Analysis of gender bias in peer review using natural language processing

Palash Pandey, Drexel University, U.S.A. pp535@drexel.edu
Erjia Yan, Drexel University, U.S.A. ey86@drexel.edu

Motivation

Publication and knowledge sharing is at the core of scientific communication and so the decisions made by conferences and journals are non-trivial in the effect that they can have on the quality of the work that is being done and as a result the quality of work that gets shared with the world. Publication venues can have profound impact on the attention and exposure a work gets and thus it is crucial that the processes of publication be as unbiased as possible. There have been some studies that analyze the available data quantitively to identify bias but most of these studies do not consider the quality of the papers themselves but rather rely on authors' attributes. These analyses can be flawed because of the huge variance in the quality of work that journals receive and making it difficult to generalize the results as substantial evidence for bias. We are using the peer read dataset of reviews and analyzing individual reviews that have been presented for manuscripts and analyzing their sentiment. We assign sentiment scores to the reviews and these scores act as a normalizing factor for the acceptance/rejection decisions so that the decisions can be analyzed with respect to authors' gender without obfuscation from other factors.

Objectives

The objective of this study is to analyze gender bias in peer review by analyzing the sentiment of the reviewers comments on the submitted manuscripts. The study aims at achieving this by using the sentiment scores for the reviews as a proxy for the quality of the papers so as normalize all the submitted manuscripts and have a fair analysis. The study attempts to eliminate the quality of papers as an attribute in the decision making process and thus compares similar papers so as to isolate the effect of authors' attributes like gender.

Data

Data for ICLR 2017 from the PeerRead dataset:
Total number of reviews 1304

A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications.

	No. of manuscripts
Accepted	172
Rejected	255
Total	427

	No. of authors
Highest	12
Lowest	1
Average	3.69

Implications

Because of the ubiquity of single blind peer review in scientific publishing it is imperative for authors and reviewers to be aware of any biases in the peer review process. Using natural language processing techniques to quantify bias can serve this very purpose because of the objective nature of the measure and its uniformity across different publications, it can serve as a benchmark for comparing gender bias across different publication venues in multitude of fields of study.

References and Future Work

We aim to extend this study to nationality bias, we can divide the authors by the nationality of their institution of affiliation and analyze the data similarly to gender bias method to identify any nationality bias.

[1] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

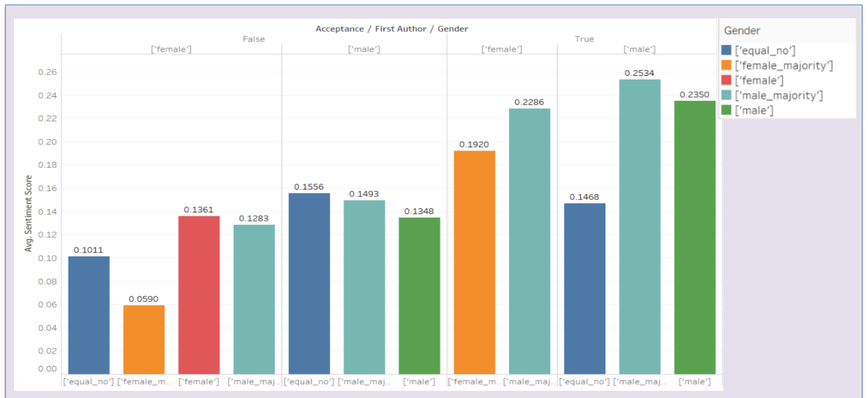
[2] A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, Roy Schwart

[3] Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

Methods

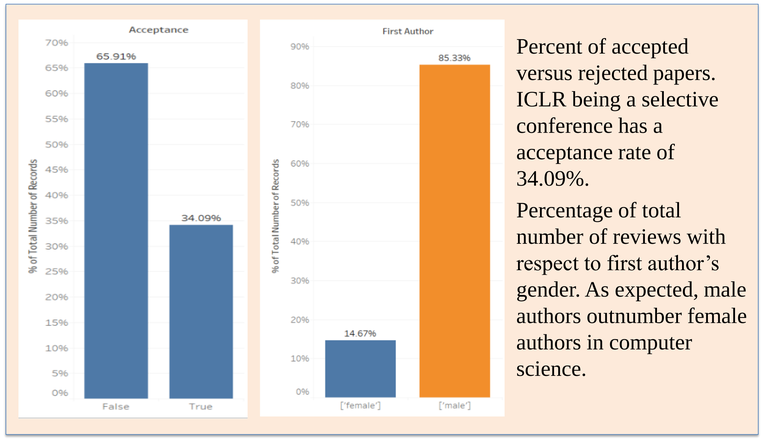
The data for ICLR 2017 was used and processed for analysis throughout the study. The reviewers' comments were processed through Vader sentiment classifier and sentiment scores -1 – 1 were assigned to every review. The Vader sentiment classifier is trained on social media data i.e. tweets, IMDB movie reviews, Amazon product reviews, New York Times editorials which makes it perfect for this application because it can detect more subtle language patterns than a classifier trained on any other dataset. Furthermore, reviewers' comments are very similar to movie or product reviews but are more subtle making Vader sentiment perfect for this application. For the gender classification of authors, we use Genderize API, which utilizes a massive database of names from social networks. We assign the papers into 5 categories based on gender: male majority, female majority, male, female, equal number. We also take into account the first author for each of these categories, we assume as per common practice that the first author is the lead author.

Results



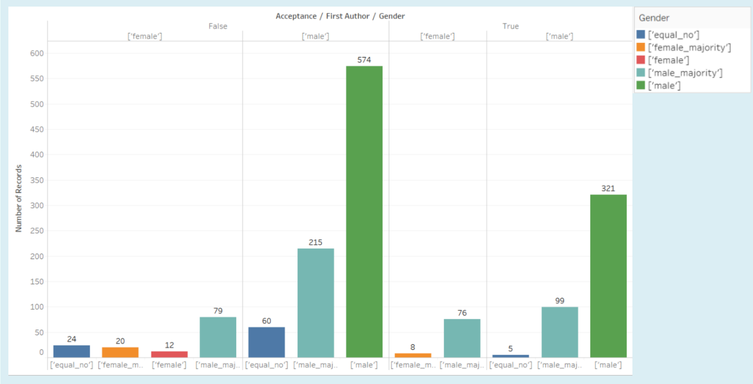
Average sentiment score against genders of lead authors on x-axis labeled by acceptance/rejection decisions and gender composition of authors. One key point to note is that 'equal_no' category has no lead female authors and that female_majority category has no male lead authors.

There are multiple interesting observations that can be made from this chart. Comparing the average sentiment scores for gender categories in accepted and rejected categories can give insights about any gender bias. For instance, if there is a substantial difference between the average sentiment score for two gender categories in either accepted or rejected categories might be indicative of bias. We did not find any substantial gender bias for ICLR 2017 because the average sentiment scores are consistent with the pattern we would expect in an unbiased situation.

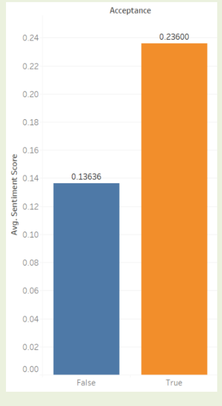


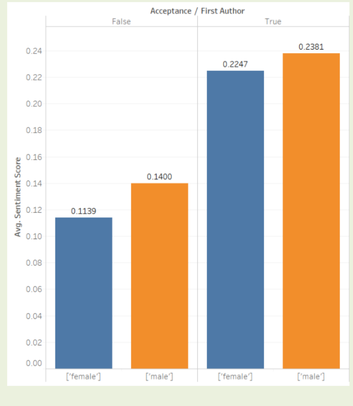
Percent of accepted versus rejected papers. ICLR being a selective conference has a acceptance rate of 34.09%.

Percentage of total number of reviews with respect to first author's gender. As expected, male authors outnumber female authors in computer science.



Number of records of reviews against genders of lead authors on x-axis labeled by acceptance/rejection decisions and gender composition of authors. One key point to note is that 'equal_no' category has no lead female authors and that female_majority category has no male lead authors. As expected, the most numerous group is male category authors with male lead authors.





Average sentiment score vs acceptance/ rejection and first author gender. As expected, accepted papers have higher sentiment score than rejected papers. There is a slight difference in the average sentiment score for male and female authors with male authors getting higher sentiment scores, this can indicate that male authors need higher scores to get accepted as compared to female authors.