Revealing Relevant Factors that Prompt Gendered Language in Job Descriptions

He Zhang
Georgia Tech
charles.zhang@gatech.edu

Miasia Jones
Georgia Tech
miasia.jones@gatech.edu

Haomin Lin
Georgia Tech
humaslin97@gatech.edu

ABSTRACT

This project investigates if gendered language used in online job descriptions vary by job salary, location, and company size. The dataset included job descriptions from Indeed.com and state demographic information from the U.S. Census Bureau. To investigate job salary, we proposed a salary prediction model, where we utilized an ensemble comprised of the k-nearest neighbors, random forest, and support-vector machine models. Gendered language based on job location was analyzed by conducting analysis of variance tests and Tukey's range tests. T-test were used to investigate the statistical difference in gendered language among companies of different sizes. Results showed that there is evident variation among the three aforementioned factors.

KEYWORDS

gendered language, online job descriptions, k-nearest neighbor, random forest, support-vector machine, ANOVA, Tukey's range test, T-test

ACM Reference Format:
He Zhang, Miasia Jones, and Haomin Lin. 2020. Revealing Relevant Factors that Prompt Gendered Language in Job Descriptions. In *Web Search & Text Mining*. CSE 6240, Atlanta, GA, USA, 6 pages.

1 INTRODUCTION

Gender impacts the candidate journey. Companies can influence candidate journeys in order to attract a certain crowd of job seekers, but they can inadvertently discriminate candidates before they even apply for the job because of use of gendered language in their job descriptions. Gendered language refers to language that stereotypically assigns nouns and adjectives to distinct sex-based categories, masculine and feminine. In this project, we aimed to discover if gendered language used in job descriptions vary by job salary, location, and company size.The dataset included job descriptions from Indeed.com and state demographic information from the U.S. Census Bureau.

To investigate job salary, We trained the base models, our three baseline models, k-nearest neighbors (KNN), random forest, and support-vector machine (SVM) using job descriptions and their corresponding salaries as features, and determined that the random forest model performed the best among the other models with a 86% accuracy. We then utilized an ensemble comprised of the three aforementioned models and was able to achieve 87% accuracy. The sigmoid gendered language score measure was used as baseline for the next two studies, and both incorporated the tanh gendered language score measure in the proposed methods. Gendered language based on job location was analyzed by conducting analysis of variance tests to compare the gendered wording in the job descriptions among different job locations and Tukey's range tests to know the pairs of significantly different locations. From the results of this study, we determined that there were significant statistical differences in gendered wording in job descriptions based on job location. T-test were used to investigate the statistical difference in gendered language among companies of different size. The results showed behavioral differences between the large and small firms tested. We also found that big firms tend to use more masculine words in job descriptions than the small firms. Using the two different gendered language score measures, we found that both gave different results partly dues to the fact that the tanh measure exhibited higher variation than the sigmoid measure.

Web Search & Text Mining, Spring 2020, Atlanta, GA, USA

2 LITERATURE REVIEW

On job markets, even though regulations such as Title VII of the Civil Rights Act[1] have been established to prohibit employment discrimination and promote equal opportunity for all employees, discrimination based on gender still cannot be eliminated from employees. Sandra Bem and Daryl Bem[2] provide a quantitative study demonstrating that sex-biased or gendered language in job advertisements and their placement in sex-segregated newspaper columns stops qualified men or women from applying. Gaucher, Friesen, and Kay[4] use archival and experimental analyses to show that gendered language is still being used in job recruitment materials. Both studies provide useful historical and present-day perspectives regarding this matter. However, these studies do not seek to explain relevant factors that could prompt the use of gendered language in job advertisements. And recent study on job ads by Afra R Chowdhury et al.[3] in Indian and Peter Kuhn[7] in China using regression analysis also reveal that firms show gender preference when advertising their jobs. However, those researches might not be convincing enough as well for only adopting regression analysis in the papers.

Recently, Newman et al.[8] reveal that one factor that could that could prompt the use of gendered language in job advertisements could be that men and women communicate differently. They use multivariate analyses of variance to weight language features differently and achieve maximum discrimination between the genders. In their results, they discuss how women tend to communicate their internal processes, such as thoughts and emotions, while men spoke

about possession of objects and used words that described quantity. This research provides a great technique for content analysis but lacks psychological analysis. We will try to examine descriptive.

3 DATA DESCRIPTION

3.1 Data Preparation

3.1.1 Data Source

. In this project, we use two types of data. The first stream of data that we use is the job posting data. The job posting data are scraped from Indeed.com[1] directly by using Python scripts. The reason for using the data from Indeed.com is that Indeed.com is such a popular and comprehensive job search website that it covers almost all the positions on other job searching websites (such as Linkedin, Monster.com and so on). The other dataset that we use is the demographic information from the U.S. Census Bureau. The U.S. Census Bureau provides the demographic estimation based on its 2010 census and 2012 business census statistics[2]. The U.S. Census Bureau data is the one and only official demographic data published by the government, and we believe it's the most reliable source of the demographic information.

3.1.2 Data preprocessing steps and explanations

. We crawled the job posting records for 100 biggest cities in the U.S. which are posted since the beginning of the year 2020. More specifically, we scraped the company name, job location, posting url, salary, and position description contained in each job posting to get the raw dataset. The raw dataset contains about 290k records. In order to get the gendered scores of the job postings based on the description which is the key element to answer our main research questions, we conducted the following steps, following the gendered word analysis research by Gaucher et al. [4]:

1. We removed all html tags and all non-letter string syntaxes from the description by using the NLTK, bs4 and re modules.
2. We split the cleaned descriptions into lists by using blanks as the separators to get lists of tokens for the job descriptions.
3. We compared the word in the lists of tokens to the list of gendered words used by Gaucher et al.[4], and counted the numbers of masculine words and feminine words in each list of tokens of job description.
4. We followed Gaucher et al.[4] and computed the differences between the number of masculine words and that of feminine words for each job description by subtracting the number of feminine words from the number of masculine words.
5. We followed Gaucher et al.[4] to generate the gendered scores by doing a sigmoid transformation on the difference obtained from step 4. Since we also proposed a new gender score measurement by using the tanh transformation. Therefore, we also generated the gendered scores by doing a tanh transformation .

Besides, since we also want to explore whether the location and the local demographics have impacts on the gendered scores of the local job postings, we do the following:

1. We obtained the population, female population percentage, white population, minor ethnicity percentages, Internet usage percentages, high school graduate percentage, college graduate percentages, number of firms owned by women, and the total number of firms in the city from The U.S. Census Bureau website (see footnote 2).
2. We merge the data obtained from step 1 with the dataset with the gendered scores.

In addition, because we also try to explore the relationship between the generated words used in the description and the salary level, we further processed the dataset by doing the following steps:

1. We classified the salary types into "hourly", "daily", "weekly", "monthly", and "yearly" categories, and we labeled the records accordingly.
2. For the salary data with a range, like "$ 20 - $ 30 an hour", we extracted the first number as the minimum salary and the second number as the maximum salary. We also computed the mean salary of the maximum and minimum and treat it as a fixed salary.
3. For the salary data with a fixed number, like " $ 60000 a year", we extracted the one and only number as the fixed salary.
4. For the salary data with a upper bound or lower bound, like "Up to $ 5000 a month" or "From $ 200 per day", we extracted the numbers as the minimum salary or maximum salary.
5. We annualized the salary based on the label obtained from step 1. More precisely, we multiplied the hourly rate with 8*253 to get annualized salary, we multiplied the weekly rate with 52 to get annualized salary, and we multiplied the monthly rate with 12 to get annualized salary.

3.2 Raw Data Statistics

The raw dataset contains 282718 observations. All records in the raw dataset contain the count of feminine words, the count of masculine words, the difference between the word count, the sigmoid gendered score, and the tanh gendered score. However, only 68845 records in the dataset contain the salary information.

On average, the descriptions in the sample contain 481 words and 23 sentences, which means each sentence contains around 21 words on average. The longest description contains 3732 words and 623 sentences.

The most important features of the job postings are number of feminine words, number of masculine words, the difference between the counts, the gendered scores, minimum annualized salary, maximum annualized salary, fixed annualized salaries. Table 1 is the summary statistics of these key variables.

3.3 Raw Data Analysis

The key takeaways from the summary statistics of the raw data include the following:

- On average, the new job posting of this year in the 100 biggest cities of the U.S. are leaning more towards femininity with a sigmoid score of 0.43 and a tanh score of -0.16.
- On average, the new job posting of this year in the 100 biggest cities of the U.S. are using more feminine words than masculine words.

[1]Please see https://www.indeed.com/

[2]Please see https://www.census.gov/quickfacts/fact/table/houstoncitytexas, seattlecitywashington,losangelescitycalifornia,indianapoliscitybalanceindiana, atlantacitygeorgia,newyorkcitynewyork/PST045219

Table 1: Summary statistics of these key variables.

Variable	Mean	Std. Deviation	Min	Max
Number of masculine words	5.55	5.49	0	70
Number of feminine words	6.25	5.43	0	97
Difference	-0.70	5.12	-76	55
Sigmoid gendered score	0.43	0.38	0	1
Tanh gendered score	-0.16	0.86	-1	1
Minimum annualized salary	25698.44	36321.3	0	$500000
Maximum annualized salary	39191.92	59723.03	0	$500000
Fixed annualized salaries	36326.98	47150.74	0	$500000

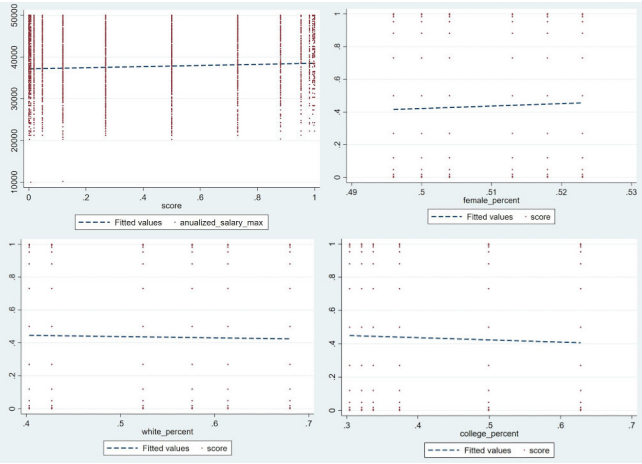

Figure 1: Scatter and fitted line graphs.

From Figure 1, we can spot some interesting preliminary findings:

- The annualized salary is positively correlated with the gendered score, which means if the job posting is more masculine, the salary is likely to have a higher upper-bound.
- The gendered score is positively correlated with the percentage of females in the city, which means if a city has more females, the job postings in that city is likely to be more masculine. This finding is sort of counter-intuitive.
- The gendered score is negatively correlated with the percentage of white and the percentage of college grads, which means if a city has a higher percentage of college degree holders or white population, the job postings are likely to be more feminine.

4 EXPERIMENTAL SETTINGS

For model building and experiments, we set test_size value as 70%, which means that 70% of the whole dataset is used to train our model and the rest for validation. We chose to not use cross validation to save time, with accuracy score as the evaluation metric. The computing is run on Google Colab, with RAM=25.51GB.

5 SALARY PREDICTION WITH JOB DESCRIPTIONS

5.1 Baselines

As the number of samples is still limited for analysis on the relationship of salary and gender score, we decided to train a model to produce more samples. Since the regression methods in the milestone report performs not as expected and lacks significance in interpreting the relationship between salary and gender scores, we then turned to classification. And a three level division approach is taken to label each salary value as high, medium, or low salary level.

To choose the values for salary level classification, we investigated the distribution of all existing salary values, then chose 60000 and 120000 as two values that split the dataset.

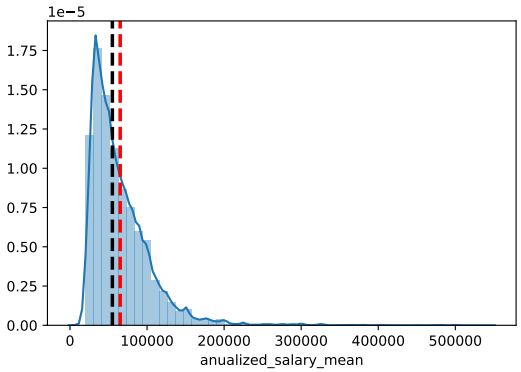

Figure 2: Distribution of all existing salary values

To prepare features for prediction, we joined information of each job posting including job title, location, company, description as a new string, then used the Tf-idf algorithm to vectorize each string. With features ready, according to previous work from Zhang et al.[10], Jackman et al.[5], and Khongchai et al.[6], we implemented three baseline models to predict the labels, including KNN (k=2), Random Forest (n_estimator = 650), and SVM (with a polynomial kernel). The performances of them are respectively 0.7985, 0.8638, 0.8542.

5.2 Proposed Method

Then, to improve the performance of prediction, we proposed a new model called ensemble learning[9] . In this method, we make predictions in two layers. In the first layer, we train three baseline models with hyperparameters that have the best performance, then use them to predict on training sets to create new labels. These new labels will be merged into a new dataset as the new features. And the test sets will also be predicted for scoring the ensemble learning model. In the second layer, we will train a new model

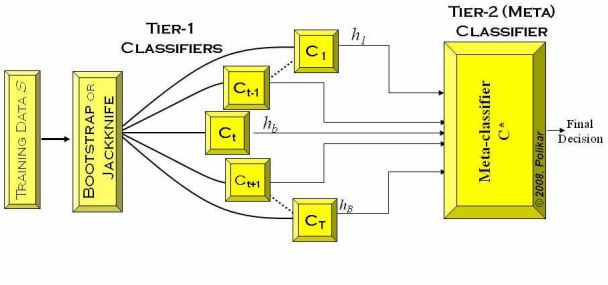

Figure 3: Illustration of ensemble learning process

called combiner. This combiner will be used to predict final results using features predicted and merged in the first layer, which will be scored by the new test set generated from the first layer.

In our proposed method, we chose KNN (k=2), Random Forest (n_estimator = 650), and SVM (with a polynomial kernel) from baselines as the first layer predictors. Then, in the second layer, we tune a KNN model on the k value as well as the n_estimator of a Random Forest model to get the best performance of them. Comparing their prediction results, we chose the Random Forest model with n_estimator = 195 as the combiner.

This ensemble learning is more specifically called stack learning. It can help to correct some errors made by models in the first layer and have better performance.

To make sure the baseline models and ensemble learning is tested on the same context, we split the dataset before testing all the models, so that we can score the baseline models in the first layer prediction, then use the same splitting to score the ensemble learning method. Using the same dataset With ensemble learning implemented, we got an accuracy score of 0.87, which is slightly improved from what baseline models got.

Table 2: Comparing Result of ensemble learning with baselines

Model	KNN	Random Forest	SVM	Ensemble Learning
Accuracy Score	0.7985	0.8638	0.8542	0.8705

As we can see, the results are promoted by ensemble learning. It demonstrates that this method is able to correct some errors made by models in the first layer when they conduct prediction independently.

5.3 Experiment results

With the prediction of salary level done, we applied the result to interpret gender scores in different industries.

The method is similar to Countvectorizer. However, we count the occurrences of keywords from different industries in the description in a job posting, and assign that job to a category with its keywords having most occurrences. And we can see that the distribution of industries in the dataset:

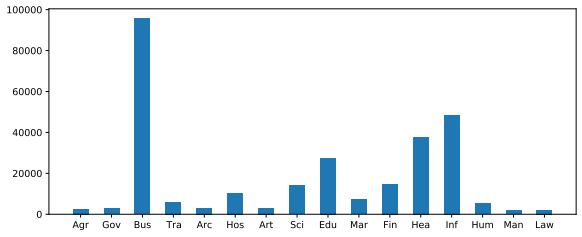

Figure 4: Distribution of industries in the dataset

With the prediction and classification done, we then studied the relationship between gender scores and industry, as well as salary level. As is shown in the graph, we can see that the Financial and IT industry are the industries that tend to post more masculine job descriptions, while Human services, along with Hospitality and Tourism, are the most feminine ones.

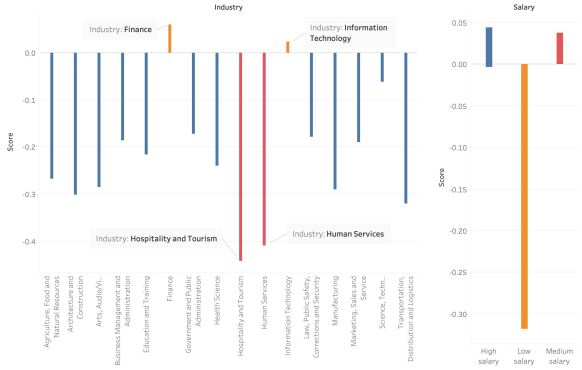

Figure 5: Gender scores in different industries and salary levels

As for salary level, we can see the high and medium salary levels are associated with higher gender scores, which means more masculine job postings. While job postings with lower salary levels are more feminine. This result is consistent with what we found in the data analysis part.

6 ANALYZING GENDERED LANGUAGE BASED ON JOB LOCATION

This section will explore if job location influences the use of gendered language in job descriptions. The job locations that will be examined are Atlanta, Georgia (ATL), New York City, New York (NYC), Los Angeles, California (LA), Indianapolis, Indiana (IND), Seattle, Washington (SEA), Houston, Texas (HOU), Denver, Colorado (DEN), and Minneapolis, Minnesota (MIN). These specific cities were selected so different regions of the United States could be represented in the analysis. From each job location, 2,400 job postings were arbitrarily chosen and included in the analysis.

6.1 Statistical Analysis

Figure 6 shows the distributions of both sigmoid and tanh gendered scores for each job location. The sigmoid scores fall into the range of 0 to 1, while the tanh scores fall into the range of -1 to 1. Scores that fall in the lower bound of the range correspond to male targeted job postings, scores that fall in the upper bound of the range correspond to female targeted job postings, and scores that fall in the middle are considered gender neutral job postings.

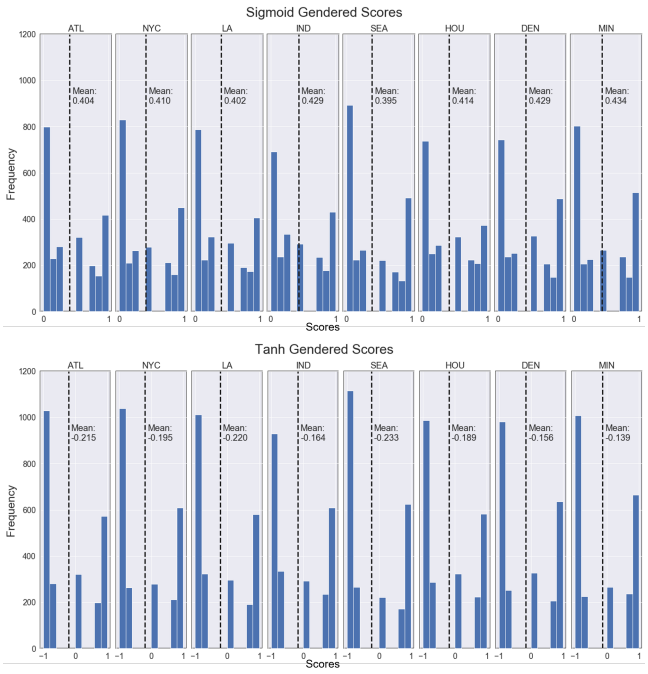

Figure 6: Distributions of gendered scores by city.

6.2 Analysis of Variance Testing

Gaucher et al. [1] conducted a study on the wording differences in a public sample of online job advertisements. For this study, they collected job advertisements from male-dominated and female-dominated occupations and calculated the percentage of masculine and feminine words for each job advertisement. To test the impact of gendered wording, they conducted a 2 (occupation: male dominated vs. female dominated) x 2 (wording: masculine vs. feminine) mixed model analysis of variance (ANOVA), with wording as the repeated measure. The main effect of gendered wording that was found by Gaucher et al. was F(1, 491) = 24.51 and p-value = .001 indicating that advertisements contained more masculine words than feminine words.

As a similar approach, we will conduct One-Way ANOVA tests to compare the gendered wording mean scores between job locations using statistical significance. The ANOVA tests will be obtained using the Statsmodels Python library[3]. There will be two separate ANOVA tests, one using sigmoid gendered wording scores, and the other using tanh gendered wording scores. The gendered wording scores will be the dependent variables and Ordinary Least Squares regression will be the model used. The alpha the ANOVA will be tested at is 0.05. The hypotheses to test include the following:

H_0: There is no significant variation in the gendered wording mean scores of the job locations.
H_a: There is significant variation in the gendered wording mean scores of the job locations.

6.3 Results and Post-hoc Testing

Table 4 contains the results of the ANOVA tests. These results indicate that we reject our null hypothesis (P-value < 0.05) and conclude that there are significant statistical differences among the job locations and their gendered wording scores.

Table 3: ANOVA tests results.

Score Type	F(7,19192)	P-value
Sigmoid	2.5294	0.0270
Tanh	2.1078	0.0614

From the ANOVA testing, we know that the differences of gendered wording scores of the job locations are statistically significant, but we did not identify which locations are significantly different from each other. To know the pairs of significant different locations, we will perform multiple pairwise comparison analysis using Tukey's range test from Statsmodels[4]. The gendered wording scores will be the independent data samples and the group labels correspond to each job location.

The results of the Tukey's range test using the sigmoid gendered wording scores show that all pairwise comparisons of job locations, except DEN-SEA, IND-SEA, and MIN-SEA, rejects the null hypothesis and indicates statistically significant differences. Seattle shows the most statistical difference in sigmoid scores than any other location. The results of the Tukey test using the tanh scores show that all pairwise comparisons of job locations, except ATL-MIN, DEN-SEA, LA-MIN, and MIN-SEA, rejects the null hypothesis and indicates statistically significant differences. Minneapolis shows the most statistical difference in tanh gendered wording scores than any other location. Overall, the two different gendered score measures output different results in the range Tukey tests, which could be due to how the tanh scores are more skewed in both directions than the sigmoid scores.

7 BIG FRIMS VS. SMALL FIRMS

As perceived by many, big firms are usually offering higher salaries. Since we identified that there exists positive correlation between gendered scores of the job descriptions and salary levels, testing whether big firms offering higher salaries are associated with higher gendered scores will help us to confirm the robustness of the conclusion.

[3]See https://www.statsmodels.org/stable/anova.html

[4]See https://www.statsmodels.org/stable/generated/statsmodels.stats.multicomp.pairwise$_tukeyhsd.html$

7.1 Classification of Firms and Firm Matching

We define the big firms to be the F500 firms[5] and small firms to be non-F500 firms. As indicated by the data, company names can vary even they are referring to the same company. For instance, 'J.P. Morgan' and 'JP Morgan' can refer to the same firm. 'IBM' and 'International Business Machine' can refer to the same firm as well. In order to overcome this difficulty and match the firms in the F500 list, we use a fuzzy match and set the threshold to 0.85. More precisely, we use the 'fuzzywuzzy.py' module to compute the fuzzy similarity scores of the pairs of company names. If the score is higher than 0.85, we think they are referring to the same firm. If a firm can be matched with any firms in the F500 list with a similarity score above 0.85, we classify it into the group of big firms, otherwise we classify it into the group of small firms.

7.2 Proposed Method of Statistical Analysis

We conducted t-tests on both the sigmoid scores and the tanh scores: The baseline measure that we use is the sigmoid score, but because it lacks variation and most values concentrate between 0.25-0.75, the t-test might not give an accurate result. Therefore, we also use the tanh score as another measure for this test. The tanh scores have higher variation and can increase the statistical significance.

7.3 Results

The results of t-tests are shown in Table 5.

Table 4: Table 5:Results of t-tests.

Gendered Score Type	Mean of F500 firms	Mean of non-F500 firms	p-value
Sigmoid Scores	0.432	0.430	0.54
Tanh Scores	-0.139	-0.156	0.01

The confidence level is set to 0.05 in these tests. The hypotheses to test include the following:
H_0: There is no significant difference in the gendered wording mean scores between big firms and small firms.
H_a: There is significant difference in the gendered wording mean scores between big firms and small firms.

As we can see, by using the two different gendered score measures, the p-values are different. By using the sigmoid score, the difference between the F500 firms and the non-F500 firms is statistically insignificant. However, when using the tanh gendered score, the difference becomes significant and the F500 firms tend to use more masculine words when posting jobs online. Please note that This is nothing but a simple and preliminary analysis which generates some high-level insights. In the future, the tests can be conducted for different job types within the big and small firms, which may provide more precise information.

[5]The 2019 F500 list was obtained from https://www.someka.net/excel-template/fortune-500-excel-list/

8 CONCLUSION

This project investigated if gendered language used in job descriptions varied by job salary, location, and company. Through testing and modelling, we found that there is variation among those three factors. We revealed that more masculine job descriptions are associated with higher salaries with our proposed salary prediction model. We concluded that there are statistical differences in gendered wording in job descriptions based on job location with our ANOVA testing. With our T-test, we found behavioral differences between the large and small firms. We found that there exist behavioral differences between the big firms and the small firms, and that the big firms, especially those of the Fortune 500, tend to use more masculine words in online job postings than the small firms.

For future work in the investigation of job salary, we want build a dataset with labels as the validation set and find a model that can most accurately classify each point in the set. Also, we plan on being more strategic in what data is used in the investigations on job locations and the size of firms by collecting data for specific occupations, instead of randomly selecting job descriptions.

9 CONTRIBUTION

For this report, He Zhang contributed mostly to the data processing and analysis, and conducted the analysis of firm size. Haomin Lin contributed mostly to build the prediction baselines. Miasia Jones contributed mostly to analyzing gendered language based on job location.

REFERENCES

[1] 1964. Civil Rights Act of 1964. § 7, 42 U.S.C. § 2000e et seq (1964).
[2] Sandra L Bem and Daryl J. Bem. 1973. Does Sex-biased Job Advertising "Aid and Abet" Sex Discrimination?1. *Journal of Applied Social Psychology* 3, 1 (1973), 6–18. https://doi.org/10.1111/j.1559-1816.1973.tb01290.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1973.tb01290.x
[3] Afra R. Chowdhury, Ana C. Areias, Saori Imaizumi, Shinsaku Nomura, and Futoshi Yamauchi. 2018. Reflections of employers' gender preferences in job ads in India : an analysis of online job portal data. *Policy Research working paper* 8379 (2018), 1–24.
[4] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology* 101, 1 (2011), 109-28. https://doi.org/10.1037/a0022530
[5] Shaun Jackman and Graham Reid. 2013. *Predicting Job Salaries from Text Descriptions.* Ph.D. Dissertation. University of British Columbia.
[6] P. Khongchai and P. Songmuang. 2016. Random Forest for Salary Prediction System to Improve Students' Motivation. In *2016 12th International Conference on Signal-Image Technology Internet-Based Systems (SITIS).* 637–642.
[7] Peter Kuhn and Kailing Shen. 2018. Gender Discrimination in Job Ads: Evidence from China. *Quarterly Journal of Economics* 128, 1 (2018), 287–336.
[8] Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes* 45, 3 (2008), 211–236. https://doi.org/10.1080/01638530802073712
[9] Cha Zhang and Yunqian Ma. 2012. *Ensemble Machine Learning: Methods and Applications.* Springer Publishing Company, Incorporated.
[10] Junyu Zhang and Jinyong Cheng. 2019/08. Study of Employment Salary Forecast using KNN Algorithm. In *2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019).* Atlantis Press. https://doi.org/10.2991/msbda-19.2019.26