Mitigating Demographic Bias in AI-based Resume Filtering

Ketki V. Deshpande
ketkid1@umbc.edu
University of Maryland,
Baltimore County
Baltimore, MD

Shimei Pan
shimei@umbc.edu
University of Maryland,
Baltimore County
Baltimore, MD

James R. Foulds
jfoulds@umbc.edu
University of Maryland,
Baltimore County
Baltimore, MD

ABSTRACT

With increasing diversity in the labor market as well as the work force, employers receive resumes from an increasingly diverse population. However, studies and field experiments have confirmed the presence of bias in the labor market based on gender, race, and ethnicity. Many employers use automated resume screening to filter the many possible matches. Depending on how the automated screening algorithm is trained it can potentially exhibit 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 such as ethnicity. A biased dataset is often translated into biased AI algorithms and de-biasing algorithms are being contemplated. In this work, we study the effects of socio-linguistic bias on resume to job description matching algorithms. We develop a simple technique, called fair-tf-idf, to match resumes with job descriptions in a fair way by mitigating the socio-linguistic bias.

CCS CONCEPTS

• **Information systems**; • **Computing methodologies** → **Machine learning approaches**;

KEYWORDS

tf-idf, fair machine learning, job recommendation, term weighting

ACM Reference Format:
Ketki V. Deshpande, Shimei Pan, and James R. Foulds. 2020. Mitigating Demographic Bias in AI-based Resume Filtering. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3386392.3399569

1 INTRODUCTION

According to Glassdoor, a popular job search engine and a review site, on an average a company receives around 250 resumes for each job posting, and the number can be even higher for Fortune 500 companies. Out of these, only four to six qualified candidates are called for an interview [11]. As the internet became the preferred place for posting as well as accepting job applications, an increasing number of companies have started using software for filtering out the resumes. Only a handful of qualified candidate resumes are actually seen by recruiters or hiring managers. According to a study published in 2015, almost 75% of the recruiters and/or talent managers use software for recruiting or applicant tracking [14]. The recruiting software is trained on a training dataset that is subject to biases in the actual process of hiring, which may be transferred [1]. Studies in the past have proven that discrimination based on gender [5], race [4], and ethnicity [16] is prevalent in job market even when the employer claims to be an "Equal Opportunity Employer." If the training data, which consists of details of actual hired candidates, is not diverse enough, the resulting software can produce biased recommendations [8]. Linguistic and accent differences are a major contribution in generating prejudiced impressions against women, minority groups or non-native speakers [6].

We center our study on discriminatory behavior in the automated matching component of the hiring process regarding the origin country of the applicants, potentially arising from socio-linguistic tendencies. To mitigate the socio-linguistic bias in resume screening process, we propose de-biasing methods that penalize matching keywords that are typical to one section of society while encouraging matching keywords that are common among all demographics. Five different methods are evaluated based on fairness and accuracy measures. The experiments show that our proposed method, *fair-tf-idf with Sigmoid Transformation*, provides an adjustable balance between accuracy and fairness, and leads to the most desirable fairness/accuracy result according to our criteria.

2 RELATED WORK

In the past, many studies have been conducted which demonstrated the bias in recruitment with respect to gender, race, ethnicity or the accents of the applicants. Some of the studies focused on racial discrimination in the recruitment process, e.g. [4], which confirmed the presence of bias in labor market based on race.

[4] demonstrated that interview calls received were almost 50% higher for white sounding names compared African-American sounding names. They further show that by increasing the quality of resumes the interview calls were increased by 30% for white sounding names as compared to a marginal increase for African-American sounding names. Another such study conducted in 2009 targeted racial as well as ethnic discrimination in a job market in Canada [16].

In the book "Weapons of Math Destruction" [15], O'Neil discusses a personality test conducted by companies before hiring the candidates which was designed by experts that may not take into account any feedback. Candidates who get red-lighted as a result of this automated screening test suffer without knowing the reason. She also explains in detail how the bias in training data

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy

ACM ISBN 978-1-4503-7950-2/20/07…$15.00
https://doi.org/10.1145/3386392.3399569

(past human decisions in this case) is inherited by machine learning algorithms. O'Neil illustrates this with an example of the screening process of a medical school which was highly biased [13]. When the school decided to build a computer model for screening the applicants, they captured the bias from previous human decisions and the resulting algorithm discriminated against non-native English speaking candidates and female candidates [13].

Rudinger et al., in 2017 [17] studied social bias in natural language inference corpora and how it is susceptible to amplification. They demonstrated that the hypotheses of Stanford Natural Language Inference (SNLI) consist of various types of stereotypes including gender-based, age-based, racial or religious bias.

While the previous studies were mostly field experiments and did not talk about their application to training a machine learning algorithm, numerous recent studies have been conducted to find out the impact of biased data on the trained machine learning models [1]. Numerous debiasing methods have been proposed, several of which are summarized by [3]. Notable studies that have addressed "fairness" focused on two different approaches. Some studies like Feldman et. al. [9] proposed methods to make the the dataset unbiased, whereas others like Bolukbasi et. al. [5] proposed a debiasing algorithm for achieving fairness. [5] showed that word embeddings trained on very general input data like news articles display gender-biased word associations. They proposed an algorithm to "de-bias" the word embeddings to reduce gender stereotypical associations.

While previous works including [2, 5] have proposed debiasing algorithms for text data, they mainly focused on word embeddings. Our method instead focuses on debiasing tf-idf representations.

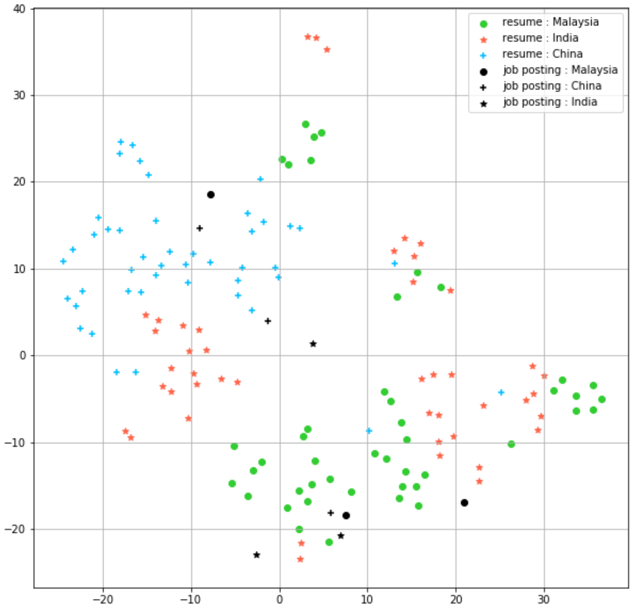

Figure 1: A *t*-SNE plot showing all the collected resumes and job postings, based on tf-idf features.

3 DATASET

We studied nationality bias in automated resume filtering in the context of a relatively pluralistic nation, Singapore. The population of Singapore consists of three major ethnic groups: Chinese, Malaysian and Indian. Around 43% of its population consists of foreign born people. As we wanted to study the socio-linguistic bias, which can be a result of the different linguistic styles being used in different countries, we selected the resumes of candidates born in China, Malaysia and India. We used 135 resumes of candidates applying to accounting and finance jobs in Singapore collected by Jai Janyani,[1] consisting of 45 candidates each from Chinese, Malaysian and Indian origin, the primary demographics in Singapore. We manually selected the resumes from the said dataset based on carefully reading and understanding the origin of candidates. Only the resumes where origin of the candidates can be clearly inferred were selected for this study. The categorization of resumes into Chinese, Indian or Malaysian origin was done based on the education or initial employment records. For example, candidates were categorized as Chinese origin if they have completed all or initial education in China and/or have started their employment in China and then moved to Singapore or applied for a job in Singapore.

To assess and mitigate the bias in matching job descriptions and resumes, we manually collected 9 finance and accounting job postings from a popular job site with 3 postings from each country (China, India and Malaysia). Although our focus is on nationality-based AI hiring discrimination in Singapore, we used the job ads posted by people from those countries since we could use the national origin as a ground truth label for the nationality of the people who posted the ads. Finally, all job-resume pairs were annotated according to whether the candidate was qualified for the job (a binary label), by three annotators. The majority vote was used to determine the final ground truth label regarding whether the candidate was a "match" for the job or not. The aim of the study is to match the resumes with the job postings in an unbiased way irrespective of the country in which they were posted. For example, for a job posting from China, the possibility of resumes getting selected from all the three countries should be equal and only depend on the qualifications of candidates rather than their country of origin.

4 THE BIAS EXPLAINED

An AI-based resume filtering algorithm aims to find resumes that match the job postings. The standard text-based document retrieval approach [12] uses tf-idf vectors to represent each query and each document [10]. In our case, a job advertisement is considered a "query" and a resume is a document to retrieve. We rank the match of a resume and a job posting based on the cosine similarity of the job posting vector and the resume vector. In our experiments, we selected the top 5 resumes for a particular job posting based on the cosine similarity between them. Our first goal is to examine the fairness behavior of this standard approach.

It was observed that out of all the resumes selected by our algorithm, around 46.66% were from Malaysia, 42.22% were from India, while Chinese resumes formed only 11.11% of the total selected resumes. To formally define and quantify fairness, we used the following *p% Fairness Measure* in our study:

$$FairnessMeasure = \frac{P(match \mid demographic_1)}{P(match \mid demographic_2)}, \qquad (1)$$

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

Table 1: Fairness Measure and Accuracy for Tf-idf

Country (of Job Posting)	Fairness Measure	Accuracy
China	0.7272	90
India	0.0526	80
Malaysia	0.0526	100
Overall	0.0723	90

where $P(match \mid demographic_1)$ is the value from the demographic with the lowest match probability value given a fixed demographic of job posting and $P(match \mid demographic_2)$ is the value from the demographic with the highest match probability value given a fixed demographic of job posting.

This widely used fairness metric is adapted from the 80% rule for disparate impact analysis, a legal criterion which identifies discriminatory behavior if the ratio is less than 0.8 [7]. Here, the *p*% is expressed as a ratio between 0 and 1, rather than a percentage. The lower the value of Fairness Measure, the more "unfair" it is in terms of disparity. We found that the fairness properties varied greatly with respect to the country of the job posting (Table 1). The overall fairness measure across all job postings came out to be 0.0723, very far below the legal threshold of 0.8.

To evaluate the accuracy of the matching process, we labeled all possible job-candidate matches using the majority vote from our three annotators. For this part of the experiment, we considered only the top 5 resumes retrieved. The accuracy of the tf-idf method varied by country but was generally high, being 90% accurate overall (Table 1).

We used a *t*-SNE plot [18] to visualize all the tf-idf representations of resumes and job postings together in two dimensions in order to better understand the tf-idf algorithm's disparate behavior (Figure 1). We observe substantial clustering of resumes by national origin, especially for the Chinese resumes. It is visible that many job postings are closer to Malaysian-origin resumes, whereas fewer postings were near Chinese origin resumes. Also, it can be seen that Chinese job postings are closer to Chinese origin resumes than any other job postings, making it difficult to match Chinese resumes to non-Chinese ads. These factors may partly explain the large differences in the fairness metrics between countries.

To more closely examine the reasons for these demographic differences, we plotted word clouds to show the top words used by applicants from each country of origin. Figures 2, 3 and 4 show word clouds for China, India and Malaysia respectively. It is visible from the word clouds that, apart from the common words "financial," "management," "business" and "accounting," location words like "China," "India" and "Malaysia" are also the most frequent terms. The tf-idf weighting did not strongly down-weight those terms, as the terms typically appear in only those resumes which belong to the candidates from that particular country (and similarly for the job ads). Such differences in language patterns between demographics could explain a substantial fraction of the disparity.

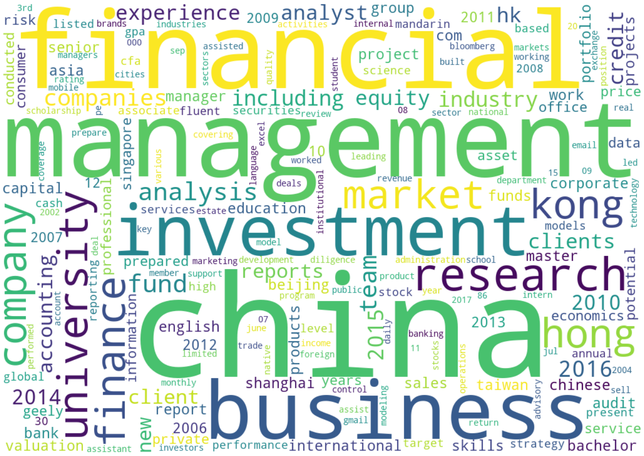

Figure 2: Word cloud for resumes from China.

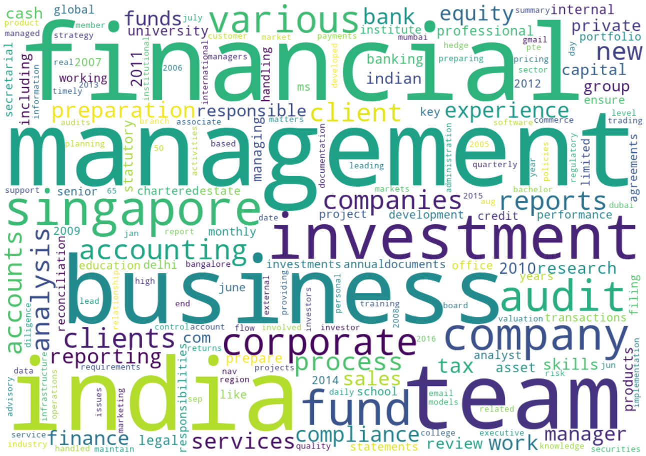

Figure 3: Word cloud for resumes from India.

5 METHODOLOGY

The above results strongly suggest that systematic differences in word usage between demographic groups can substantially contribute to disparate behavior in job-resume matching methods such as tf-idf-based document retrieval. This phenomenon is consistent with studies on the impact of sociolinguistic bias in human hiring decisions [6]. Accordingly, we proposed and evaluated various approaches for fairly matching resumes to job descriptions, by mitigating bias due to sociolinguistic behavior. Our overall approach is to modify tf-idf to correct for demographic bias in word usage.

5.1 Fair-tf-idf

We propose a new method called *fair-tf-idf*, where we re-weight the previously calculated tf-idf values with an extra fairness term to make the word features fair for all demographics. Analogously to the *p% Fairness Measure* (Equation 1), we perform the re-weighting of the tf-idf values in a manner inspired by the legal criterion for discrimination, the *p*-% rule [7]. We calculate a fairness term for

Figure 4: Word cloud for resumes from Malaysia.

each term t, which we call the *'p-ratio(t)'* . For each term, the *'p-ratio(t)'* is calculated as:

$$p\text{-}ratio(t) = \frac{P(t \mid demographic_1)}{P(t \mid demographic_2)} \;, \tag{2}$$

where $demographic_1$ is the demographic with the lowest $P(t \mid demographic)$ and $demographic_2$ is the demographic with the highest $P(t \mid demographic)$. To calculate the fairness weight, we first calculate $P(t \mid demographic)$, where $P(t \mid demographic)$ represents the probability that a word t occurs in documents which come from one demographic group. For example, if a word occurs in 20 out 45 documents, where all the 45 documents are from a same demographic group (country of origin in this case), then $P(t \mid demographic)$ is $\frac{20}{45}$.

We then obtain the *'fair-tf-idf'* by multiplying the *tf-idf* value of every term t by its *'p-ratio'*:

$$\text{fair-tf-idf}(t) = \text{tf}(t) \times p\text{-ratio}(t) \;.$$

The p-ratio(t) is always between 0 and 1. The effect of *fair-tf-idf* essentially is that, if a word occurs equally in resumes from all the 3 demographics, then the values $P(t \mid demographic_1)$ and $P(t \mid demographic_2)$ would be same, and the *fair-tf-idf* value will be equal to its *tf-idf*. For those words which never occur in one of the demographics, however, the value of *fair-tf-idf* becomes zero.

The tf-idf vectors are thus converted into a reweighted *"fair-tf-idf"* vectors for all the 135 resumes and job postings. We normalize the tf-idf vectors to unit length and use the cosine similarity with a given job posting to rank the resumes, then select, e.g., the top-5 resumes.

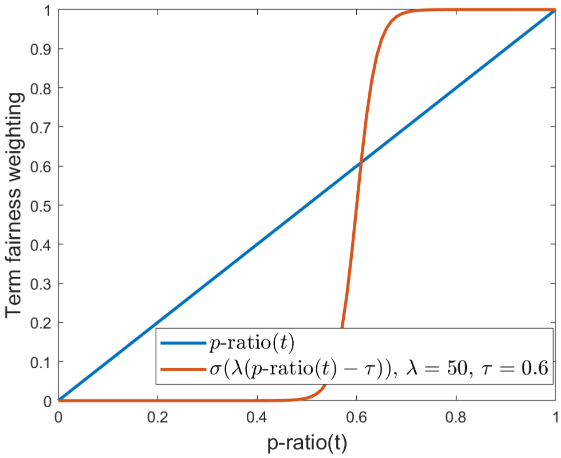

Figure 5: The Effect of the Sigmoid Transformation on the *p*-ratio.

5.2 Fair-tf-idf with Sigmoid Transformation

Although intuitive, the direct use of *fair-tf-idf*, as described above, has a few limitations. Fair-tf-idf negated the effect of the idf weighting for certain stopwords, giving them more weight again (see experiments below). This is because fair-tf-idf increased the relative weights of words that are common among all the demographics, which typically includes stopwords. A second concern is that the method does not provide users with the ability to control the fairness/fidelity trade-off. To address both of these concerns, we further propose an enhanced method in which the *p-ratio*'s are passed through a *sigmoid transformation* before multiplying them with the tf-idf:

$$\text{fair-tf-idf}(t) = \text{tf}(t) \times \text{idf}(t) \times \text{transformation}(p\text{-ratio}(t)) \text{ , where}$$

$$transformation(p\text{-ratio}(t)) = \sigma(\lambda(p\text{-ratio}(t) - \tau)) \tag{3}$$

and σ is the sigmoid function,

$$\sigma = \frac{1}{1 + e^{-x}} \;. \tag{4}$$

The τ in the above transformation function is a translation parameter which roughly determines the cut-off point in terms of the p-ratio to keep the tf-idf unchanged, i.e. it encodes a degree of tolerance in the demographic differences of the terms (see Figure 5). The value of τ should generally be set between 0 and 1. The hyperparameter $\lambda > 0$ is a scaling parameter which determines how sharply the weights drop off. The effect of this transformation is roughly that, for $\lambda \geq 50$ or so, the terms whose *p-ratio* is at least around τ are kept with approximately their original tf-idf weight, and the remaining *"unfair"* words below the cut-off point are sharply discounted. This keeps the relative weight of stop words and other relatively *"fair"* words unaffected, which helps to solve the stop-words issue.

5.3 Term Frequency (TF) Baseline

Tf-idf was introduced as an enhancement in information retrieval by adding a term weighting system along with term frequency (tf) to de-emphasize common words [10]. But in our case, for matching resumes to job postings, we suspected that tf-idf might down-weight the keywords such as educational degrees or skills. We therefore also considered a simple term frequency method as a baseline. The tf baseline is performed analogously to the other methods, using cosine similarity for ranking the resumes.

Table 2: Comparison — Percentage of resumes selected from each demographic

	Percentage of Selected Resumes		
	Country of Candidate		
Method	China	India	Malaysia
TF-IDF	11.11	42.22	46.66
TF	8.88	43.33	47.77
Fair TF-IDF	**25.5**	**33.33**	**41.11**
Fair TF	17.77	36.66	45.55
Fair TF-IDF Sigmoid ($\lambda = 50$, $\tau = 0.6$)	**26.66**	**31.11**	**42.22**
Limiting Number of Resumes (LNR)	33.33	33.33	33.33

Table 3: Comparison — Fairness Measure for each demographic

	Fairness Measure			
	Country of Job Posting			
Method	China	India	Malaysia	Overall
TF-IDF	0.7272	0.0526	0.0526	0.0723
TF	0.5833	0	0.0583	0
Fair TF-IDF	0.3846	0.2142	0.5833	0.3673
Fair TF	0.4	0.1428	0.3846	0.3571
Fair TF-IDF Sigmoid ($\lambda = 50$, $\tau = 0.6$)	**0.5**	**0.4615**	**0.4615**	**0.923**
Limiting Number of Resumes (LNR)	1	1	1	1

5.4 Fair TF Baseline

For completeness, we consider a baseline where we multiply the term frequency with our p-ratio(t) term:

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

where $demographic_1$ is the demographic with the lowest $P(t \mid demographic)$ and $demographic_2$ is the demographic with the highest $P(t \mid demographic)$.

5.5 Limiting Number of Resumes (LNR) Baseline

In this baseline method, limiting the number of resumes (LNR), we used simple tf-idf terms, but fixed the number of resumes that can be matched from each demographic with the given job posting to be proportional to the number of applicants from that demographic, to ensure parity. As the proportions of resumes that are matched from each demographic, $P(match \mid demographic)$, are equal, the results are always fair in the sense that the $p\%$ Fairness Measure is 1, which is the optimal value.

The major drawback with this method, however, is that we *need to have demographic labels for every resume at test time* while matching the resumes with job postings. Since this demographic information is legally protected in many contexts, it may be difficult to obtain for test users. In contrast, our approach only needs *demographic labels for the training data*, which are often easier to obtain. The parity constraint can also be very harmful to the accuracy of the matching when there are demographic skews in the qualified candidates per individual positions, as we will see in our experiments below.

6 EVALUATION

We evaluated all five methods based on both *accuracy* and the *p% Fairness Measure*, using all 135 resumes and 9 job postings. We also reported t-SNE plots showing the position of resumes and job postings in two-dimensional space after performing de-biasing.

6.1 p% Fairness Measure

Table 2 compares the percentage of total resumes from each demographic that were selected as a match to all the 9 job postings for each of the methods. We observe that the results are relatively fair with the fair-tf-idf and fair-tf-idf with Sigmoid Transformation methods. The Sigmoid Transformation method has two hyperparameters which affect the results. For Table 2, we used default untuned hyperparameter values, $\lambda = 50$ and $\tau = 0.6$, which leaves the tf-idf weights of words with p-ratio$(t) > 0.6$ relatively undisturbed, while setting the rest to approximately 0 (see Figure 5). We study the impact of these hyperparameters in the next section. Note that the baseline method which limits the number of resumes (LNR) always selects an equal proportion of the resumes per demographic. This means that the *Fairness Measure* will always be 1, which is the maximum value. Table 3 compares the $p\%$ Fairness Measures of each of the demographic as well as the overall Fairness Measure. The results show that the Fairness Measure value is best when using fair-tf-idf with the Sigmoid Transformation method, except for the LNR baseline.

6.2 Impact of Hyperparameters on Fairness

The fair tf-idf with Sigmoid Transformation method has two hyperparameters to set, λ and τ. We performed experiments to study the

Table 4: Comparison — Accuracy of each Method

	Accuracy (Percentage)			
	Country of Job Posting			
Method	China	India	Malaysia	Overall
TF-IDF	90	80	100	90
TF	100	90	90	93.33
Fair TF-IDF	70	80	70	73.33
Fair TF	80	60	80	73.33
Fair TF-IDF Sigmoid ($\lambda = 55$, $\tau = 0.32$)	80	80	100	86.66
Limiting Number of Resumes (LNR)	80	60	90	76.66

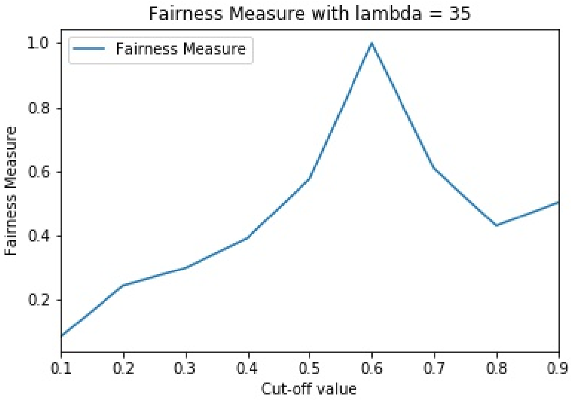

Figure 6: p% Fairness Measure for $\lambda = 35$, varying τ.

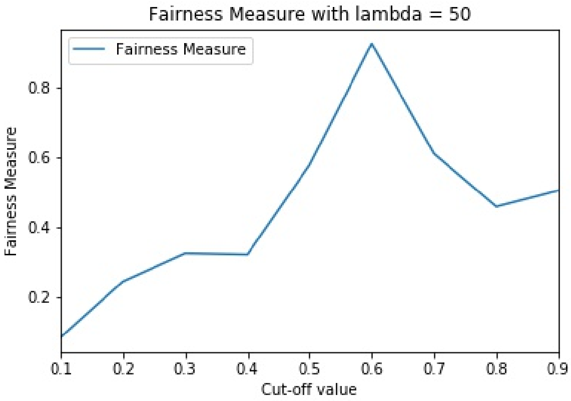

Figure 7: p% Fairness Measure for $\lambda = 50$, varying τ.

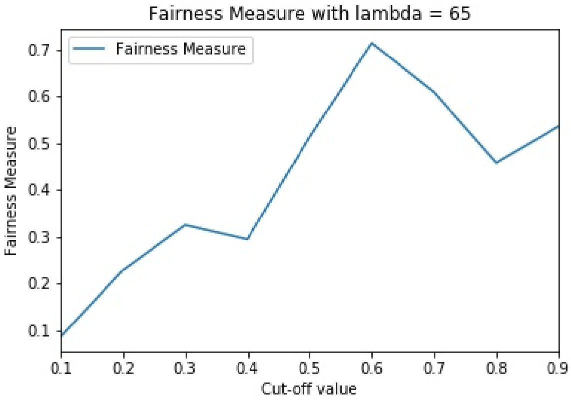

Figure 8: p% Fairness Measure for $\lambda = 65$, varying τ.

impact of the hyper-parameter values, reporting few graphs which display how the p% Fairness Measure changes with the values of λ and τ.

The graphs plotted in the Figures 6 – 8 show that the p% Fairness Measure changes drastically with the cut-off value τ. The fairness generally dips after around $\tau = 0.6$, which is due to the top of the "S" in the sigmoid curve being shifted past the maximum value of p-ratio(t) = 1.0, which results in a "J-shaped" weighting curve, instead of an "S-shaped" curve. The λ hyperparameter also significantly impact the performance. Among the different λs we tested, the peak fairness measure was achieved when λ=35.

6.3 Accuracy

The accuracy is calculated by the percentage of correctly matched resumes among the top N resumes selected by a system. We used three annotators to determine resume matches. We provided annotators with anonymous resumes, where we masked the demographic information. They were asked to annotate a job-resume pair as a match if the required education and/or experience correctly matches the job description. A candidate was considered a match even if he or she was overqualified for the job. We disregarded language requirements in the annotations. Annotators were asked to annotate all the other cases as "no-match." We used a majority vote by the annotators to determine the final label and used this in system evaluation. We calculated average inter-annotator agreement and got the average Kappa value of 0.55.

Our aim is to make selection of resumes fair for each demographic without losing much of the accuracy.

Table 4 compares the total accuracy of each method along with per-demographic accuracy. It can be observed from Table 4 that the baseline which limits the number of resumes from each nationality (LNR) has an accuracy rate of 76.66% overall, substantially lower

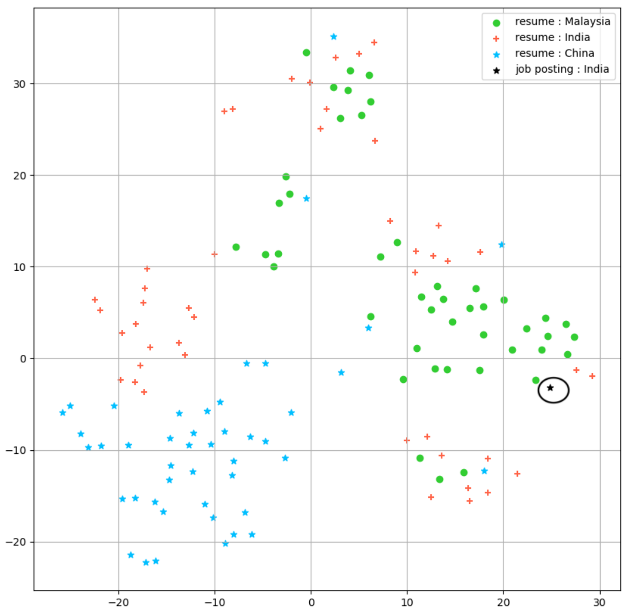

Figure 9: *t*-SNE plot for Tf Method.

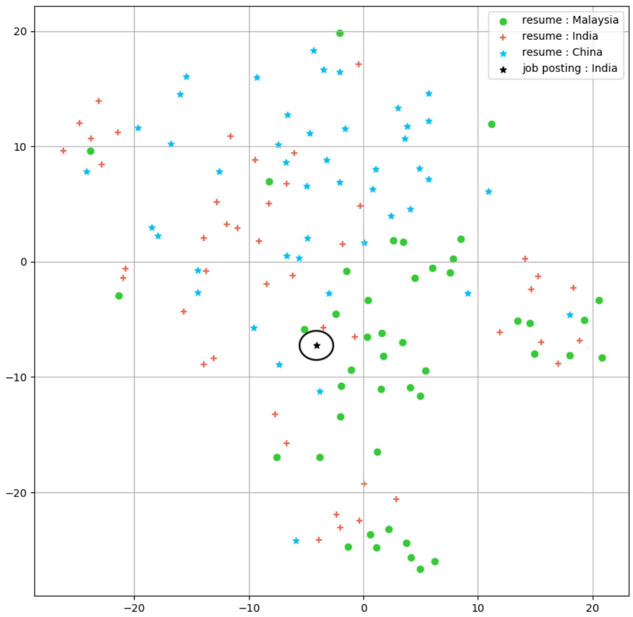

Figure 10: *t*-SNE plot for Fair-tf-idf Method.

than the original TF-IDF method. This method has a perfect 1 *p*% Fairness Measure. However, the LNR method needs nationality labels for each resume during testing and predicting time so that the equal proportion of resumes can be selected. Having labels for nationality is not always practical in many contexts, since applicants may not be willing to provide this sensitive information.

The accuracy for fair-tf-idf with Sigmoid Transformation, at 86.66%, was the best of the fair algorithms, while simultaneously achieving excellent fairness compared to the baselines other than LNR. The total accuracy values for all the other fair baseline methods lie between 73% and 77%.

6.4 Visualization

We also created some visualizations to demonstrate how our proposed de-biasing methods were effective in selecting the resumes fairly. The *t*-SNE plots in Figures 9– 11 represent all the resumes plotted in 2-dimensional space using three of the methods mentioned in the previous section. For clarity and ease of comparison, only one job description has been plotted along with the 135 resumes. The plots illustrate the matching behavior due to any demographic clustering behavior in the resumes, and similarities between the job posting and the resumes.

In the plot for TF method (Figure 9), it can be seen that most of the Malaysian and Indian resumes are close to the job posting. So, when resumes were selected using cosine similarity it picked most of the Malaysian and Indian resumes. In contrast, the Chinese resumes are further away from the job posting. Most of the resumes are also clustered according to their demographics, which is likely to cause disparity in the matching.

Figures 10 and 11 represent the *t*-SNE plots of all the resumes and the same job posting after applying fair-tf-idf and fair-tf-idf with the Sigmoid Transformation respectively. After applying these techniques, it can be observed that the job posting is close to several resumes from each demographic. It can also be observed that resumes with the same country of origin are no longer strongly clustered into different regions, but have substantial overlap. This was

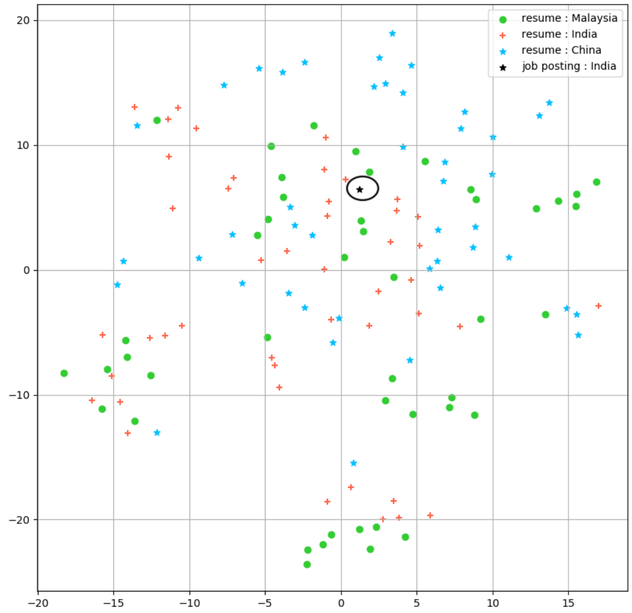

Figure 11: *t*-SNE plot for Fair-tf-idf with Sigmoid Transformation Method.

Table 5: Top Words from China before and after de-biasing

Before de-biasing	After de-biasing
china	management
management	financial
financial	research
business	investment
investment	business
research	credit
university	finance
finance	company
company	university
kong	fund

especially the case when applying fair-tf-idf with the Sigmoid Transformation, making improved demographic parity in the matching much more likely to occur.

6.5 Top Words

We also compared the 10 top words from each demographic before and after applying fair-tf-idf with the Sigmoid Transformation in Table 5. The main differences are that the words *"China"* and "Kong" (as in "Hong Kong"), common words in Chinese resumes, are no longer top word after applying the de-biasing method. It seems that our method down-weights nationality-specific proper nouns, which can be a source of bias. In contrast, the word "credit", which was not present before de-biasing, was promoted to the top word list after de-biasing. Other job-related keywords such as "financial," "management," and "business," remain in the top word list. This is important since these words are very relevant for matching.

7 CONCLUSION AND FUTURE WORK

We have studied how socio-linguistic patterns can lead to demographic bias in text-based algorithms for matching resumes to job postings, and we proposed several fair modifications to the tf-idf matching method to correct for these issues. It can be seen from our experimental results that our proposed fair-tf-idf with Sigmoid Transformation method performs well in de-biasing the data and selecting resumes fairly and accurately. The parity in the distribution of selected resumes was improved compared to the results when no de-biasing was performed, as quantified by the $p\%$ Fairness Measure, with relatively little loss in accuracy compared to traditional tf-idf matching.

Along with the fair matching of resumes to job postings, we expect that this method can be extended to provide debiased career recommendations to applicants. Thus the model is not only useful for recruiters but also for the applicants.

We also believe this method can be generalized and used in other contexts. For example, in addition to nationality, the same method can be used to mitigate gender bias in text retrieval. In fact, the proposed method can be used to improve the fairness of a wide range of text processing applications where tf-idf text representations are used.

The main limitation of this study was that only a small number of resumes was collected due to the burden of data annotation. In the future, we plan to collect and annotate a larger resume dataset to further validate our results. This will facilitate the development and use of fair deep neural network models for matching job posts and resumes.

ACKNOWLEDGMENTS

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant Nos IIS 1850023; IIS1927486. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] S. Barocas and A.D. Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[2] Schmidt Ben. 2015. *Rejecting the gender binary: a vector-space operation.* Retrieved April 12, 2020 from http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *In Sociological Methods and Research* 1050 (2018), 28.

[4] Marianne Bertrand and Sendhil Mullainathan. 2003. ARE EMILY AND GREG MORE EMPLOYABLE THAN LAKISHA AND JAMAL? A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION. *American Economic Review* 94, 9873 (July 2003), 991–1013. https://doi.org/10.3386/w9873

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

[6] Faye K Cocchiara, Myrtle P Bell, and Wendy J Casper. 2014. Sounding "Different": The Role of Sociolinguistic Cues in Evaluating Job Candidates. *Human Resource Management* 55, 3 (November 2014), 463–477. https://doi.org/10.1002/hrm.21675

[7] Equal Employment Opportunity Commission. 1978. Guidelines on employee selection procedures. *C.F.R.* 29 (1978), 1607.

[8] J. Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (2018).

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[10] Salton Gerard and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24 (1988), 513–523. Issue 5. https://doi.org/10.1016/0306-4573(88)90021-0

[11] Glassdoor. 2019. *50 HR and Recruiting Stats for 2019.* Glassdoor. Retrieved January 12, 2020 from https://www.glassdoor.com/employers/resources/hr-and-recruiting-stats-2019

[12] Anna Huang. 2008. Similarity Measures for Text Document Clustering. New Zealand Computer Science Research Student Conference, Christchurch, New Zealand.

[13] Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. *British medical journal (Clinical research ed.)* 296, 6623 (1988), 657.

[14] J.P. Medved. 2015. *Recruiting Software Impact Report.* Capterra. Retrieved January 12, 2020 from https://www.capterra.com/recruiting-software/impact-of-recruiting-software-on-businesses

[15] Cathy O'Neil. 2016. *Weapons of Math Destruction* (1st. ed.). Crown Publishing Group, New York, Chapter Getting a Job.

[16] Philip Oreopoulos. 2011. Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment With Thirteen Thousand Resumes. *American Economic Journal: Economic Policy* 3 (2011), 148–71. https://doi.org/10.1257/pol.3.4.148

[17] Rachel Rudinger, 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*. 74–79.

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.