Imbalanced gendered wording in data scientists' JD: a computational research by LDA and t-SNE

Abstract

Gender inequality in workplaces has always been the elephant in the room, and a popular social science topic. There are lots of academic accomplishments explaining the economic, historic, cultural, and hierarchical reasons of this phenomenon, one of which is imbalanced gendered wording used in job advertising blocked women out of labor market. Based on this idea, we analyzed over 10000 data scientists' job descriptions based in the U.S from online job-hunting platforms, and applied computational methods, mainly LDA and t-SNE to visualize the proportion gap between the usage of masculine coded words versus feminine ones.

Agenda

Work separation:

- **Toni Law (SID_20748828)**
  Topic design, Dataset searching, Data cleaning, Data classification, Paper writing (Part 1, Reference).

- **Yifan Cao (SID_20746894)**
  Schematic design, Algorithm implementation, Visualization, Paper writing (Part 2, 3, 4, 5, Reference).

- **Together**
  Paper revision

1. Background

Gender inequality in employment has been a recurring issue ever since women entered the labor force. Gender wage inequality is one of the most addressed social issues in the United States. There are two main reasons that contribute to the national gender wage gap. The first reason is known as the “motherhood penalty” (Correll et al., 2007). Women without children earns twice as much as women with children (Kleven et al., 2019), which partially contributes to the current gender wage inequality gap of a woman earning 81 cents to every 1-dollar a man earns (Payscale, 2020). The second reason is the percentage of women working in high paying and

prestigious sectors such as IT is much lower than men (Sassler et al., 2016). Our study focuses on the gender disproportion among the IT field. Gender disproportion in IT companies has raised so much awareness that companies such as Apple have pledged to hire more women to encourage diversity (Titcomb, 2018). Representation and diversity are beneficial to the culture and atmosphere of a company. Women representation matters because it empowers and encourages other women to join. Female students who major in STEM studies are more persistent and less likely to switch to other majors when there is a higher percentage of female representation among STEM field graduate students (Griffith, 2010). Diversity allows different perspectives to be seen, enhances team performance, and facilitates innovation (Ellemers and Rink, 2016), which is especially valuable to IT companies.

According to a report conducted by McKinsey & Company (2020), over 70 percent of senior leaders consider gender diversity as a priority in their company (Coury et al., 2020). Women have outnumbered men in obtaining Bachelor Degree and Doctoral Degree (Duffin, 2020) and women's labor participation rate have increased over the past 30 years (World bank data, *Labor participation rate,* 2020). However, women continue to be underrepresented in the IT field.

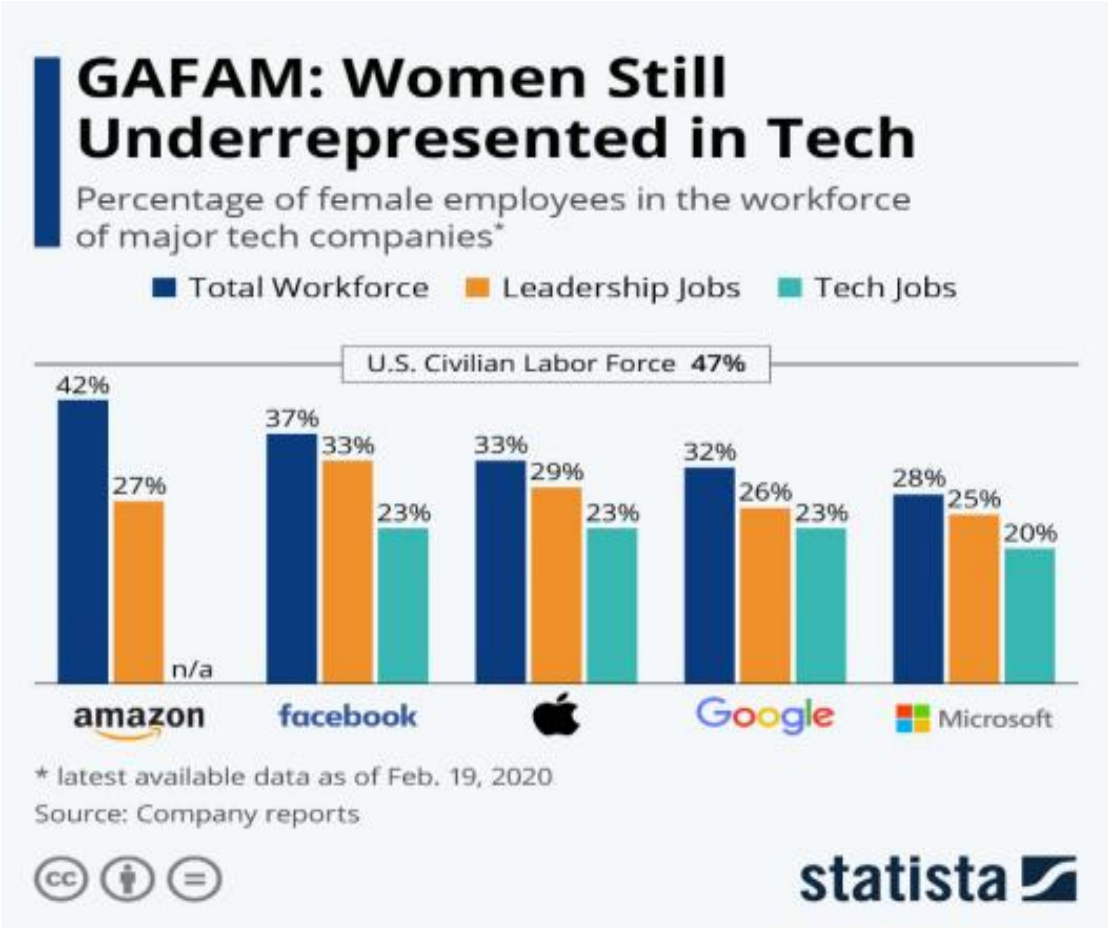

Fig 1. Public data collected from Statista

Although companies, schools, and government have encouraged women to join the IT fields, the percentage of women participates in IT-related jobs have declined since the 2000s (Sassler et al., 2016). Although there is a menial pay gap between women and men that obtained the same entry-level position (Kleven et al., 2019), less women with STEM majors end up entering the STEM field after they graduate than men with STEM majors (Sassler et al., 2016). Women with STEM majors are more likely to shift career paths and enter education or healthcare sector (Beede et al., 2011), which reinforces the social stereotype that women are better caregivers so they should work

in particular fields. With less women major in STEM studies and even less women entering the STEM field when they graduate, gender disproportion arises even in entry-level IT positions. The gender proportion becomes even more imbalanced in senior and managerial IT positions (Coury et al., 2020).

The process of recruitment in the digital era has increased in efficiency and reduced cost and time. However, the bias in recruitment have not reduced. The response rate for a résumé with a man's name on top is significantly higher than an identical résumé with a woman's name on top (Watts, 2014 & Rivera and Tilcsik, 2016). Men were also more likely to be offered managerial positions and were offered a higher salary in those positions (Watts, 2014). Women, especially women with children, still face negative cultural stereotypes while applying for job (Blau and Kahn, 2017 & Kleven et al., 2019). Besides gender bias in the recruiting process, gender bias is also embedded in job advertisements. One of the contributing factors to gender disproportion in male-dominated fields is that only small fraction of the applications these companies received are from women. Scholars have found that job advertisements for male-dominated positions often contain gender biased language or words (Gaucher, Friesen, & Kay, 2011). The Gaucher et al. used traditional social science methods such as posting job posts that contains different percentage of feminine coded and masculine coded words on campus and examining the response rate. Their study found that gender coded words did not significantly affect men but job advertisement that includes more feminine-coded words are more appealing to women and increase their sense of belongingness to the job. Job posts with job

descriptions that contains gender biased words may further discourage women to apply even if they have the capacity to perform the job duties of the position (Collier and Zhang, 2016). A vicious cycle is formed when gender bias in job advertisement deters women, less women applying for IT-related jobs leads to underrepresentation in the IT field, and underrepresentation leads to less women willing to major in IT in education.

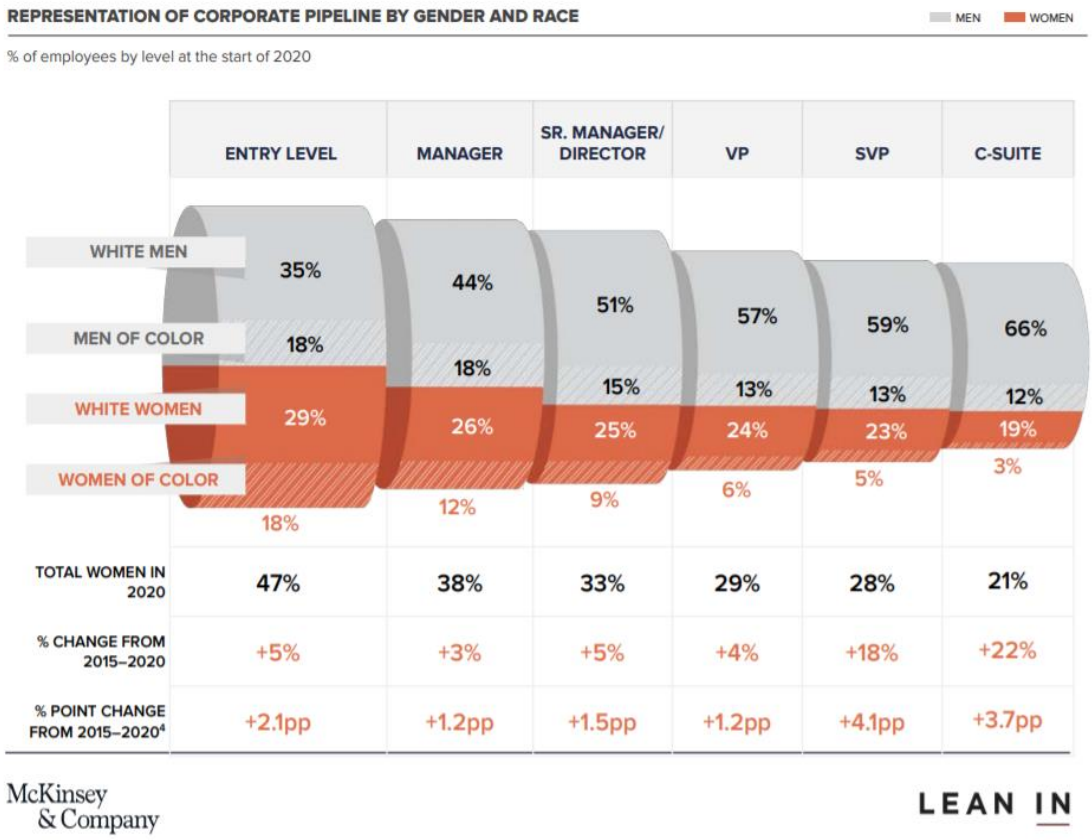

Fig 2. Public data from Report *Women in the Workplace 2020* by McKinsey & Company

Objective

There are three objectives in our study. The first objective is to examine whether gender biased job posts is a contributing factor to gender proportion in the IT field.

The second objective is to build a model that can detect gender bias within the job description of recruiting advertisements in IT related positions. The third objective is to apply our model and illustrate the gap of gendered wording usage in the job descriptions between junior data scientists' positions and senior data scientists' positions.

2. Schematic program and Dataset Status

According to research results from Gaucher et al., gendered wording (i.e., masculine- and feminine-themed words) applied in job advertisements will reinforce gender inequality in workplaces (Gaucher, Friesen, & Kay, 2011). This kind of mechanism fits in a traditional social psychological theory, SDT, which asserted that "human societies tend to organize as group-based social hierarchies" (Pratto, Sidanius, & Levin, 2010). Particularly, in employment conditions, unacknowledged, subtle gendered language use in job advertisements will block women out of areas that men typically occupy or dominate, by making female candidates feel less affiliated or inferior (Gaucher, Friesen, & Kay, 2011).

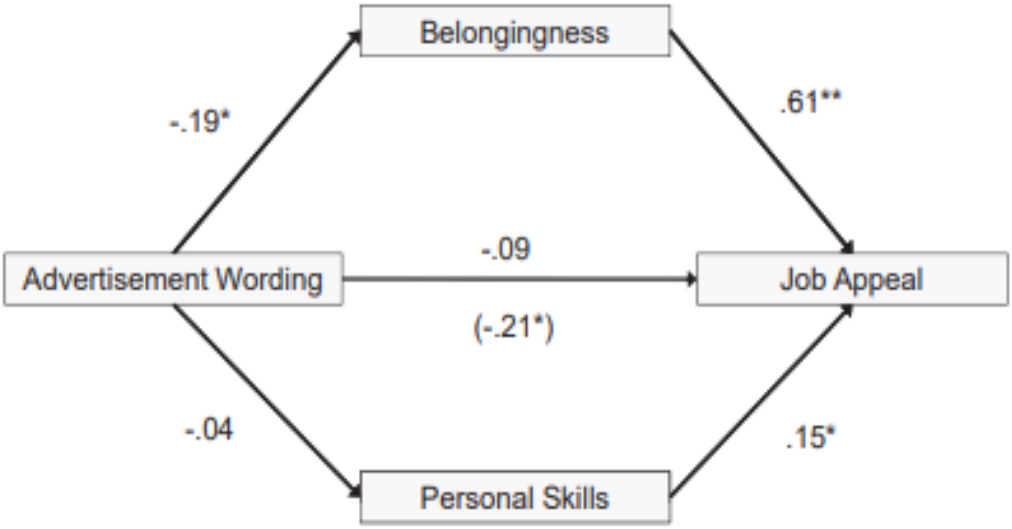

Fig 3. Standardized path estimates from Gaucher et al. (Advertisement wording coded as 1=masculine wording and 0=feminine wording. *p <.05. **p <.001.)

Gaucher et al. drew this conclusion in 2011 from a perspective of traditional social science methodologies, based on relatively small sample data. Therefore, we try to adopt data analysis methodologies, including supervised and unsupervised machine learning algorithms, from computational social science to validate if imbalanced gendered wording usage is still noticeable in workplaces, specifically in a newly emerged occupation——data scientist. Then we will apply dimension reduction strategies as well as data visualization algorithms to illustrate the gap between gendered wording proportion within different levels of data scientist recruitment advertisements posted in recent years.

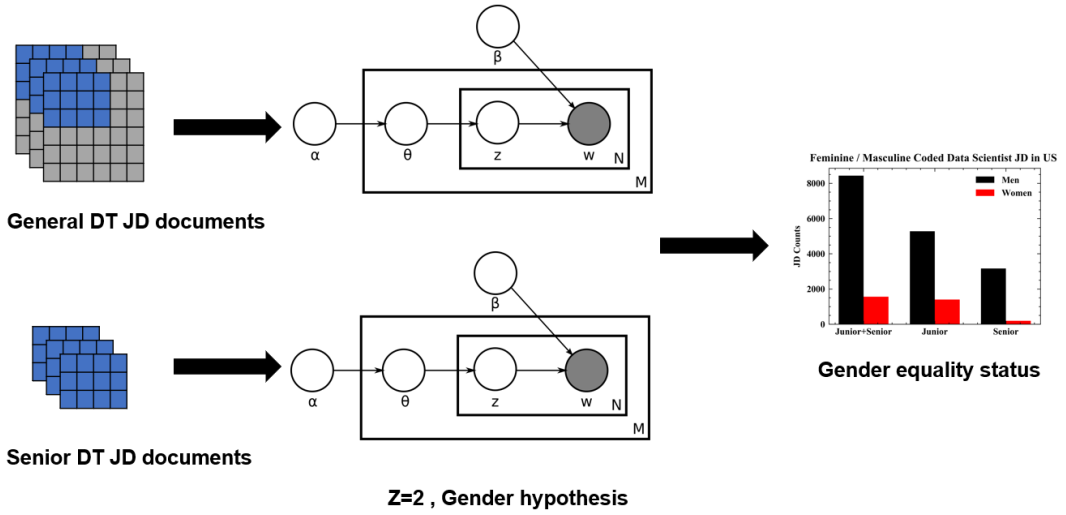

Fig 4. Schematic program

Fig 4 shows our schematic program of this gender equality research of data scientist job description (JD). The first step is to preprocess the document with general nonstructural text preprocessing such as filtering stop words, removing missing value and tokenization. Regarding the objective of observing gender equality, we labeled the senior and entry level data scientist JD with 0 and 1 according to the original description of job level which will be talked about latter. Then, we trained two Latent Dirichlet allocation (LDA) topic models for observing masculine and feminine words frequency with the hypothesis of gender status of men and women. Finally, we got the masculine and feminine words frequency proportion in general and senior data scientist JD with the help of unsupervised generative methods like LDA and visualized by t-distributed stochastic neighbor embedding (T-SNE). The implementation details and results will be discussed in section 3 and section 4.

The data set we adapted for this project is constructed by 2 subsets. The first one contains 10000 online job descriptions collected from global online employment platforms, including Indeed, Monster, Dice, and CareerBuilder. This one is open source which we got from Github.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	crawl_timest	url	job_title	category	company_name	city	state	country	inferred_city	inferred_state	inferred_country	post_date	job_description	job_type	salary_offered	job_board	geo
2	2019-02	https	Enterprise	Accountin	Farmers Insurance	Woodland Hills	CA	Usa	Woodland hills	California	Usa	2019/2/6	Read what people are	Undefined		indeed	usa
3	2019-02	https	Data Scientist		Luxoft USA Inc	Middletown	NJ	Usa	Middletown	New jersey	Usa	2019/2/5	We have an imme	Undefined		dice	usa
4	2019-02	https	Data Scientist		Cincinnati Bell T	New York	NY	Usa	New york	New york	Usa	2019/2/5	Candidates shou	Full Time		dice	usa
5	2019-02	https	Data Scien	Accountin	BlackRock	New York	NY 100	Usa	New york	New york	Usa	2019/2/6	Read what	Undefined		indeed	usa
6	2019-02	https	Senior Dat	biotech	CyberCoders	Charlotte	NC	Usa	Charlotte	North carolina	Usa	2019/2/5	We are seeking	Full Time		monster	usa
7	2019-02	https	CIB - Fix	Accountin	JP Morgan Chase	New York	NY 101	Usa	New york	New york	Usa	2019/2/5	Read what people are	Undefined		indeed	usa

Fig 5. raw data examples of the first data set

According to direct observation of raw data, we found out that although belonging to various kinds of industries, all these data scientists' occupations located in the U.S and the job_title could be further divided into 2 subgenres, one is entry level, another is senior level. Based on this primary observation, we narrowed our research to focus on the imbalanced gendered wording usage during data scientists' employment and promotion in the U.S within recent 2 years.

Although we could compare the percentage of masculine words with feminine ones from this data set, the salary data is missing which may obstruct further observation about consequences of gender inequality in workplaces, such as the income gap between men and women.

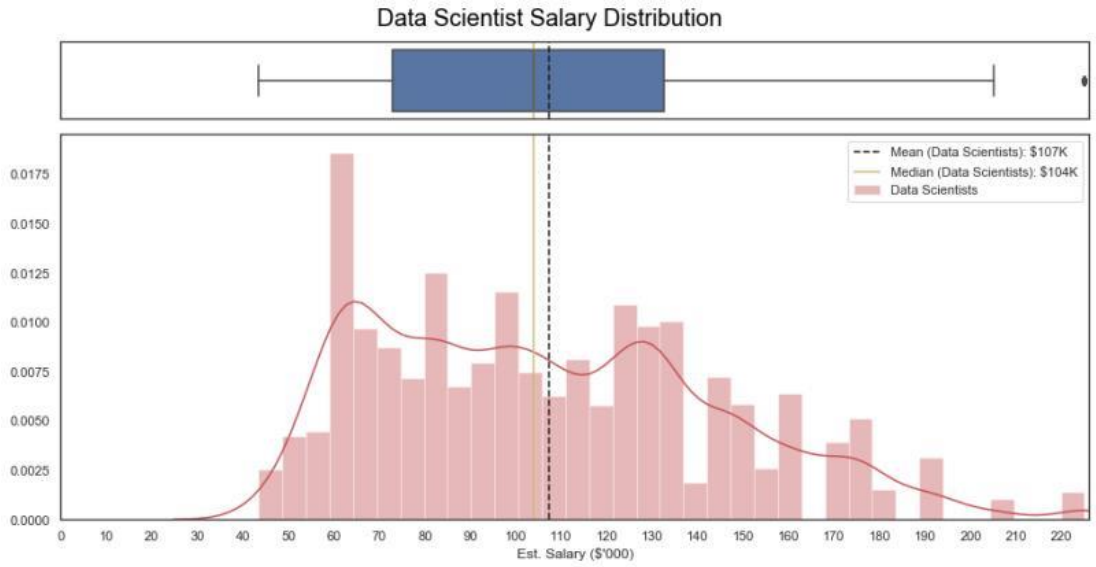

Fig 6. distribution of data scientists' salary

Therefore, we constructed another supplementary data set, which contains 3000 online job descriptions collected from online data scientist recruitment job advertisements by ourselves.

These 2 data sets share similar core information as well as data structures. Both highlight post_date, URL, job_title, industry_category, company_name, city, state, country, detailed job description. Specifically, the second one has concrete salary data. Based on the observations introduced above, we cleaned these 2 data sets to prepare for further exploration.

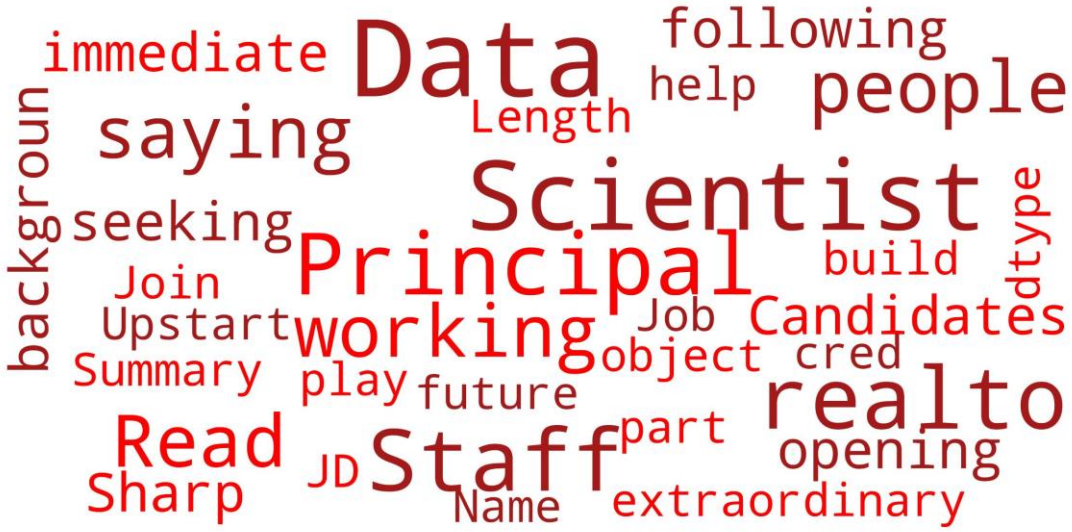

Fig 7. word frequency based on all texts of the data set

Besides text data of job description, we also constructed a list of masculine and feminine words which was adapted from the research outcomes of Gaucher et al. According to social linguists and social psychologists, different genders possess different language usage preferences. Women tend to be more communal and interpersonally, while men exhibit more sense of leadership and agency (Eagly & Karau, 1991; Heilman, 1983; Rudman & Kilianski, 2000).

List of Masculine and Feminine Words Coded in Studies 1 and 2

Masculine words	Feminine words
Active	Affectionate
Adventurous	Child*
Aggress*	Cheer*
Ambitio*	Commit*
Analy*	Communal
Assert*	Compassion*
Athlet*	Connect*
Autonom*	Considerate
Boast*	Cooperat*
Challeng*	Depend*
Compet*	Emotiona*
Confident	Empath*
Courag*	Feminine
Decide	Flatterable
Decisive	Gentle
Decision*	Honest
Determin*	Interpersonal
Dominant	Interdependen*
Domina*	Interpersona*
Force*	Kind
Greedy	Kinship
Headstrong	Loyal*
Hierarch*	Modesty
Hostil*	Nag
Implusive	Nurtur*
Independen*	Pleasant*
Individual*	Polite
Intellect*	Quiet*
Lead*	Respon*
Logic	Sensitiv*
Masculine	Submissive
Objective	Support*
Opinion	Sympath*
Outspoken	Tender*
Persist	Together*
Principle*	Trust*
Reckless	Understand*
Stubborn	Warm*
Superior	Whin*
Self-confiden*	Yield*
Self-sufficien*	
Self-relian*	

Note. The asterisk denotes the acceptance of all letters, hyphens, or numbers following its appearance.

Table 1. gender-biased wording list by Gaucher et al.

However, this list was constructed 9 years ago, and we believe that modification is necessary to make this list more applicable in present conditions. Therefore, we asked 10 postgraduate students who major in linguistic or gender studies to help us adjust the word list. We made a questionnaire containing 10 controversial words/roots in this list, which are Confident, Courag*, Independen*, Intellect*, Self-confiden*,

Persist, Honest, Loyal*, Modesty and Trust*. We applied Likert Scale to measure from strong masculine, masculine, neutral, feminine to strong feminine, which was correspondingly assigned as 5-1. Besides, we asked these students to pick out words which are obviously gender-biased to them from job descriptions within our data sets. Then we respectively calculated the mean values of each word, the higher the value the more masculine the word is. Combining all these results, we reconstructed the list of masculine and feminine words.

Feminine_coded_words		Masculine_coded_words	
"agree",	"pleasant",	"active",	"impulsive",
"affectionate",	"polite",	"adventurous",	"independen",
"child",	"quiet",	"aggress",	"individual",
"cheer",	"respon",	"ambitio",	"intellect",
"collab",	"sensitiv",	"analy",	"lead",
"commit",	"submissive",	"assert",	"logic",
"communal",	"support",	"athlet",	"objective",
"compassion",	"sympath",	"autonom",	"opinion",
"connect",	"tender",	"battle",	"outspoken",
"considerate",	"together",	"boast",	"persist",
"cooperat",	"trust",	"challeng",	"principle",
"co-operat",	"understand",	"champion",	"reckless",
"depend",	"warm",	"compet",	"self-confiden",
"emotiona",	"whin",	"confident",	"self-relian",
"empath",	"enthusias",	"courag",	"self-sufficien",
"feel",	"inclusive",	"decid",	"selfconfiden",
"flatterable",	"yield",	"decision",	"selfrelian",
"gentle",	"share",	"decisive",	"selfsufficien",
"honest",	"sharin"	"defend",	"stub
"interpersonal",		"determin",	
"interdependen",		"domina",	
"interpersona",		"dominant",	
"inter-personal",		"driven",	
"inter-dependen",		"fearless",	
"inter-persona",		"fight",	
"kind",		"force",	
"kinship",		"greedy",	
"loyal",		"head-strong",	
"modesty",		"headstrong",	
"nag",		"hierarch",	
"nurtur",		"hostil",	

Table 2. gender-biased wording list by our field investigation.

3. Computational methodology

Just as mentioned in the background before, gender inequality exists in all levels of career paths from various walks of life and deteriorates in higher-level positions. Specifically, women remained dramatically underrepresented even in developed countries, take America for example, between 2015 and 2020, the share of women grew from 23 to 28 percent in SVP roles—and from 17 to 21 percent in the C-suite, according to the latest research accomplished by McKinsey & Company. Needless to mention conditions in less developed regions or countries.

There are many reasons to explain gender inequality in workplaces, one of which is the male-oriented language used in job descriptions. Considering gender stereotypes reflected by the language used, words can have a subtle but significant influence on the employment process. On the one hand, candidates whose self-presentation contained more communal traits were less likely to be hired especially when applying for male-dominated jobs (Gaucher, Friesen, & Kay, 2011). On the other hand, female candidates will probably give up on fighting for vacancies when they perceived a strong masculine language style in job descriptions (Gaucher, Friesen, & Kay, 2011).

Thus, we decided to conduct this project in 2 steps. Firstly, we will analyze the general conditions of gendered wording usage within the whole data set; Secondly, we will try to verify whether differences exist between entry level and senior level.

3.1 Generate topics within job descriptions by LDA topic modeling

3.1.1 Why topic modeling?

Since the original dataset is not including the target candidate gender labels, we cannot make a supervised learning method to achieve our objective. At the same time, the general unsupervised statistical learning methods like principal components analysis (PCA), K-means clustering can't provide the text data related insights rather the numerical dimension reduction results. The reasons for implementing LDA topic modeling as shown below:

Topic modeling offers approaches for storing, comprehending, scanning, and summarizing large electronic collections automatically.

With the following, it will assist:

1. Discovering the themes underground.
2. Classifying the records into the patterns found.
3. The grouping is used to organize/summarize/search the data.

Therefore, we can refine our search process by annotating the text, based on the topics expected by the modeling approach, and notice that gender inequality occurs in the data scientist's JD. The implementation of the LDA in our assignment started with those assumptions:

3.1.2 Assumptions:

1. Each document is just a collection of words or a "bag of words".

2. StopWords such as am/is/are/of/a/the/but/... do not contain any "topic" information and should therefore be omitted as a preprocessing stage from the papers. In fact, without missing any details, we can exclude phrases that appear in at least 80% ~ 90% of the papers.
3. We hypothesis beforehand how many topics we want. ‘k’ is pre-decided. In this case, k is equal to 2 since the gender bias is between women and men.
4. All subject assignments are accurate, except for the current word in question, and then the assignment of the current word is modified using our model of how documents are created.

3.1.3 How does LDA work?

There are 2 parts in LDA:

1. The words are from the document.
2. The words that belong to a topic or the probability of regarding a topic, that need to be calculated.

3.1.4 The Algorithm process

Firstly, go through each document and randomly assign each word feature in the d to one of k is chosen beforehand. Then, for each document d, go through each word weight and compute:

1. p (t | d)
2. p (w| t)

3. Update the final p for the weight (w) which is belonging to topic t, as:

$$p\ (w * t) = p\ (t|d) * p(w|t)$$

In this case, we mentioned before that the LDA model has been trained by 2 times for detecting the gender style coded words difference between the entry and senior level of data scientist. The final results of the 4 topics as shown in the table 3 and will be discussed comprehensively in the section 4 of this paper. In addition, the topics generated results have been validated by some linguists manually.

General Topic Masculine_coded JD		General Topic Feminine_coded JD		Senior Topic Masculine_coded JD		Senior Topic Feminine_coded JD	
deep_learning	0.004	please_apply	0.007	this_role	0.003	reference_code	0.004
predictive_modeling	0.002	please_read	0.005	predictive_modeling	0.003	will_doing	0.004
artificial_intelligence	0.002	reference_code	0.005	artificial_intelligence	0.003	please_apply	0.004
programming_language	0.002	must_authorized	0.005	operation_research	0.002	posted_today	0.004
natural_language	0.002	deep_learning	0.005	team_member	0.002	what_need	0.004
cross_functional	0.002	protected_veteran	0.004	cross_functional	0.002	least_year	0.003
decision_making	0.002	other_characteristic	0.004	natural_language	0.002	characteristic_protected	0.003
quantitative_field	0.002	eligibility	0.004	programming_language	0.002	summary_location	0.003
operation_research	0.002	employer_qualified	0.004	more_than	0.002	employer_qualified	0.003
more_than	0.002	required_verify	0.003	decision_making	0.002	please_read	0.003
best_practice	0.002	form_upon	0.003	best_practice	0.002	will_receive	0.003
health	0.002	complete_required	0.003	health	0.002	large_scale	0.003
large_scale	0.002	identity_eligibility	0.003	security	0.002	must_authorized	0.003
summary_location	0.002	federal_person	0.003	quantitative_field	0.002	consideration_employment	0.003
fast_paced	0.002	hired_will	0.003	preferred_qualification	0.002	doing	0.002
national_origin	0.002	verification_document	0.003	structured_unstructured	0.002	competitive_salary	0.002
work_closely	0.002	cybercoders_proud	0.003	fast_paced	0.002	origin_disability	0.002
marketing	0.002	employment_eligibility	0.003	internal_external	0.002	religion_national	0.002
real_world	0.002	origin_disability	0.003	sexual_orientation	0.002	status_other	0.002
veteran_status	0.002	least_year	0.003	financial	0.002	work_united	0.002
demonstrated	0.002	will_receive	0.003	marketing	0.002	protected_veteran	0.002
internal_external	0.002	religion_national	0.003	must_have	0.002	right	0.002
open_source	0.002	summary_location	0.003	large_scale	0.002	medical_dental	0.002
time_series	0.001	united_state	0.003	gender_identity	0.002	feature	0.002
cutting_edge	0.001	structured_unstructured	0.002	software_development	0.002	employment_eligibility	0.002

Table 3. Gender bias coded word observation by LDA .

It is very interesting that the JD topic modeling result of senior data scientist occurred some new terms like sexual_orientation and gender_identity together with the decision_makeing which is a strong masculine token. These results are telling us if

there are existing some gender tendency in job description which means this job is more intended to hire the men candidates.

3.2 Differentiate masculine and feminine wording by t-SNE

We get some gender bias insights from the topic modeling results and some quantitative visualization by t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a method for dimension reduction which is outperform for visualizing the high-dimensional nonlinear relationship datasets. The pseudo code of t-SNE is showed in figure x and implemented by sklearn repository with python. In this case, we utilize t-SNE for visualizing the 2 topics modeled by LDA. It is easy to observe that the difference between Fig.9 and 10, the red color cluster represents a feminine coded JD, on the other hand, the black represents a masculine coded JD.

The absolute numbers of feminine coded JD in senior level is quiet small which need to be seen by zoom in/out. Besides, this qualitative visualization results prove that the strong gender inequality phenomenon existed. Moreover, these results can indirectly illustrate that the gender inequality problem is even worse. The reason for this estimation is that if the company has more bias for recruiting potential candidates with advertisements which will make more women candidate afraid to apply. In the end, the gap of gender inequality will be larger than before.

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$,
cost function parameters: perplexity *Perp*,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, \ldots, y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity *Perp* (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, \ldots, y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for $t=1$ **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

Fig 8. t-SNE pseudo code

Fig 9. t-SNE visualization of general data scientist JD LDA topics

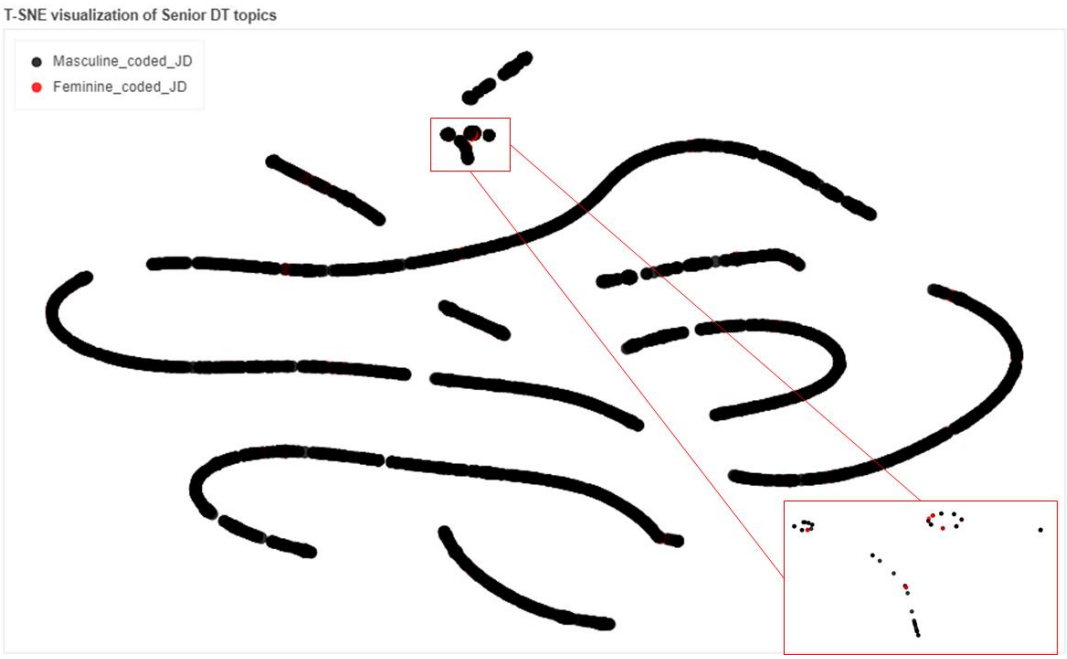

Fig 10. t-SNE visualization of senior data scientist JD LDA topics

4. Research results

4.1 Results

Combining our dataset of job descriptions and revised gendered wording list, we firstly filtered stop words, removed missing value and applied tokenization to get a comprehensive proportion of masculine words versus feminine ones in data scientists' recruitment advertisements in the U.S. As a primary result, we found that masculine oriented job descriptions (8437) are over 5 times of feminine ones (1563), which implies a strongly imbalanced as well as masculine oriented language usage of data scientists' recruitment in the U.S.

With this general results in mind, we dug further into different levels of data scientists' recruitment advertisements to figure out whether gendered wording intensified inequality promotion in workplaces. To fulfill this hypothesis, we labeled entry level and senior level job advertisements, and found out 6629 entry-level job descriptions, compared with 3356 senior-level ones.

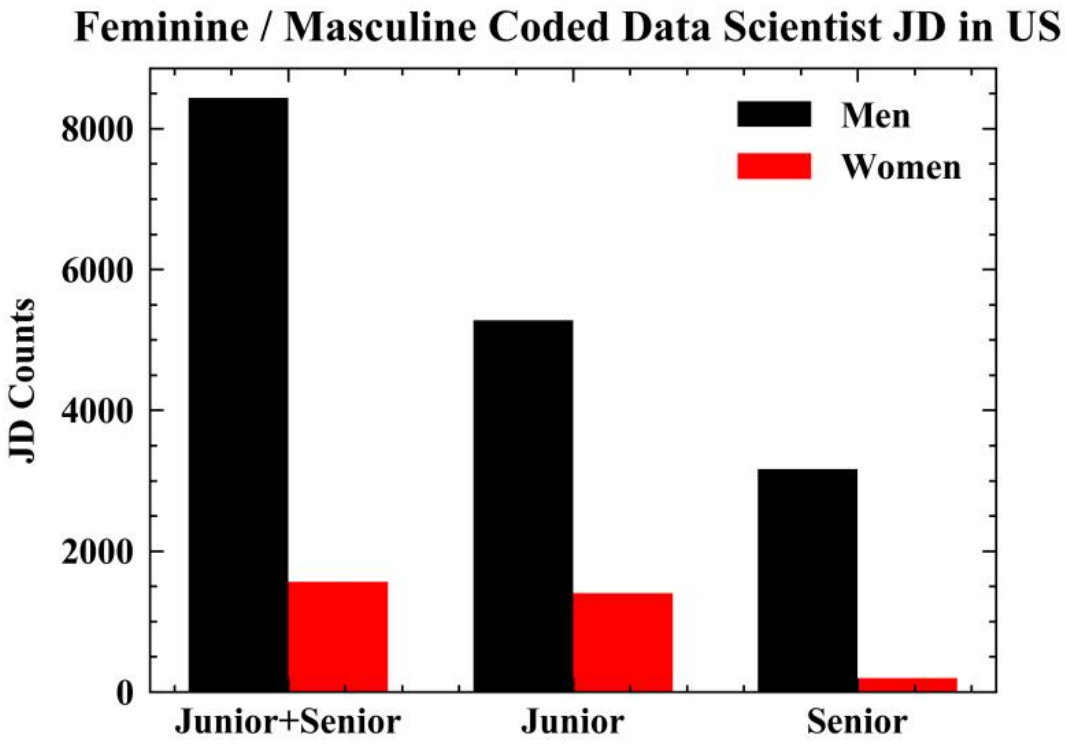

Fig 11. Femine / Masculine coded data scientist JD comparison

Then, we applied LDA algorithm to verify whether topic differentiation exists. As shown in Fig 11 and Table 3, gap indeed existed, and this gap was generated based on different topics selection towards different levels of job recruitment.

Then, we compared masculine and feminine topics frequency appeared in general dataset, entry level job advertisements and senior level job advertisements respectively. According to Fig 11, the higher the job title, the more masculine oriented the language used in the job descriptions, which increased from approximate 80% (5274/6674) to 96% (3163/ 3302). Also, we visualized this result by t-SNE algorithm.

According to social linguists and social psychologists, imbalanced gendered wording usage hugely influenced candidates' perception of self-qualification. Our comparisons between gendered topics frequencies in different levels of job positions are correspond with public reports and previous academic research focusing on gender inequality in IT industries, which we mentioned before. Furthermore, stronger masculine oriented job descriptions in senior-level just implies a rough condition of women's career paths, as long as they desire higher positions as a data scientist.

4.2 Discussions and Future direction

Although we've fulfilled 3 objectives proposed at the beginning of this paper, several dimensions need more discussions.

Firstly, all the job descriptions within our dataset located in the U.S, and we only collected recruitment advertisements in recent 2 years, which is a little bit limited. Although we collected supplementary data, yet not sufficient. If we want to draw a

more comprehensive conclusion, we need dataset covering a larger range of time and space.

Secondly, we used traditional social science method (questionnaire) to modify the gendered wording list, which could be challenged as arbitrary. Actually, the notions and tones of words keep changing, and there is no universal standard to classify masculine coded words and feminine coded words. However, we could use cosine distance to quantitatively illustrate the opposite gender orient of those words. Also, we could ask authoritative experts for help instead of postgraduate students.

Last but not the least, we directly accepted previous research results asserted that gender coded words applied in job advertisements will reinforce gender inequality in workplaces. However, these opinions could be validated by more interpersonal surveys and quantitative research.

5. Conclusion

It's universally acknowledged that gender inequality existed in all walks of life. There are alleged male-dominated type pf works, and female-dominated type of works. However, regardless impact of some special events like war or financial depression, women are always regarded as inferior human resources intentionally or unintentionally. This was proved by the lower proportion of women in workplaces, relatively lower salary and less opportunity of promotion or appointed to higher positions.

Based on this reality we tried to find out whether any innovation generated from technology development. Thus, we choose data scientists as our main focus and the currently most developed country, America as the representative region. Then, we adopted previous social science research that language usage will have huge subconscious influence on job hunters' self-evaluation, which may strengthen or alleviate gender inequality in workplaces.

Therefore, we choose job descriptions of data scientists' recruitment advertisements as data, and applied tokenization、LDA、t-SNE along with many other computational methods to figure out whether gender inequality is noticeable in newly generated occupation, that is data scientists.

Through accomplishing the objectives, we proposed before, we find out that masculine oriented descriptions are still mainstream in recruitment advertisements, and the higher the position, the higher the proportion of masculine coded language usage. These results are corresponding with public report about female situation in high tech industries. With big data and computational methods, we believe that the topic of gender inequality in workplaces could be better illustrated and further studied.

Reference:

Beede, D. N., Julian, T. A., Langdon, D., Mckittrick, G., Khan, B., & Doms, M. E. (2011). Women in STEM: A Gender Gap to Innovation. *SSRN Electronic Journal*, *04*(11). https://doi.org/10.2139/ssrn.1964782

Blau, F. D., & Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, *55*(3), 789–865. https://doi.org/10.1257/jel.20160995

Collier, D., & Zhang, C. (2016, October 1). Can We Reduce Bias in the Recruiting Process and Diversify Pools of Candidates by Using Different Types of Words in Job Descriptions? https://ecommons.cornell.edu/handle/1813/74363.

Correll, S. J., Benard, S., & Paik, I. (2007, March). Getting a Job: Is There a Motherhood Penalty? *American Journal of Sociology*, *112*(5), 1297–1339. https://doi.org/10.1086/511799

Coury, S., Kumar, A., Huang, J., Prince, S., Krivkovich, A., & Yee, L. (2020, September 30). *Women in the Workplace 2020*. McKinsey & Company. https://www.mckinsey.com/featured-insights/diversity-and-inclusion/women-in-the-workplace.

Duffin, E. (2020, July 28). *U.S. higher education - number of bachelor's degrees 2030*. Statista. https://www.statista.com/statistics/185157/number-of-bachelor-degrees-by-gender-since-1950/.

Eagly, A. H., & Karau, S. J. (1991). Gender and the emergence of leaders: A meta-analysis. Journal of Personality and Social Psychology, 60, 685–710. doi:10.1037/0022-3514.60.5.685

Ellemers, N., & Rink, F. (2016). Diversity in work groups. *Current Opinion in Psychology*, *11*, 49–53. https://doi.org/http://dx.doi.org/10.1016/j.copsyc.2016.06.001

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. https://doi.org/10.1037/a0022530

Heilman, M. E. (1983). Sex bias in work settings: The lack of fit model. Research in Organizational Behavior, 5, 269 –298.

Rudman, L. A., & Kilianski, S. T. (2000). Implicit and explicit attitudes toward female authority. Personality and Social Psychology Bulletin, 26, 1315–1328. doi:10.1177/0146167200263001

Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate). World Bank Data. (2020, September 20). https://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS?end=2020.

Pratto, F., Sidanius, J., & Levin, S. (2010, 7 29). Social dominance theory and the dynamics. European Association of Experimental Social Psychology, pp. 271 – 320.

Rivera, L. A., & Tilcsik, A. (2016). Class Advantage, Commitment Penalty: The Gendered Effect of Social Class Signals in an Elite Labor Market. *American Sociological Review, 81*, 1097–1131. https://doi.org/10.31235/osf.io/ywp93

Titcomb, J. (2018, April 3). *Apple pledges to hire more women as it reveals UK gender pay gap*. The Telegraph. https://www.telegraph.co.uk/technology/2018/04/03/apple-pledges-hire-women-reveals-uk-gender-pay-gap/.

Watts, A. W. (2018, March 28). *Why does John get the STEM job rather than Jennifer?* The Clayman Institute for Gender Research. https://gender.stanford.edu/news-publications/gender-news/why-does-john-get-stem-job-rather-jennifer.