# Analysis of Employee Reviews from Top Tech Companies

**Bhargav Reddy Godala**
Princeton University
bgodala@princeton.edu

**Greg Chan**
Princeton University
gc14@princeton.edu

**Teague Tomesh**
Princeton University
ttomesh@princeton.edu

**Ziyang Xu**
Princeton University
ziyangx@princeton.edu

## Abstract

In the modern world, technology is king. Many aspects of our daily lives, including the news and entertainment we consume, the devices and applications we use, and even many of the social interactions we have are products of one of the six major tech companies: Facebook, Apple, Amazon, Microsoft, Google, and Netflix. As such, it would be advantageous for us computer science students to understand the advantages and disadvantages of possibly working at each of these companies. In this paper we take a closer look at these companies by leveraging a dataset composed of employee reviews on Glassdoor.com [1]. We use natural language processing techniques to automatically extract the concerns of the employees from their reviews, and we then use these concerns in combination with knowledge of employee satisfaction to characterize each of the companies. Finally, we propose an algorithm to match a prospective employee with the company which best addresses his/her concerns. This work can help future employees to choose his/her best matching companies as well as help tech companies address the concerns of their employees.

## 1 Introduction

In this paper, we provide an analysis of a dataset consisting of approximately 67,000 reviews obtained from employees of Facebook, Apple, Amazon, Microsoft, Google, and Netflix [2]. We begin with a survey of some related work on similar topics in Section 2, and then move to Section 3 where we address the particular questions we are interested in and the motivation behind the analysis performed. In Section 4 we describe our methods for first, processing the raw data into a machine-analyzable format, and second, building a model to investigate the concerns of different employees. Additionally, we present our matching algorithm in Section 5 and discuss how this model was used to characterize each of the companies and how this can be used to recommend the best fit for potential employees choosing between these companies. Finally, we discuss in Section 6 and Section 7 their impact within the framework of the questions we are interested in and several conclusions.

## 2 Related Work

The dataset that we use in this analysis has been downloaded from `Kaggle` over 5,000 times by people all over the world. After downloading the data, some of those users return to `Kaggle` to publish the results of their analyses for other users to see and build upon [3]. A flavor of those analyses includes: “Classification of Satisfied/Unsatisfied Employees”, “Which corporation is worth working for?", and "Natural Language Processing (NLP) Project". We read through many of these analyses to get a sense for what had already been done and what was left to do. It soon became apparent that most of the posted results were simple explorations of the dataset and generally did not include machine learning models applied to the data. Those who did take a machine learning approach focused on the classification of satisfied or unsatisfied employees. This was done by generating a new binary feature which was given a value of 1 for satisfied employees and a value of 0 for those who were unsatisfied. An employee was deemed satisfied if all of the scores they gave to the five metrics: `work balance`, `culture values`, `opportunities`, `benefits`, and `management` were above their means. Classification of the satisfied/unsatisfied feature was then performed using linear classifier (with stochastic gradient descent training), random forest, and support vector machine models. They found that both the random forest model and support vector machine performed extremely well while the linear classifier struggled to classify every employee. The majority of these works also did not take into account the text responses in their classification, something which we use as the main source of input to our models.

## 3 Motivation

After familiarizing ourselves with the previous work on this dataset we decided we would apply various machine learning techniques to the data in an effort to uncover any hidden latent structure. Rather than focusing on employee satisfaction, we wondered whether the reviews given by the employees could be used to characterize the companies they work for. In addition, we investigate how the application of Natural Language Processing (NLP) and classification models can be used to predict where an employee is currently working and which company would be the best fit for them given their concerns and preferences. Therefore, this analysis aims to explore and provide answers to the following four questions:

1. Are latent structures present in the employee reviews?
2. Can these structures be used to characterize these tech companies?
3. Is it possible to predict where an employee is/was working at based on his/her reviews?
4. Can we construct an algorithm to produce recommendations based on matching employees to those companies that best address their concerns?

To answer the first two questions, we first need a way to obtain the latent structures in the data, which we can later analyze either manually or through unsupervised learning to find patterns that would explain an employee's satisfaction or dissatisfaction with a company. Then, from the same inputs of a new employee review, can we use the same latent structures across all companies to find the one that fits the employee of a new review? Lastly, given a prospective employee's concerns, how do we rank the companies that fit his/her concerns without a model explicitly trained with employee concerns?

Based on these motivations, we did several explorations and proposed a matching algorithm to generate a recommendation report for a user. The structure of this project is shown in Figure.1.

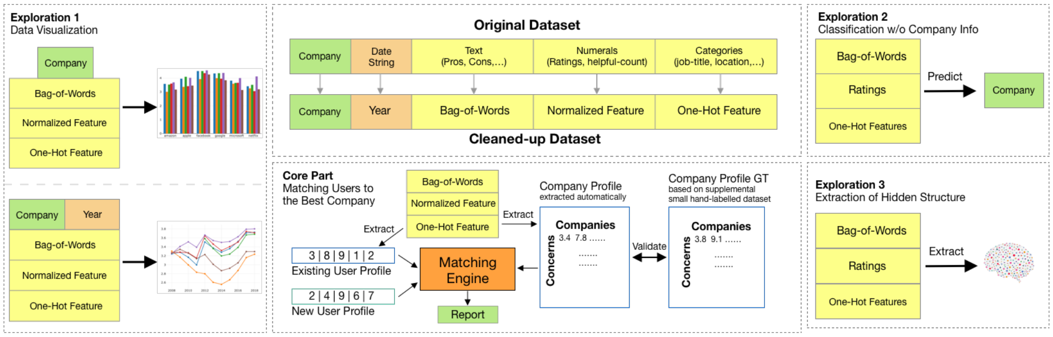


Figure 1: Main Structure of This Project

# 4 Exploration

The start of this project includes some data exploration. As the goal of the project is discovering hidden structures of the dataset to help users find the best matching companies, we tried various methods to gain insights that could guide us to build the model. As the first exploration, we performed data visualization on the rating information with and without the axis of time to look for distributions or trends of ratings. Then we used supervised machine learning models to predict the company using only the text responses to discover any correlation between the comments and the companies. Unsupervised models including, but not limited to, *PCA[4], LDA, and K-Means* are also used to extract the hidden structure of the features.

Our data exploration gives us many exciting insights and leads to the conclusion that we need to define some explicit concerns and map the latent variables to the concerns. It is the most efficient way to reason about the performance of our matching algorithm and generate a report with non-trivial and helpful facts instead of just one single company name. Also, this can aid us in providing multiple companies that fit a user well, with reasons for choosing each.

## 4.1 Data Processing

Our data set comes from a collection of public Glassdoor.com employee reviews of several top technology companies, available at kaggle.com [2]. It includes approximately 67,000 employee reviews from Google, Amazon, Facebook, Apple, Microsoft, and Netflix. From this dataset, we are able to extract 16 raw features, which include categorical, numerical, and text responses. Note that some of these features contain invalid inputs, which we drop as we have more than enough samples to satisfy our needs.

We left the numerical values as they were and performed one-hot encoding on the categorical value. One challenge for categorical data is that the feature `job-title` consisted of combinations of "Current Employee - " or "Former Employee - " and the actual job position. Furthermore, the majority of the responses to this question were "Anonymous Employee" (>70%). As such, we decided to use only the "current" or "former" information to simplify the encoding of this feature. We also did not use the "location" feature because we believe it is irrelevant to the concerns we extract later.

For the text responses (`pros, cons, summary, advice-to-mgmt`), we performed some *natural language processing (NLP)* including bag-of-words[5] and polarity analysis to analyze hot words and the sentiment of a comment. We applied variance threshold to the bag-of-words to reduce the vocabulary. This required us to manually classify some of the reviews as positive or negative for our training set; we performed this classification on 200 randomly chosen reviews for a fairer and hopefully more normalized training set.

## 4.2 Exploration 1: Data Visualization

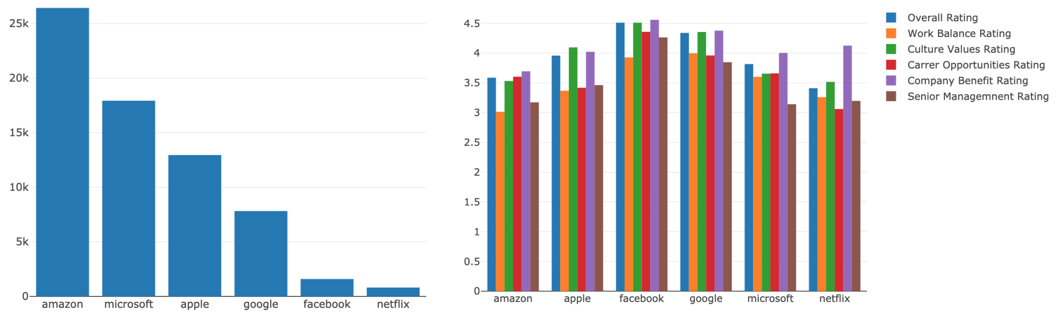
Data visualization helps us understand data in an intuitive way. As the first step, we applied multiple data visualization techniques to the dataset, leading to several interesting facts about the dataset. However, the information we can extract from these numbers (ratings) is not enough. It is an excellent way to see the whole picture, but it cannot show us the hidden structure such as clusters that may exist. Thus, we need to incorporate in the later explorations the text features as well as the ratings.

### 4.2.1 Basic Visualization

After filtering out invalid or unanswered responses, we can naively say that Facebook is the best company to work for and Netflix is the worst, which can be seen in Figure 2b. However, in Figure 2a, we also note that the number of reviews for Netflix and Facebook were the lowest amongst all, so we should keep in mind to not give the average ratings of every company the same weight.

### 4.2.2 Time Series Visualization

By adding the axis of time into the visualization, we can observe trends of each company over time, leading to some interesting correlations between certain events occurring and the employees'

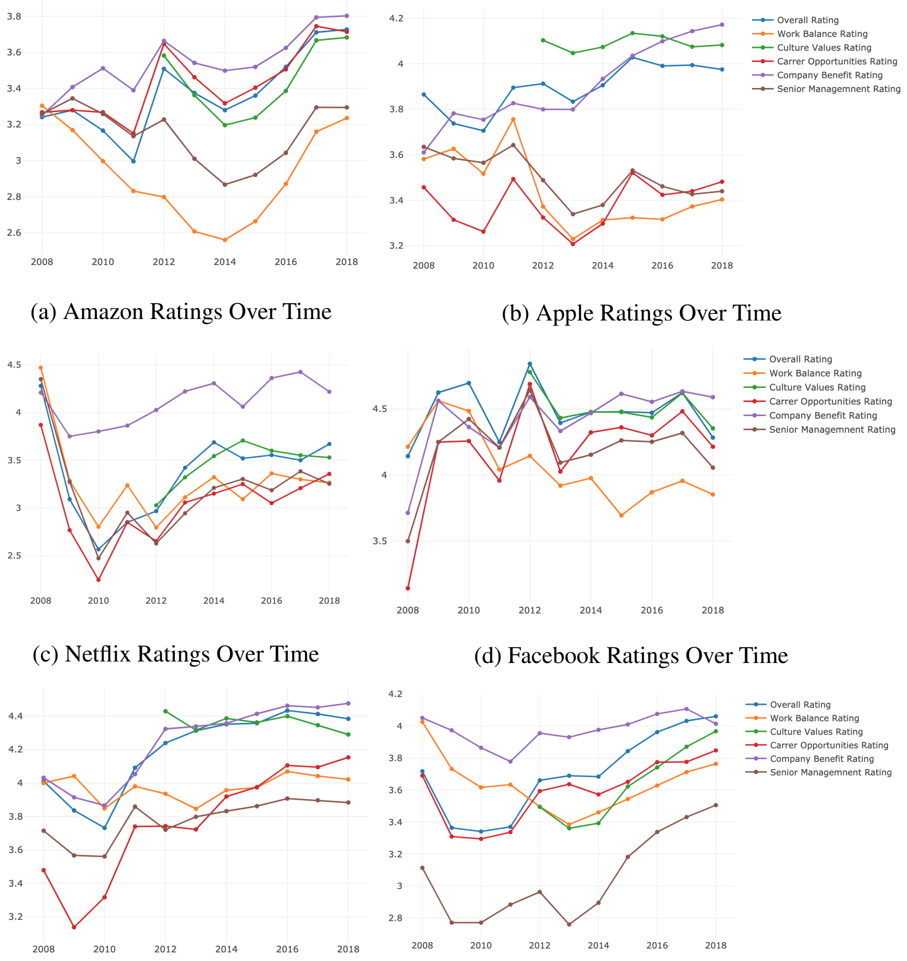


(a) Review Counts for Each Company

(b) Rating Comparison

Figure 2: Basic Visualization for the Dataset

opinions. For example, Steve Jobs passed away in October of 2011, and we see a sharp drop in multiple ratings, as shown in Figure 3b; Microsoft announced a new CEO in February of 2014, and we see an apparent rise in senior management rating, as shown in Figure 3f.



(a) Amazon Ratings Over Time

(b) Apple Ratings Over Time

(c) Netflix Ratings Over Time

(d) Facebook Ratings Over Time

(e) Google Ratings Over Time

(f) Microsoft Ratings Over Time

Figure 3: Time Series Analysis

### 4.3 Exploration 2: Finding Latent Structure

Above, we discuss how the latent information in the dataset will be useful for further analysis. Here, we show the results of this latent analysis on the employee review dataset.

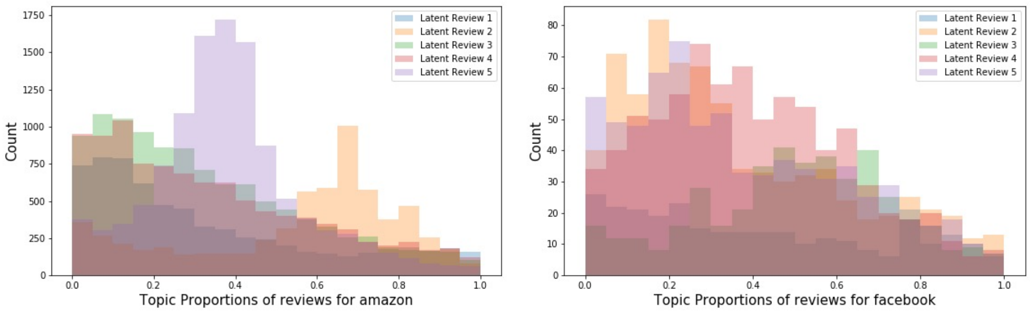
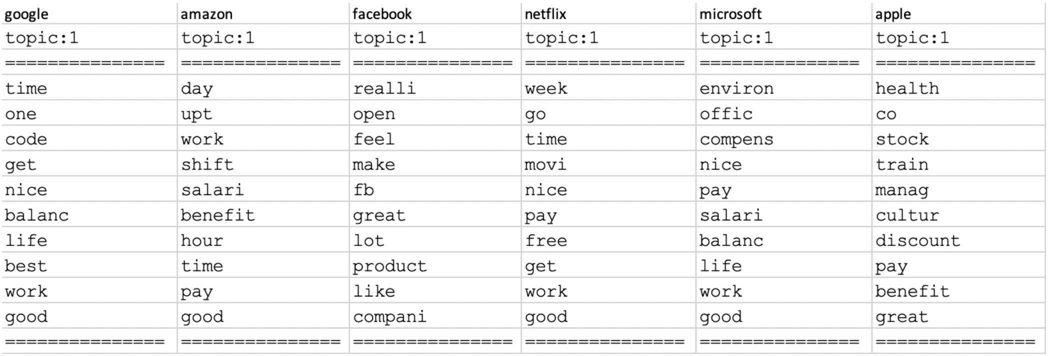


Figure 4: Example of company composition for Amazon and Apple after using PCA with 5 components. Our final analysis leveraged PCA using 100 components. Similar profiles for the remaining companies are shown in Appendix A, Figure 10.

First, we perform PCA on the entirety of the raw reviews, and are left with a matrix describing a reduced set of "latent reviews." We also obtain a matrix which tells us the degree to which each raw review is composed of members of the set of latent reviews (essentially a weighting of each latent review). By grouping the PCA data based on the company we can produce a characterization (shown in Figure 4) of each company based on the latent users who tend to work there. Using these constructed profiles, we can investigate similarities and differences across companies.

As another method of characterization, we perform topic modeling on our bag-of-words representation on a per-company basis. We group employee reviews based on which company they work for and then analyze each of these groups independently. Figure 5 shows a slice of a single topic for each of the companies. This shows that each company can be characterized based on the words that their employees use to describe them. As an example, the employees at Apple seem to enjoy the health benefits and company culture and so we can characterize Apple as a company which excels at addressing these concerns. Additional topic slices can be found in Appendix A, Figure 11.



| google | amazon | facebook | netflix | microsoft | apple |
|---|---|---|---|---|---|
| topic:1 | topic:1 | topic:1 | topic:1 | topic:1 | topic:1 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| time | day | realli | week | environ | health |
| one | upt | open | go | offic | co |
| code | work | feel | time | compens | stock |
| get | shift | make | movi | nice | train |
| nice | salari | fb | nice | pay | manag |
| balanc | benefit | great | pay | salari | cultur |
| life | hour | lot | free | balanc | discount |
| best | time | product | get | life | pay |
| work | pay | like | work | work | benefit |
| good | good | compani | good | good | great |
| =============== | =============== | =============== | =============== | =============== | =============== |

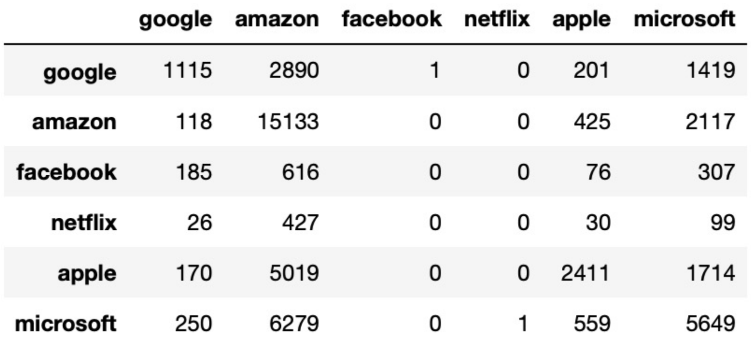
Figure 5: A slice of a single latent topic from each company.

### 4.4 Exploration 3: Employer Classification

One of the questions we hoped to answer in this analysis was the possibility of predicting where an employee is currently working based on the review he/she gave. The answer to this question is also interesting concerning the other questions presented in Section 3 because of what it would reveal about the structures uncovered in PCA. If we found we were able to make predictions with high accuracy, then the latent features extracted by PCA must be strongly related to the information concerning an employee's current company. On the other hand, if we were unable to do so then we may assume that PCA found latent structure within the reviews that were not particularly correlated with this information. If the latent structure turns out to be representative of the concerns and preferences expressed by the employees in their reviews, then we would have a solid basis for constructing a recommendation algorithm.

We chose to approach this problem from a classification perspective and applied both unsupervised and supervised techniques. For the unsupervised classification, we first represented each company in the latent space using the data shown in Figure 4. Since each of the companies shows a different latent user profile, we hypothesized that each company would occupy a different region of the latent space and employees that reside near a region could be classified as currently working at its corresponding company. The transformation from Figure 4 to the latent space was accomplished by forming a unit vector for each company where the component of the vector along each latent feature axis is proportional to that latent feature's presence in reviews given by employees working for that company. We then tried two different methods for classification. First, we computed the dot product between an employee's latent vector and each company latent vector; the company whose dot product is the largest is chosen as that user's current company. Second, we used the $l^2$-norm in a similar fashion, where the company that produced the smallest $l^2$-norm with the employee was chosen. In practice, these methods performed quite poorly, with prediction accuracies of **15%** and **14%**, respectively.

Our supervised classification was performed using a Random Forest Classifier (RFC) trained on the latent review data generated by PCA. Using an RFC with 100 estimators and a max depth of 15, we trained and tested on an 80/20 split of the data. We achieved training accuracy of **92%** and testing accuracy of **51.45%**.



| | google | amazon | facebook | netflix | apple | microsoft |
|---|---|---|---|---|---|---|
| **google** | 1115 | 2890 | 1 | 0 | 201 | 1419 |
| **amazon** | 118 | 15133 | 0 | 0 | 425 | 2117 |
| **facebook** | 185 | 616 | 0 | 0 | 76 | 307 |
| **netflix** | 26 | 427 | 0 | 0 | 30 | 99 |
| **apple** | 170 | 5019 | 0 | 0 | 2411 | 1714 |
| **microsoft** | 250 | 6279 | 0 | 1 | 559 | 5649 |

Figure 6: Confusion matrix for RFC predictions. Columns are the predicted class and rows are the true class.

Figure 6 shows the confusion matrix for the predictions obtained using Random Forest Classification. It is evident from this data that Amazon has the highest prediction accuracy among all. This is not unexpected, since Figure 2a clearly shows that reviews from Amazon employees dominate the dataset. The number of reviews available for Facebook and Netflix are almost negligible in comparison and hence most of the time these are misclassified.

## 5 The Matching Algorithm

One of our goals for this project is to find a way to predict suitable matches between employee and company. In this section, we will introduce our method for matching an existing or new user with the best-matching company.

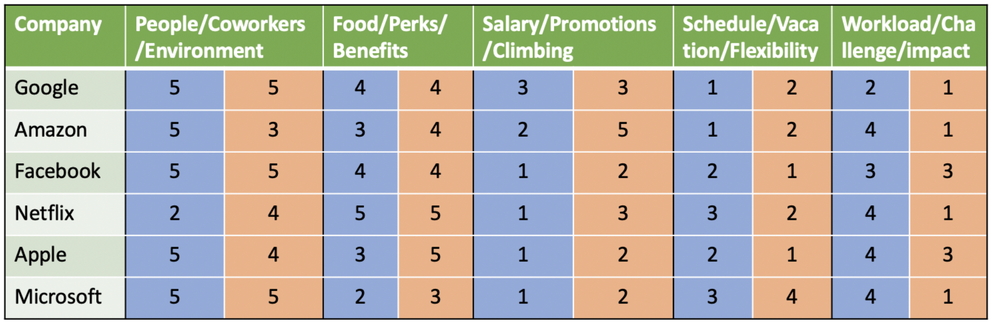
As we discovered from previous explorations, there are limitations for both supervised learning and unsupervised learning. To better reason about the matching result, we proposed the concept of "concerns" and used hand-labeling to generate a ground-truth dataset with 200 users.

With the concept of concerns in mind, we extracted the companies' profiles, i.e., tables consisting of the score of each concern for each company, from both hand-labeled data and the whole dataset. The profile table for the entire dataset was obtained by PCA analysis on the Bag of Words of the text data coupled with polarity (sentiment) analysis using the TextBlob Python module [6]. The topics from PCA were used to get the top vocabulary that matches with the set of concerns we built manually.

Once we obtained the company profiles, we can use the same method to extract concerns for an existing user from his/her comments and ratings. We can use the user's concerns and the companies' profiles to generate a report for the user. If a new user wants to use this system, he/she can directly rank his/her top concerns from the list and use the generated report to choose the next company wisely.

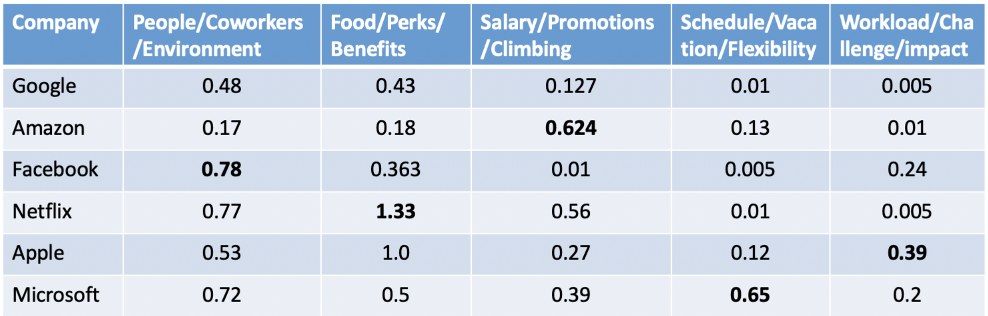
### 5.1 Ranking Concerns

We manually grouped all the concerns we found into five categories: *"People/Coworkers/Environment"*, *"Food/Perks/Benefits"*, *"Salary/Promotions/Climbing"*, *"Schedule/Vacation/Flexibility"*, and *"Workload/Challenge/Impact"*. We then hand-labeled 200 randomly chosen users, giving each concern a score from one to five with five being "very concerned" and one being "not concerned", based on the content of their reviews. We also generated an extra field – "satisfied". The "satisfied" field is important when we try to generate the ground truth profiles for each company. We assume that when a user is satisfied with the company, the concerns he/she has have been met, and vice versa. Figure 7 shows the rankings of each concern for each company with the manually labeled data in the blue columns. The numbers in the red columns also give rankings for each concern, but these rankings were generated using the PCA component scores shown in Figure 8. In general, the rankings given by the manually labelled data match the rankings produced through PCA. This is an interesting result, and it shows that the latent topics extracted by PCA can be used as the basis for a matching algorithm.



| Company | People/Coworkers/Environment | | Food/Perks/Benefits | | Salary/Promotions/Climbing | | Schedule/Vacation/Flexibility | | Workload/Challenge/impact | |
|---|---|---|---|---|---|---|---|---|---|---|
| Google | 5 | 5 | 4 | 4 | 3 | 3 | 1 | 2 | 2 | 1 |
| Amazon | 5 | 3 | 3 | 4 | 2 | 5 | 1 | 2 | 4 | 1 |
| Facebook | 5 | 5 | 4 | 4 | 1 | 2 | 2 | 1 | 3 | 3 |
| Netflix | 2 | 4 | 5 | 5 | 1 | 3 | 3 | 2 | 4 | 1 |
| Apple | 5 | 4 | 3 | 5 | 1 | 2 | 2 | 1 | 4 | 3 |
| Microsoft | 5 | 5 | 2 | 3 | 1 | 2 | 3 | 4 | 4 | 1 |

Figure 7: Concern rankings based on manually labeled (Blue) and PCA generated (Red) data. A ranking of 5 means this is the concern best met by this company, while the least met concern is given a ranking of 1. The rankings generated by PCA closely mirror those that were manually generated.



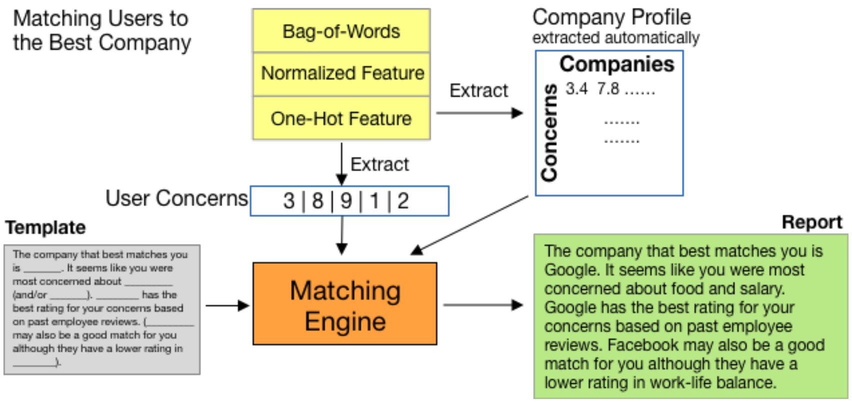
| Company | People/Coworkers/Environment | Food/Perks/Benefits | Salary/Promotions/Climbing | Schedule/Vacation/Flexibility | Workload/Challenge/impact |
|---|---|---|---|---|---|
| Google | 0.48 | 0.43 | 0.127 | 0.01 | 0.005 |
| Amazon | 0.17 | 0.18 | **0.624** | 0.13 | 0.01 |
| Facebook | **0.78** | 0.363 | 0.01 | 0.005 | 0.24 |
| Netflix | 0.77 | **1.33** | 0.56 | 0.01 | 0.005 |
| Apple | 0.53 | 1.0 | 0.27 | 0.12 | **0.39** |
| Microsoft | 0.72 | 0.5 | 0.39 | **0.65** | 0.2 |

Figure 8: Scores of the top concerns obtained using PCA analysis per company.

### 5.2 Matching Engine

Our analysis above shows that it is possible to build an algorithm which can match a potential employee to the company which best meets their specific concerns and preferences. We are able to extract latent topics from the dataset of employee concerns that can be used to characterize each company on the basis of the concerns they satisfy (Figure 7).

Since a user may want the detailed information of the recommendation, our analysis also enables a template-based report generating process tailored to each individual. In Figure 9, we put together a template and the possible result for one user.



The company that best matches you is ________. It seems like you were most concerned about ________ (and/or ________). ________ has the best rating for your concerns based on past employee reviews. (________ may also be a good match for you although they have a lower rating in ________).

The company that best matches you is Google. It seems like you were most concerned about food and salary. Google has the best rating for your concerns based on past employee reviews. Facebook may also be a good match for you although they have a lower rating in work-life balance.

Figure 9: Generating report from template, user concerns, and company profile.

## 6 Discussion

We can see in Sections 5.1 and 5.2 that the two tables generated manually and automatically are pretty close in regards to the score each company was assigned to each concern. This shows that although not perfect, our method for automatically extracting concerns from reviews provides us with a relatively good characterization.

To deal with Facebook's and Netflix's sparseness of data, we used an alternate approach. The motivation for the new approach was that we want to predict the top three choices for a user. The total number of permutations for picking three out of six companies with the order mattering is 120. With this intuition that the total number of possibilities is relatively small, we used a K-Means clustering unsupervised learning model on top of the PCA extracted features with 120 clusters, and achieved close to 30% accuracy on the training set. Figure 12 shows the confusion matrix. Netflix and Facebook achieve better accuracy, but at the cost of overall accuracy. In the interest of time and space, we have not explored extending this approach further. We do have one interesting observation though, namely as the number of clusters increases so does the prediction accuracy.

We were able to predict an employee's current employer with a maximum accuracy of 51%. Random classification would only be expected to obtain an accuracy of 16.5% in this case, so our results are better than random guessing at the very least. However, it is interesting that we were unable to reach higher accuracy, and we view this as an indication that the latent structure found by PCA may serve as the basis for the construction of an employee-employer matching algorithm.

## 7 Conclusion

In this work, we've seen how machine learning techniques can be leveraged to better understand the companies that run many aspects of our lives. One of the challenges we had dealing with the dataset is that the reviews available per company is very much skewed (Figure 2a). The contribution of reviews to Netflix and Facebook are not that significant. Although we were able to extract latent concerns from the data, assigning a strong score to these companies based on a new user's concern is very challenging.

With a larger dataset with more information, this work can be extended beyond just predicting employee satisfaction, which company an employee works for, and predicting which company best fits a potential employee. We could, for example, predict how a company's stock will do if we had employee opinions on the quality of the latest product release – the possibilities are endless with modern machine learning techniques.

## References

[1] Glassdoor, Inc. https://www.glassdoor.com/index.html.

[2] Sunga, P. (2018; December). Google Amazon and more Employee Reviews, Version 2. Retrieved 04/22/2019 from https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews.

[3] Previous Work. https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews/kernels

[4] Arthur Gonsales. An approach to choosing the number of components in a principal componentanalysis (pca), 2018. Retrieved April 2019 from https://towardsdatascience.com/an-approach-to-choosing-the-number-of-components-in-a-principal-component-analysis-pca-3b9f3d6e73fe.

[5] Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical-framework.International Journal of Machine Learning and Cybernetics, 1(1-4):4352, 2010.

[6] TextBlob Python Module. https://textblob.readthedocs.io/en/dev/

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret-tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Per-rot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:28252830, 2011.

# A Appendix A

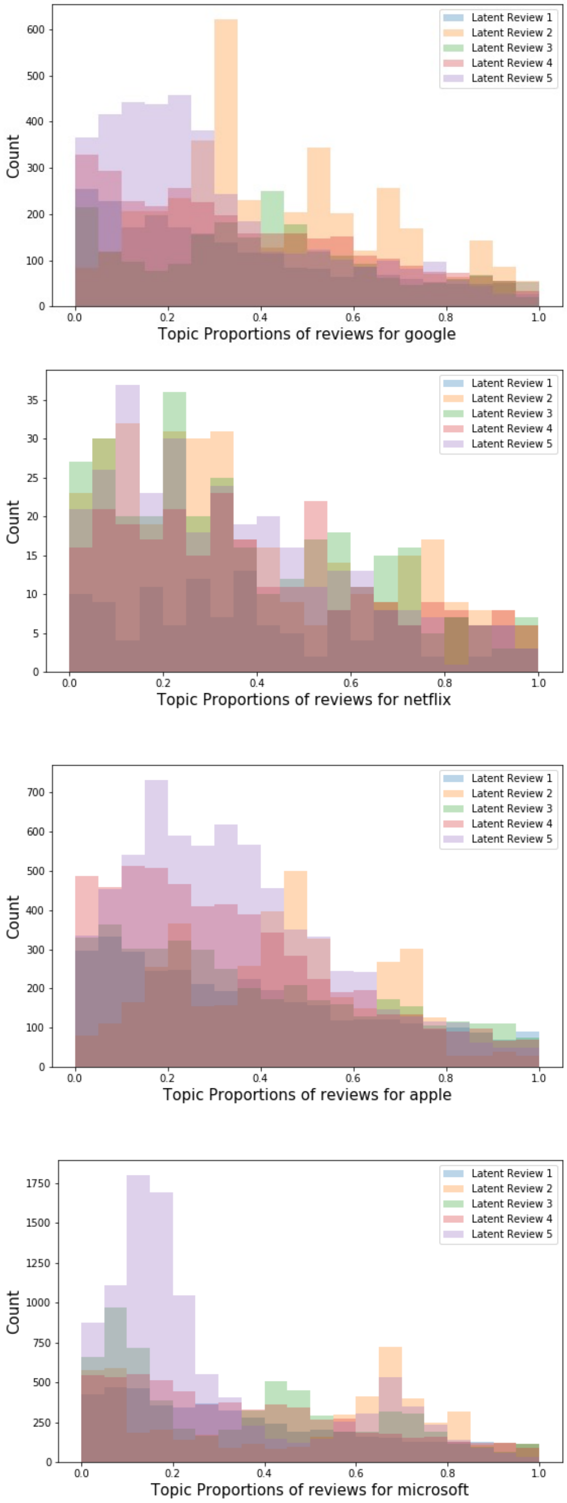
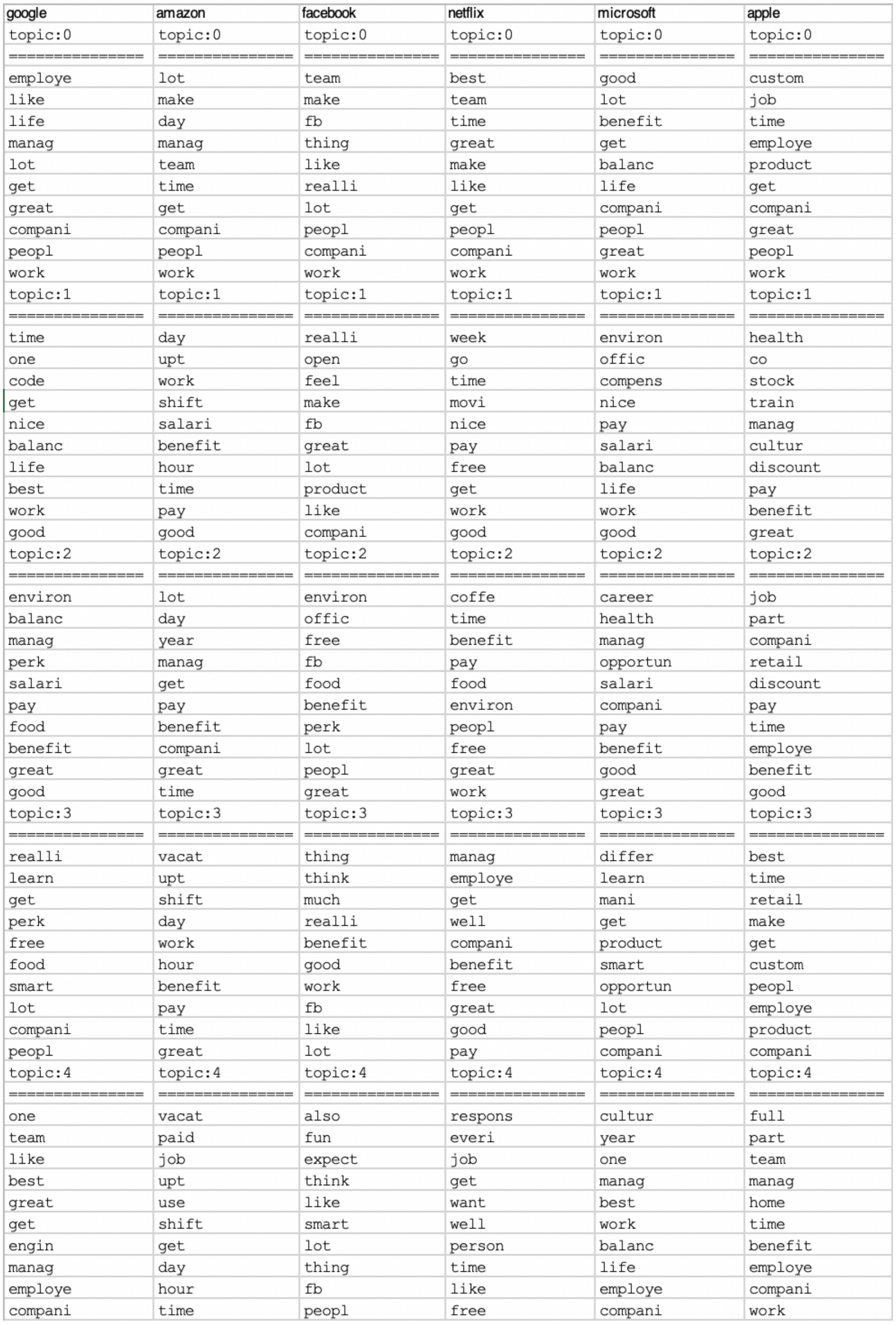
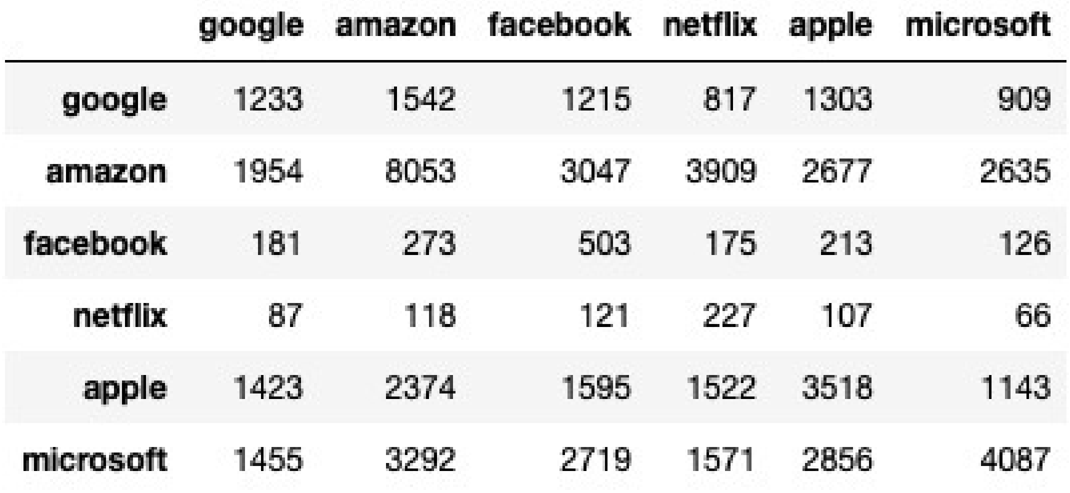


Figure 10: PCA proportions for other companies.



| google | amazon | facebook | netflix | microsoft | apple |
|---|---|---|---|---|---|
| topic:0 | topic:0 | topic:0 | topic:0 | topic:0 | topic:0 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| employe | lot | team | best | good | custom |
| like | make | make | team | lot | job |
| life | day | fb | time | benefit | time |
| manag | manag | thing | great | get | employe |
| lot | team | like | make | balanc | product |
| get | time | realli | like | life | get |
| great | get | lot | get | compani | compani |
| compani | compani | peopl | peopl | peopl | great |
| peopl | peopl | compani | compani | great | peopl |
| work | work | work | work | work | work |
| topic:1 | topic:1 | topic:1 | topic:1 | topic:1 | topic:1 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| time | day | realli | week | environ | health |
| one | upt | open | go | offic | co |
| code | work | feel | time | compens | stock |
| get | shift | make | movi | nice | train |
| nice | salari | fb | nice | pay | manag |
| balanc | benefit | great | pay | salari | cultur |
| life | hour | lot | free | balanc | discount |
| best | time | product | get | life | pay |
| work | pay | like | work | work | benefit |
| good | good | compani | good | good | great |
| topic:2 | topic:2 | topic:2 | topic:2 | topic:2 | topic:2 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| environ | lot | environ | coffe | career | job |
| balanc | day | offic | time | health | part |
| manag | year | free | benefit | manag | compani |
| perk | manag | fb | pay | opportun | retail |
| salari | get | food | food | salari | discount |
| pay | pay | benefit | environ | compani | pay |
| food | benefit | perk | peopl | pay | time |
| benefit | compani | lot | free | benefit | employe |
| great | great | peopl | great | good | benefit |
| good | time | great | work | great | good |
| topic:3 | topic:3 | topic:3 | topic:3 | topic:3 | topic:3 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| realli | vacat | thing | manag | differ | best |
| learn | upt | think | employe | learn | time |
| get | shift | much | get | mani | retail |
| perk | day | realli | well | get | make |
| free | work | benefit | compani | product | get |
| food | hour | good | benefit | smart | custom |
| smart | benefit | work | free | opportun | peopl |
| lot | pay | fb | great | lot | employe |
| compani | time | like | good | peopl | product |
| peopl | great | lot | pay | compani | compani |
| topic:4 | topic:4 | topic:4 | topic:4 | topic:4 | topic:4 |
| =============== | =============== | =============== | =============== | =============== | =============== |
| one | vacat | also | respons | cultur | full |
| team | paid | fun | everi | year | part |
| like | job | expect | job | one | team |
| best | upt | think | get | manag | manag |
| great | use | like | want | best | home |
| get | shift | smart | well | work | time |
| engin | get | lot | person | balanc | benefit |
| manag | day | thing | time | life | employe |
| employe | hour | fb | like | employe | compani |
| compani | time | peopl | free | compani | work |

Figure 11: Latent topics generated by BoW analysis on employee reviews from each individual company.



| | google | amazon | facebook | netflix | apple | microsoft |
|---|---|---|---|---|---|---|
| google | 1233 | 1542 | 1215 | 817 | 1303 | 909 |
| amazon | 1954 | 8053 | 3047 | 3909 | 2677 | 2635 |
| facebook | 181 | 273 | 503 | 175 | 213 | 126 |
| netflix | 87 | 118 | 121 | 227 | 107 | 66 |
| apple | 1423 | 2374 | 1595 | 1522 | 3518 | 1143 |
| microsoft | 1455 | 3292 | 2719 | 1571 | 2856 | 4087 |

Figure 12: K-means cluster (120 clusters mapped to company using the highest contribution of a company) confusion matrix