Revealing Relevant Factors that Prompt Gendered Language in Job Descriptions

Charles Zhang
Haomin Lin
Miasia Jones

Introduction

- Aim to discover if gendered language used in job descriptions vary by job title, location, and company.
- Investigate if the use of gendered language in job descriptions is still prominent today and if it explains why the male and female employee ratio is imbalanced in certain occupations.
- Explore if gendered language changes with job title and salary, and whether salary can be even predicted from the job descriptions.
- The results of this project would be meaningful to job seekers and employers and, if successful, would reveal relevant factors that can prompt the use of gendered language in job descriptions.

Dataset Description and Analysis

Data Source

1. Indeed.com (scraping job descriptions)
2. U.S. Census Bureau (demographics)

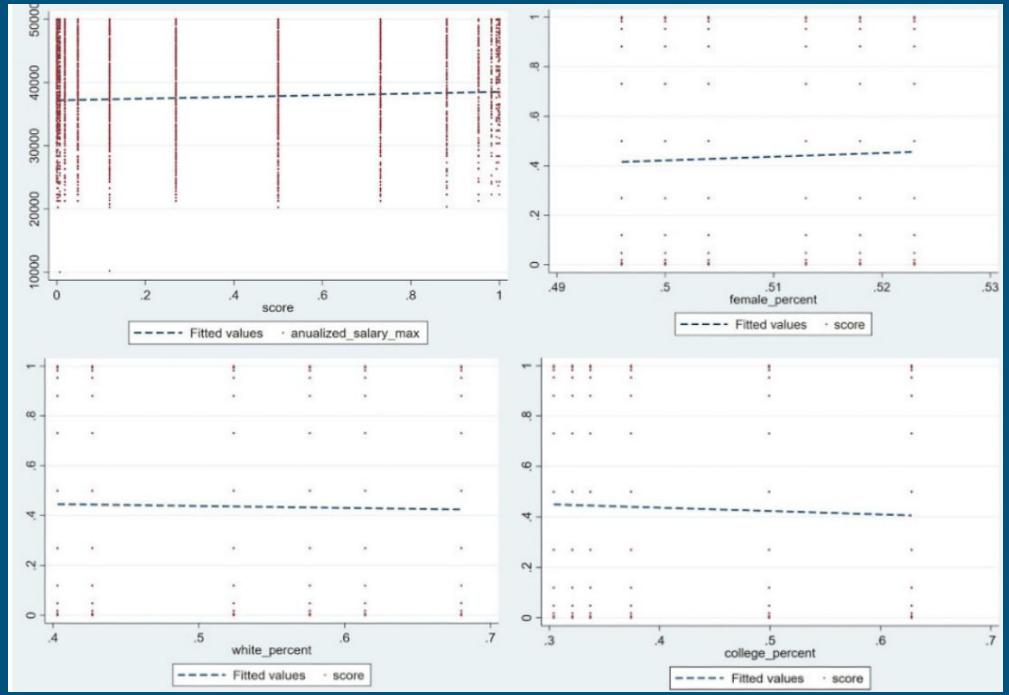
Data Overview

- The raw dataset contains **282718** observations.
- Only **68845** records in the dataset contain the salary information, we annualized the hourly, daily, weekly and monthly salaries.
- 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.

Dataset Description and Analysis

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



Salary Prediction with Usage of Gender Language

Experiment setting and baselines

Experiment environment: The data parsing, model training and prediction are all run on Google Colab, with RAM=25.51GB.

Experiment data: 70% training/30% testing

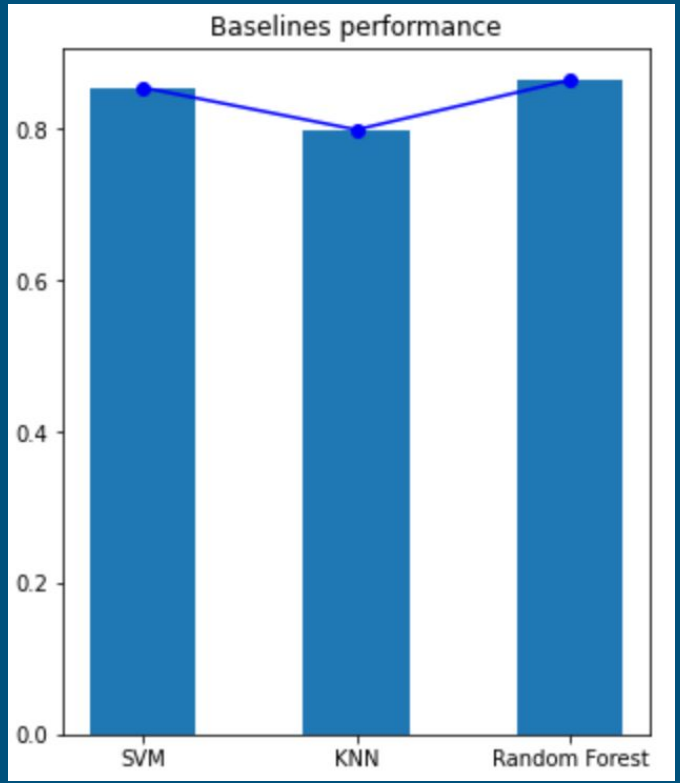
Baseline models:

- Tang et al. (2017): sigmoid gendered scores
- Junyu Zhang & Jinyong Cheng (2019): Random Forest
- Shaun Jackman & Graham Reid (2013): KNN

Baselines

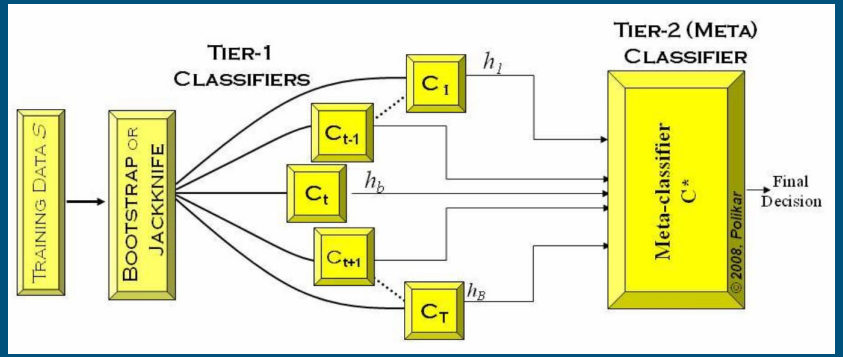
The Salary prediction model

- Evaluation metrics: Accuracy of prediction (classification)
- Baseline model:
  - KNN, Random Forest, and SVM classification model
- Baseline results:
  - (Accuracy) KNN: 0.7985, Random Forest:0.8639, SVM: 0.8542



Proposed Method

The Salary prediction model



- Model: Ensemble Learning（Stacked generalization）
  - Train a set of models on a training dataset D = $\{(x_i, y_i) \mid x_i \in X, y_i \in Y\}$
  - The outputs of these classifiers on each sample along with the true label will construct a new training set $\{x'_i, y_i\}$ where $x'_i = \{h_j(x_i)$ for j = 1 to T$\}$. A new classifier is trained on the new training set, and then used to predict the results
- Innovation of the proposed model:
  - A new classifier is trained in the second layer by the new training set predicted by the first layer models and then are used to predict the values in the second layer.
- Parameter:
  - First Layer: Random Forest (n_estimator=650), KNN (n_neighbor=2), SVM (kernel='poly')
  - Second Layer: Random Forest (n_estimator=195)

Experiments

The Salary prediction model

- Proposed model results:
  - Accuracy = 0.8705
- Comparison:
  - Successfully promote the accuracy of prediction: From 0.8639 to 0.8705
- Explanation for improvement:
  - The ensemble learning model can learn and correct the wrong classification of some classifiers in the last layer. Therefore, this model leads to an higher accuracy
- Follow up: Integrated strategies
  - Build a dataset with labels as the validation set
  - Use KNN to find the closest data points of the current point in validation set
  - Find the best model for the surrounding points to predict on the currently studied point.

Experiments

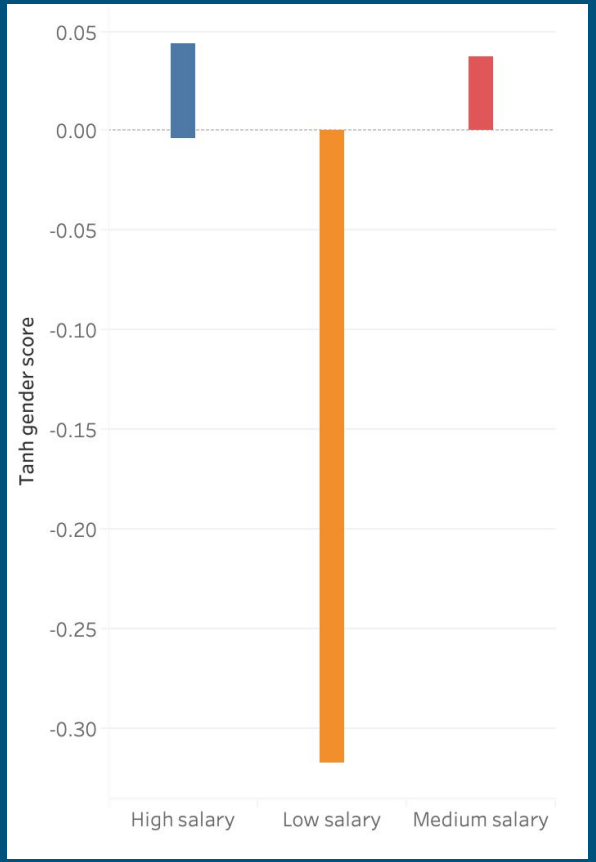
Job categories, salary level and gendered scores

- Job categories and gendered scores:
  - Finance and Information Technology have job descriptions that are most masculine
  - Human Services, Hospitality and Tourism have job descriptions that are most feminine

Experiments

Job categories, salary level and gendered scores

- Salary level and gendered scores:
  - Job descriptions with high and medium salary level are more masculine.
  - Job descriptions with low salary level are more feminine
- Actually, the gender score overall is low.



Analyzing Gendered Language Based on Job Location

Analyzing Gendered Language Based on Job Location

Proposed Method: Conduct One-Way ANOVA tests to compare gendered wording in job descriptions among different job locations.

- **Baselines:**
  - Gaucher et al. (2011): ANOVA testing done to measure statistical differences in gendered wording scores among job descriptions.
  - Tang et al. (2017) : Sigmoid function used to measure gendered wording.
- **Innovation:** Use gendered wording sigmoid and tanh scores between job locations as the dependent variables in two separate ANOVA tests.
- **ANOVA Experimental Settings*:** Ordinary Least Squares regression will be the model used. ANOVA will be tested at $\alpha=0.05$.
- **Evaluation metrics:** F-statistic and P-value

*Conducted using Python *Statsmodels* package.

Analyzing Gendered Language Based on Job Location

Results:

Baseline: The mean gendered wording scores of 2,400 randomly selected job descriptions from each location were successfully obtained.

Location	Sigmoid Mean Scores	Tanh Mean Scores
ATL	0.403591	-0.214981
NYC	0.409991	-0.195151
LA	0.402435	-0.220219
IND	0.428838	-0.164387
SEA	0.394856	-0.233321
HOU	0.413716	-0.188652
DEN	0.429376	-0.156212
MIN	0.433974	-0.138544

Gendered wording mean scores by job location*.

*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); Minneapolis, Minnesota (MIN).

Our Method: ANOVA testing concluded that there were significant statistical differences among the job locations and their gendered wording scores. (P-value < 0.05)

	F(7,19192)	P-value
ANOVA Test - Sigmoid	3.441956	0.001102
ANOVA Test - Tanh	3.629531	0.000646

ANOVA tests results.

Comparison: The distributions of the sigmoid and tanh scores are show variation for each location. ANOVA testing reveal a statistical difference between female and male targeted job descriptions.

Analyzing Gendered Language Based on Job Location

- **Post-hoc Testing:** Identify which locations are significantly different from each other using Tukey Range Tests.
  - **Tukey Range Tests Settings:** The gendered wording scores will be the independent data samples and the group labels correspond to each job location.
  - **Results:** Seattle showed the most statistical difference in sigmoid scores and Minneapolis shows the most statistical difference in tanh gendered wording scores.
- **Conclusion:** The distribution of tanh scores for every location are slightly more skewed in both directions than the sigmoid scores, which created different results.
- **Future Work:** Collect data for specific occupations, instead of randomly selecting job descriptions for each location.

Big Firms v.s. Small Firms

Proposed Method

Firm size and gendered scores

- T- test on the tanh scores of F500 firms and non-F500 firms
- Shortcomings overcomed: The tanh score introduces more variation and improve the statistical significance of the difference between big firms and small firms.

Experiments

Firm size and gendered scores

- ❏ Evaluation metrics: t-statistics
- ❏ Baseline: t-statistics on the sigmoid gendered scores
- ❏ Innovation of the proposed model: test on the tanh gendered scores instead
- ❏ Results:

	Mean of F500 firms	Mean of non-F500 firms	p-value under the **H0: The means are the same**
Sigmoid Score	0.432	0.430	0.54
Tanh Score	-0.139	-0.156	0.01

- ❏ Comparison: The statistical significance gets much larger when using the tanh scores by introducing more variation.

Conclusions

- The annualized salary is positively correlated with the gendered scores: More masculine job descriptions are associated with higher salaries.
- The ensemble modeling can improve prediction accuracy.
- The financial and IT industries relatively have the most masculine job descriptions while Human Services industries have the most feminine job descriptions.
- There are significant statistical differences in gendered wording in job descriptions among different job locations.
- There does exist behavioral differences between the big firms and the small firms: Big firms, especially the F500 firms tend to use more masculine words in online job postings than the small firms.

Reference

- 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)
- Shaun Jackman and Graham Reid. 2013. Predicting Job Salaries from Text Descriptions . Ph.D. Dissertation. University of British Columbia.
- C. Zhang and Y. Ma. 2012. Ensemble machine learning: methods and applications. Springer.
- 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
- Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. 2017. Gender Bias in the Job Market: A Longitudinal Analysis. *Proc. ACM Hum.-Comput. Interact*. 1, 2, Article 99 (November 2017), 19 pages.

Thank you for watching!