Word Embedding Gender Bias in Resume Parsing

Hannah Burns

May 7, 2022

Abstract

This project aims to call attention to the matter of gender bias in AI and its problematic implications. The focus will rest upon the natural learning process: how, using mathematical connections to words through word embedding, AI can process not only human communication but the biases that exist within them. This project seeks primarily to analyze the difference between male and female word usage, specifically in resumes, and how AI is able to differentiate between them. It then moves to understand why AI draws conclusions, which may be misinformed, false, or even downright damaging to society if left unquestioned.

1 Introduction and background

Whether the vast majority of the population is aware, artificial intelligence plays a role in everyday and important aspects of connected human life. This is largely in part due to AI's ability to process human speech, language and data. This skill is accessible due to word embeddings and the advancements in natural language processing (NLP) applications. The advancements in NLP applications allows these binary systems to process a system so complex that it can be seen as one of the basis of so described intelligence.

The capabilities of natural language processing has allowed for technology to adapt closer to human life. The usefulness of this technology has created a large market for its development with NLP-based companies predicted to be valued at $26.4 billion by 2024 [HP21]. Corporations such as Apple and Amazon have already integrated this technology into everyday human interaction with personal assistance such as Siri and Alexa. With the growing market demanding new and fast developments that will further affect human life, it is necessary to pause and review the possible implications of incomplete technology.

Word embeddings work to translate human language to vectors of numerical values, something easier for computers to understand. The associated numerical values help AI to identify the word and understand it by referencing similar words that are closely represented in the vector space. Issues with word embedding begin to present themselves in the vagueness of how these associations are learned and the accuracy of these similarities.

Word embeddings and the similarities it associates are built off information it is introduced to and knows nothing of its outside world. This is what makes way for its susceptibility to biases. It is similar to a child's learned biases from their parents, due

to that being their only exposure. The difference is that humans have the capability to think abstractly and form their own moral code. The issues proposed by NLP usage occur when the word embedding model is asked to make assumptions outside of its realm.

Many applications are trained and used solely on information content without needing to make complex social assumptions. Usage of NLP's in a medical setting such as radiology can allow for accelerated and more precise extraction of information [PBHK16]. However, NLP's still process language as information and ignore the fact that people use language to achieve social goals and the behavior behind it [HY21]. This piece becomes crucial as developers look to NLP's as a solution where the layers of language dive deeper into social and historical understandings. Many models are still trained without regard for demographic aspects which becomes ever more concerning as NPL's applications are used to make important decisions about our widely diverse world such as in job hiring and mental health assessments [HP21].

The issues of this are becoming more obvious as AI implementation is becoming more heavily relied on. One large disparity is the way gender biases have appeared in NLP's. In a world and industry that has historically been dominated by males, it is not surprising that AI learned behavior also inherited these deeply rooted biases. Humans themselves deeply struggle to acknowledge their own biases, which is also why they remain so prevalent in technology. AI will always struggle with accuracy, usefulness, and acceptance if it operates on harmful biases. With the particular harm found in gender bias negatively impacting half of the population NLP's look to represent, it is necessary to search for these biases and create better adapted models.

A striking example of the presence of gender bias due to poorly trained word embedding and its negative effects were seen at Amazon in 2014. A word embedding model was trained to be used as a resume parser and rater. The goal, like much of computing, was to hand off the task of parsing through copious amounts of data from resumes Amazon received. The application not only analyzed data, but was given the task to make decisions about the fitness of each candidate. A year into the program an issue was recognized that female resumes were being wrongly rated lower when pursuing positions in engineering and technology. They concluded that this was a product of biased associations due to incomplete training of the model [Kod19]. The word embeddings used were trained what was considered 'good' by reviewing past employee resumes, a vast majority being male. This led to the embeddings association of female terms as 'bad'. The problem created and covered up by Amazon is not only an example of how biases of NLP's that are trained on incomplete data make the wrong assumptions, but also gives attention to the preexisting gender bias facing the tech industry.

This is not and will not be the only example of unintended malice from ignorant AI if these preexisting biases are not addressed. Current word embeddings, such as Google's popular pre-trained model, pull data from long-established new sources. Similarly to the tech industry's issue, this data remains a very codified domain predominantly produced by a small, homogeneous sample: typically white, middle-aged, educated, upper-middle-class men [HP21]. A domain which does not closely reflect the world it hopes to represent.

The careful choice for word embedding models has great importance on the accuracy of NLP's as the choice of data for the experimentation is the first entry point for bias in the NLP pipeline [HP21]. Corpuses have specifically been trained to combat this issue

by using embeddings trained without the use of gender specific pronouns yet studies on those models have still shown gender bias[WRAB18][CM19].

Many scholars and academic research has gone into addressing the existence of bias in NLP, a problem which is inevitable in creating technology built off of human relations [BBDIW20]. No research yet as concluded a way to end the existence of these biases just as how no one has been able to eradicate bias from the human mind. It is worth continued exploration to analyze the way in which these biases can be identified and what are the factors that are harmfully impacting the integrity of the technology. These issues have provoked the questioning of ethical AI when the root of the issue is in questioning the ethicality of the systems implementing it.

2 Approach to solve the problem

The overview of the approach is to compare the language used in a set of male and female resumes with a preexisting domain of word embeddings. The tests are to be centered around how a person describes and presents a gendered entity, themselves. The programming will uncover the gender bias that the word embedding has by a measure of association with specific words and gendered terms. It will then look to compare these gender associations on the words used in the different resumes. The process looks to approach the questions of the gendered difference in language usage especially when self describing and if it is noticeable by AI word embeddings.

The data set consists of resumes from male and female identifying undergraduates, an aggregation that has all had exposure to higher education and are entering the skilled workforce. The resumes span diverse categories such as majors of study, ethnicity, and universities attended. It should be noted though that the data set collected and used in these experiments was limited as only 21 resumes were analyzed, 12 from female identifying and 9 from male identifying students. This should be considered incomplete and the results should be used to provoke deeper exploration into the topic.

To approach this problem, the preexisting open source code provided from the GitHub repository of a 2019 study that measured gender bias across multiple domains of word embeddings was used. The original research conducted "gender association tests on a per cluster basis using means difference method (as per Caliskan et al 2017)" in their gender_associations.py program and repeated it on multiple existing word embeddings [CM19].

This specific program takes input of the word embeddings, the cluster file, originally produced in their program cluster_embeddings.py, and a list of male and female attributes. The program produces a file that gives every word in every cluster an associated gender and a gender score. The gender score is found by using the mean difference method as introduced above.

It begins by taking a given word and finding the cosine similarities between the word and all the given gender attributes. The female attributes are classical words such as woman and mother. Similarly, the male attributes are made up of words such as man and father. A complete list of the terms can be found in Table 1. The female mean is subtracted from the male mean to produce the score. If the score is negative, the similarities are greater for the female attributes and a gender association of 'F' is given to the word. If the score is positive, it is a male associated word and assigned to

'M'. This process is repeated for every word in every cluster. The code for parsing the clusters as follows:

```
for cluster in sorted(cluster2word.keys()): #parses through each cluster
          cluster_words = cluster2word[cluster] #list of all words in a
              single cluster
          for word in cluster_words: #parses through each word in a single
              cluster
              score = cosine_means_difference(wv, word, male_attrs,
                  female_attrs) #sets the words gender associated score
              gender = 'F' if score < 0 else 'M' #sets the gender
                  association
```

Where wv is the file containing the word embedding vectors, cluster2word is a list of clusters loaded in by the cluster file and male_attrs and female_attrs consisting of the list of gendered terms. The association test is as follows:

```
def cosine_means_difference(wv, word, male_attrs, female_attrs):
    male_mean = cosine_mean(wv, word, male_attrs) #mean of words
        similarities with male terms
    female_mean = cosine_mean(wv, word, female_attrs) #mean of words
        similarities with female terms
    return male_mean - female_mean #gendered score, M if > 0, F if < 0.

def cosine_mean(wv, word, attrs):
    return wv.cosine_similarities(wv[word], [wv[w] for w in attrs]).mean()
        #iterates through attribute list of a specific gender, returns the
        mean
```

The original research continues through several more tests and functions to prove their null hypothesis wrong but their programming component above was the most useful for this particular research's results. In the interest of comparing word usage and their gender associations, this research will conduct gender association tests on the words used in the collected female and male resumes. Supplementary code has been added to extract additional data more pertinent to the interest of this research such as giving each resume an overall gender score and commonly used words by each gender.

female_terms	male_terms
female, woman, girl, sister, she, her, hers, daughter, mother, aunt, grandmother	male man, boy, brother, he, him, his, son, father, uncle, grandfather

Table 1: Female and Male terms

3 The design of the experiments and results

Due to the process relying more on how the candidate is describing, inflating and presenting themselves, the resumes will first be parsed. Irrelevant data is removed from each resume to better focus on target words and ensure all resumes have respected anonymity of their data. As to not skew the data, useless items such as months, repeated words, places, headings, and stop words, provided by the NLTK library, were

removed along with names, places, and contact information. The dissected resumes are then treated as clusters to group words on a per person basis. Instead of using the clustering program from the provided code, a new program is written to create two original cluster files. One for each gender represented in which each cluster represents a person with their filtered words and a corresponding number or key. This adaptation will allow the program to centralize gender association findings on particular resumes instead of on every word in the word embedding.

This program will also only measure word associations on a single domain. The domain chosen for use is the "FastText embeddings trained on the GAP corpus [WRAB18] where GAP is a "balanced corpus of Gendered Ambiguous Pronouns" created originally to minimize gender bias. The GAP embedding was chosen due to its smaller size and training.

These cluster files will be used as input for the gender_associations.py file where every word will be tested for its association and strength of association to a particular gender. The original output file of the program will produce a tab separated file that will include every word used, the cluster it was derived from, the associated gender of the word and its association score. Further analysis of this data can help lead to conclusions about the differentiation of personal language usage between genders and individual outliers if needed.

Information will be pulled from this collection to trace what words had what gender association and what words were commonly used, at least $\frac{1}{3}$, by a certain gender. A sample of common words is provided in Figure 3. The total commonly used female words was 45 while males had 70 commonly used words on a per gender basis. The results of the gender association on a per word basis will give a physical and mathematical representation on how artificial intelligence processes and perceives natural language and how gender bias can exist within it. Broader results of common language usage on a gender basis can be used to examine differences in male and female word choice. In this instance, the way they chose to describe themselves and their skills and what is generally thought to be important information.

An overall score of all the words in a cluster will also be tracked to produce an average gender association score for each resume. All of these results can be found in Figure 1. The overall scores will be used to determine results for comparison on individuals, genders, and the populace. Individual overall scores will be able to show the outlying resumes in the data set. Their data will be further analyzed to look for causation. In the instance of outliers, the individual's major, age, and experience will be reintroduced as well as parsing through their cluster for outlying words. Outliers will be determined on their relationship to the median of their gender and the entire populace.

The overall median and mean scores for a gender will be used in comparison to the calculations of the entire population. These results will be used to test the null hypothesis that there is no differentiation in female and male resumes on the basis of AI gender bias in word embedding. The results of the medians from each gender are represented in Figure 2.

The results to be produced from these tests are not to conclude that different words, skills, or representations are better or worse. The purpose of these examinations is to experiment with the problem of bias in artificial intelligence. It looks to show if incomplete and ignorant training of AI can cause gender bias due to the fact that the

Clusters	Female	Male
1	0.04178796582	0.03519631947
2	0.03806733196	0.04463483142
3	0.05038329666	0.03991821062
4	0.04464844219	0.04494487271
5	0.03551607355	0.04330058326
6	0.02602025896	0.04021623411
7	0.03954565496	0.04295071089
8	0.04584256157	0.04828022096
9	0.04250536747	0.04452742071
10	0.036669101	
11	0.04522627828	
12	0.04471976573	

Figure 1: All overall gender association scores

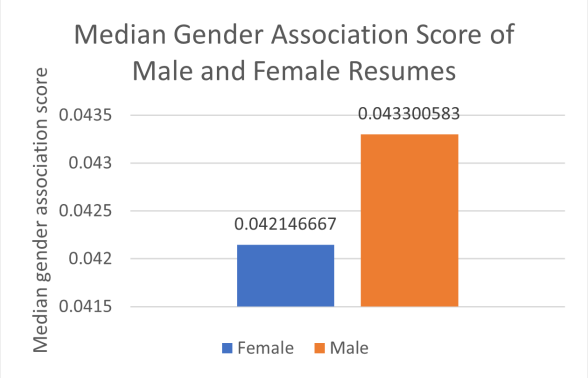

Figure 2: The median gender association score of resumes per gender.

programs and data to train it are overwhelmingly written by one gender.

4 Analysis of the results

The results of the tests provide a statistical representation of gender associations in the word embedding. The large set of data produced also allows for a deeper analysis into causation and the processing of natural language by artificial intelligence.

On the basis of the overall gender associations, each resume had an overall score that was predominantly male as they were all greater than zero. This is not to conclude that there was no differentiation between male and female resumes. Due to the nature of resumes describing professional pursuits which have been historically male. The resumes are better compared on a median basis to account for this and the presence of outliers. That being said, the data suggests the same when analyzing the means of the gender association scores. An overall median score was calculated based on all resumes to conclude that the female resume's median had a lower score, closer to zero and feminine values, than the overall median. While the males scored higher than the

Female words	Male words
assistant	analysis
assisted	assistant
children	building
community	community
computer	company
conducted	computer
created	create
design	created
development	current
environment	design
expected	designed
help	developed
helped	development
maintained	executive
managed	expected
organization	food
organize	future
organized	help
professional	information
program	issues

Figure 3: Sample of commonly used terms by gender. Blue are uniquely male, red are uniquely female words.

overall average. The conclusion to be drawn from this on a broad level is that the bias in the word embedding has the capability to recognize female written work.

When analyzing the difference in language used on a per gender basis, there is quite a difference. The female resumes produced a larger number of unique words used than the male resumes after equal word filtering. Out of the total female words, only 8.26% were commonly used. In comparison, 13.70% of words used by males were commonly used by at least $\frac{1}{3}$ of all other male resumes. The causation may range from a wider diversity of majors, experiences and skills from females. It is still to be concluded that the female resumes produced significantly less redundancy in vocabulary. Out of the commonly used words, many were the same across gender lines such as: development, expected, project, and support. This data is not surprising due to the nature of the writings.

The interesting results in comparing the commonly used words allude to the difference in how genders describe themselves. Many of the commonly used words were specific to a gender. The complete composition of words is quite large to include. For specifics, uniquely female common words consisted of words like children, maintain, organize, questions, and safety. On the contrary, common male words used were executive, lead/leader, new, provide, and quality. These terms are all also highly male associated in individual word embedding results. Although there was no usage of a program to analyze these words, there is an obvious difference in the types of commonly used words and how genders may tend to describe themselves. Further analysis on the differences in commonly used words in societal respects will be left unsaid.

To summarize the analysis of these results, there is a difference in the language usage of male and female resumes. This has been shown through the variation of vocabulary usage and the quantity of different words commonly used only by a specific gender. The conclusions of these results also help to explain the difference in gender association scores, specifically why male resumes had a smaller range.

5 Conclusions and future work

Although the results of this research help support the ongoing issue of learned gender bias in AI and it being a product of biased programming and training, it is to be noted again that the results are not conclusive. Due to the small sample size of data collection, the proposed conclusions are to inspire future research and communicate the implications of the assumptions of ignorant artificial intelligence. The results also show how word embeddings can be utilized as a tool to better understand the presence of biases, harmful or not, in data sets and in society.

Further research can and should extend itself in numerous beneficial ways to help maintain the ethicality, usefulness and societal approval of future technology. The obvious extension of this research would be to widen the scope of data collection not just for more accurate results but for more diverse representation. A data set of resumes that included graduates could tone the scope of experiences and skills. The undergraduate resumes are interesting due to the fact that none have been able to work their way into and up the professional workforce. It would be worth measuring if resumes that have been able to, might still show a differentiation in how the genders tend to describe themselves.

The entirety of the research was about gender and gender bias yet only male and female gender identities were taken into account. As this research also challenges the issues of using ignorant data that does not reflect our world, it is definitely worth mentioning the importance of further research to a more inclusive set of inclusive gender identities. For this research, data was collected from non-binary candidates as well but not used to draw conclusions. The choice to limit the represented gender identities was due to the dissimilar representation in comparison to the other male and female categories. Research, training and data sets need to properly account for the diversity of the groups they intend to represent, such as non-binary students and job applicants. Due to the limitations reached in this project, further exploration with respect to a greater diversity of gender identities is left to future research that can provide a larger and more complete data set.

Continuing on the premise of inclusive AI and research testing, further exploration into languages other than English should be conducted. Natural language processing has the capabilities to adjust uniquely in different human languages. The differentiation of gender bias particularly in Spanish language word embeddings would be useful and interesting. Although Spanish is a widely spoken language on a university, national and global level, it is not as represented in fields such as academia and tech. A deeper look could provide different results on the possible basis of it being a gendered language and having different cultural impacts on gender representation.

The objective of this research paper together with the provided ideas for further expansion are to inspire continued improvement and questioning of AI and the world it speaks for.

References

[BBDIW20] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *ACL*,

June 2020.

[CM19] Kaytlin Chaloner and Alfredo Maldonado. Measuring Gender Bias in Word Embeddings across Domains and Discovering New Gender Bias Word Categories. In *Proceedings of the 1st ACL Workshop on Gender Bias for Natural Language Processing*, Florence, 2019.

[HP21] Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

[HY21] Dirk Hovy and Diyi Yang. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online, June 2021. Association for Computational Linguistics.

[Kod19] Akhil Alfons Kodiyan. An overview of ethical issues in using ai systems in hiring with a case study of amazon's ai based hiring tool. 11 2019.

[PBHK16] Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. Natural language processing in radiology: A systematic review. *Radiology*, 279(2):329–343, 2016. PMID: 27089187.

[WRAB18] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617, 12 2018.