Research Proposal:
Investigating Algorithmic Gender Bias in Resume Ranking

Marisa Langlois
Rochester Institute of Technology
1 Lomb Memorial Dr. Rochester, NY

October 31, 2018

Table of Contents

Summary

This project, Investigating Algorithmic Gender Bias in Resume Ranking, will determine if gender bias can be observed in resume rankings obtained from the hiring website Indeed, and whether career gender demographics and gendered writing style of resumes have any effect on this bias. This project is an interdisciplinary study that will be completed by a single student researcher over a ten-week period of time. I am an undergraduate computer engineering student in the Kate Gleason College of Engineering at Rochester Institute of Technology; my experience studying computer science and working as a research assistant for MIT Lincoln Laboratory has prepared me to succeed in undertaking this project.

Introduction

Today, computer algorithms run our world. They determine everything from the path we take when commuting to work, to the online advertisements we see, to the news we read. Algorithms run the stock market and, by extension, huge swathes of the global economy. Algorithms drive risk-assessment software, the results of which are more and more frequently used to make massive, life-changing decisions regarding the educational system, the financial system, and the criminal justice system, among others. The hold that this aspect of computer science has over society warrants extensive, thoughtful interdisciplinary research.

The idea that algorithms and mathematical models can act infallibly as impartial, third-party arbiters is a common misconception, and a dangerous one. In fact, algorithms can often be biased in unexpected and obscure ways. Recently, Amazon.com Inc. came under fire when it came to light that the algorithm used by its hiring department to rank resumes exhibited bias against women. The algorithm determined which resumes were preferable by comparing them to a bank of resumes, the great majority of which came from men. In effect, the algorithm "learned" to penalize resumes that included the word "women's," included the names of all-women's colleges, or were written using language more typical of women. That is, the algorithm interpreted "gender codes" and consistently ranked female-coded resumes as less desirable. The bias of this algorithm served to perpetuate gender bias and the male dominance of the tech industry. Clearly, research is sorely needed to investigate other hidden sources of gender bias in hiring.

This project was developed to fill that need. This project will serve as an interdisciplinary study that draws upon sociology, data science, and computer science by answering the question: *Is algorithmic bias displayed in the ranking of job applicants by the hiring website Indeed?* This project will improve upon "Investigating the Impact of Gender on Rank in Resume Search Engines" by Chen, et al. by not only determining the ranking of male job applicants compared to female, but will also determine the effect, if any, of the gender coding within each applicant's resume an industry's gender dominance on the ranking of male job applicants compared to female [3].

As a Dean's List student of computer engineering at the Kate Gleason College of Engineering at Rochester Institute of Technology, I have a proficient understanding of algorithms and a

higher-than-average sensitivity to the perpetuation of bias within these algorithms. This project would allow me to transform my pet interests in sociology, data science, and social justice into expertise, while having a genuine positive effect on these fields.

Methodology

I. Data Collection

This project will largely follow the methodology of "Investigating the Impact of Gender on Rank in Resume Search Engines" by Chen, et. al, with a few changes due to time constraints [3]. That is, Chen, et. al. employed an automated web crawler to gather the results of queries for 35 job titles in 20 U.S. cities on the hiring websites Indeed, Monster, and CareerBuilder, using all possible search filter values. This data collection took six months in total. In order to expedite the data collection process, this project will use an automated web crawler to gather the results of queries for 35 job titles in 10 U.S. cities on the hiring website Indeed, using no search filters. By focusing on the 10 most populous U.S. cities and on Indeed, the most popular of the three hiring websites, the gathered data should be similar in caliber to the data gathered by Chen, et. al [1, 2]. It is estimated that this abridged dataset will require two weeks to collect. The 35 job titles that will be searched for are shown below in Table 1. The 10 most populous U.S. cities are shown below in Table 2.

1. Accountant
2. Auditor
3. Bartender
4. Business Development Manager
5. Call Center Director
6. Cashier
7. Casino Manager
8. Concierge
9. Corrections Officer
10. Customer Service
11. Electrical Engineer
12. Elevator Technician
13. Financial Analyst
14. Human Resources Specialist
15. Janitor
16. Laborer
17. Mail Carrier
18. Manufacturing Engineer
19. Marketing Manager
20. Mechanical Engineer
21. Network Engineer
22. Occupational Therapist

23. Payroll Specialist
24. Personal Trainer
25. Pharmacist
26. Physical Therapist
27. Real Estate Agent
28. Registered Nurse
29. Retail Sales
30. Speech Pathologist
31. Software Engineer
32. Tax Manager
33. Taxi Driver
34. Technical Recruiter
35. Truck Driver

Table 1: Job titles to be used in queries of Indeed.

1. New York, New York
2. Los Angeles, California
3. Chicago, Illinois
4. Houston, Texas
5. Phoenix, Arizona
6. Philadelphia, Pennsylvania
7. San Antonio, Texas
8. San Diego, California
9. Dallas, Texas
10. San Jose, California

Table 2: U.S. cities to be used in queries of Indeed.

II. Data Analysis

The raw data will first be analyzed to separate the resulting candidates by inferred gender. Gender will be inferred, as in Chen et. al., by first name based on the probability of that first name being masculine according to the U.S. baby name dataset. Unfortunately, as in Chen et. al., this project will be restricted to operating within the gender binary. However, the results of this project will still be meaningful, as they will reveal whether hiring managers exhibit bias based on inferred gender, regardless of the actual gender of the individual candidate in question.

The same analysis performed by Chen et. al. will then be performed. That is, regression tests using a Mixed Linear Model (MLM) will determine the fairness of the candidate rankings produced by Indeed. This analysis will determine if the significant bias in favor of inferred male candidates found by Chen et. al. can be confirmed.

Two additional analyses will then be performed. First, the effect of the existing demographics of each examined job title will be determined by using data from the U.S. Bureau of Labor Statistics to rank each job by ratio of men to women. The MLM regression tests will then be repeated for each job. Second, the effect of male- and female-coded words in resumes will be determined by using a Python script to analyze how "male" or "female" the writing in each resume appears. The relationship between inferred gender determined by first name and inferred gender determined by resume word choice will be determined and used to find the effect, if any, on hiring bias.

Materials

I. Direct Resources

1. Indeed Recruiter account
2. Scrapy, automated web crawler framework
3. MATLAB
4. Python

II. Reference Materials

1. "Investigating the Impact of Gender on Rank in Resume Search Engines" by Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson

III. Possible Contacts

1. Le Chen, leonchen@css.neu.edu
2. Ruijun Ma, rma@stat.rutgers.edu
3. Anikó Hannák, ancsaaa@gmail.com
4. Christo Wilson, cbw@css.neu.edu

Schedule of Work

Week	Tasks	Deliverable
Week One	● Weekly advisor meeting. ● Implement Scrapy to collect necessary data. ● Research demographics of each job title. ● Research female- and male-coded resume language.	Weekly Progress Report
Week Two	● Weekly advisor meeting. ● Automated data collection. ● Implement Python script to analyze gendered writing style of each resume.	Weekly Progress Report

Week Three	● Weekly advisor meeting. ● Sort collected data by inferred gender. ● Run MLM regression tests to determine fairness of resume rankings.	Weekly Progress Report
Week Four	● Weekly advisor meeting. ● Continue MLM regression tests. ● Analyze gendered writing style of each resume using Python script.	Weekly Progress Report
Week Five	● Weekly advisor meeting. ● Repeat MLM regression tests to determine effect of demographics.	Weekly Progress Report
Week Six	● Weekly advisor meeting. ● Continue MLM regression tests to determine effect of demographics.	Weekly Progress Report
Week Seven	● Weekly advisor meeting. ● Repeat MLM regression tests to determine effect of gendered writing style.	Weekly Progress Report
Week Eight	● Weekly advisor meeting. ● Continue MLM regression tests to determine effect of gendered writing style.	Weekly Progress Report
Week Nine	● Weekly advisor meeting, seek feedback on Research Monograph Outline. ● Begin writing Research Monograph.	Weekly Progress Report, Research Monograph Outline
Week Ten	● Weekly advisor meeting. ● Finish writing Research Monograph.	Research Monograph

Table Three. Proposed project schedule.

Budget

Material	**Cost**
Indeed Recruiter account	$0.00
Scrapy	$0.00
MATLAB license	$250.00
Python	$0.00
1TB External Hard Drive	$50.00

Total Material Costs	**$300.00**

Table Four. Material budget.

Labor	**Total Hours**	**Total Labor Cost**
One researcher, at $20/hour	400	**$8,000.00**

Table Five. Labor budget.

Material Cost	$300.00
Labor Cost	$8,000.00
Total Project Cost	**$8,300.00**

Table Six. Total combined project budget.

Biographical Background

I. Researcher

MARISA LANGLOIS is an undergraduate student of computer engineering in the Kate Gleason College of Engineering at Rochester Institute of Technology. She expects to graduate with a Bachelor's of Science in computer engineering in December 2018. Her coursework, ranging from rigorous computer science classes to upper-level mathematics courses in calculus and statistics, has provided a good foundation for successful completion of this project. Furthermore, she has developed project management skills from her experiences completing a senior design project, Board Games Over IP, and from her previous employment as a research assistant with MIT Lincoln Laboratory. Her passion for studying the intersection of algorithms and sociology provides her with a deep understanding of the nuance inherent to this project.

II. Advisor

EZEKIEL LEO is an assistant professor of management in the Saunders College of Business at Rochester Institute of Technology. His background in computer science and economics gives him a unique perspective on the widespread issues of algorithmic bias. His relevant publications include "Measuring Business Relatedness for Big Data Strategic Analysis: A Deep Learning Approach."

Conclusion

Artificial intelligence algorithms already have a huge effect on the day-to-day lives of all members of American society. As the capabilities of artificial intelligence increase, this effect will only get larger and larger as algorithms become more and more integrated into society. Algorithms, like the people who create them, are imperfect, and research is necessary now to determine the unintended effects of these algorithms before lasting damage is done.

This project will tackle the unintended gender bias perpetuated by the Indeed resume rankings. It will confirm the findings of Chen, et. al. that gender bias is shown in these rankings, and will determine the effect of existing gender demographics and gendered writing styles on this gender bias. This work will increase gender equity in American society and even address the male-female wage gap by shining a spotlight on the unconscious biases of hiring managers that continue to hold women back in the workplace.

Furthermore, this project can be funded with less than the offered grant; thus, the National Institute for Research would maximize their return on investment by funding this project.

References

1. "Indeed.com Traffic Statistics." SimilarWeb, www.similarweb.com/website/indeed.com.
2. "US City Populations 2018." World Population Review, worldpopulationreview.com/us-cities/.
3. Chen, Le, et al. "Investigating the Impact of Gender on Rank in Resume Search Engines." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18, 2018, doi:10.1145/3173574.3174225.