my process are outlined in section 3. The merging of data from the repository ([5]) was more challenging that expected, causing me some trouble in the beginning, not at least due to an oversight on my side. After properly merging the data, I employed the Dbias classifier and recognizer, writing results to a separate file. For my analysis of the biased texts, words and frequencies as well as the correlation to nationality and gender within a team I merged all data such that I had a complete dataset which I could store and open for further use without needing to rerun the Dbias package (as that took quite a while). Using logistic regression for the analysis of the correlation turned out to be more complicated than anticipated as well as the imbalanced dataset caused some issues which I had not encountered before. The final approach and reasoning for it can be found in section 3.

Doing justice to complex topics such as bias, diversity, and inclusion in a "small" course project is difficult. Next to the limitations explored in subsection 5.1 I think that more time to explore bias and language portraying bias in-depth could have allowed a more complete analysis and interesting analysis.