Gender Disparities in Tech: Evidence and Insights from Sentiment and Psycholinguistic Analysis

Abhinav Maurya
Carnegie Mellon University
ahmaurya@cmu.edu

ABSTRACT

Gender disparities and biases constitute one of the biggest problems facing the technology industry in recent times. Using an anonymized dataset of employee peer reviews and managerial performance evaluations from a large technology corporation, we study the nature of gender disparities and detrimental stereotypes that persist in the industry. We find preliminary evidence of a statistical performance ceiling whereby men are awarded a disproportionate share of the top performance outcomes compared to women. Sentiment analysis of the textual feedback provided in employee peer reviews finds weak evidence that reviews of female employees tend to be more positive than those of male employees. A multi-dimensional psycholinguistic analysis of peer reviews further reveals many of the commonly ingrained workplace stereotypes that can be detrimental to organizational culture, productivity, and equity. Our study serves to promote the strategic analysis of large-scale human resource data in technology organizations to detect and correct gender disparities and prevent such disparities from coloring the development of technologies designed for general widespread use.

ACM Reference format:
Abhinav Maurya. 2019. Gender Disparities in Tech: Evidence and Insights from Sentiment and Psycholinguistic Analysis. In *Proceedings of Anon, Anon, Anon,* 8 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent times, the technology industry has faced one of its biggest crises in the form of a growing number of revelations and accusations about systemic gender biases. According to a survey titled *Elephant in the Valley*[1] conducted shortly after the Ellen Pao versus Kleiner Perkins trial, 90% percent of the 210 surveyed women primarily from Silicon Valley said that they had witnessed sexist behavior at professional events, 60% reported unwanted sexual advances, and 87% reported being demeaned by their colleagues [1]. A self-perception of meritocracy exacerbates the situation by making people impervious to structural explanations of gender disparities [2]. A number of female workers in the tech sector have sued their employers over allegations over discrimination [3–7]. US Department of Labor has also investigated and sued technology firms in recent years over gender discrimination [8, 9].

However, employees often prefer to not report or take legal action against gender discrimination or biases because they fear retaliation, lack sufficient quantitative evidence on which to build a case, do not have the resources or support needed for protracted legal battles, or simply do not wish to revisit upsetting discriminatory experiences from the past [1]. In such cases, evidence of gendered disparate treatment takes the form of anecdotal experiences or analysis of workplace interactions. Susan Fowler's account of her experience working at Uber brought to light the toxic sexism that exists in certain parts of Silicon Valley [10]. Employees at Apple have reported misogynistic conversations and sexist harassment at the most-valued tech company on the stock exchanges [11]. An analysis of code reviews at Facebook revealed that female engineers were far more likely to have their professional contributions criticized and rejected than their male counterparts [12]. These are a few of the incidents whose discriminative and punitive implications for the affected female workers have not been fully explored.

Due to the sensitive nature of the issue and the associated legal implications for corporations, research on gender disparities has been limited by lack of concrete and complete data. As a result, much of the evidence for such disparities is descriptive and therefore inconclusive. For example, gender-based wage gaps are closely tied to occupational segregation, and therefore it is argued that occupational choices are the primary reason for the wage gap rather than gender [13]. In the case of Facebook's code reviews which were found to be overly critical of female engineers [12], an internal Facebook investigation concluded that seniority not gender explained the disparity - female engineers were facing more criticism because they were in more junior roles and therefore generally not as experienced as male engineers. Another potential factor is that certain departments or managers might be stricter and award poorer performance evaluations to everybody. If there are more women working for such departments or managers, then they will get poor scores and it might seem overall as if women are being discriminated against. However, if the men working for these departments or managers are being penalized equivalently, it cannot be deemed discriminatory. As a result, to understand the existence and nature of gender disparities, we need to control for factors related to the employee such as their occupation, seniority, organizational department, the average performance evaluation awarded by their manager to everyone in the team, etc. This process of causal inference from observational data avoids pitfalls associated with conclusions drawn from descriptive statistics, and can provide us better confidence in our findings.

In this work, we make the following contributions to research on gender disparities and biases in the workplace:

- We analyze managerial performance evaluations and find statistically significant evidence of a performance "glass ceiling," a phenomenon in which it is difficult for minorities to break an invisible barrier that doesn't apply to the majority class

[1] https://www.elephantinthevalley.com/

Anon, Anon, Anon
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[14, 15]. In some evaluation periods, we also find evidence for the "sticky-floor" hypothesis [16] wherein minorities are far more likely to be assigned the poorest professional outcomes e.g. promotions or performance evaluations. The discovered gender disparities persist even after we control for many of the commonly correlated factors.

- An automated sentiment analysis of peer reviews does not reveal negative sentiment toward female coworkers. In fact, the sentiment is on average slightly more positive toward female coworkers than male coworkers, contrary to reports of negative sentiment toward female workers [17].
- However, a multi-dimensional psycholinguistic analysis using the widely adopted LIWC scoring system [18] and its proprietary Receptiviti extension [19] reveals many of the commonly held gender stereotypes that can be detrimental to the professional success of female employees.

Often, cases of bias against protected minority groups are difficult to ascertain on an individual basis because the bias is subtle However, weak signals of bias against members of a protected group can be pooled together to effectively detect pervasive bias toward protected minority groups even if the bias is subtle at the individual level. In other words, the proper use of "big data" can help detect issues of gender disparity in a robust way.

2 DATA

In this paper, we study qualitative free-form text reviews and quantitative managerial performance evaluations provided to employees of a large technology corporation during their half-yearly evaluation cycles, denoted as MY for mid-year and YE for year-end. Figure 1 shows a bird's eye view of the technology organization being studied in this paper. Each employee is connected to their manager. The shape of the nodes in the graph indicates the gender. The color of a node is used to indicate the average performance score of employees in the subtree on a scale of 1-6. For leaves, the average performance is the performance of the single employee at the node. For a non-leaf node, the average performance is calculated using the performance scores of all employees in the subtree rooted at the said node. The inner ring indicates the department of the employees in a color-coded fashion. The outer ring indicates the tenure of the employees through the height of the ring.

The performance evaluation data spans five half-yearly cycles from 2014 YE to 2016 YE. However, text reviews are available only for 2016 MY and 2016 YE, since the organization did not use text reviews as part of their employee evaluation process before 2016 MY. Table 1 lists some salient summary statistics to highlight the scale of the dataset and the underlying performance evaluation process. Table 2 lists the top-level departments and their employee counts. Table 3 shows the distribution of employees across various geographical regions.

Figure 2a traces the average performance of employees through past evaluation periods. The plots are stratified by employees' most recent performance score vertically and employee gender horizontally. In each graph, the three curves indicate the average performances of employees who have been at the organization for different number of evaluation periods. As a result, each curve has a different number of datapoints. For example, the green curves

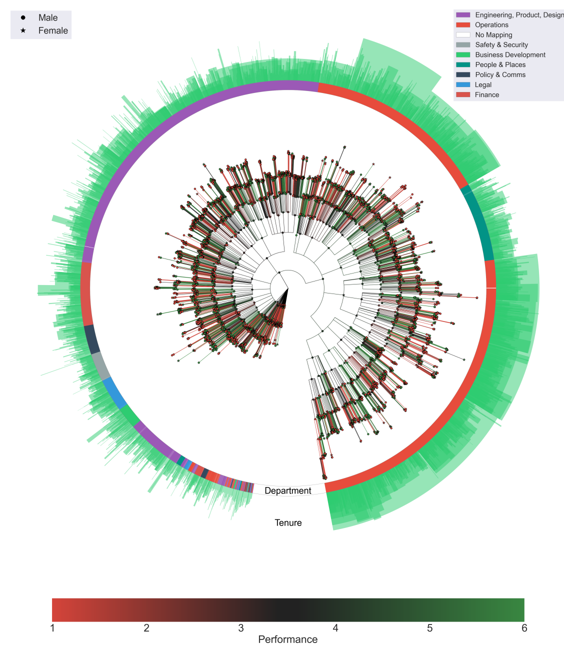


Figure 1: Infographic of employee-manager hierarchy in the last evaluation period. The shape of the nodes in the graph indicates the gender. The color of a node is used to indicate the average performance score of employees in the subtree on a scale of 1-6. The inner ring indicates the department of the employees in a color-coded fashion. See legend for details of the departments. The outer ring indicates the tenure of the employees through the height of the ring.

include employees who have been at the organization since the 2015 YE evaluation period. Hence, it has two data values - one for 2015 YE and another for 2016 MY. Figure 2b shows corresponding barplots for the number of employees associated with each of the curves in figure 2a. We see that there are a fair number of employees who have joined since more recent evaluation periods and therefore did not have ratings in earlier evaluation half-years.

During an evaluation, each employee writes peer reviews for around five other employees, mostly for people in their own team. They also write one review evaluating themselves, one review evaluating their immediate manager in the corporate hierarchy, and one review for each employee that directly reports to them. Thus, there are four types of reviews: self, peer, manager, and direct report.

Each review is structured into a list of positive and negative feedbacks i.e. pros and cons. A reviewer can list upto 3 pros (P1, P2, P3), and upto three cons (C1, C2, C3) in each review. Thus, each positive/negative feedback can be identified using `<feedback-id, reviewer-id, reviewee-id, P1/P2/P3/C1/C2/C3, {words}>`, and a review consists of all six feedbacks that share the same `<(reviewer-id, reviewee-id)>` pair. Such a structured review helps an employee understand their best professional qualities as well as provides constructive criticism which the employee can use to improve their work and professional conduct. With around 8, 000

Count Statistic	2016 MY	2016 YE
Unique Reviewers	6302	8272
Unique Reviewees	6080	8330
Unique Employees	6451	8460
Unique Reviews	37921	54294
Self Reviews	4969	7593
Peer Reviews	21993	31648
Manager Reviews	5266	7342
Direct Report Reviews	5693	7711
Non-empty feedbacks (P1)	37775	54078
Non-empty feedbacks (P2)	37497	53005
Non-empty feedbacks (P3)	36273	49078
Non-empty feedbacks (C1)	37009	52340
Non-empty feedbacks (C2)	34198	44833
Non-empty feedbacks (C3)	30456	34795

Table 1: Summary statistics of the employee performance reviews dataset.

Department	Count
Ops and Marketing	3597
Engineering and Product	2814
People and Places	521
Finance	503
Policy and Comms	229
Legal	209
Safety and Security	199
Business Development	156
No Mapping	700

Table 2: Number of employees in each of the top-level departments of the studied tech organization.

Region	Count
North America	5275
Asia-Pacific	1482
Europe, Middle East, and Africa	1040
Latin America	442

Table 3: Count of employees in major geographical regions.

employees, the dataset contains nearly half a million individual pieces of feedback.

A manager also provides a performance evaluation to each of their immediate reports on an ordinal scale of 1-6. This quantitative evaluation is linked to an employee's promotions and bonuses, and is available for all evaluation periods.

Since the data was anonymized before being handed over to us, employee names were not revealed during any analysis reported in this paper. In order to perform gender-specific analysis, we inferred an employee's gender using the occurrence of gendered pronouns in their reviews. If the pronouns `she` or `her` appeared more times in all peer reviews associated with an employee than `he` or `him`, the employee was considered female, and male otherwise.

3 PERFORMANCE EVALUATION DISPARITIES

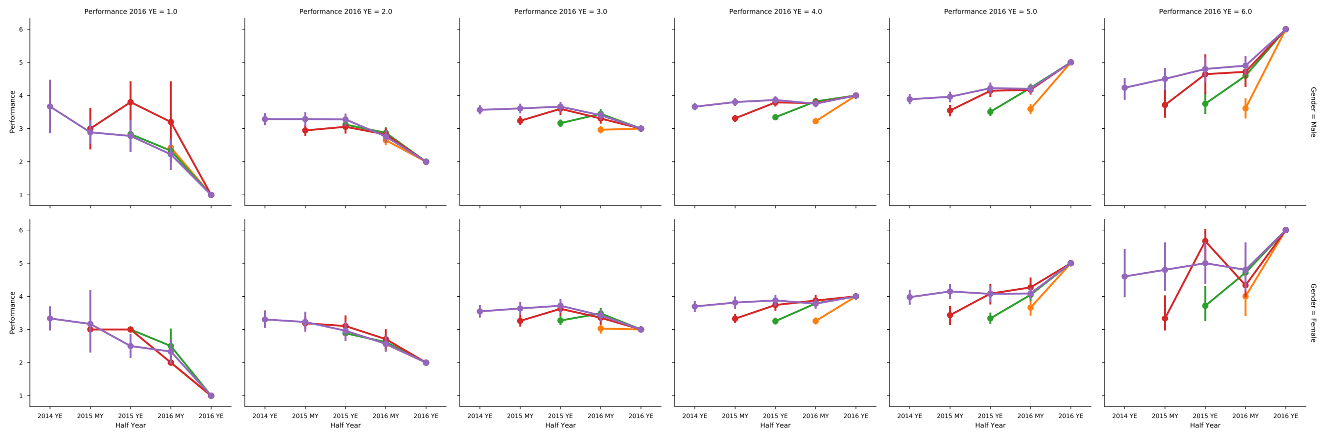
Since employees are assigned performance evaluation ratings by their managers on an ordinal scale of 1-6, we can compare the empirical distribution of performance ratings between the two genders. Figure 3 shows the gender ratio stratified by performance evaluation period and performance level i.e. the ratio of females to males who received a particular performance score in a particular evaluation period. The horizontal orange lines indicate the overall gender ratio in an evaluation period irrespective of the performance level. The ratio of females to males who received a performance score of 6 can be seen deteriorating since 2015 YE.

The ratio of females to males receiving performance score of 1 increased in 2015 YE and 2016 MY before dipping again in 2016 YE. Data revealed that this change from 2016 MY to 2016 YE was not caused by the company firing the lowest performing females (rating=1) from 2016 MY. Most of them had transitioned to higher ratings of 2 and 3. It is possible that some managers noticed the disproportionate number of poor ratings being given to female employees in 2016 MY and countered this disparity by eliciting and recognizing their contributions. Employee performance evaluations on an ordinal scale require subjective judgment calls from the managers in contextualizing and quantifying complex non-ordinal employee contributions to the team and company. Such evaluations at technology companies are even more informal and unstructured than at some of the more established corporate sectors. As such, simple quantitative realizations of gender disparities can be very effective in helping managers counter any unconscious biases when assigning performance ratings.

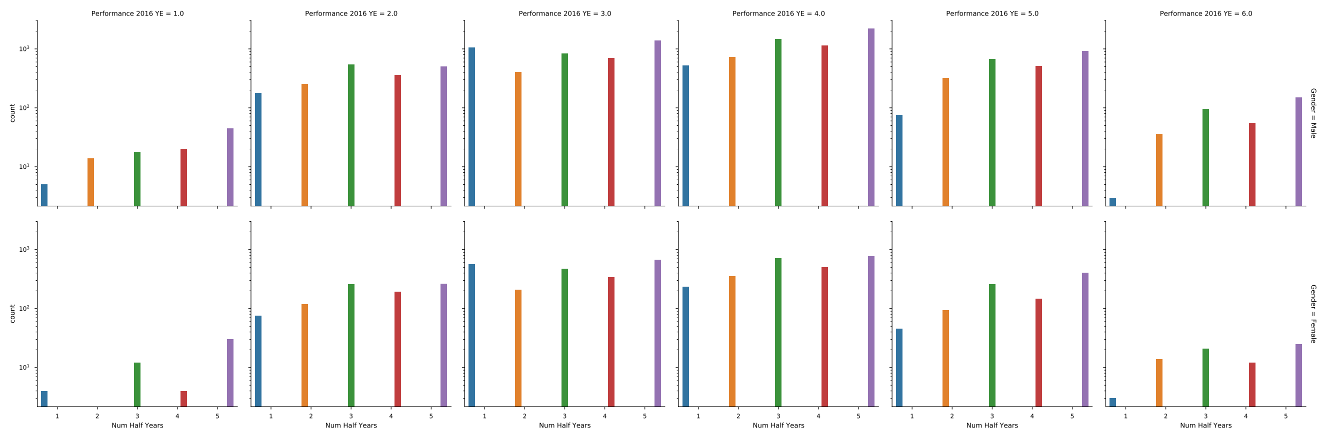
Half Year	Count	Males	Females	Test Statistic	p-value
2014 YE	1520	1063	457	1.859886	0.868173
2015 MY	2577	1788	789	1.699509	0.888962
2015 YE	4557	3097	1460	10.308789	0.066944
2016 MY	6064	4080	1984	15.162205	0.009691
2016 YE	8342	5690	2652	25.915173	0.000093

Table 4: Chi-squared test for independence of performance and gender in each of the five evaluation periods.

In order to determine whether the visual disparities depicted in figure 3 are statistically significant, we treat the six-dimensional count vector of male employees receiving performance ratings 1-6 as a draw from a multinomial distribution P_M. The corresponding count vector for female employees is considered drawn from another multinomial distribution P_F. We then perform a chi-squared two-sample test to determine if the two count vectors could have been sampled from the same multinomial distribution i.e. the distributions P_M and P_F are the same. Results of the hypothesis test for each of the evaluation periods is shown in table 4. In 2014 YE and 2015 MY, the difference between the performance count vectors for male and female employees was not statistically significant. However, from 2015 YE to 2016 YE, evidence for statistical significance



(a) Average performance in previous evaluation periods stratified by most recent performance and gender.



(b) Number of employees in previous evaluation periods stratified by most recent performance and gender.

Figure 2: Stratifying performance and employee counts in previous evaluation periods based on most recent performance and gender. Each curve in (a) shows employees who have been at the company for different number of evaluation periods. Similarly color-coded bars in (b) show the number of employees by how long in terms of half-years they have been at the company.

of disparity between the performance count vectors for the two genders rapidly increases.

3.1 An Ordinal Probit Analysis

In order to further analyze the gender disparities in performance while controlling for observed factors that may explain the disparities, we perform an ordinal probit regression since the output variable `performance` is an ordinal one. We use performance data from the last evaluation period 2016 YE as the output variable since this period has the most employees as seen in table 4.

The ordered probit model can be described using a latent variable model. If y is an ordinal response variable taking values from $\{0, 1, 2, \ldots, J-1\}$, there is an associated latent variable y^* such that $y^* = \mathbf{x}\beta + \epsilon$. Here $\mathbf{x}$ denotes the covariates that y is being regressed on. Threshold parameters $\alpha_1 < \alpha_2 < \ldots < \alpha_{J-1}$ determine the response y as follows:

$$
\begin{aligned}
y &= 0 && \text{if } y^* \le \alpha_1 \\
y &= 1 && \text{if } \alpha_1 < y^* \le \alpha_2 \\
& && \vdots \\
y &= J-1 && \text{if } y^* > \alpha_{J-1}.
\end{aligned}
$$

After setting $\alpha_0 = -\infty$ and $\alpha_J = \infty$ we have:

$$
\begin{aligned}
P(y = j) &= P\left(\epsilon \le \alpha_{j+1} - \mathbf{x}\beta\right) - P\left(\epsilon \le \alpha_j - \mathbf{x}\beta\right) \\
&= F\left(\alpha_{j+1} - \mathbf{x}\beta\right) - F\left(\alpha_j - \mathbf{x}\beta\right)
\end{aligned}
$$

where F is the cumulative distribution function for ϵ. Ordered probit assumes that the error term ϵ has a normal distribution.

Results of the ordinal probit regression are shown in table 5. We have controlled for the following factors: employee department, their seniority indicating by an ordinal level assigned to each employee by the company, the geographical location where each

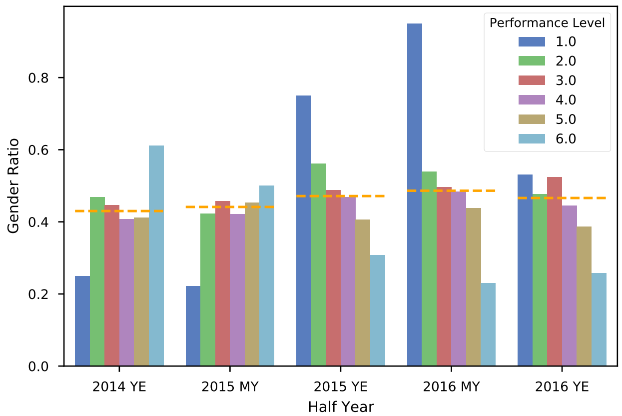


Figure 3: Ratio of female to male employees stratified first by performance evaluation period and second by performance level within each period. The horizontal orange bars show the overall gender ratio for each evaluation period.

```
Ordered Probit Regression
Log-Likelihood: -10263.22
No. Iterations: 6
McFadden's R2: 0.005015283
AIC: 20566.44
```

	Estimate	Std. error	t value	Pr(>|t|)
Gender	0.08	0.03	2.95	0.00
deptEng. & Product	-0.08	0.09	-0.89	0.37
deptFinance	-0.03	0.10	-0.33	0.74
deptLegal	-0.07	0.12	-0.55	0.58
deptNo Mapping	-1.26	0.39	-3.23	0.00
deptOps & Marketing	0.08	0.10	0.80	0.42
deptPeople & Places	0.01	0.11	0.10	0.92
deptPolicy & Comms	0.06	0.12	0.50	0.62
deptSafety & Security	-0.14	0.12	-1.14	0.25
employee-level	0.04	0.01	3.59	0.00
regionAPAC	-0.45	0.13	-3.45	0.00
regionEMEA	0.03	0.05	0.66	0.51
regionLatAm	0.05	0.06	0.79	0.43
regionNorth America	0.07	0.04	1.66	0.10
num_direct_reports	0.01	0.00	4.56	0.00
Threshold (1->2)	-2.26	0.12	-18.83	0.00
Threshold (2->3)	-0.89	0.11	-8.07	0.00
Threshold (3->4)	0.13	0.11	1.23	0.22
Threshold (4->5)	1.28	0.11	11.62	0.00
Threshold (5->6)	2.43	0.11	21.10	0.00

Table 5: Ordinal Probit Regression Summary

employee works, and the number of direct reports of an employee. The effect of gender on performance is statistically significant. An analysis of marginal effects of gender on performance provided in table 6 indicates that $Pr(Y = 1, 2, 3)$ decreases by 3.22% when employee gender is male i.e. 3.22% of the risk of obtaining a performance rating less than or equal to 3 is attributable to the employee being female, even after controlling for a substantial number of factors.

	Marg. Eff.	Std. error	t value	Pr(>|t|)
Pr(Y=1 | Male)	-0.0014	0.0005	-2.6607	0.0078
Pr(Y=2 | Male)	-0.0156	0.0054	-2.8995	0.0037
Pr(Y=3 | Male)	-0.0152	0.0051	-2.9908	0.0028
Pr(Y=4 | Male)	0.0131	0.0046	2.8623	0.0042
Pr(Y=5 | Male)	0.0161	0.0054	2.9807	0.0029
Pr(Y=6 | Male)	0.0030	0.0010	2.9730	0.0029

Table 6: Ordinal Probit Marginal Effects of Gender

While the regression indicates that gender disparities exist in the performance evaluations, it is not a conclusive evidence of gender discrimination. It is possible that the company aggressively hired more females to improve their diversity metrics. This can cause the quality of hires to differ between genders. Without an objective metric of work productivity and quality, it is difficult to conclude that the gender disparities revealed in our analysis constitute gender discrimination.

4 SENTIMENT AND PSYCHOLINGUISTIC DISPARITIES

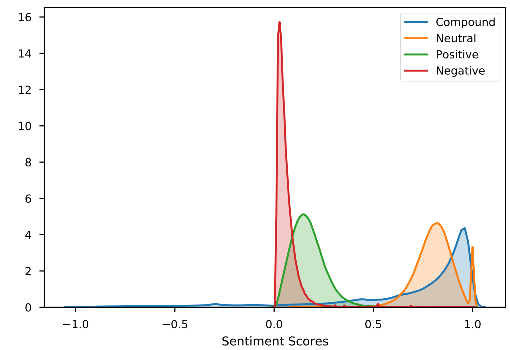
While the analysis discussed in the previous section uses structured data about the employees and their performance ratings, we have not yet analyzed the rich trove of unstructured textual peer feedback which can often reveal substantial gender disparities [17]. We describe two types of text analysis we performed on the peer reviews.

4.1 Sentiment Analysis

Sentiment analysis is a common natural language processing technique for analyzing whether a piece of text is positive or negative and the extent of this sentiment. Using the `nltk.sentiment` Python package with the in-built rule-based *Vader* sentiment scorer, we scored each feedback on four related dimensions. The *neutral*, *positive*, and *negative* scores lie between 0 and 1 and add up to 1. The *compound* score provides a single scalar representation of the overall sentiment contained in a piece of text and lies between -1 and 1.

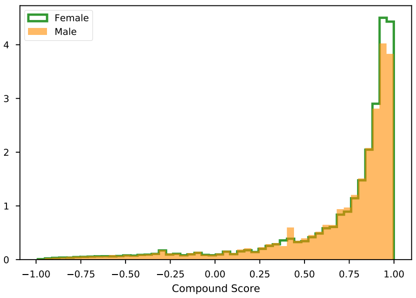
Kernel density estimates of each of the four different sentiment scores on peer feedbacks are shown in figure 4. Compound scores are rarely below 0, indicating that feedbacks are usually positive. This is evidenced in the corpus where the cons are mostly framed as constructive criticism and suggestions for improvement. This phenomenon also shows in the distribution of positive and negative scores. The distribution of negative sentiment is highly skewed toward zero, whereas the distribution of positive sentiment is spread out further away from 0.

In figure 5, we plot the distribution of each of the four scores split by the gender of the employee who received the feedback. The compound scores for female employees are more skewed toward 1 than for male employees. On the contrary, the negative and neutral sentiment scores for feedbacks of female employees are slightly skewed toward 0 compared to the male employees. The sentiment

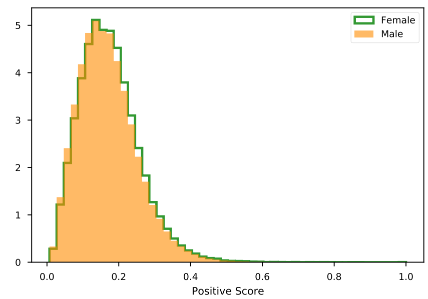


(a) Kernel density plots of sentiment scores for feedback texts.

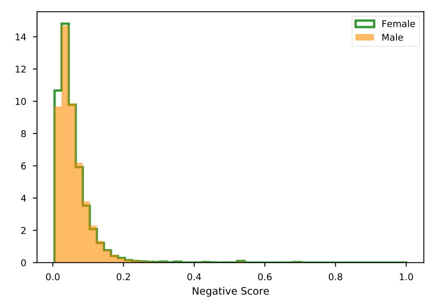
Figure 4: Density plots of sentiment scores for feedback texts using `nltk.sentiment` Python package.



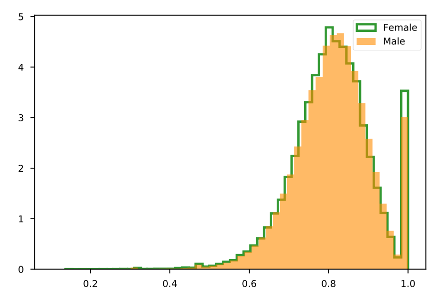
(a) Normalized histograms of compound sentiment scores split by gender.



(b) Normalized histograms of positive sentiment scores split by gender.



(c) Normalized histograms of negative sentiment scores split by gender.



(d) Normalized histograms of neutral sentiment scores split by gender.

Figure 5: Density plots of sentiment scores for feedback texts using `nltk.sentiment` Python package.

distributions for the two genders are comparable with a slight skew toward positive sentiment for female employee feedback.

4.2 Psycholinguistic Analysis

Compared to sentiment analysis, psycholinguistic analysis is a deeper multi-dimensional analysis of text to identify the subtle psychological connotations hidden within a piece of text. It can reveal if a text describing a person implies they are aggressive, agreeable, independent, disciplined, etc. LIWC is a widely used lexicon-based psycholinguistic scoring mechanism for text analysis [18]. Receptiviti [19] is a proprietary psycholinguistic scoring system developed by one of the authors of [18] and provided as an online API. We scored each peer review feedback using both LIWC and Receptiviti systems. LIWC provides only raw scores whereas Receptiviti provides both raw and percentile scores.

4.2.1 LIWC Scores. LIWC (Linguistic Inquiry and Word Count) is a popular psycholinguistic scoring mechanism. Its lexicon includes over 86% of the words used commonly in written and spoken English. A detailed list of LIWC scores can be found in [18]. We choose a subset of the LIWC scores for our regression analysis, because highly correlated inputs lead to estimation difficulties in standard OLS/Logit regressions. The scores considered in our analysis include the following:

- liwc-anxiety: indicates how anxious the person being described is.
- liwc-body: indicates a focus on body image in the feedback
- liwc-cognitive-processing: indicates a discussion of cognitive processing abilities
- liwc-negate: indicates that negations are used frequently in the feedback
- liwc-quant: indicates that the feedback is focused on quantitative mentions of performance
- liwc-risk: indicates the risk-taking behavior of the employee
- liwc-wordcount: word count of the feedback

From figure 6, these scores are not severely correlated and can be used together in a regression. We use them as inputs and employee gender as the output in a logistic regression to ascertain which of these scores are associated with gender disparities. From table 7, all these LIWC scores are associated with employee gender in a statistically significant manner. The results indicate that female peers were generally perceived as being more anxious and less risk-taking than male coworkers. Their reviews focused more physical appearance, had lesser quantitative discussions of their work contributions, and were phrased with more negations than reviews for male workers.

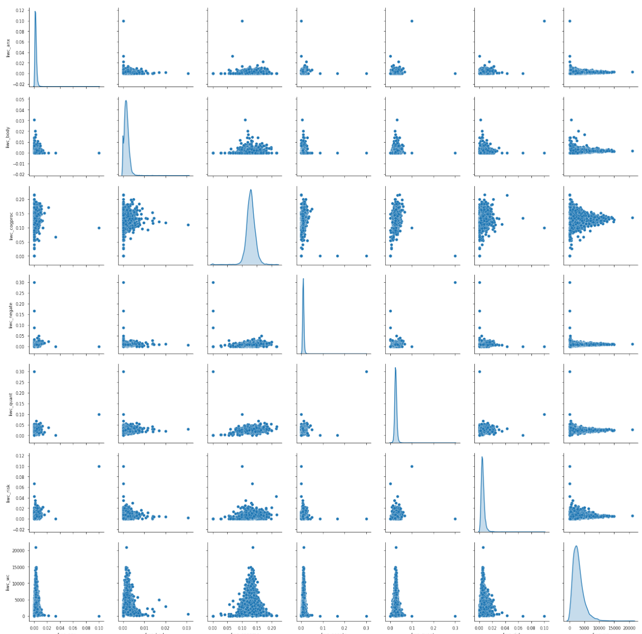


Figure 6: Pairwise plots of psycholinguistic LIWC features used in Logit regression in the analysis.

4.2.2 Receptiviti Scores. Since Receptiviti scores are available both in the raw format as well as percentiles, we resort to using percentiles in our analysis because the scores are more evenly spread out as percentiles. Similar to LIWC, we do not wish to use all Receptiviti scores in the regression analysis because many of them are correlated with each other leading to model estimation difficulty. Instead, we choose a subset of scores which are not heavily correlated with each other. The following Receptiviti scores chosen for the analysis are self-explanatory:

- receptiviti-aggressive
- receptiviti-ambitious
- receptiviti-disciplined
- receptiviti-independent
- receptiviti-insecure
- receptiviti-social-skills
- receptiviti-work-oriented

These scores are shown in figure 7, and are not particularly strongly correlated with each other. Similar to LIWC, we regress

Dep. Variable:	Gender	**No. Observations:**	6050
Model:	Logit	**Df Residuals:**	6042
Method:	MLE	**Df Model:**	7
Converged:	True	**Log-Likelihood:**	-3777.3
Pseudo R-squ.:	0.02334	**LL-Null:**	-3867.6

	coef	**std err**	**z**	**P>|z|**
const	-0.9095	0.262	-3.475	0.001
liwc-anxiety	-128.6584	34.922	-3.684	0.000
liwc-body	-97.8537	21.941	-4.460	0.000
liwc-cognitive-processing	7.8361	2.027	3.865	0.000
liwc-negate	-19.0293	7.211	-2.639	0.008
liwc-quant	23.9727	5.795	4.137	0.000
liwc-risk	75.9337	15.231	4.985	0.000
liwc-wordcount	7.336e-05	1.54e-05	4.760	0.000

Table 7: Logistic regression results with LIWC features

employee gender on the receptiviti percentile scores to understand associations between the receptiviti scores and gender disparities. All the input scores are associated with gender in a statistically significant manner, according to the regression summary in 8. According to the results, male workers were considered more aggressive, ambitious, independent and work-oriented than female workers. The latter were deemed to be more disciplined but also more insecure and lacking social skills required in the workplace.

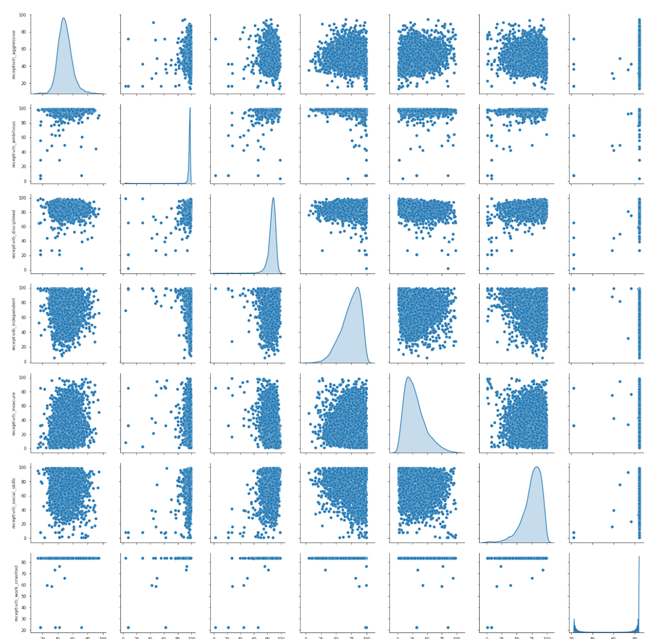


Figure 7: Pairwise plots of proprietary Receptiviti features used in Logit regression in the analysis.

Dep. Variable:	Gender	**No. Observations:**	6050
Model:	Logit	**Df Residuals:**	6042
Method:	MLE	**Df Model:**	7
Converged:	True	**Log-Likelihood:**	-3660.9
Pseudo R-squ.:	0.05343	**LL-Null:**	-3867.6

	coef	**std err**	**z**	**P>|z|**
const	-8.0985	4.762	-1.701	0.089
receptiviti-aggressive	0.9149	0.341	2.687	0.007
receptiviti-ambitious	4.2273	1.310	3.228	0.001
receptiviti-disciplined	-4.6886	0.720	-6.515	0.000
receptiviti-independent	0.7484	0.195	3.832	0.000
receptiviti-insecure	-2.2481	0.181	-12.424	0.000
receptiviti-social-skills	-2.5661	0.228	-11.265	0.000
receptiviti-work-oriented	12.3951	5.788	2.141	0.032

Table 8: Logistic regression results with Receptiviti features

5 DISCUSSION

Gender disparity in the performance evaluation process does not necessarily imply discrimination. Such disparities can arise due to many reasons including employees self-sorting themselves into specific job roles, disparities in the quality of employees hired, etc. However, it is necessary to monitor for such disparities because they can be symptomatic of deeper organizational issues that may need correction such as a toxic workplace culture or unconscious biases in the performance evaluation process. Sentiment and psycholinguistic analysis of peer reviews provides another glimpse into the subconscious biases that employees may be unaware of. We saw persistent gender stereotypes emerge from the text corpus of peer reviews. Many of these stereotypes can be actively detrimental in the professional success of female employees and executives.

One of the advantages of our analysis is that it is simple, not heavily customized to the dataset, and relies on readily available software. Such an analysis could therefore easily be carried out by data scientists embedded in human resource departments to detect any emerging disparities before they become moral or legal liabilities for the executives of the organization.

A key limitation of our research is that we do not have an objective measure of productivity for the employees. If such a metric exists, it can serve as a very good control for the causal inference of gender discrimination. An example of such a metric in a non-tech domain is the volume of sales accomplished by a salesperson. However, measuring productivity in technology organizations is difficult since contributions are multifaceted and not measured by a single number such as lines of code or number of meetings. Lack of such a metric prevents us from inferring if the disparities discussed here are due to discrimination or other causes.

6 CONCLUSION

In this paper, we presented our findings from analyzing the qualitative and quantitative data associated with the half-yearly employee performance evaluations at a large technology firm. We found evidence of gender disparities in managerial performance evaluations. Since these evaluations directly affect promotions and rewards such as bonuses, persistent disparities can enforce workplace irregularities such as *glass ceilings* and *sticky floors* [16]. A psycholinguistic analysis of peer reviews further revealed many common stereotypes about female workers that can make it difficult for them to have access to the same level of opportunities and support as their male colleagues. We hope that our analysis will serve to promote the strategic analysis of large-scale human resource data in technology organizations to detect and correct gender disparities and prevent such disparities from coloring the development of technologies designed for widespread use.

REFERENCES

[1] Jana Kasperkevic. Sexism valley: 60% of women in silicon valley experience harassment. https://www.theguardian.com/technology/2016/jan/12/silicon-valley-women-harassment-gender-discrimination, 2016. Accessed: 2018-08-01.

[2] Erin A Cech and Mary Blair-Loy. Perceiving glass ceilings? meritocratic versus structural explanations of gender inequality among women in science and technology. *Social Problems*, 57(3):371–397, 2010.

[3] Clare O'Connor. Google sued for gender discrimination by female former employees. https://www.forbes.com/sites/clareoconnor/2017/09/14/google-sued-for-gender-discrimination-by-female-former-employees/#688a948050c9, 2017. Accessed: 2018-08-01.

[4] Madeline Farber. Qualcomm is paying almost $20 million after claims it didn't pay women equally. http://fortune.com/2016/07/27/qualcomm-settlement-equal-pay/, 2016. Accessed: 2018-08-01.

[5] Jack Linshi. Twitter faces gender discrimination lawsuit by former female engineer. http://time.com/3753458/twitter-gender-lawsuit/, 2015. Accessed: 2018-08-01.

[6] Heather Kelly. Facebook gets sued for gender discrimination. https://money.cnn.com/2015/03/18/technology/facebook-discrimination-suit/index.html, 2015. Accessed: 2018-08-01.

[7] Dan Levine. Microsoft women filed 238 discrimination and harassment complaints. https://reut.rs/2Nb0iSe, 2018. Accessed: 2018-08-01.

[8] Leo Kay and Jose Carnevali. US Department of Labor sues Oracle America Inc. for Discriminatory Employment Practices. https://www.dol.gov/newsroom/releases/ofccp/ofccp20170118-0, 2017. Accessed: 2018-08-01.

[9] Sam Levin. Google accused of 'extreme' gender pay discrimination by US labor department. https://www.theguardian.com/technology/2017/apr/07/google-pay-disparities-women-labor-department-lawsuit, 2017. Accessed: 2018-08-01.

[10] Susan Fowler. Reflecting on one very, very strange year at uber. https://www.susanjfowler.com/blog/2017/2/19/reflecting-on-one-very-strange-year-at-uber, 2017. Accessed: 2018-08-01.

[11] Melanie Ehrenkranz. Leaked apple emails reveal employees' complaints about sexist, toxic work environment. https://bit.ly/2wiTPx8, 2016. Accessed: 2018-08-01.

[12] Deepa Seetharaman. Facebook's female engineers claim gender bias. https://www.wsj.com/articles/facebooks-female-engineers-claim-gender-bias-1493737116, 2017. Accessed: 2018-08-01.

[13] Ariane Hegewisch and Heidi Hartmann. Occupational segregation and the gender wage gap: A job half done. 2014.

[14] Roberto M Fernandez and Santiago Campero. Gender sorting and the glass ceiling in high-tech firms. *ILR Review*, 70(1):73–104, 2017.

[15] An-Ju R Tai and Randi L Sims. The perception of the glass ceiling in high technology companies. *Journal of Leadership & Organizational Studies*, 12(1):16–23, 2005.

[16] Margaret Yap and Alison M Konrad. Gender and racial differentials in promotions: Is there a sticky floor, a mid-level bottleneck, or a glass ceiling? *Relations Industrielles/Industrial Relations*, 64(4):593–619, 2009.

[17] David G Smith, Judith E Rosenstein, Margaret C Nikolov, and Darby A Chaney. The power of language: Gender, status, and agency in performance evaluations. *Sex Roles*, pages 1–13, 2018.

[18] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[19] About Receptiviti: The science of psychology and language. https://www.receptiviti.com/about/. Accessed: 2018-08-01.