# Glassdoor Company Review Analysis

------A case study on a large company's review data

Project notebook:
https://github.com/teresanan/capstone2-employee-review-analysis

Teresa Nan
July 2020

# Table of Contents

# Executive Summary

This project is an analysis on a large company's Glassdoor review using Natural Language Processing.The goal is to help employers gain real insights on their employee engagement and create a structure for future similar tasks. The analysis aimed to answer these questions: what employees like and dislike about their company? Has the company's reputation gotten better or worse over the years? What are the keywords that employees are talking about this company? What can this company do to improve employee engagement?

We analyzed the numerical rating data, performed sentiment analysis on the text data, conducted topic modeling to get the key words from the comment, and then dove into the groups who are unhappy so we could figure out the root cause.

From this study, we discovered these insights. In general, this employer has a very high rating. The employees love this company because it offers good benefits, generous profit sharing, and employees enjoy the work-life balance. However, people in their call center seem to have the most complaints and this negatively affects the company's overall rating. Call center employee's complaints are focused on low pay, long hours. Other employees who submit negative reviews choose to be anonymous and their complaints are about senior management, red tape of a large company and frequent organizational changes.

This analysis has built a standard workflow for similar Glassdoor company review tasks. Although the dataset is about one company's data, the analysis process and techniques can be applied to any company's Glassdoor review data.

# 1. Introduction

## 1.1 Background

People Analytics is a newly emerging field.Organizations are embracing this new trend and focusing on using data to build retention models. “While most People Analytics teams seem to study that particular issue, they quickly find that employee retention is really a surrogate for dozens or hundreds of other issues: management, leadership, work environment, rewards, job fit, and more”[1]. Without knowing the real underlying issues, retention models can’t be meaningful.

For those who conducted employee surveys to figure out the root cause, often faced with the problem that data doest represent the true picture. It is understandable that incumbent employees tend to hide their complaints if the survey is conducted within the organization.

However, employee feedback submitted to outside platforms (i.e. Glassdoor) anonymously generally is more genuine than feedback shared within the organization.Therefore, this project will use a dataset that comes from Glassdoor. This analysis will help employers get a full picture of their employee engagement and enable them to make decisions based on data.

This analysis aims to build a workflow for Glassdoor company review tasks. Although the data in this project is about one company, the techniques can be applied to any mid-large sized company that has enough Glassdoor reviews.

## 1.2 Method Overview

- **Data**:The data is webscrapped from Glassdoor with Glassdoor’s written permission.To protect the company’s privacy, we have replaced the company name with ‘Anonymous’.

- **Process:** First, data cleaning and feature engineering by using string manipulation.Then we conducted exploratory data analysis, sentiment analysis and finally topic modeling.

- **Tools:**The programming is done in Python. For web scraping, we used requests and Beautiful Soup. The Python libraries used are NLTK, TextBlob, Gensim and Spacy. For data analysis and visualization, we used Numpy, Pandas, matplotlib and seaborn.

---

[1] https://medium.com/@josh_bersin/people-analytics-market-tsunami-ten-things-you-need-to-know-7d1a78db559e

# 2. Data Wrangling

Ideally, we should scrape the data using Glassdoor API, but it is only available to companies and Glassdoor partners. Therefore we had to write more codes to scrape the data.The file consists of 6675 company reviews(rows).The features(column) include both numerical data (ratings from 1 to 5) and text data(comments).

The features are:

- Reviewer information: Comment Datetime,Author Title, Author Years and Location
- Text columns: Summary, Recommendation, Pro, Con
- Numerical columns:Overall Rating, Career Opportunities,Senior Management, Culture & Values, Compensation and Benefits, Work/Life Balance

## 2.1 Data Cleaning

This dataset is unstructured and untidy, it requires extensive cleaning and preprocessing.

**Drop duplicates**

When glancing at the data, we see some employees submitted twice, so we need to drop the duplicate records. That leaves us with 6386 records.

**Handle missing values**

There are 8 columns with missing values. These columns include both numerical rating columns and categorical string columns. We could have filled the missing ratings with each column's median value, however, in order to make this analysis more accurate, our strategy is to leave NaN numerical rating as is, and fill the missing categorical information with 'Unknown', so we can differentiate it from the originally existing data.

## 2.2 Feature Engineering

In the original dataset, some columns include multiple features which is against the tidy data principle. Fortunately, some missing information can be found in other text columns. We can extract text from these columns to create new features.

**Extract information of Current/Former Employee from column Author Year.**

There is a column "Author Title" which has "Current Employee" or "Former Employee" information, but there are more than 2000 null values. The other column "Author Years" column has no null value and the text contains relevant information about current/former employee flags, notice the string in column "Author Years" always starts with either "I have been working" or "I worked at". Hence, we can use "Author Year" to extract employee type data.

**Extract employee Tenure from column Author Years.**

Below are some sample data:

- "I have been working at Anonymous Investments full-time for more than a year"
- "I have been working at Anonymous Investments full-time for more than 3 years"
- "I worked at Anonymous Investments full-time for more than 5 years"

Notice there are 3 different patterns.There could be a number before year(s)(i.e. 3 years) or a letter before year (i.e. 'a year') or 'more than/less than'. First, we replace 'a year' as '1 year', then find all 'more than/less than'. If there is 'more than', we add 0.5 year to the number of years, if there is 'less than', we minus 0.5 year from the number of years. If no tenure is specified, we set it to NaN.

**Extract information of Full-time/Part-time Employee from column Author Year.**

For full-time / part-time flags, we can get that information from the "Author Years" column as well. If this information is not specified, we default it to NaN.

**Extract Recommended, Positive Outlook, Approves of CEO from columnRecommendation.**

Since the column 'Recommendation' consists of 3 pieces of information, we would need to separate them and make each feature as one column.

**Extract employee location, job title from column Author Title.**

As a national company, employees in different states can have different engagement, therefore location information is meaningful. However, we don't need to get the specific city, only the state information is sufficient. As of job titles, most job titles have this string format "Current Employee - Financial Associate", so we need to extract the job title after the "-". However, there are more than 1000 records missing this information, so we fill in the information with "Unknown Title". After that, we remove "senior" and "principal" from the job titles to get fewer job categories.

More details about the data cleaning and feature engineering process can be found in this Jupyter Notebook. After these steps, the dataset is clean and each column only represents one feature. Next, we will explore the data with visualization.

# 3. Exploratory Data Analysis

As mentioned earlier, this dataset includes both numerical rating columns and string columns. In this initial exploratory data analysis phase, we only focus on visualizing the numerical data. For string columns, we will visualize them in sentiment analysis and topic modeling steps.

**The main findings from exploratory analysis are as following:**

- This company's overall employee satisfaction has been increasing since 2008.
- 60% of employees would recommend this company to others.
- Employees in KS and WI have the highest employee satisfaction and employees in NE are the unhappiest.
- Part-time employees have higher satisfaction than full-time employees.
- Former employees give relatively lower ratings than current employees.
- Financial Associate group has the lowest satisfaction and Relationship Managers give the highest ratings.
- In general, this company has a very high employee satisfaction rate.

## 3.1 Overall Rating

From below line chat, we can see the company's overall average rating has been increasing since 2008 with small dips in 2011, 2013 and 2017. This indicates this company has been improving.

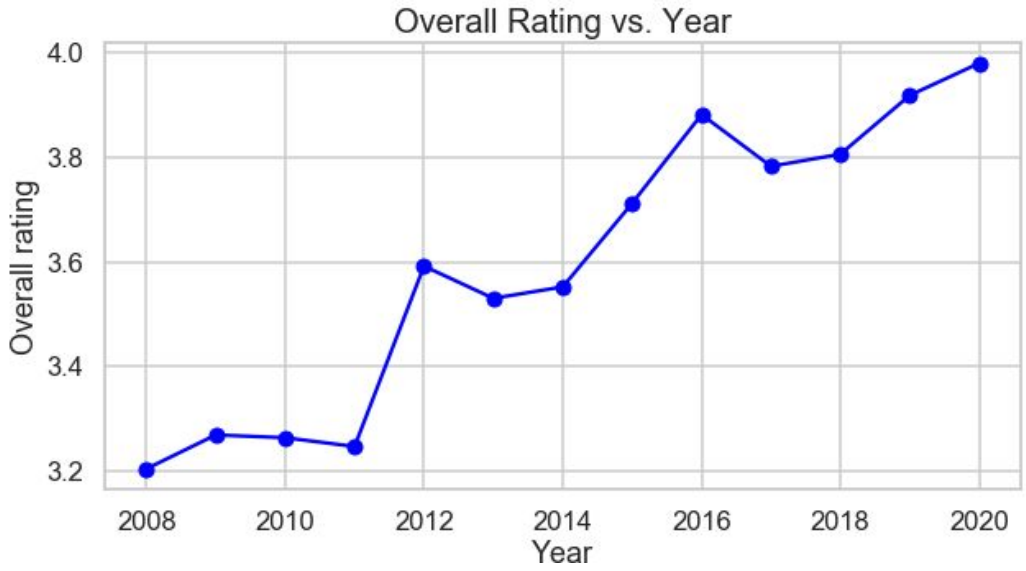



*Figure 1: Company overall average rating over 12 years*

## 3.2 Recommendation

Among all reviewers, 60% of them would recommend this company to others, 20% do not recommend it, another 20% do not answer this question.

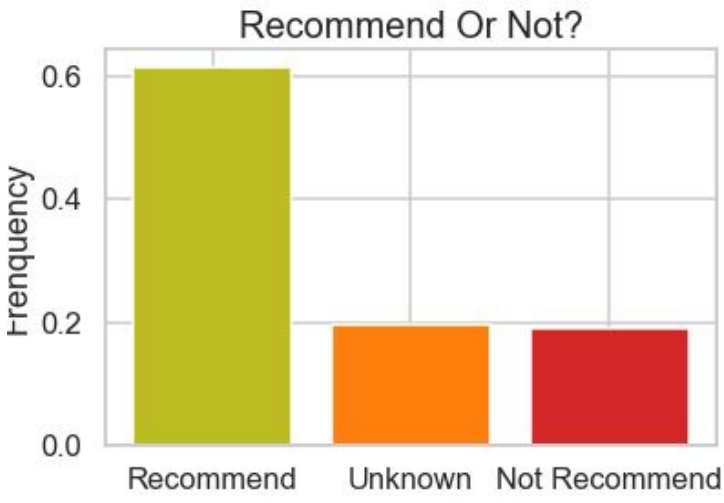



*Figure 2: Overall recommendation*

## 3.3 Sub-category Rating

In addition to overall rating, there are specific ratings on 5 sub-categories. For this company, every category has an average rating higher than 3. The two relative lower ratings are ‘career opportunities’ and ‘ senior management’.

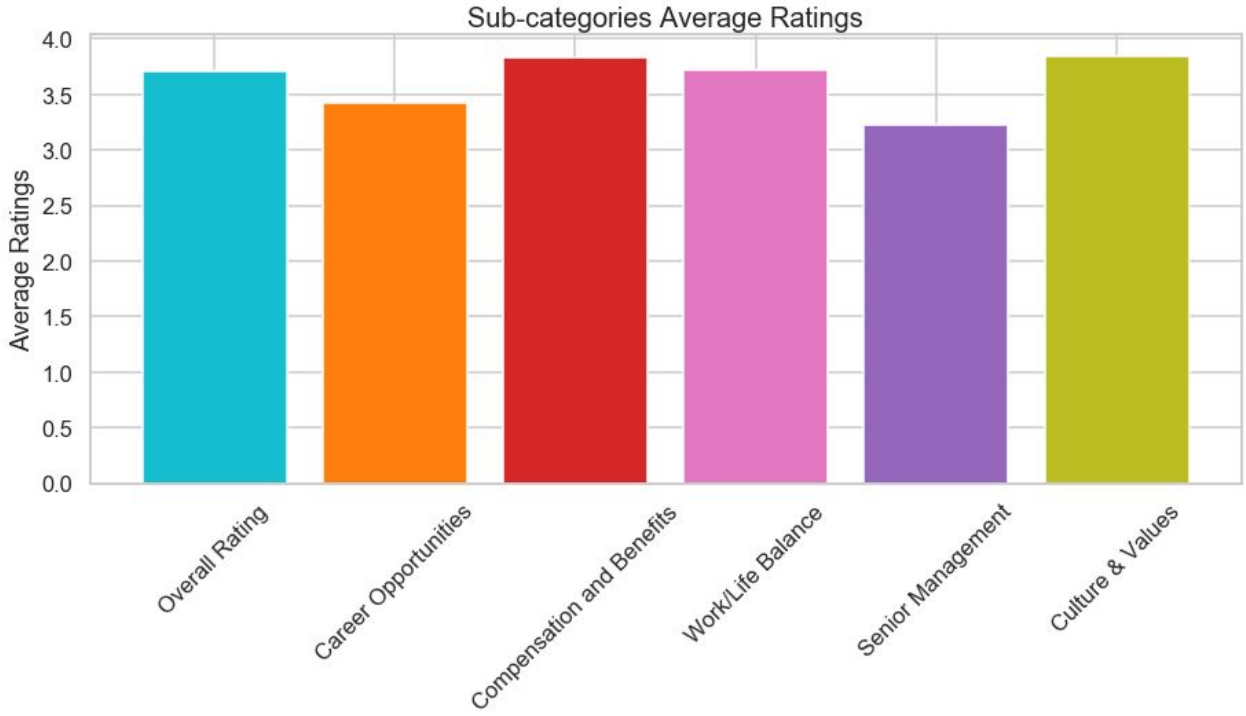



*Figure 3: Subcategories overall average rating*

## 3.4 Rating by States

There are 38 states identified from the data. From the below 2 plots, employees in KS and WI have the highest employee satisfaction and employees in NE (only 1 submitted review) are the unhappiest. Rating for all states is included in Appendix. Of note, the location data indicates 'Author Location', we assumed these employees are based locally in the company's branches in that particular state. If there are substantial remote employees, this conclusion may not be accurate.

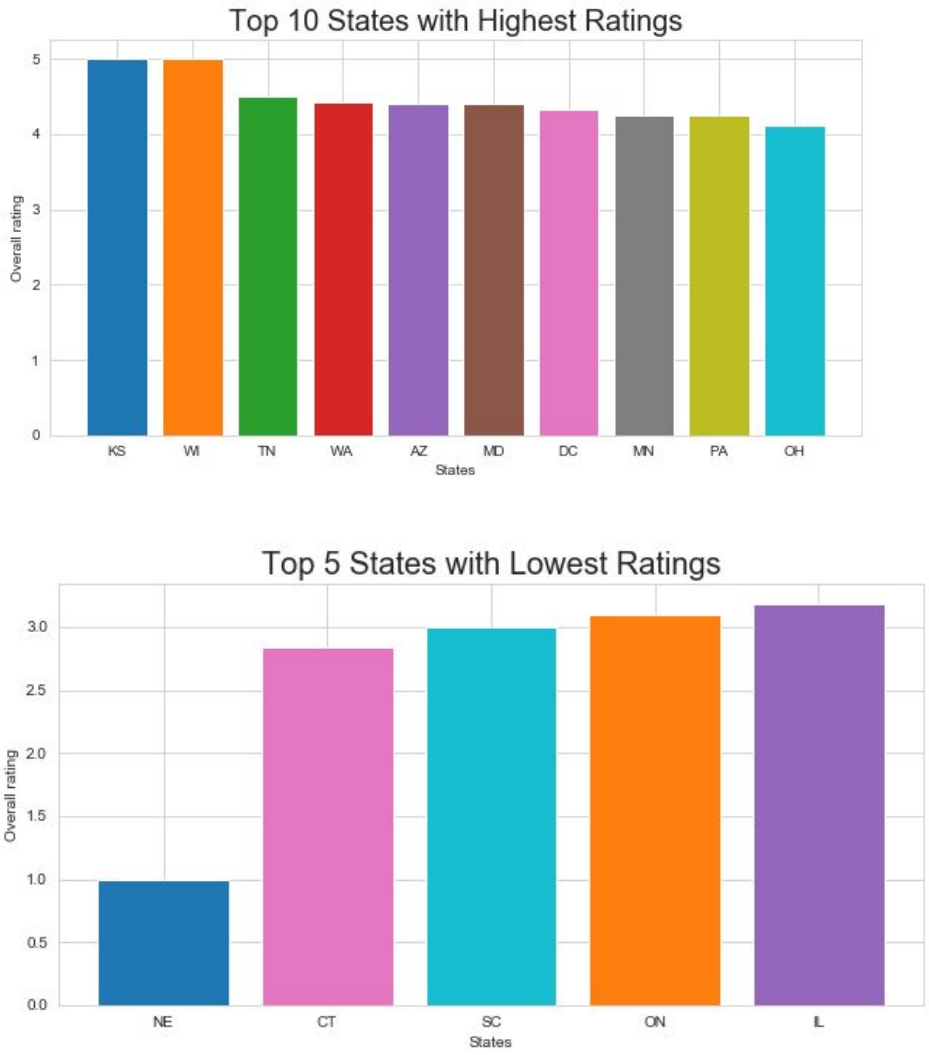



*Figure 4: Top 10 states with the highest rating and top 5 states with the lowest ratings.*

## 3.5 Rating by Employee Types

It is obvious that former employees give lower ratings than current employees and part time employees give higher ratings than full time employees.

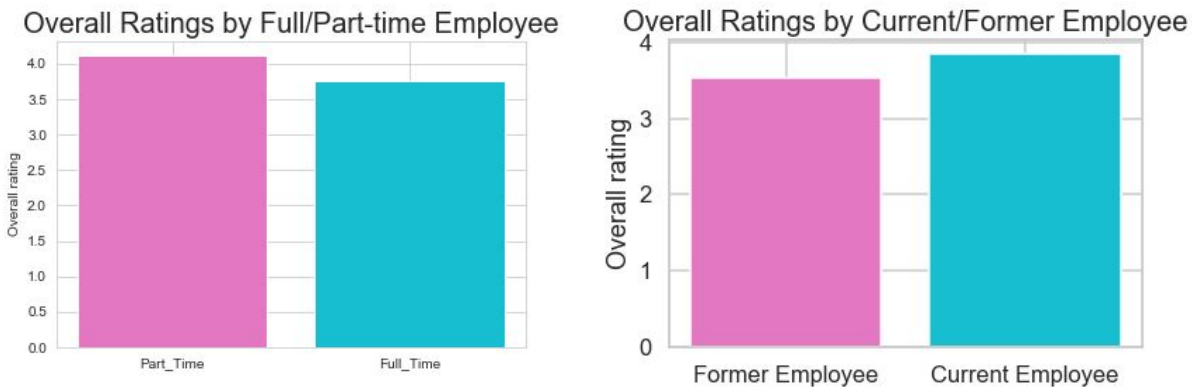



*Figure 5: Overall rating by employee types*

## 3.6 Frequent Job Titles

Most reviewers choose not to disclose their job titles, for those who specified their titles, they are either Software Engineers or Financial Representatives. From the bar chart in Figure 6, we notice there is a Financial Representative and Financial Service Representatives, a Software Engineer and Software Engineer/Developer, assuming these are the same job titles, we are going to combine those 2 into 1.

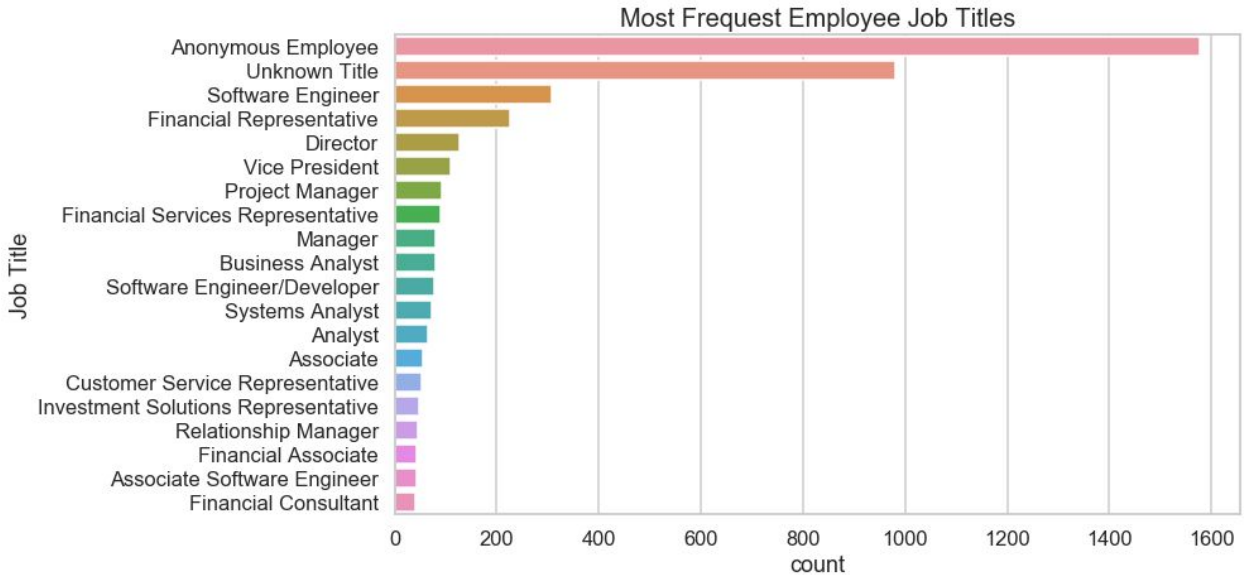



*Figure 6: Most frequent employee job titles*

## 3.7 Rating by Job Families

Among the most frequent job titles, Financial Associate group has the lowest satisfaction and Relationship Managers give the highest ratings.

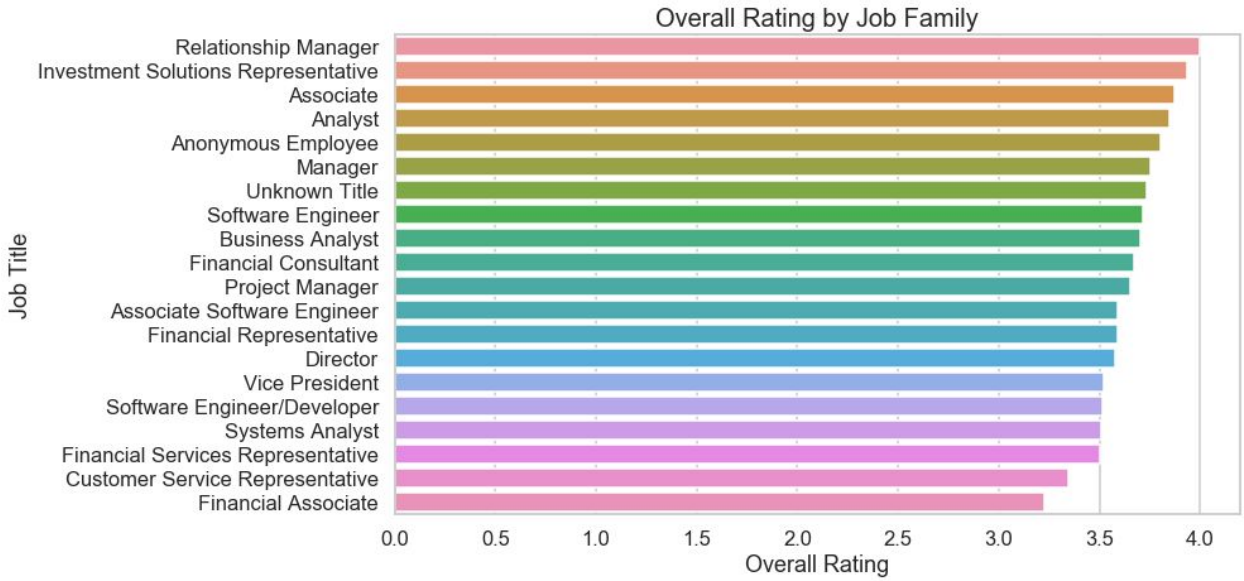



*Figure 7: Overall rating by most frequent job families*

# 4. Sentiment Analysis

In this section, we work on the text columns and perform sentiment analysis. Sentiment of words can vary based on where it is in a sentence.The TextBlob module allows us to take advantage of these labels. A little introduction on sentiment labels: each word in a corpus is labeled in terms of polarity and subjectivity (there are more labels but we will focus on these 2 in this analysis). A corpus' sentiment is the average of these 2 labels.

- Polarity: how positive or negative a word is. -1 is very negative. +1 is very positive.
- Subjectivity: how subjective, or opinionated a word is. 0 is fact. +1 is very much an opinion.

## 4.1 Summary Column

There are 3 columns that have text - ‘Summary’, ‘Pro’ and ‘Con’. First, let’s look at the column ‘Summary’.

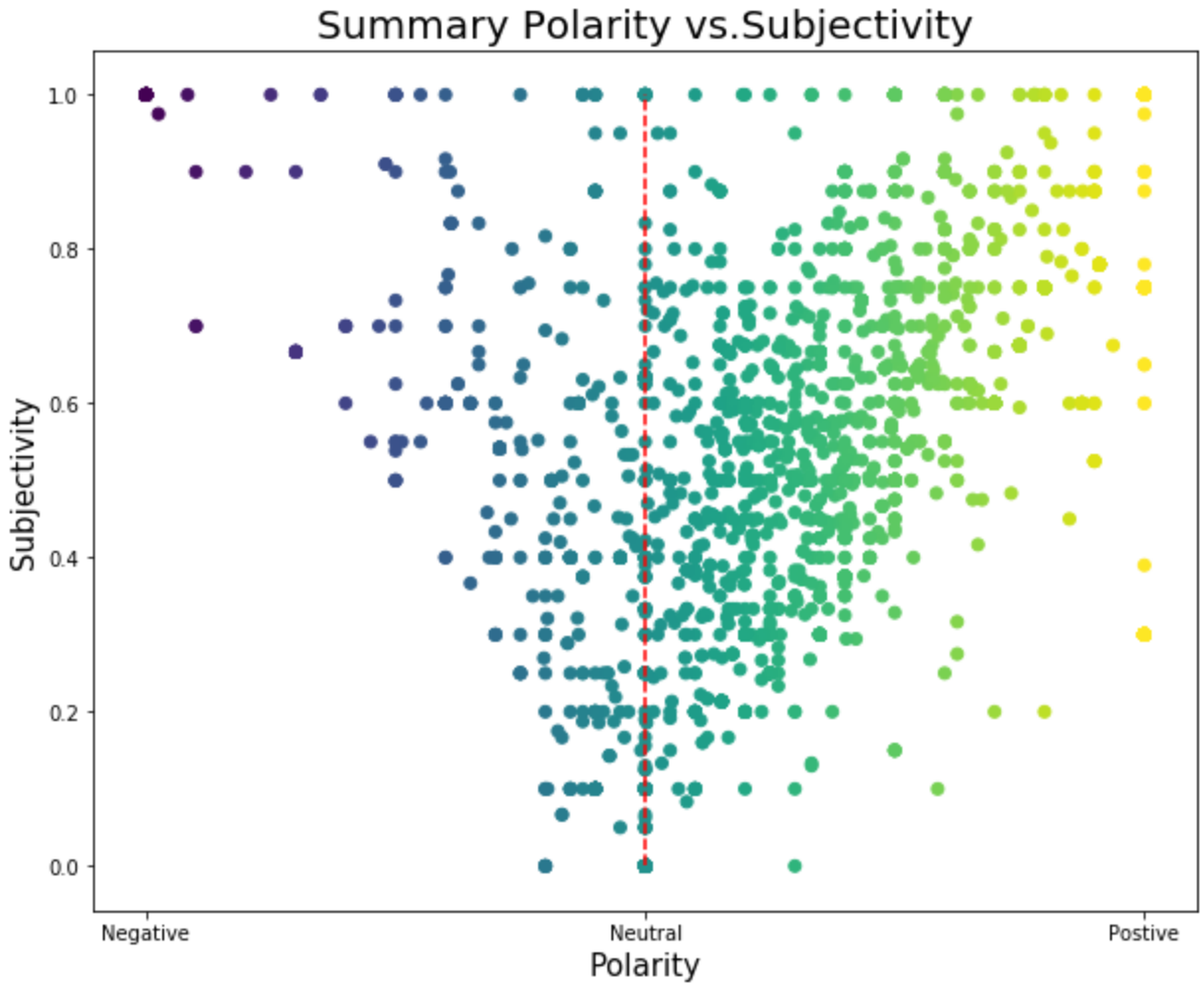



*Figure 8: Summary polarity and subjectivity*

**Findings:**

- It is clear that there are more positive comments than negative ones as we see more data points on the right side of the red vertical line. This verifies what we have observed from the precious EDA. People are generally happy with this company.
- The more polarized the comment is (positive or negative), the more subjective it is.
- Reviewers who give positive comments based more on facts (lower subjectivity) and reviewers who give negative comments based more on opinion.

We are also interested in knowing which job families have the most and the least positive comments, so we plot the most frequent 10 job titles' sentiment scatter plot in Figure 9.

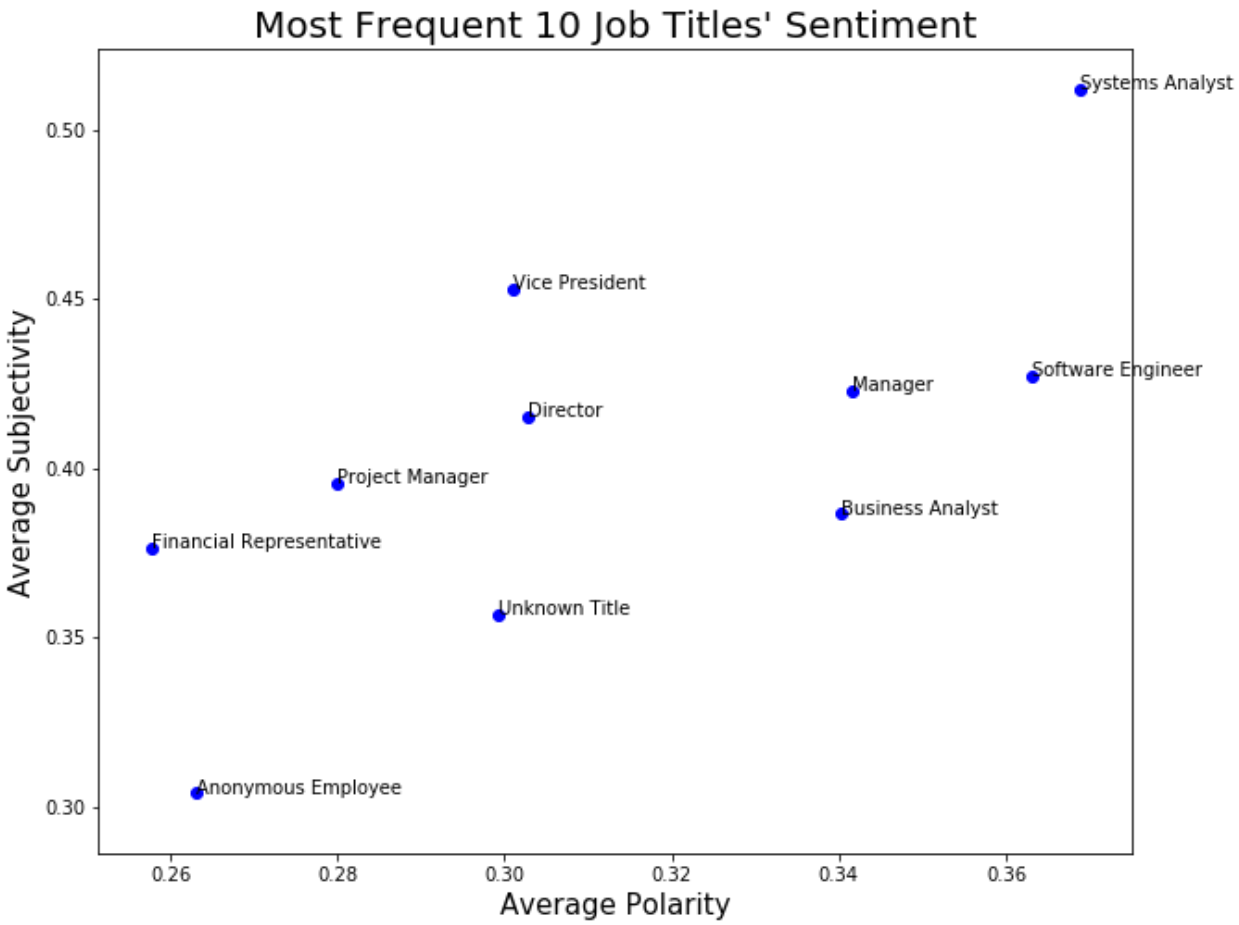



*Figure 9: Most frequent job titles' sentiment.The higher average polarity score, the more positive the comment.*

**Findings:**

- System Analysts and Software Engineers have the most positive comments.
- Financial Representatives, Project Managers and reviewers who do not disclose their titles (Anonymous Employee) tend to give the most negative comments in 'Summary'. In the next section (keywords extraction and topic modeling), we will dive into these groups and see what makes them unhappy.

## 4.2 Pro and Con Columns

Now, let's look at column 'Pro' and 'Con'.

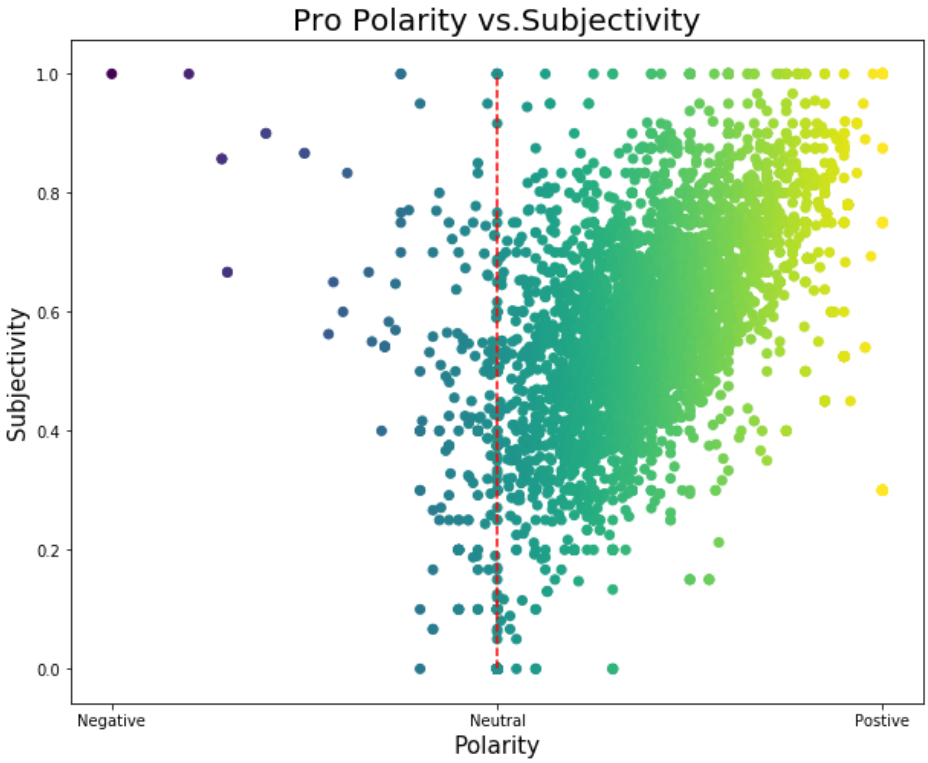



*Figure 10: Column 'Pro''s polarity and subjectivity.*

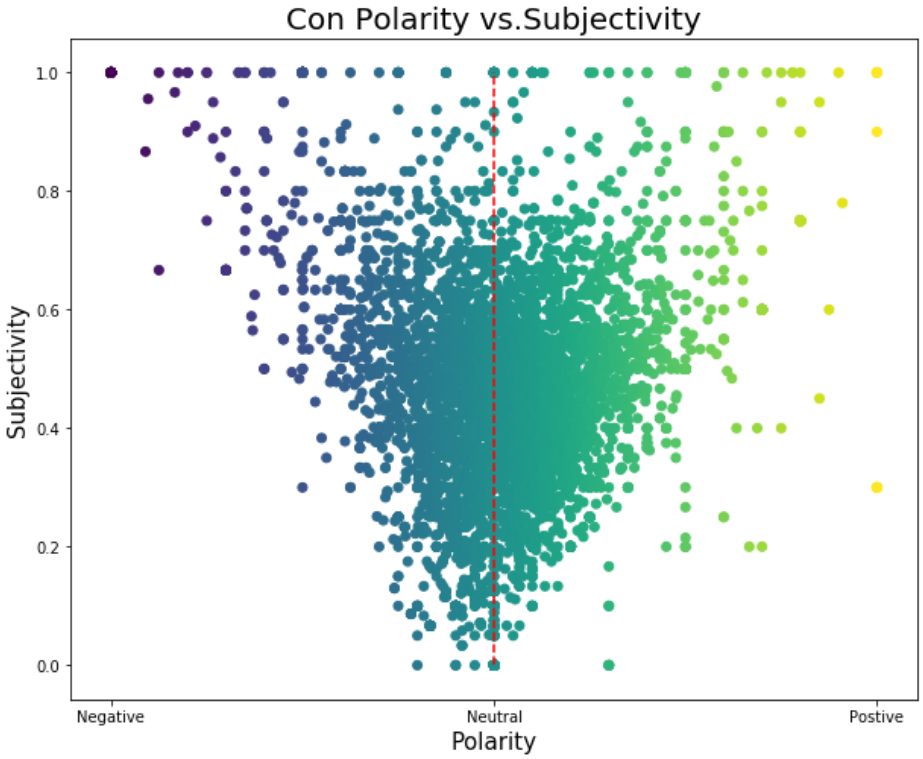



Figure 11: *Column 'Con''s polarity and subjectivity.*

**Findings:**

- It is not surprising to see the polarity of ‘Pro’ mostly falls in the positive part because we know these are the nice things employees say about this company.
- It is surprising to see even in the ‘Con’ column, half of the data points' polarity falls on the positive side. This is strong evidence that most employees love this company and they don't have many complaints.

# 5. Topic Modeling

In this section, we extract keywords from the text data and perform topic modeling. The Python NLP libraries we will be using are TextBlob, NLTK, Gensim and Spacy.

## 5.1 Keywords Extraction

### 5.1.1 Uni-gram Words

First of all, let’s take a look at the top uni-gram words in the text data.


*Figure 12: Summary column word cloud*


*Figure 13: Pro column word cloud*


*Figure 14: Con column word cloud*

**Findings:**

- The most common words in Summary is positive, which is consistent with our previous plot where 60% of employees recommend this company.
- Employees who love this company due to its great benefits and good work-life balance.

- It appears employees' complaints are mostly about management and the job nature of call centers.

### 5.1.2 Bi-gram Words

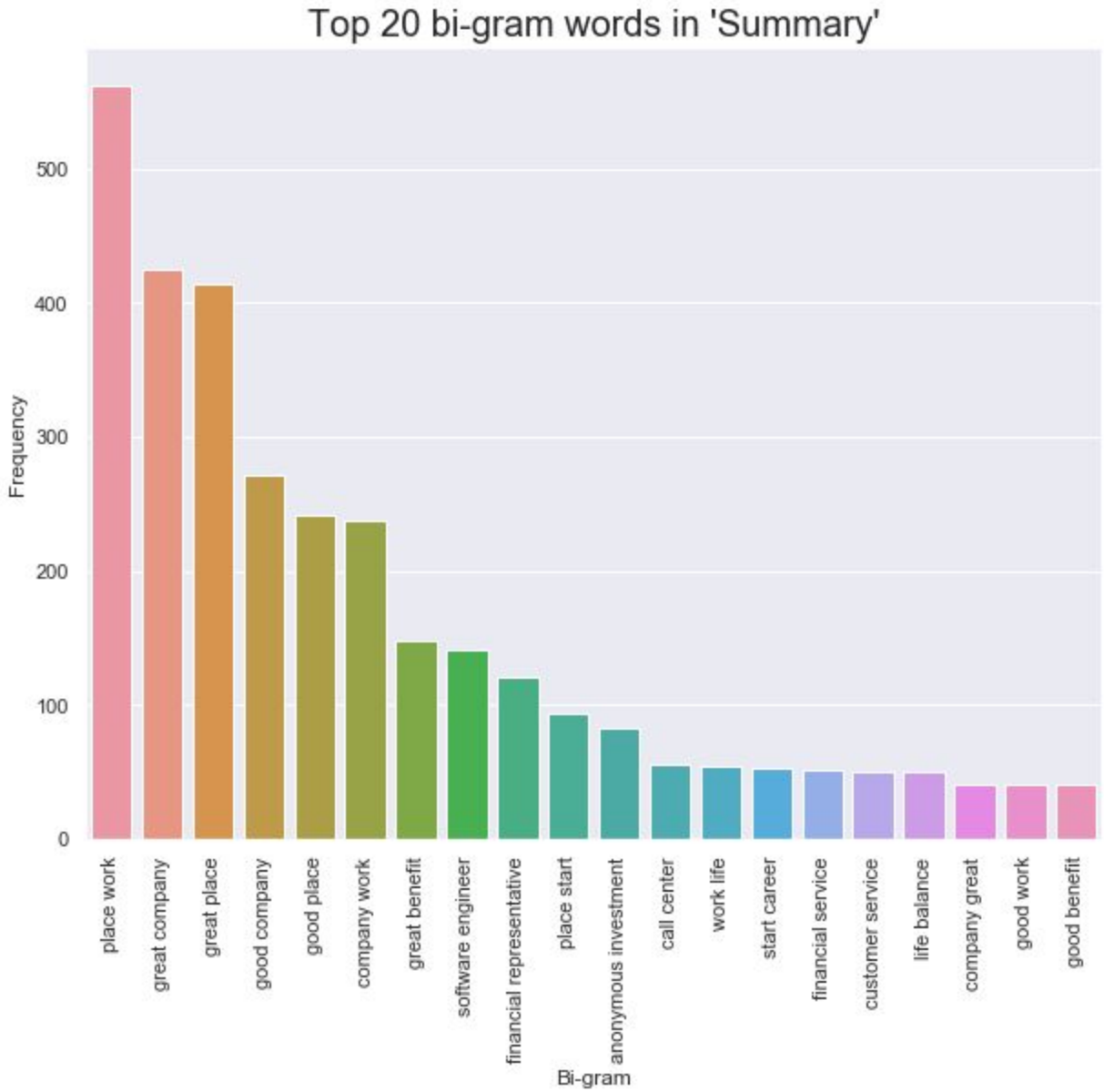



*Figure 15: Top 20 bi-gram words in 'Summary'*

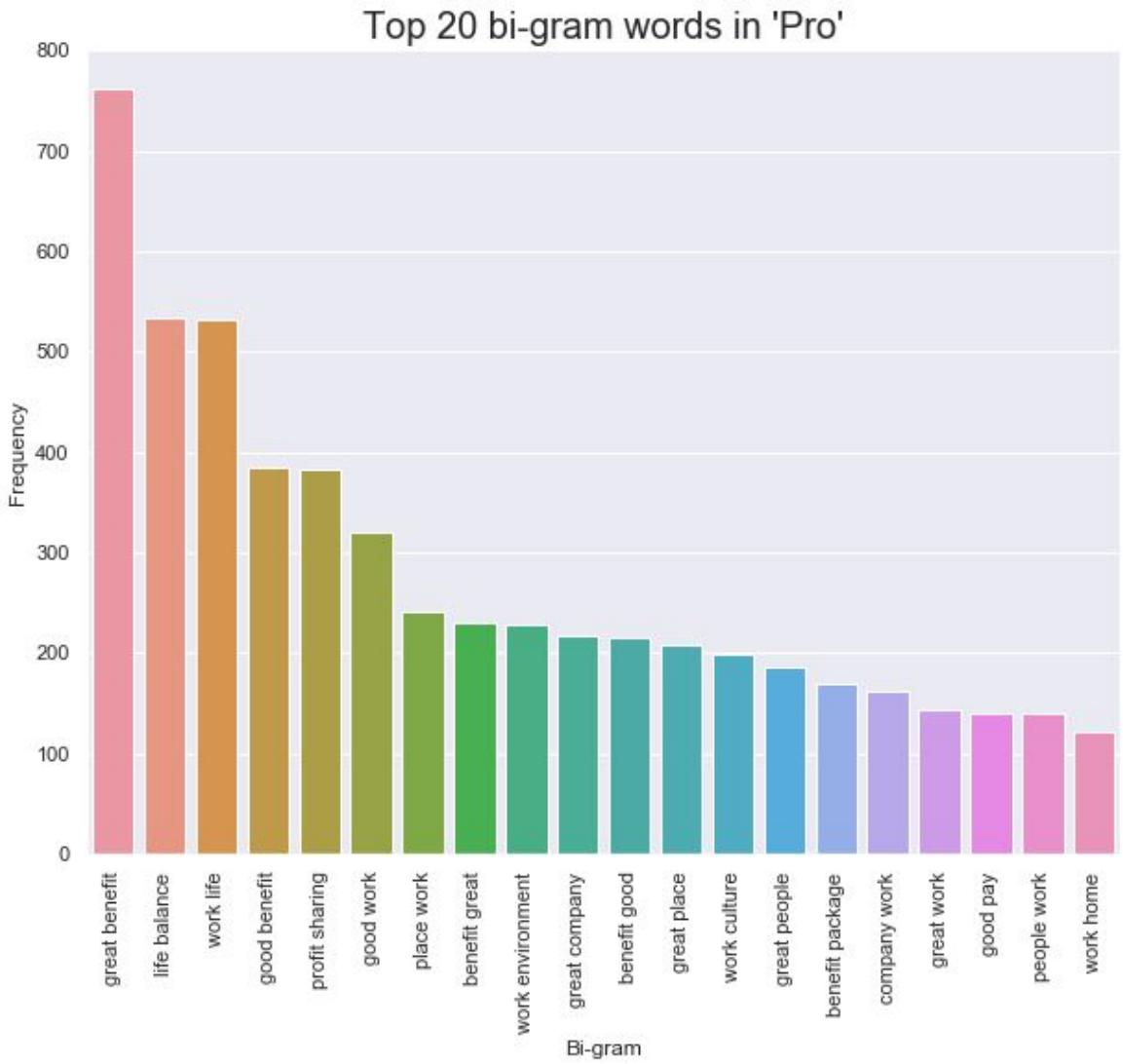



*Figure 16: Top 20 bi-gram words in 'Pro'*

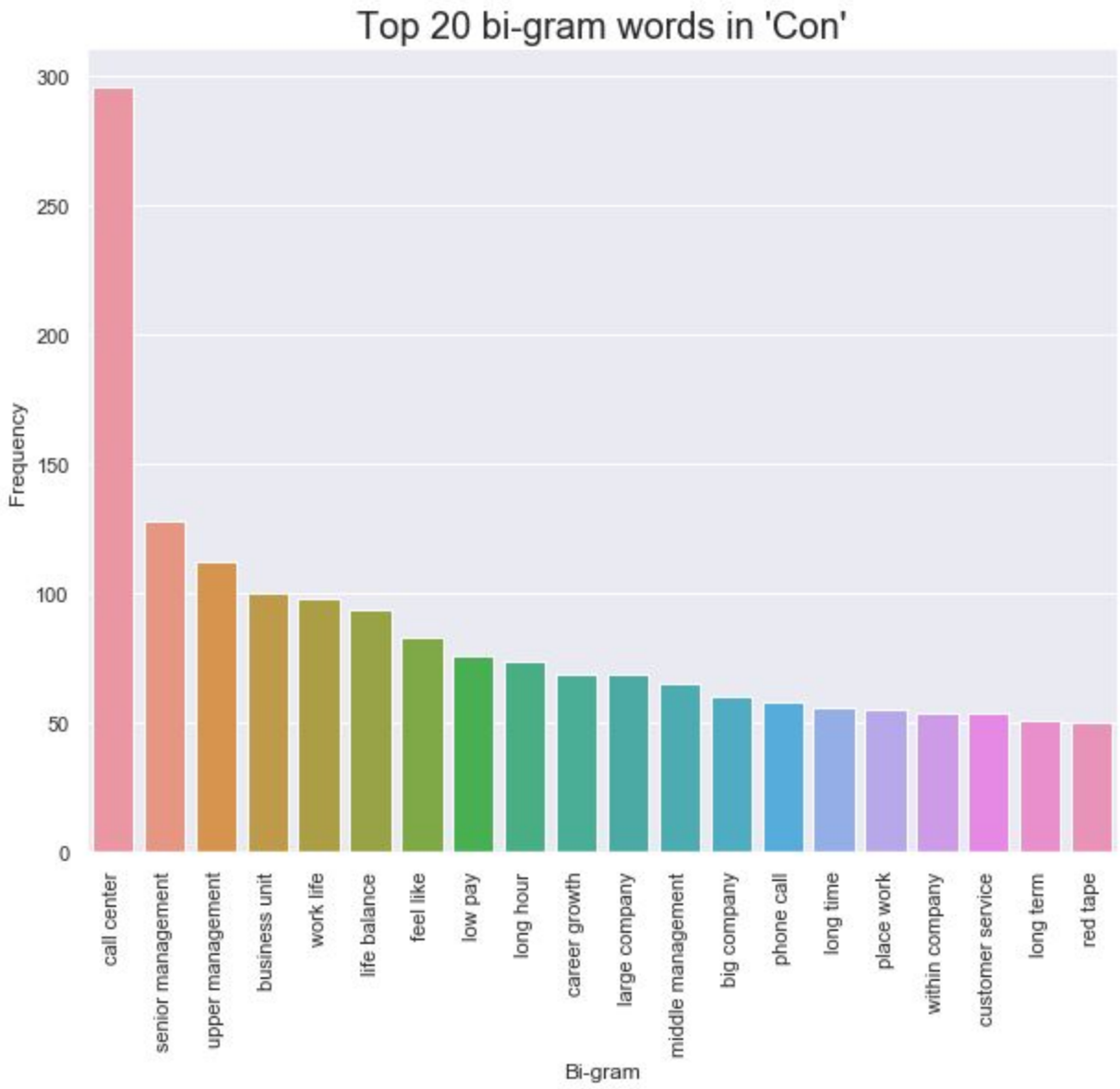



*Figure 17: Top 20 bi-gram words in 'Con'*

**Findings:**

- Bi-gram bar plots give us great insights on the topic discussed. It is clear people love this company mainly because of the company's generous benefits.
- The negative comments mostly come from employees in the call center. They seem to be complaining about answering phone calls for long hours and with low pay. Other issues include: people do not like the red tape culture in a large company, they don't approve senior management, some employees don't think there are great career advancement opportunities.
- Recall from previous sentiment analysis, we noticed employees with the job title 'Financial Representative' tend to give negative comments, if Financial Representatives

are working in call centers, then this is a clear message that this company needs to make some improvements in the call center.
- This bi-gram chart is a great starting point in top-modeling. We will explore more with Latent Dirichlet Allocation (LDA) technique and verify the modeling results with these plots.

## 5.2 Topic Modeling

The ultimate goal of topic modeling is to find various topics that are present in the corpus.For topic modeling, we will use Gensim's Latent Dirichlet Allocation(LDA) model. To use a topic modeling technique, we need to provide a corpus and the number of topics we would like the algorithm to pick up.

There are 2 ways to define the number of topics. One is from the previous key-words visualizations, another way is to use K-means clustering. We tried K-Means clustering, but this algorithm is sensitive to initiations, each rerun gives me very different results, therefore I decide to define the number of topics from the visualizations using my best judgement: 3 topics in 'Summary', 5 topics in 'Pro' and 'Con'.

Usually, LDA uses the bag-of-words feature representation to represent a document, meaning order does not matter. However, in this section, we apply 2 feature extraction techniques to perform LDA:

- Bag-of-words
- TF-IDF

For LDA model's outputs, it is hard to identify topics based on the modeling output because these are all unigram words. We will need to refer back to previous plots and use our best judgement to extract the topics.

The topics in Summary are:

- Great place
- Benefits
- Call center

The topics in Pro are:

- Work-life balance
- Good pay
- Career opportunity
- Good people
- Good benefits (profit-sharing,bonus,401k match)

The topics in Con are:

- Management
- Call center/Phone call
- Low pay
- Slow
- Culture and politics

Finally we tested the models using one sample from the existing corpus and a new unseen comment, both tests showed the Bag-of-Words feature selection method performed better for this dataset.

## 5.3 Unhappy Groups

In the previous sentiment analysis, we noticed that employees with the job title 'Financial Representative',' Anonymous Employee' or 'Project Manager' submitted the most negative comments. We use Spacy's rule-based pattern matching to get the keywords from these 3 groups. These are what we find out.

Financial Representatives are complaining about:

- Low pay
- Long hours
- Phone calls
- Management

Project Managers are complaining about:

- Management. We see a heavy focus on this from the Project Managers group
- Annual performance review.They could be unhappy with the review process

Anonymous employees could come from any job family, so it can be a combination from the above two groups or include other groups. Anonymous employees are complaining about:

- Low pay
- Long hours
- Upper management/autocratic management
- Red tape of a large company
- Constant changes/frequent reorganization

# 6. Conclusions

At the beginning of this project, we have defined a problem statement and now we have answers to all of the questions.

## 6.1 Key Findings

**What employees like and dislike about this company?**

Employees love the good pay, generous benefits, profit sharing and a good work-life balance. Employees from call centers complain about the low pay, long hours handling phone calls and the management team. Project managers complain about the upper management team. Other anonymous employees which could consist of any job family complain about the aforementioned issues plus red tape of large corporations and constant changes.

**Has the company's reputation gotten better or worse in the recent year?**

This company did great compared to other companies' average ratings on Glassdoor. The majority of employees are satisfied with the company. Its rating has been increasing since 2008 with small dips in 2011, 2013 and 2017.

**Which job families have the highest and lowest satisfaction rates?**

Among the most frequent job titles, Financial Associate group has the lowest satisfaction and Relationship Managers give the highest ratings. Since we know the negative comments are from call center employees, our best guess is Financial Representatives work in the call center.

**What are the keywords that people say about this company?**

The keywords are most clear in the top bi-gram plots. They are: good salary, good benefits, generous profit sharing, big company, work-life balance, upper management, low pay, long hours.

**What can this company do to improve employee engagement?**

This analysis identified the negative comments are mainly from call center employees. We suggest the company look into the salary, working hours and find solutions to address these issues. If call center employee engagement increases, the total employee engagement will increase significantly.

## 6.2 Project Contribution

In addition to answering the above questions on that specific company, this project has created a standard workflow for any similar tasks. Given any company's Glassdoor review data, the same structure and code can be applied with little tweaks, especially the data cleaning code blocks can be used without change since all Glassdoor review data follow the format.

## 6.3 Limitations and Future Work

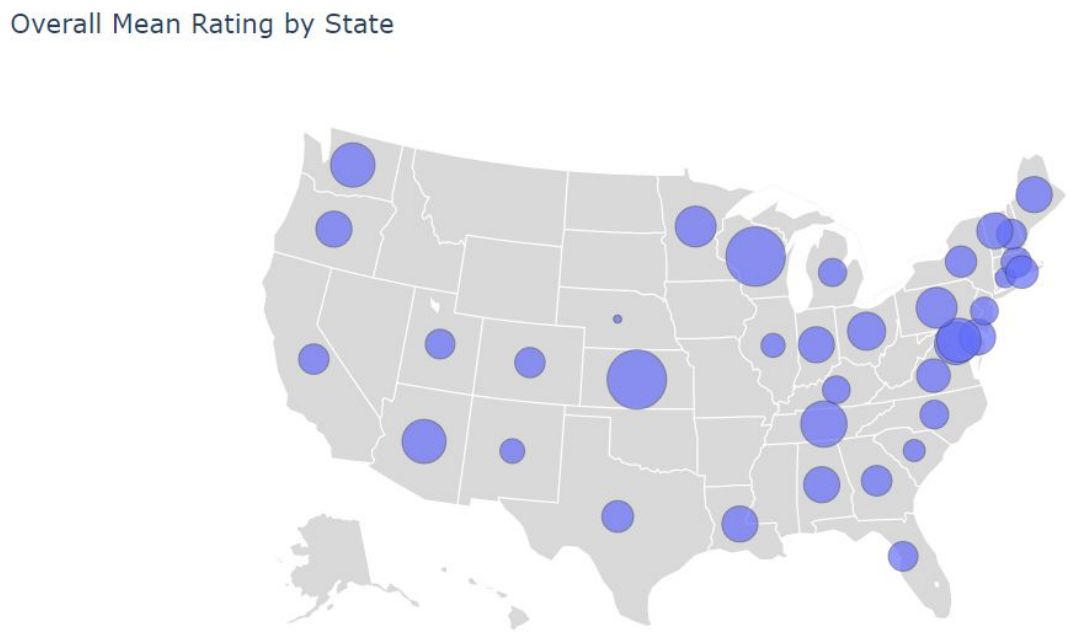
- **Limitations:** Since there are 1578 unhappy employees submitted anonymously, we can't do a more granular analysis on specific job families. It is our assumption that Financial Representatives work in call centers, if this is untrue, we would need to analyze this further.

- **Future Work:** We could do a similar analysis on this company's competitor and see how they compare and what they can learn from each other.

# Acknowledgement

I would like to extend my appreciation to Glassdoor for allowing me web scraping the data, so I could conduct this analysis.

# Appendix




# References

1. Gensim website: https://radimrehurek.com/gensim/auto_examples/index.html
2. Alice Zhao: nlp-in-python-tutorial https://github.com/adashofdata/nlp-in-python-tutorial