Glassdoor Company Review Analysis with NLP

A case study on review data of a national company

Teresa Nan
August, 2020

Problems to Resolve

Project Purpose

- Build a reusable workflow for Glassdoor employer review tasks.
- Help employers gain insights on employee engagement.

Potential Clients

- Mid-large sized company that has enough Glassdoor reviews.

Dataset Introduction

Numerical Columns (rating) ★ ★ ★ ★ ★

- Overall Rating
- Career Opportunities
- Senior Management
- Culture & Values
- Compensation and Benefits
- Work/Life Balance

Text Columns

- Summary : subject of the review
- Pro: what they like about this company
- Con: what they dislike about this company

Data Cleaning

Get data ready for exploration

How do we get the unstructured, untidy data ready for exploration?

Dataset Processing & Cleaning

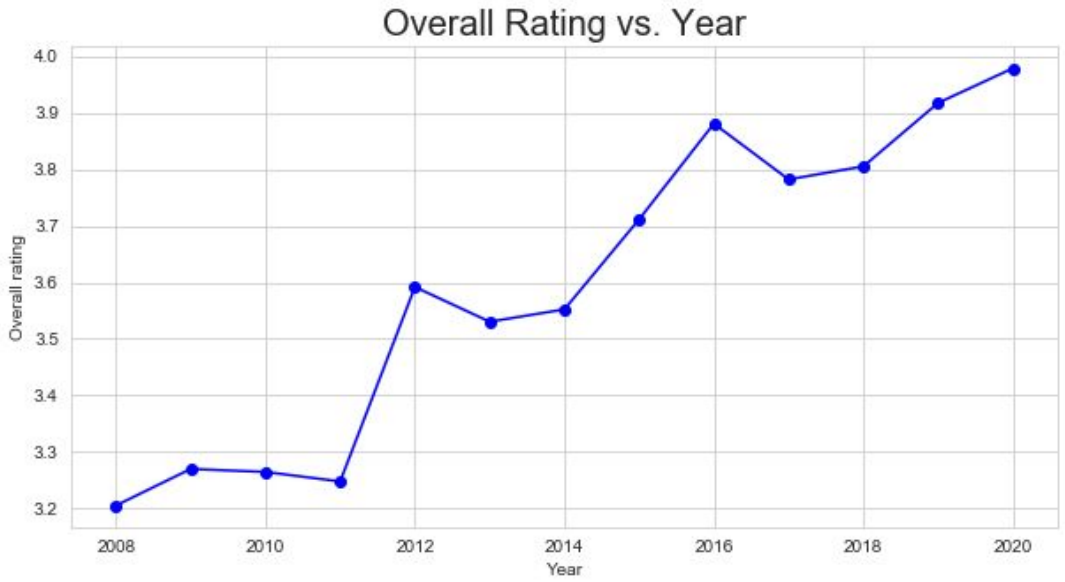
- Web scraping Glassdoor data
- Anonymized the company name
- Drop duplicates, 6386 reviews left
- Leave missing numerical values as is
- Fill missing categorical values with ‘unknown’
- Extract 5 features from existing columns
- Turn unstructured data into structured data

Data Exploration

Explore on rating columns

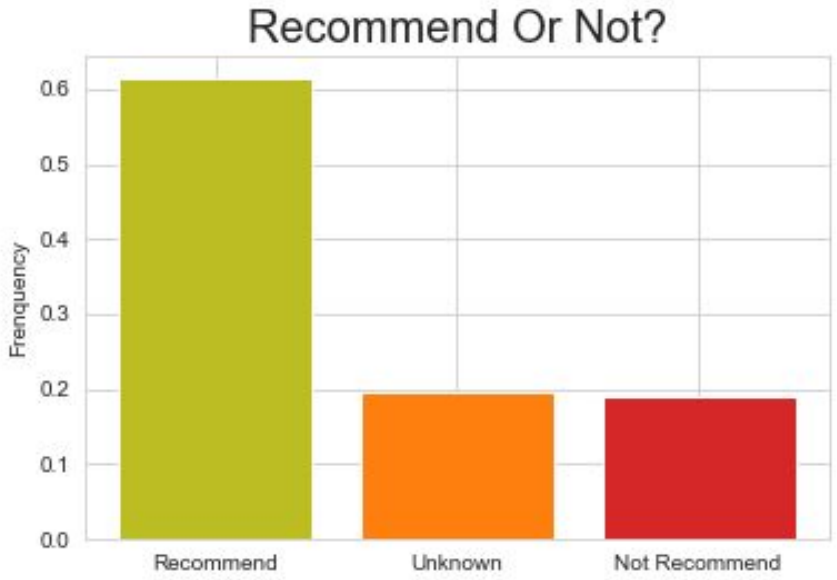
What patterns are discovered from exploring the numerical rating data?

Overall Rating Over the Years



Average overall rating has been increasing since 2008 with small dips in 2011, 2013 and 2017.

Employee Recommendation Rate

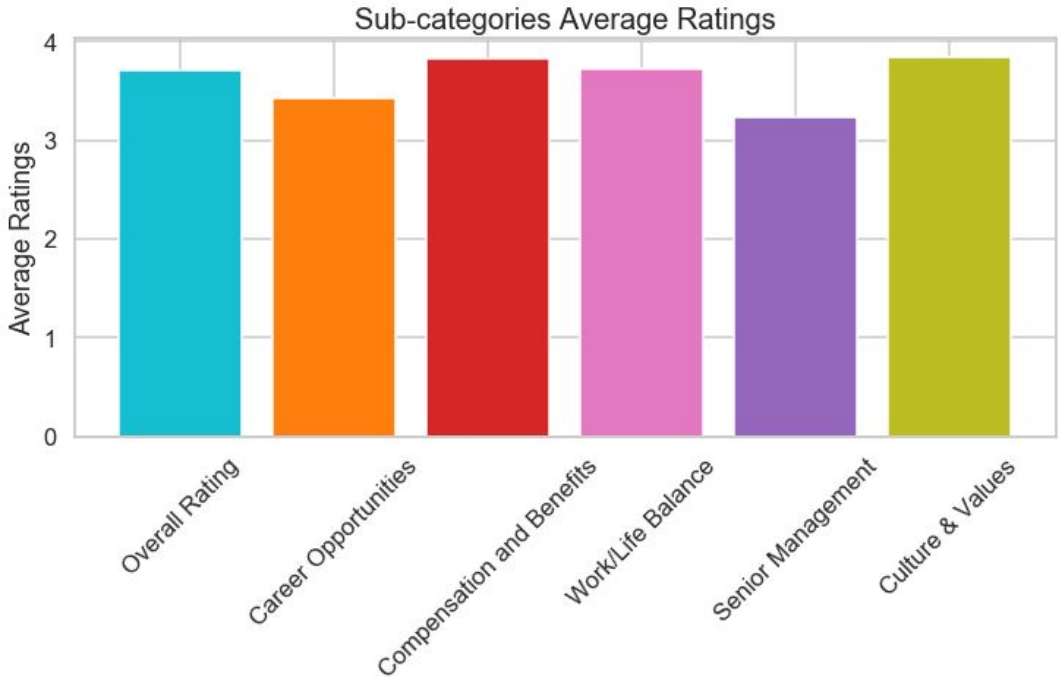


- **60%** recommend
- 20% do not recommend
- 20% do not answer

Sub-category Rating

Every category has average rating > 3 stars

Two lower categories
- career opportunities
- senior management



Other Findings from EDA

- Most employees who submit reviews choose to be **anonymous.**
- Current employees give higher ratings than former employees.
- Part-time employees give higher ratings than full-time employees.
- Employees in KS and WI have the highest employee satisfaction.

Sentiment Analysis

Analyze text data using TextBlob

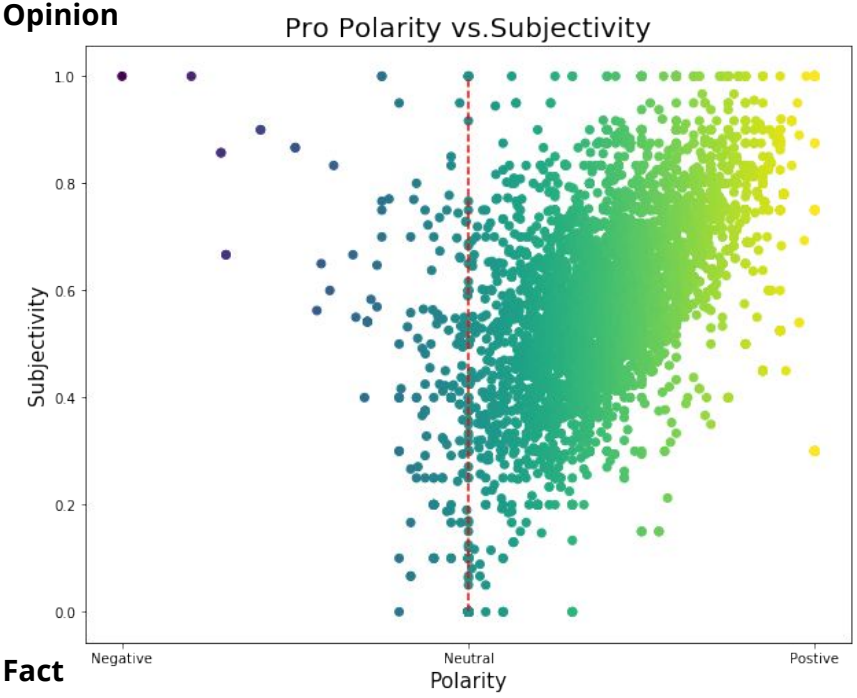
How positive or negative the comments are?

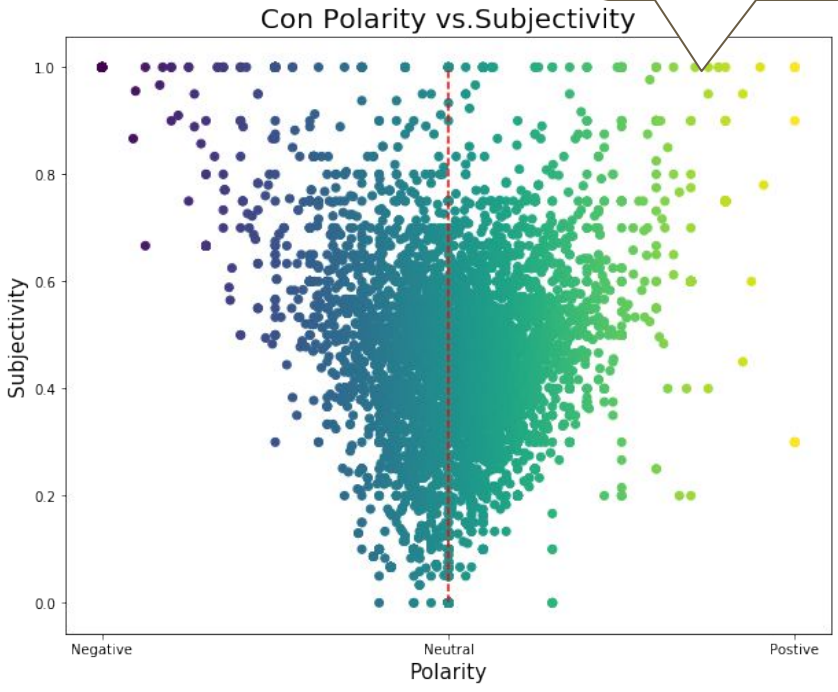
Are comments based on facts or personal opinions?

‘Pro’ and ‘Con’ Polarity and Subjectivity

Employees don’t say many negative words even in ‘Con’

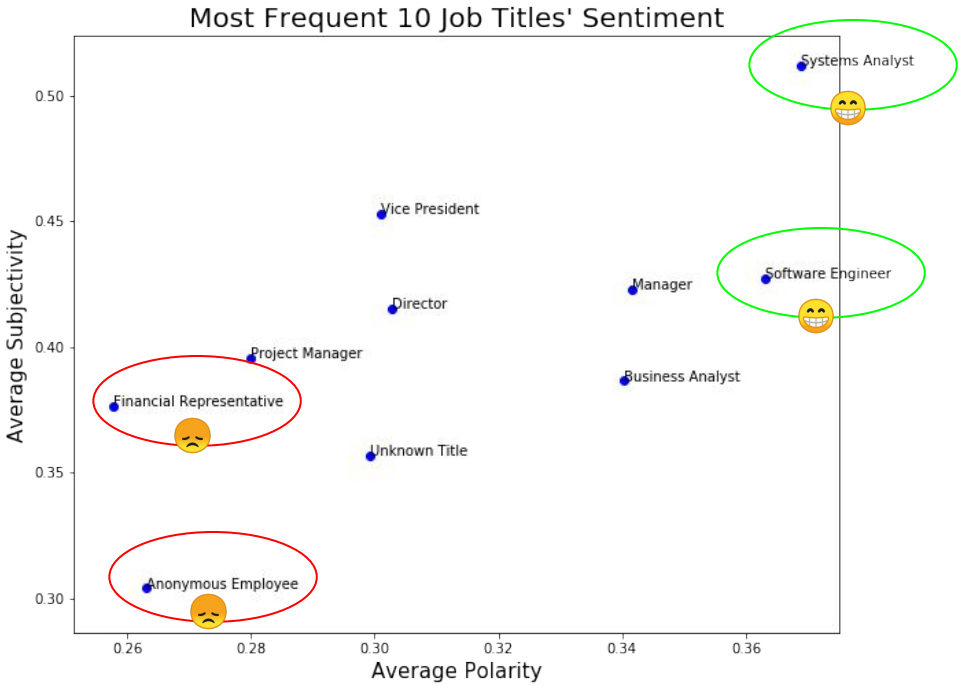
Opinion





Fact

Most Frequent Job Titles' Sentiment



The most positive group 😁

- System Analyst
- Software Engineers

The most negative group 😞

- Financial Representatives
- Project Manager
- Anonymous Employee

Keywords Extraction

Extract unigram and bigram words on text columns

What are the key words that employees say about this company?

Top Bi-gram Words

Pro

- Great benefit
- Work-life balance
- Great company
- Great people

Con

- Call center
- Senior management
- Long hours
- Low pay

Unhappy Group's Key Complaints

*This could consist of any job title

Financial Rep.

- Low pay
- Long hours
- Phone calls
- Management

Project Manager

- Management
- Performance review

Anonymous*

- Low pay
- Long hours
- Management
- Red tape
- Constant change

Topic Modeling

LDA modeling on text columns

What are the topics that NLP models extract from the text columns?

Topic Modeling

LDA Topic Modeling
Bag-of-Words
TF-IDF

Topics in 'Pro'

- Work-life balance
- Good pay
- Career opportunity
- Good people
- Good benefits

Topics in 'Con'

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

Conclusions

- The majority of employees are very satisfied with the company.
- The company's rating has been increasing since 2008.
- Employees love the good benefits, nice people and work-life balance.
- Employees dislike the long hours and low pay in call center, upper management, red tape and constant change.
- Negative comments are mainly from call center employees.

Recommendations

- We suggest this company look into its call center employees' salary, working hours and find solutions to address these issues.
- We suggest the company to check project managers' performance review process.

Project Contribution

- This project has created a **reusable standard workflow** for 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 same format.

Limitation & Future Work

- **Limitations:** 1578 unhappy employees submitted anonymously, we couldn’t do a more granular analysis on specific job families. It is our assumption that Financial Representatives work in call centers.
- **Future Work:** A similar analysis on this company's competitor and compare the results.

Thank You for Reading!

Acknowledgement:

I extend my appreciation to Glassdoor for allowing me web scraping the data. A huge thanks to my Springboard mentor Tommy Blanchard for guiding me through this project.

Project Notebook:

https://github.com/teresanan/glassdoor-company-review-analysis

Teresa Nan LinkedIn:

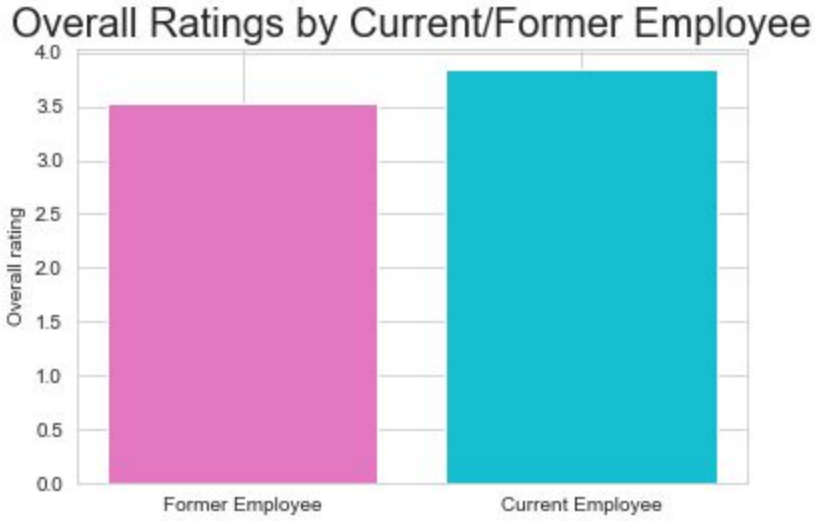
https://www.linkedin.com/in/teresa-n-39287042/

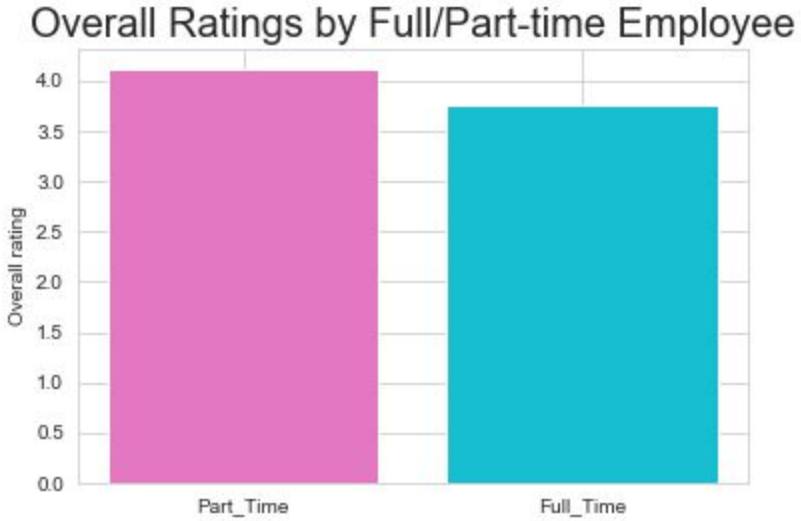
Supplemental Slides

Figures and details not included in the previous slides

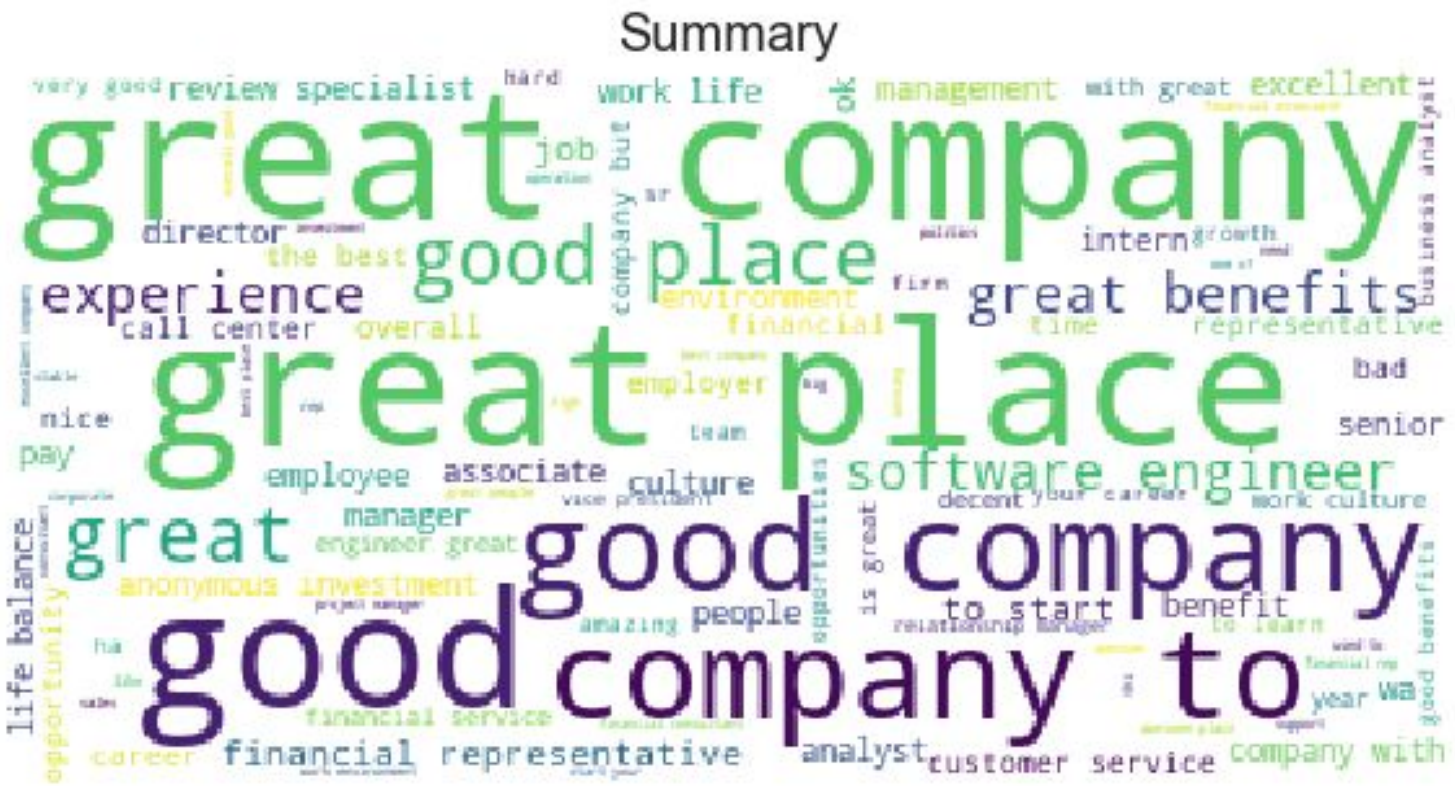
This subsection includes supplemental figures and information about the analysis.

Overall Rating by Employee Type

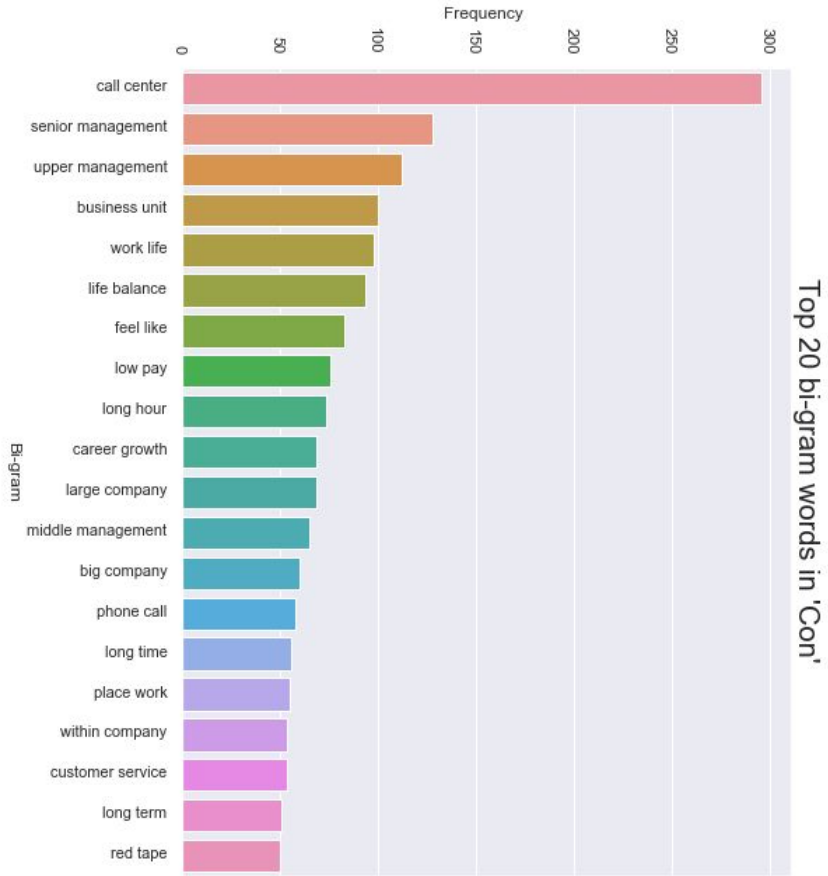




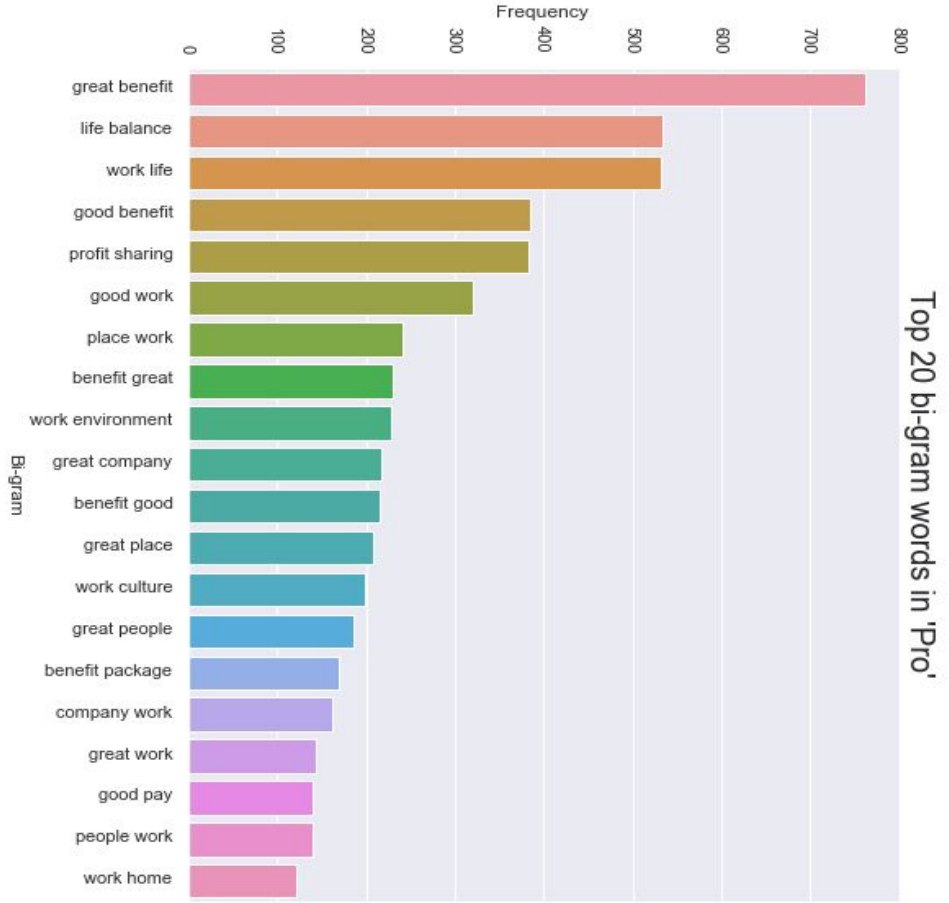
Word Cloud of 'Summary'



Top 20 Bi-gram in Con



Top 20 Bi-gram in Pro



Topic Modeling Details

- **Model Used**: Latent Dirichlet Allocation(LDA)
- **Define # of topics:** K-Means clustering did not yield reliable results since it is high dimensional data. Chose numbers of topics based on previous EDA worked better.
- **Feature Representations**: Bag-of-words and TF-IDF
- **Challenge Faced**: Hard to identify topics based on the LDA modeling output alone since there are overlaps. Previously bi-gram words chart helped in the model output interpretation.
- **Model Evaluation**: Tested model using one corpus from the data and an unseen sample, both tests showed the model was correct.