An Investigation of Ethical Bias in Embeddings

Will Landin
Lyle School of Engineering
Southern Methodist University
Los Angeles, California
wlandin@smu.edu

Rick Lattin
Lyle School of Engineering
Southern Methodist University
Houston, Texas
rlattin@smu.edu

Ryan Schaefer
Lyle School of Engineering
Southern Methodist University
Dallas, Texas
rdschaefer@smu.edu

***Abstract*—This study investigates ethical biases in word embeddings, focusing on GloVe and ConceptNet Numberbatch, and their implications in resume application contexts. With the advent of online job applications, applicant tracking systems (ATS) are widely used, often inadvertently filtering out qualified candidates due to embedded biases. This paper explores the extent of gender, racial, and educational biases encoded in these embeddings by analyzing their impact on sentiment analysis of resume excerpts of our creation. We trained sentiment classifiers using both aforementioned embeddings on the Sentiment140 dataset, then applied these models as a proxy to analyze the sentiment of modified resume excerpts representing various genders, races, and educational backgrounds. The analysis revealed that both embeddings do exhibit biases, but GloVe showed significantly higher bias levels compared to ConceptNet Numberbatch. This paper does not aim to say that having a lower sentiment score means that your chances of getting hired are lower, but instead aims to identify some weaknesses found in GloVe and ConceptNet Numberbatch.**

I. Motivation

Entering the job market in 2024 presents many challenges, with online applications becoming a new standard and larger companies like Google seeing upwards of 3 million applications a year and rising [1]. Recent university graduates are often told that finding the right job for you is not only hard work but sometimes a little bit of luck as well [2]. But how much of the perceived "luck" is actually a result of bias?

To handle large applicant pools more efficiently, many companies employ machine learning models called applicant tracking systems (ATS) to screen applicants' resumes prior to a manual review [3]. While these models can save employers time and resources, they can be problematic. People with great qualifications for a job position can, be filtered out for what seems to be no reason at all. If a resume screening model is biased against someone based on something in their resume, that could significantly harm their chances of getting the job, even if they are the perfect candidate [4].

Another motivation stems from the experience of a chemistry professor, whom we met at Santa Monica College who as a brilliant Persian student named Mohammed graduated with a master's degree at 17, yet faced difficulties securing employment from 2001 to 2008, likely due to post-9/11 prejudices against his name. Only after changing his resume name to "Marcus" did he begin receiving responses for relevant positions despite his outstanding qualifications. This anecdote highlights how implicit biases can manifest in hiring practices, underscoring the importance of identifying and mitigating such biases embedded within language models, resume screening tools, and word embeddings.

An important part of many models that could be a source of such biases is the embeddings used to help models understand the meanings of words. This paper aims to analyze potential biases found in resumes, by comparing sentiment classifiers trained with two of the more popular static embeddings: GloVe and ConceptNet Numberbatch. The goal is to identify and explore any shortcomings given the resume context to see the extent and limitations of using these embeddings given specific race, education, and gender contexts.

This study aims to address the following research questions: How do sentiment classifiers trained on the GloVe and ConceptNet Numberbatch embeddings differ in how the inclusion/exclusion of potentially biased words affects the predicted sentiment? Which types of bias potentially found in resumes, such as gender, race, or educational background, have the most influence on the classifiers?

We hypothesize that classifiers trained on both GloVe and ConceptNet Numberbatch will be biased, but the ConceptNet classifier will be less biased than the GloVe classifier. This hypothesis is based on how ConceptNet was designed with bias mitigation in mind, whereas GloVe was solely trained on statistical co-occurrences of words. We also hypothesize that resumes with male-gendered words will produce more positive results than female-gendered words due to historic gender bias.

II. Related Work

There has been a plethora of research done in the field of investigating bias within resumes and adjacent fields. The primary differences that our research holds when compared to the other papers within the field, is the specific content we are investigating and the width of the analysis we are doing. Our research consists of an investigation of bias in multiple different pre-trained embeddings through the lens of race, gender and education within the scope of real-life resumes. The investigation of bias in embeddings has been done before, [5] [6] [7] but throughout all of the embedding bias papers we investigated, none of them performed any form of bias analysis on ConceptNet. The most common embedding analysis we found was on BERT [8], but we instead performed bias analysis on the ConceptNet and GloVe embeddings. As a result, we decided to leave BERT out of our research, as we would be providing nothing new to the conversation.

Most of the relevant papers solely look at gender bias [9] [10] [11], but in contrast, we want to investigate race and education in addition to gender. This is already more of a wide-spanning investigation than the state of the field, but we also perform cross-comparisons using these different aspects of bias, further providing results to support our hypotheses. Finally, we could not find a single paper that talked about bias in embeddings within the context of resumes. The majority of the existing papers regarding bias in resumes discuss AI models that are tangential to the text semantic investigation that we are doing, but are by no means the same experiment. The existing research focuses much more on investigating real-life hiring situations while using resumes as the tool of investigation [12] [13]. We are using resumes as a tool for investigation as well, but the target of our investigation is still very different.

III. Methods

Our methodology to identify the potential biases in the embeddings lies in the experiments we ran, as visualized in Figure 1.

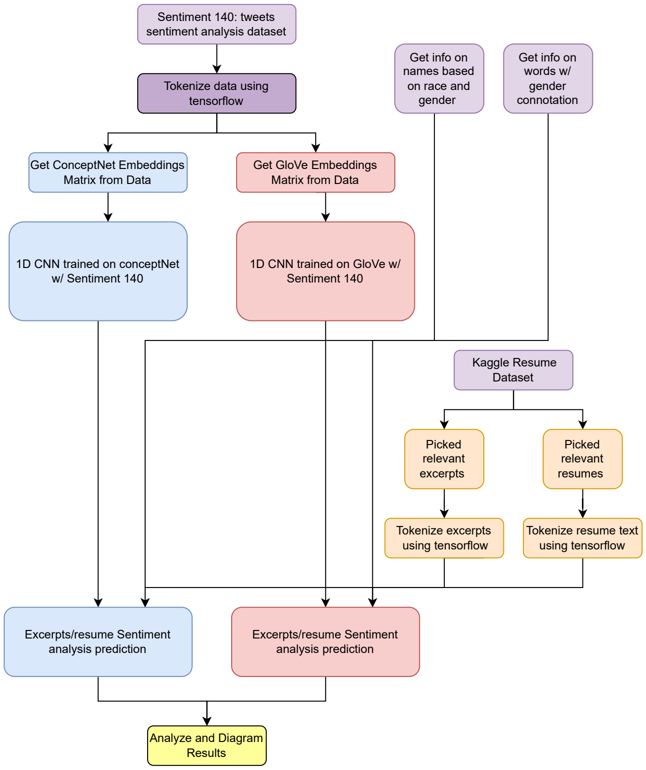
Sentiment 140: tweets sentiment analysis dataset
Tokenize data using tensorflow
Get info on names based on race and gender
Get info on words w/ gender connotation
Get ConceptNet Embeddings Matrix from Data
Get GloVe Embeddings Matrix from Data
1D CNN trained on conceptNet w/ Sentiment 140
1D CNN trained on GloVe w/ Sentiment 140
Kaggle Resume Dataset
Picked relevant excerpts
Picked relevant resumes
Tokenize excerpts using tensorflow
Tokenize resume text using tensorflow
Excerpts/resume Sentiment analysis prediction
Excerpts/resume Sentiment analysis prediction
Analyze and Diagram Results

Fig. 1. Project Flow Chart

We begin by training two sentiment classifiers on the Sentiment140 tweets dataset [14], one with the GloVe embedding and the other with the ConceptNet embedding. These classifiers will both be using a 1-dimensional convolutional neural network (CNN) architecture. These models achieve around a 78.9% validation accuracy where the hypothetical best accuracy for this dataset (please see the limitations section of this paper) is 80%. The exact model architecture is seen in Figure 2. Hyperparameter tuning was done to bring our model as close to 80% as we could achieve. We chose to use CNNs because their limited context will not dilute the embeddings too much and will allow us to see the limitations of the two embeddings when given unfavorable situations. Because we aim to mainly target the limitations of the embeddings, the architectures for both models will be exactly the same. The embedding layers are the only layers that are different between these architectures. To conduct our experiments our models will not be classifying tweets but will instead predict the sentiment of resume excerpts found using a resume dataset [15].

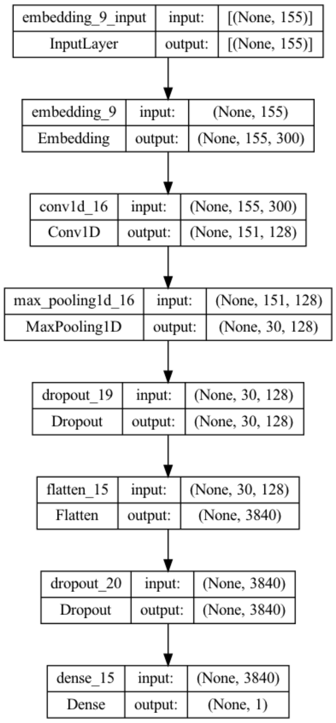
embedding_9_input | input: | [(None, 155)]
InputLayer | output: | [(None, 155)]
embedding_9 | input: | (None, 155)
Embedding | output: | (None, 155, 300)
conv1d_16 | input: | (None, 155, 300)
Conv1D | output: | (None, 151, 128)
max_pooling1d_16 | input: | (None, 151, 128)
MaxPooling1D | output: | (None, 30, 128)
dropout_19 | input: | (None, 30, 128)
Dropout | output: | (None, 30, 128)
flatten_15 | input: | (None, 30, 128)
Flatten | output: | (None, 3840)
dropout_20 | input: | (None, 3840)
Dropout | output: | (None, 3840)
dense_15 | input: | (None, 3840)
Dense | output: | (None, 1)

Fig. 2. Model Architecture

When evaluating if an embedding exhibits bias, our key criteria is whether all groups, be it race, educational institution, or gender have comparable sentiment scores with minimum deviation from one another. In an unbiased embedding, changing a name from ”Mohammed” to ”Marcus” should have a negligible impact on the sentiment score. Ideally, any change in sentiment, whether positive or negative, would be minimal or consistent across all categories. However, if certain groups in our experiments demonstrate a significant increase or decrease relative to other groups, this indicates a level of bias present within that embedding.

We have run 5 experiments:

- Experiment 1: Testing racial bias on smaller excerpts.
- Experiment 2: Testing racial bias on larger excerpts.
- Experiment 3: Testing university bias.
- Experiment 4: Testing university + racial bias.
- Experiment 5: Testing gender bias.

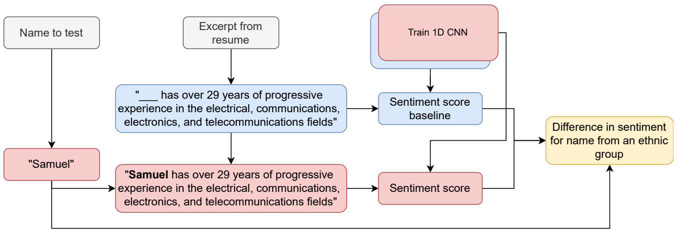
Name to test
Excerpt from resume
Train 1D CNN
"___ has over 29 years of progressive experience in the electrical, communications, electronics, and telecommunications fields"
Sentiment score baseline
Difference in sentiment for name from an ethnic group
"Samuel"
"Samuel has over 29 years of progressive experience in the electrical, communications, electronics, and telecommunications fields"
Sentiment score

Fig. 3. Race Experiment Flow

Experiments 1 and 2 see us adding to the research done in [16] which has grouped 24 or more ”common” names of 4 race categories that we have grouped into ”White”, ”Black”, ”Hispanic”, and ”Arab/Muslim” names. We have gathered 10 smaller excerpts (200 characters or less), and 8 larger excerpts (500 characters or more) found in the resume dataset [15] manually and will be slightly modifying it to fit our needs.

This process can be seen in Figure 3, where we take an unmodified excerpt's sentiment from the two models, then we modify the sentence with a name belonging to the 4 race types, get its new sentiment score from the two models, and subtract the difference of the two to see if adding a specific name belonging to a perceived race into this resume excerpt would influence its sentiment. Our results will be plotted which you will see in Figures 4 5 and 7 in the Results section.

In experiment 3 we conduct our own new experiment which is a slight variation to the race experiment. Instead of adding a person's name of a particular racial background to a resume excerpt we instead add the name of a University. Universities will show up on every resume a person submits and we were curious if we add 8 universities of 4 different types of: ”Public”, ”Ivy League”, ”Private”, and ”International” to see if a person's type of university would have an impact on their sentiment in a resume. This follows a similar pattern to the flow found in Figure 3 but instead of a name, it would be a line of ”Attends _ University” followed by the resume excerpt. These results will be found in Figures 6 and Figure 7 in our Results section.

In experiment 4 we combine our experiments 1, 2, and 3 to have a combined race + educational bias impact observation. In this experiment, we take a neutral prompt missing both the name and university and take its sentiment score. We modify this prompt to go through every name of each racial group to every university name of each university type and take its modified sentiment score. We then calculate the difference between the neutral prompt and the modified prompt sentiment score to see if a combination of race and Education has an impact on the sentiment scores in our two embeddings. These results can be found in Figure 7.

Lastly, our experiment 5 we conduct a gender experiment. This experiment was inspired by the work found in [11]. We plan to take the masking ideology of changing masculine words to feminine words to additional heights. This experiment consisted of us finding three existing resumes in the resumes dataset which consisted of people referring to themselves by their pronouns. We selected three excerpts where the person started every sentence with ”he” and inside the excerpt were instances of him, his, and he. We decided to locate the instances of gender pronouns in these excerpts, remove them all, and then take the sentiment of this sentence. We will then intentionally add masculine, feminine, and gender-neutral pronouns to those instances in the original excerpt, take the new sentiment score, calculate the difference, and see the results. These results can be found in Figure 8.

IV. Results

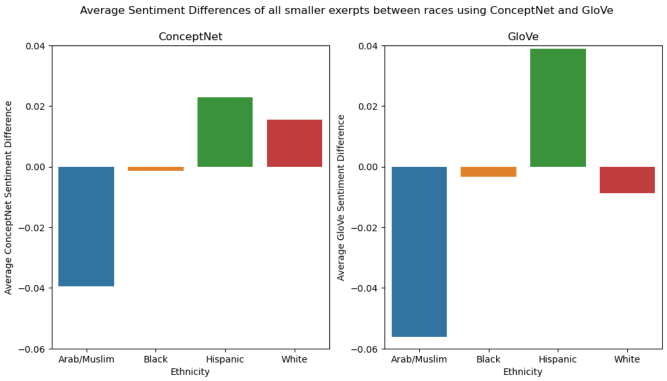

Fig. 4. Experiment 1: Smaller Excerpt Racial Bias

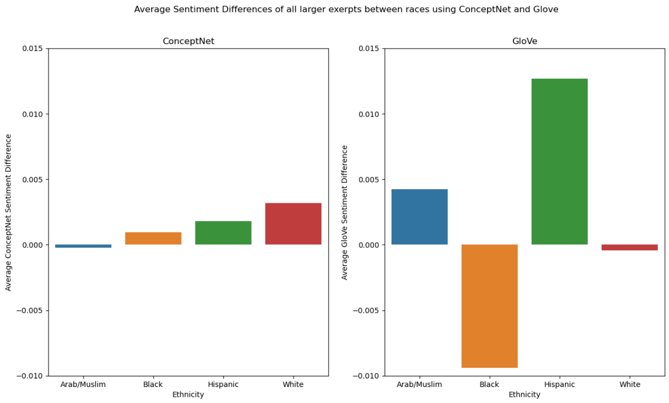

Fig. 5. Experiment 2: Larger Excerpt Racial Bias

Figure 4 compares the sentiment scores for each group of names by race within smaller-sized excerpts of resumes. This is done across both the ConceptNet and GloVe embeddings. Both embedding representations do appear to express bias across the four categories, but it does become apparent that the bias is slightly more pronounced in the GloVe embedding than the ConceptNet embedding. The most negative sentiment racial group in both embeddings, Arab/Muslim, was shown as more negative in GloVe. The highest sentiment group in both, Hispanic, is shown as being more positive in GloVe. This supports our hypothesis of GloVe having more intrinsic bias than ConceptNet by displaying that GloVe had a higher variance across racial groups. Regardless, we were concerned that this experiment would not accurately represent how the data would be seen within a complete resume, and as a result, might have an impact on the sentiment scores recorded. To rectify this, we decided to perform another experiment with the same racial groups by names while using much larger excerpts from the resume dataset.

The results shown in Figure 5 reinforce the previous findings of stronger racial bias exhibited by the GloVe embeddings compared to ConceptNet. The variance across the four categories is more pronounced in GloVe, as evidenced by the larger deviations in bias observed, particularly for the Black and Hispanic racial groups. ConceptNet continues to perform very well as every racial group saw a positive increase in sentiment with comparable magnitudes. Despite these changes in sentiment, the changes in ConceptNet are all less than 0.005.

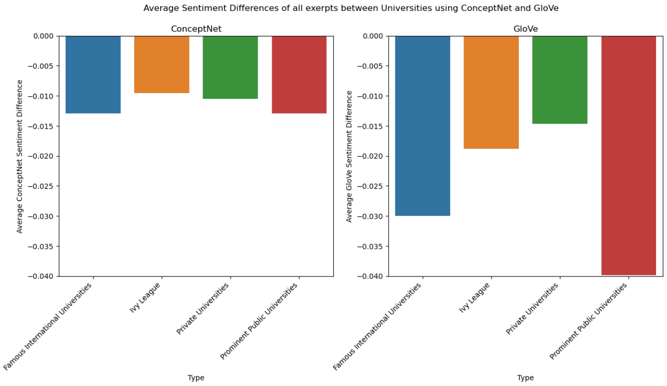

Fig. 6. Experiment 3: University Type Impacts

Figure 6 compares the sentiment scores across each different university groups of universities within excerpts of resumes. The GloVe chart reveals a continued high variance in sentiment scores, further corroborating our hypothesis of GloVe exhibiting more intrinsic bias compared to ConceptNet. An interesting observation is that both embeddings demonstrate this negative sentiment impact score on all categories. We posit that this can be attributed to limitations within the Sentiment140 tweet dataset which we used for training.

The dataset may contain a disproportionate number of negative tweets associated with colleges and universities, which shows every university group negatively impacting sentiment score. Notably in the GloVe embedding, both the Ivy League and Private Universities appeared much more positively than the remaining university groups, where ConceptNet has minimal variance between the groups.

To gain more insight into the interplay between race and educational backgrounds which would be both present in a formal resume, we decided to cross-analyze the university group analysis done in Experiment 3 with the previously examined racial group sentiment impact in Experiments 1 and 2. This combined approach aims to uncover if any compounded effects would have an impact on the sentiment scoring of the two embeddings. Figure 7 compares each group of names by race to each group of universities. For both embeddings, it becomes immediately apparent that the university has little to no impact on the sentiment returned for any of the racial groups present. It seems that Ivy League schools are slightly more positive in ConceptNet and slightly more negative in GloVe. Nonetheless, the differences in bias between the university groups are negligible. For GloVe the combination of race and university has a negative sentiment impact compared to the positive impact shown in ConceptNet. ConceptNet is not free from bias, as we see that the White racial group has the highest increase in sentiment. This cross-analysis continues to show a stark contrast in variance across racial groups and educational groups between the two embeddings, once again reinforcing our hypothesis that GloVe encodes more pronounced intrinsic biases compared to ConceptNet.

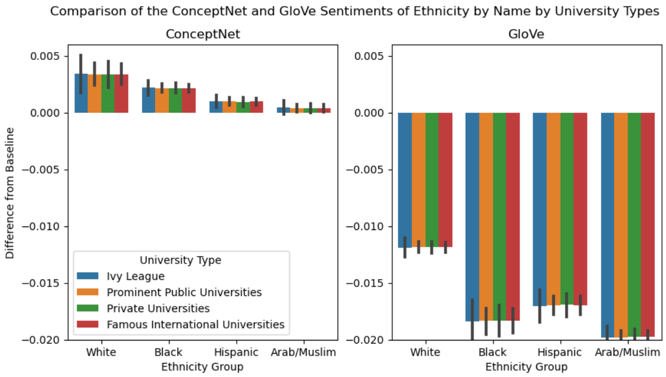

Fig. 7. Experiment 4: Race and University Bias

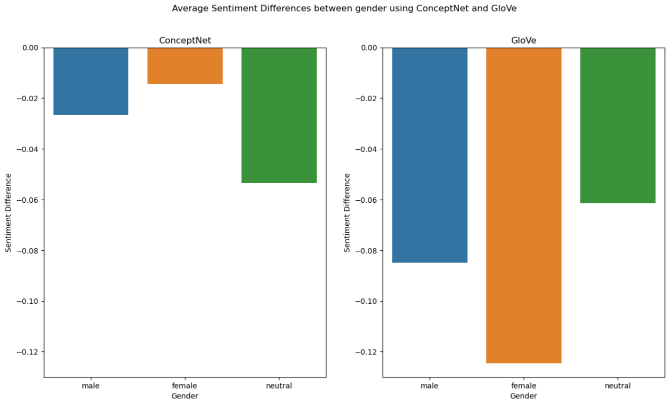

Fig. 8. Experiment 5: Gender Bias

Figure 8 compares the sentiment scores for male, female and neutral gender pronouns within excerpts of resumes. Both embedding representations do appear to express bias across the three categories, but it does become apparent that the bias is much more pronounced in the GloVe embedding than the ConceptNet embedding. This further supports our hypothesis of GloVe having more intrinsic bias than ConceptNet. Regarding our second hypothesis of words referring to women generally having a lower sentiment than words referring to men within resumes, the figure does not seem to support this conclusion.

Although GloVe seems to provide a higher sentiment score for male pronouns than female pronouns, ConceptNet gives slightly more positive sentiment to female pronouns and is overall more balanced between the three groups than GloVe. This implies that there is not a consistent trend of both the embeddings favoring males but does show the continuing trend of high variance in the GloVe embedding as mentioned above.

V. Limitations

Despite the benefits, there are a few limitations to the experiments we conducted in this paper. The primary limitation is that the CNNs we created were trained off of a sentiment analysis dataset that is composed solely of tweets. This does not invalidate any of the experiments we have done, but it does make them less generalizable. The text of Twitter is very different from the text seen in our resume dataset, resulting in our models having a more limited ability to understand the text of the resumes. As mentioned in the results section, given a tweet context, we've seen in our experiment 3 and 4 resulting in a more negative change in sentiment. This leads us to believe that a tweet context in making a sentiment classifier would be more negative to university names than a resume context.

The Sentiment140 dataset itself has a limited accuracy threshold, which is explained by the following breakdown of the dataset's labels:

- 60% of the data points have either a positive or negative sentiment and are correctly labeled. We will correctly classify all of these with a ”perfect” classifier.
- 20% of the data have neutral sentiment, but are labeled as positive. We will correctly ”guess” the classification of about half of these on average, because our classifier can only classify positive and negative tweets.
- 20% of the data have neutral sentiment, but are labeled as negative. We will correctly ”guess” the classification of about half of these on average, because our classifier can only classify positive and negative tweets.

Thus, the creator of the dataset concluded that there is a hypothetical maximum accuracy of 80% that can be achieved [14].

Another limitation of this study is the lack of a pre-trained ATS and labeled resume data. A more interesting version of this study may have been to take an ATS classifier that is used to determine if people get jobs or not and see how potentially biased words change their classifications. We did not perform this study, as we were unable to find a pre-trained ATS to use and there does not appear to be a publicly available resume dataset that contains labels for if they got the job or not. This made this version of the study impossible for us, so we used a sentiment classifier as an approximation for how bias might affect an ATS. We do not claim that sentiment and hireability are equivalent measures. However, names from different races should have the same sentiment, just as they should has the same hireability.

Our test cases are constrained by the limited vocabularies of the tokenizer trained on the tweets dataset and the embeddings themselves. If a name or school used in a test case is not found in either the embedding or the tokenizer, it will be zero-padded. This means it's meaning will not be properly conveyed to the CNN when it makes sentiment predictions. We did our best to remove names from the experiment that did not exist in the embedding space as preprocessing.

VI. Conclusion

The results provided above support our initial hypothesis that ConceptNet would show less bias than GloVe on average. This is not to say that ConceptNet completely eliminates bias, but it has been shown to have a more equivalent change of sentiment given race, gender, and educational institution in comparison to GloVe. However, the results do not fully support our other hypothesis that female-gendered words would result in a more negative sentiment than male words. While this study did not aim to correlate sentiment scores with hiring outcomes, the identified biases raise ethical concerns and highlight the potential pitfalls of relying on older word embeddings.

By shedding light on these biases, we hope this research will contribute to the ongoing efforts toward unbiased natural language processing systems, which can allow more college students, of any race, alma mater, or gender identity from being disadvantaged in the hiring process due to characteristics beyond their control.

Future Work

Future work should focus on developing debiasing strategies for existing word embeddings or developing new embeddings with a better understanding of context, such as BERT, to promote fair and equitable practices in applicant tracking systems.

Acknowledgements

He has asked to keep his last name anonymous but we would like to thank Professor Mohammed *** at Santa Monica College for his testimonial.

References

1 Popomaronis, Tom, “Here's how many Google interviews it takes to hire a Googler,” https://www.cnbc.com/2019/04/17/heres-how-many-google-job-interviews-it-takes-to-hire-a-googler.html, CNBC, 04 2019.

2 Larson, Eric, “CS 8321 Lecture,” 2024.

3 Henderson, Robert, “What Is An ATS? 8 Things You Need To Know About Applicant Tracking Systems,” https://www.jobscan.co/blog/8-things-you-need-to-know-about-applicant-tracking-systems/, Jobscan Blog, 03 2024.

4 Dastin, Jeffrey, “Insight - Amazon scraps secret AI recruiting tool that showed bias against women,” *Reuters*, 10 2018.

5 Papakyriakopoulos, Orestis and Hegelich, Simon and Serrano, Juan Carlos Medina and Marco, Fabienne, “Bias in word embeddings,” *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 01 2020.

6 Bhardwaj, Rishabh and Majumder, Navonil and Poria, Soujanya, “Investigating Gender Bias in BERT,” *Cognitive Computation*, vol. 13, pp. 1008–1018, 05 2021.

7 Basta, Christine and Costa-jussà, Marta R and Casas, Noe, “Evaluating the Underlying Gender Bias in Contextualized Word Embeddings,” *arXiv (Cornell University)*, 04 2019.

8 Kumar, Vaibhav and Bhotia, Tenzin Singhay and Kumar, Vaibhav, “Fair Embedding Engine: A Library for Analyzing and Mitigating Gender Bias in Word Embeddings,” *arXiv (Cornell University)*, 01 2020.

9 Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., Eds., *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*, vol. 29. Curran Associates, Inc., 2016.

10 Wang, Tianlu and Zhao, Jieyu and Yatskar, Mark and Chang, Kai-Wei and Ordonez, Vicente, "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations," 10 2019.

11 Gagandeep and Kaur, Jaskirat and Mathur, S M and Kaur, Sukhpreet and Nayyar, Anand and Singh, Sukhwinder and Mathur, Sandeep, "Evaluating and mitigating gender bias in machine learning based resume filtering," *Multimedia Tools and Applications*, vol. 83, 09 2023.

12 Deshpande, Ketki V. and Pan, Shimei and Foulds, James R., "Mitigating Demographic Bias in AI-based Resume Filtering," *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 07 2020.

13 Bertrand, Marianne and Mullainathan, Sendhil, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," http://www2.econ.iastate.edu/classes/econ321/Orazem/bertrand_emily.pdf, 2003.

14 Go, Alec and Bhayani, Richa and Huang, Lei, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, p. 12, 2009.

15 Bhawal, Snehaan, "Resume Dataset," https://www.kaggle.com/datasets/snehaanbhawal/resume-dataset, www.kaggle.com, 2021.

16 Speer, Robyn, "How to make a racist AI without really trying," https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/, ConceptNet blog, 07 2017.