Mitigating Socio-lingustic Bias in Job Recommendation

Ketki V. Deshpande **Shimei Pan** **James R. Foulds**
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore MD 21250
{ketkid1, jfoulds}@umbc.edu

With increasing diversity in the job market as well as the workforce, employers receive resumes from an increasingly diverse population. Many employers have started using automated resume screening to filter the many possible matches. Depending on how the automated screening algorithm is trained it may show bias towards a particular population by favoring certain socio-linguistic characteristics. The resume writing style and socio-linguistics are a potential source of bias as they correlate with protected characteristics. Studies and field experiments in the past have confirmed the presence of bias in the labor market based on gender, race (Bertrand and Mullainathan, 2004), and ethnicity (Oreopoulos, 2011). A biased dataset is often translated into biased AI algorithms (Rudinger et al., 2017) and de-biasing algorithms are being contemplated (Bolukbasi et al., 2016). In this work, we aim to identify and mitigate the effects of socio-linguistic bias on resume to job description matching algorithms.

We selected a total of 135 resumes of candidates applying in Singapore from a dataset due to Jai Janyani,[1] consisting of 45 each from Chinese, Malaysian and Indian origin candidates. We also manually collected job postings from each of the three countries. Following a popular approach in information retrieval (Tata and Patel, 2007), we converted all the resumes along with the job postings into a *tf-idf* matrix and found the similarity between the resumes and job postings by evaluating the cosine similarity score. We examined the top-10 matched resumes for each job posting, and found that bias was apparent towards Chinese resumes. Only 15% of the total matched resumes were of Chinese origin while more than 50% of the matched resumes were of Malaysian origin. A *t-SNE* plot of all resumes and job postings revealed that most of the job postings and Malaysian origin resumes were quite close together in the plot indicating document similarity while Chinese origin resumes were far from the others.

To mitigate this bias, we have developed a new *'fair tf-idf'* method, where we re-weighted all the *tf-idfs* with an extra fairness term, based on the p-% rule, a legal criterion for discrimination (Equal Employment Opportunity Commission, 1978). To calculate the fairness term, we first calculate $P(t \mid demographic)$, which represents the probability that a word t occurs in documents which come from one demographic group (eg. chinese resumes or job postings), was calculated for every demographic. Inspired by the p-% rule, the fairness term for each term t, which we call the *'p-ratio'*, is the ratio $P(t \mid demographic_1)$ / $P(t \mid demographic_2)$, where $demographic_1$ is the demographic with lowest $P(t \mid demographic)$ and $demographic_2$ is the demographic with highest $P(t \mid demographic)$. We then obtain the *'fair-tf-idf'* by multiplying the *tf-idf* value of every term t by its *'p-ratio'*.

$$\text{fair-tf-idf}(t) = \text{tf}(t) \times \text{idf}(t) \times \frac{P(t \mid demographic_1)}{P(t \mid demographic_2)}$$

This essentially means that the *fair-tf-idf* of words having same chance of occurring in any of the demographics is its *tf-idf*, while for those words which never occur in one of the demographics, the value of *fair-tf-idf* becomes zero. We again matched the resumes with job postings using *'fair-tf-idf'* and observed that the resumes were fairly matched with job postings from Malaysia and India. For the Chinese job postings, 90% of the Chinese origin resumes were matched, i.e. it overcompensated and became biased *towards* the Chinese origin resumes. We are currently working on annotating the matches manually to calculate matching accuracy, and we are investigating transformations of the *'p-ratio'* to control the fairness-accuracy trade-off.

[1] https://github.com/JAIJANYANI/Automated-Resume-Screening-System

References

Bertrand, Marianne and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, July 2003:991–1013.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *NIPS*.

Equal Employment Opportunity Commission. 1978. Guidelines on employee selection procedures. *C.F.R.*, 29.1607.

Oreopoulos, Philip. 2011. Why do skilled immigrants struggle in the labor market? a field experiment with thirteen thousand resumes. *American Economic Journal: Economic Policy*, 3:148–71.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Ethics in Natural Language Processing, Proceedings of the First ACL Workshop*, pages 74–79.

Tata, Sandeep and Jignesh Patel. 2007. Estimating the selectivity of tf-idf based cosine similarity predicates. *Sigmod Record*, 36.