Vidya Sagar Gunnala, Rohan Sunki, Sagar Kulkarni, Vijay Venkata Kambhampati

5/1/2020

TEXT MINING EMPLOYER BRAND AND EMPLOYEE SATISFACTION OF 4 MAJOR EMPLOYERS IN US

Contents

Executive Summary

Text Analytics is gaining importance with each passing day and individuals and organizations are exploring ways to use this to their own advantage. We too attempt to leverage the benefits of text analytics and implement it in the area of human resources to gain some meaningful insights useful for both the organizations and candidates looking for jobs. Social media data has been exploited to understand the pulse of general populace for a product or service, but we try to look at the employees view of employers. We extract employee reviews from Glassdoor for textual analysis to understand the employer brand and employee's satisfaction with their employers.

The organizations we analyze in this study are Microsoft, Amazon, Google and IBM. The reviews are anonymous, and we extract all the reviews from 2008 to 2020. We applied well researched text analytical methods like sentiment analysis and similarity score analysis in this study. We also experiment with different ways of doing the analysis to ensure we use what is suitable for the corpus and task we have at hand.

We found that the employee satisfaction with all the companies have declined from 2008 and Google and Microsoft are doing better on employee satisfaction metric compared to Amazon and IBM. We also identify the characteristics which are fetching highly positive and highly negative reviews for each of these companies. The results of this study are aggregated into a dashboard for easy reference.

Problem Definition and Significance

Every organization has a reputation, which includes thoughts attached to its products and services, history and people involved. People, over time develop this reputation and associate certain characteristics with an organization. This becomes the brand of a company. In a similar way, every organization has a secondary brand called Employer brand, which is the emotional reaction people have to the idea of working for an employer. This brand lives in the minds of employees - former, current and future.

Companies with positive employer brands stand to gain in more than one way. The most significant impact a positive employer brand has on a company is in the area of recruitment. It is noted that companies with positive employer brands can get up to twice as many applicants as companies with negative brands and the costs drop by almost 43%. We don't have to stress further how important this is, especially in the era when it is getting difficult to hire the right talent. Also, to gauge the significance of employer brand, one can look no further than the costs incurred by companies who fail to invest in their reputation as an employer, which is on an average $5,000 per employee[1].

In addition, 78% of candidates look at the company's employer brand before applying for a job and 88% of millennials[2] believe that being part of the right company is of at most importance. We all know how important salary is for job applicants, but a CareerBuilder report found that 67% of candidates are willing accept lower pay to be part of a company that has positive

reviews online. It is becoming increasingly important for companies to build a positive employer brand on social media platforms as 79% of candidate pool are likely to use of social media in their job search.

Keeping these details in mind, we set out to explore employer brand and employee satisfaction of four major employers in the U.S. We intend to find the key characteristics of these companies which are fetching high positive and negative reviews and chart out the employee satisfaction from the reviews provided on social media platform Glassdoor. A collective dashboard with these findings will help employers make decisions to either improve or alter employer brand and potential job candidates to identify a preferred employer.

Prior Literature

The areas of employer branding and employee satisfaction have always dominated analytics domain in Human Resources function and are looked at extensively by many people and organizations. Although there has been lot of research on these topics, there was limited work which looked at Text mining employee reviews on social media to obtain employer brand characteristics and employee satisfaction.

Employee Satisfaction and Corporate Performance[3] looked at mining employee reviews from Glassdoor to understand the relation between employee satisfaction and organizational performance across industries. The study found that there is a positive correlation between general employee satisfaction and corporate performance along with the categories which correlated positively and negatively with performance. *Improving Employee Satisfaction Through Text Analytics*[4] explored the employee satisfaction of six technology companies using predictive rule-based model on text portions of the reviews from Glassdoor. The study found the words from text which make the review positive or negative using classification techniques. *Text Mining Company Reviews*[5] tried to find what employees write the most about their workplace experience and if they viewed the organization positively, negatively or neutrally. The study looked the frequency of single words and two words together to look at what was most talked about and ran sentiment analysis on the reviews to understand the variation of sentiment across the time range of reviews.

However, these studies did not go on to make a comprehensive analysis on employer brand and were limited to sentiment analysis and extracting the most frequent words. Thus, we try to take inspiration from these works and use them as foundation to provide both organizations and employees a comprehensive analysis on employer characteristics that are viewed favorably and unfavorably along with the variation in employee satisfaction over the years.

Data Source and Preparation

Social media has had large impact on multiple areas including in the domain of Analytics. The abundant of social media data made the data gathering easier for all types of analysis. One such social media platform is the Glassdoor website. Glassdoor.com was founded by Robert Hohman, Rich Barton and Tim Besse in 2007 (Glassdoor: About Us 2016). It is a website where current and former employees anonymously provide review for their respective companies. Glassdoor covers a diverse user population where users' profiles are fairly distributed across different sectors such as age, income, education[6]. Glassdoor contains 475,000+ companies with more than 8 million company reviews, company ratings along with salary reports from worldwide locations.

We used Web crawling technique to get the data from Glassdoor.com. We scraped the reviews for four major employers in US – Microsoft, Google, Amazon and IBM and crawl a total of 100,000+ reviews from 2008 to Mar 2020. Figure 1 below shows the final generated corpus.

	AuthorLocation	Pros	Cons	ReviewDate	ReviewMainText	ReviewTitle	OverallRating	company
0	Current Employee - Amazon Warehouse Worker in ...	Its a great job to have	Not much bad about it	27-Feb-20	I have been working at Amazon part-time	"FC Associate"	5	Amazon
1	Current Employee - Anonymous Employee in Seatt...	Really smart people, a lot of opportunity for ...	You have to be self motivated. NO ONE will hol...	10-Jan-16	I have been working at Amazon full-time	"You Get What You Put In"	5	Amazon
2	Current Employee - Senior Engineering Manager ...	Jeff Bezos and his "S-Team" are brilliant and ...	The management process is abusive, and I'm cur...	20-Feb-16	I have been working at Amazon full-time for mo...	"Exciting Work, Abusive Culture"	3	Amazon
3	Current Employee - Software Development Manage...	- You can learn a lot very quickly in a very s...	- Can be overwhelming, very steep learning cur...	3-Dec-18	I have been working at Amazon full-time for le...	"Amazing company and the most driven and smart...	5	Amazon
4	Current Employee - Software Development Manage...	I've been at Amazon for a month now, and I've ...	No cons, so far - seriously. Like I said, I'm ...	23-Feb-18	I have been working at Amazon full-time for le...	"An Amazing Place to Work"	5	Amazon

Figure 1: Sample of the data scraped from Glassdoor

The text data under Pros and Cons was then cleaned and preprocessed for analysis. As the analysis is only as good as data we provide, we spent some considerable time in preprocessing the data. The process involved replacing null values, fixing contractions, removing special characters and numbers, removing stop words, tokenization using word ninja and lemmatization using NLTK.

Text Analytics Workflow

Once the data from Glassdoor is extracted, we must consider (or construct) a systematic methodology for analyzing the text. In figure 2 below, we provide the framework that we follow for this analysis. Armed with this framework, we perform text processing for data cleaning as described in previous section, which is in turn used as source for subsequent analysis. Once the basic exploration is done on the prepared corpus, sentiment analysis will be performed to assign the polarity of each review.

Later, n-grams are extracted from text (both pros and cons) along with its corresponding sentiment score. Top bi-grams and tri-grams are identified by the frequency of occurrence in the corpus for each company as these are the most talked in the reviews. Next, we perform Similarity score analysis to identify unique n-grams and zero in top 5 positive and negative n-grams by average sentiment score. The results are then aggregated on a comprehensive dashboard.

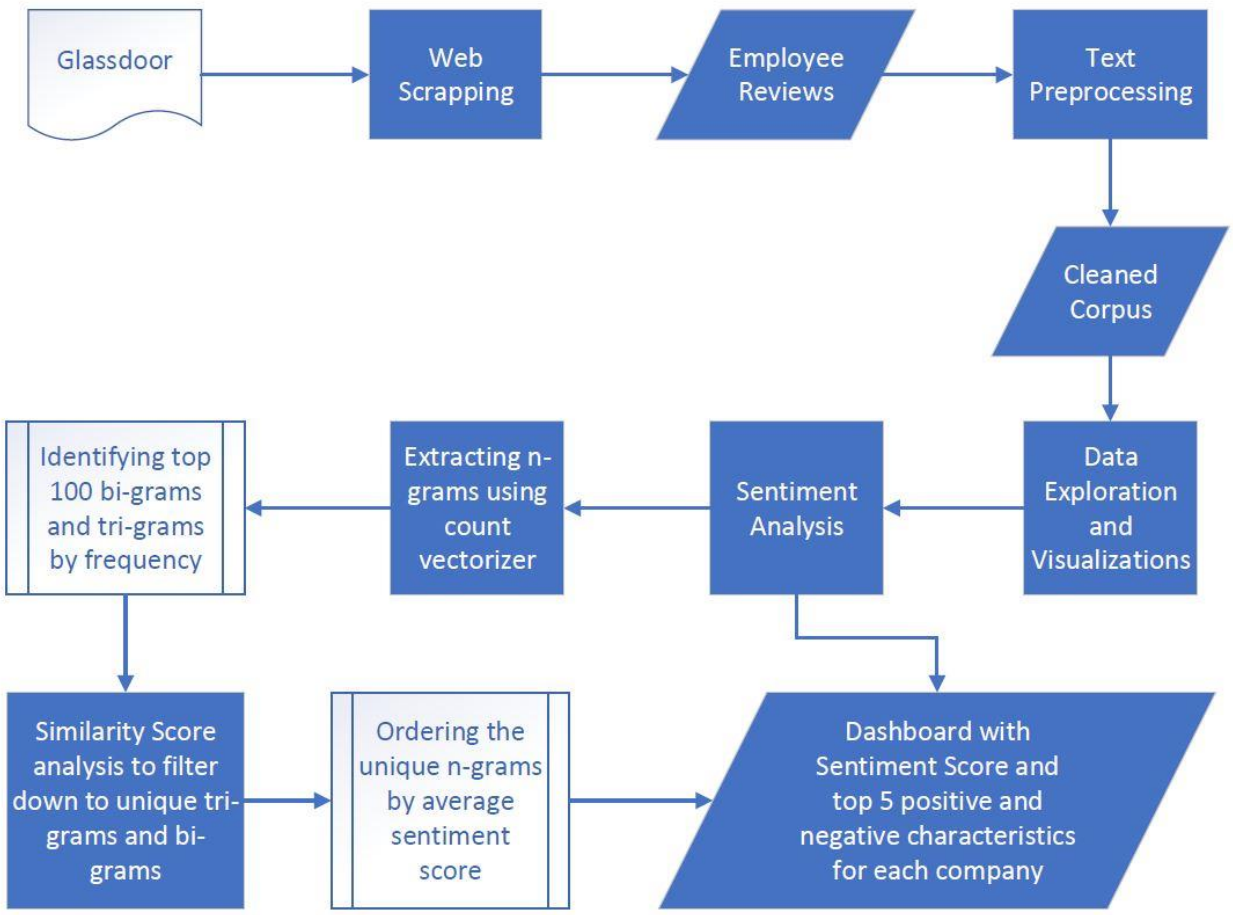


Figure 2: Flow chart depicting the process flow for the study

Exploratory Data Analysis

We tried to explore the data we have gathered to gain more insights into the data before doing any textual analysis. In this process, we tried to get answers for below questions;

- Number of reviews per company
- Rating distribution by company
- Word count in reviews
- Polarity score by company

1. Number of reviews per company

The figure 3 below shows the total number of reviews by company along with the distribution between Current, former and employees with no status. We can observe that Amazon has a lot more reviews compared to other firms followed by IBM, Microsoft and Google. Also, among those categorized as current and former employees, current employees have more reviews for all the companies.

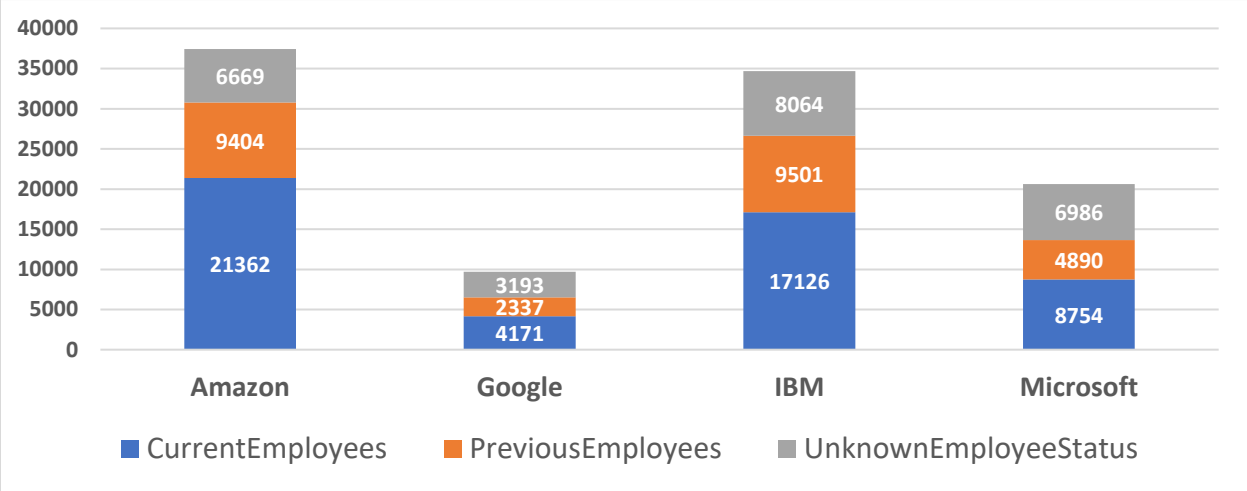


Figure 3: Distribution of reviews by company

2. Rating distribution in the companies

Looking at the rating distribution by company on a scale of 1-5 (5 being the best) in figure 4 below, we can conclude that all the companies were predominantly rated either 4 or 5. Interestingly, Microsoft is the only firm with more 4 ratings than 5 ratings.

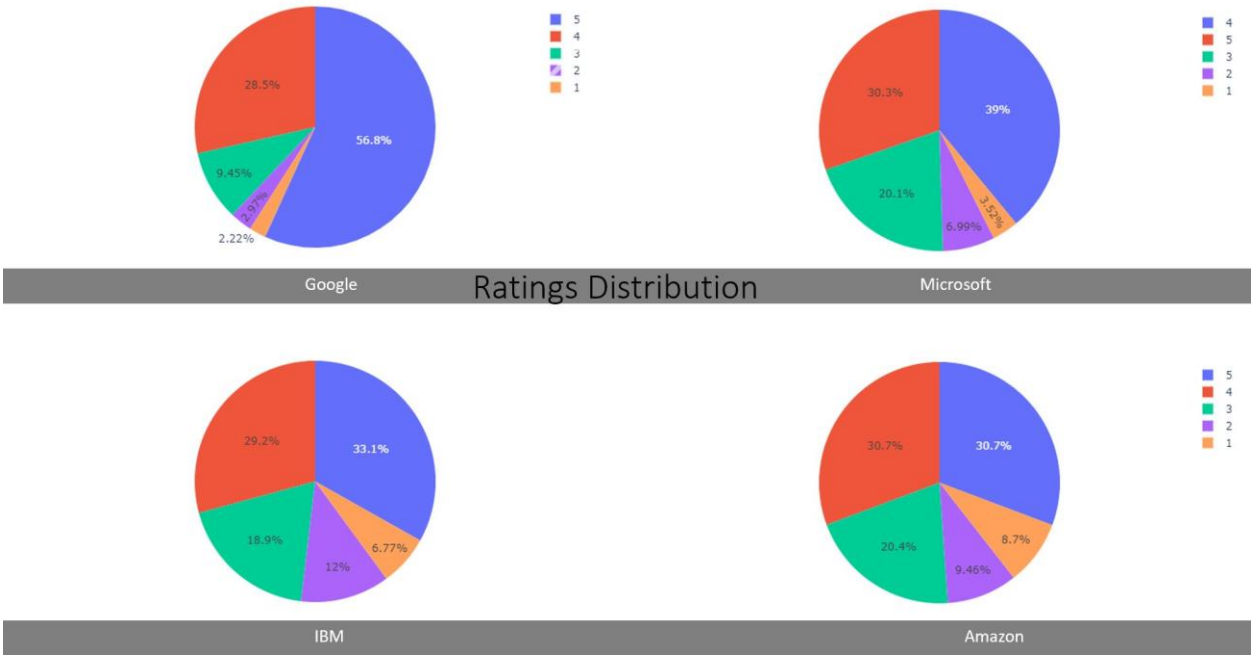


Figure 4: Rating distribution by company

3. Word count in reviews

We wanted to see how long (in words) the reviews were typically to understand if we can infer some meaning from the text. A review with just few words would not have helped the analysis as most of the words would be stop words or words with no real meaning. But much to our comfort, the reviews were not too short with most falling under the range of 5-15 words for both pros and cons. Figure 5 and 6 show the distribution of words in pros and cons respectively.

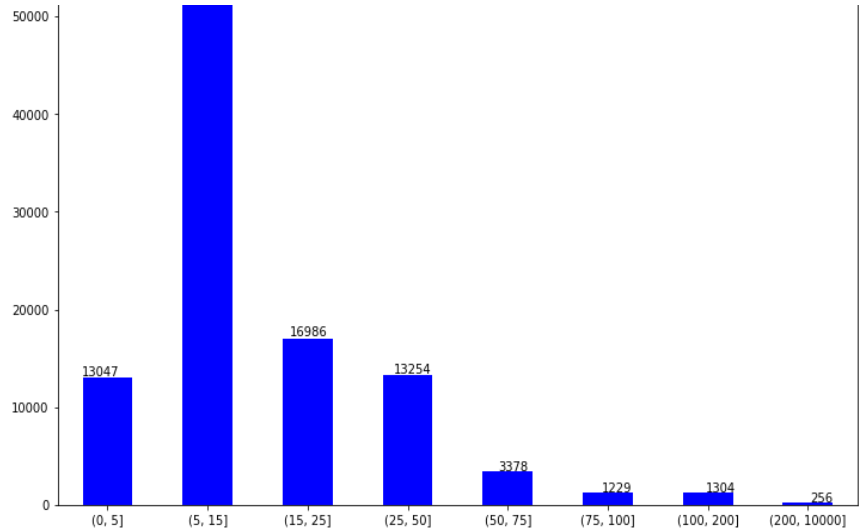


Figure 5: Word count in pros text

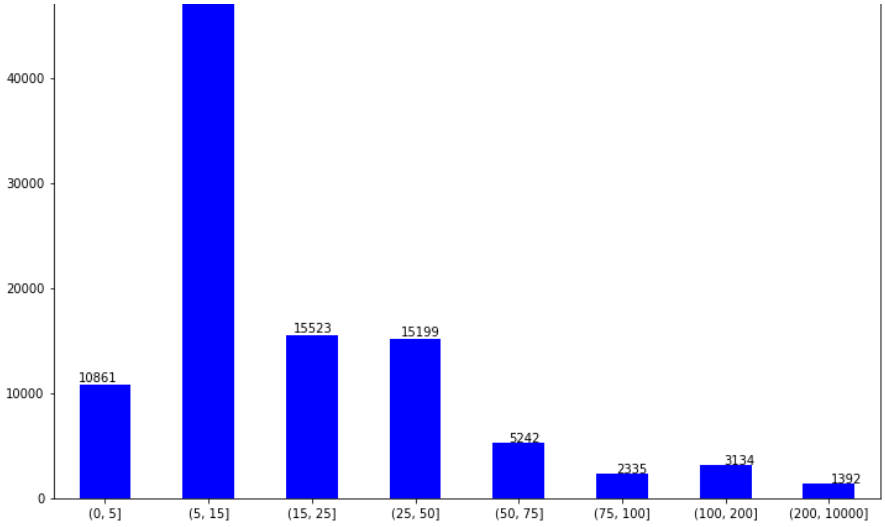
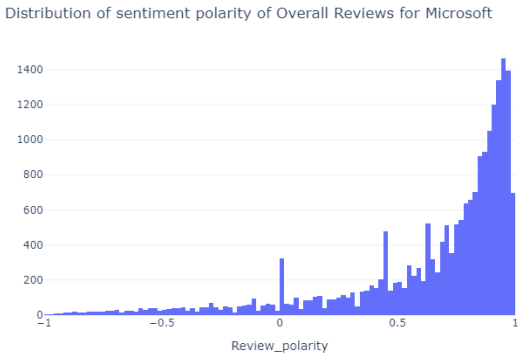
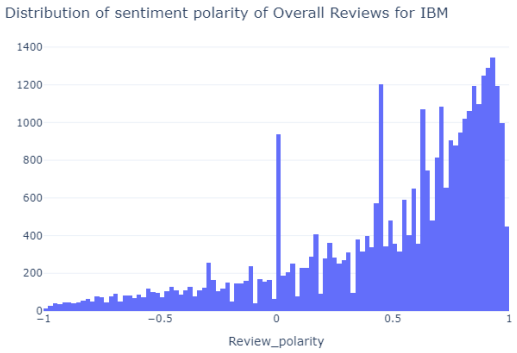


Figure 6: Word count in cons text

4. Polarity score distribution by company

Another important analysis we wanted to explore was the distribution of sentiment score for each company. Consistent to the ratings, the sentiment score was predominantly on the positive side (Figure 7). Although, there were some anomalies, the ratings were consistent with the polarity distribution. In other words, reviews which had good ratings (4 or 5) had high positive polarity score. Figure 8 shows the distribution of polarity score for reviews that were rated 3 or below. Ideally, we would have liked to see a lower polarity score for these but almost 19,000+ reviews show high polarity score of >0.6.





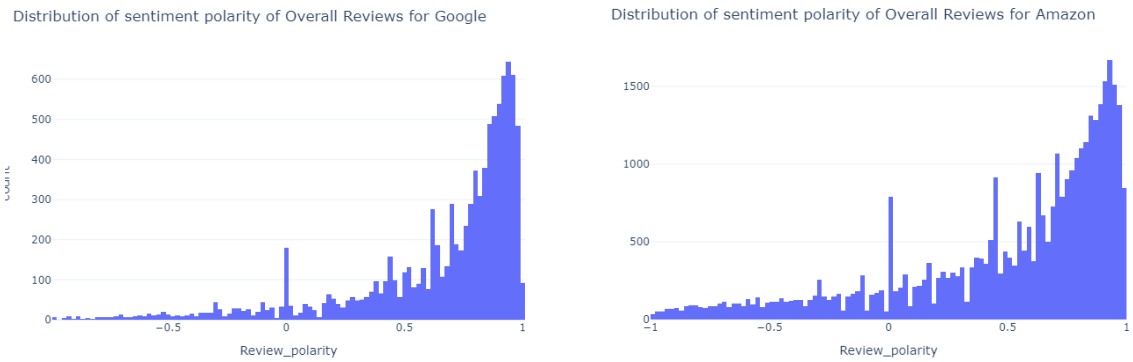


Figure 7: Polarity score distribution for each review by company

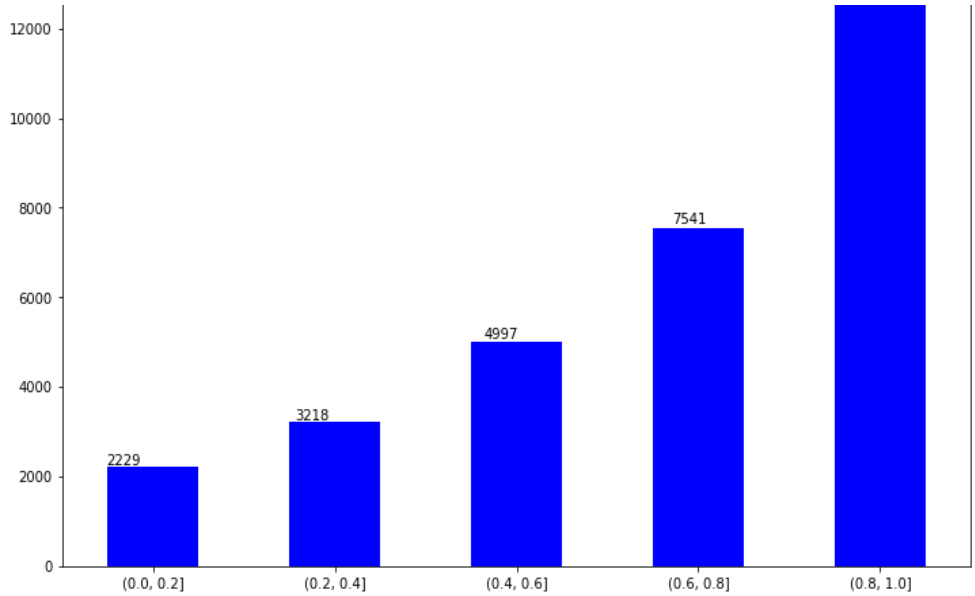


Figure 8: Polarity score of reviews that were rated 3 or below.

Choice and Rationale for Text Analytics Methods and Results

1. Sentiment Analysis

After cleaning the text and performing exploratory analysis, we decided to do sentiment analysis to achieve two purposes. The primary purpose was to estimate the employee satisfaction with the organization from the reviews. Secondly, to check if the ratings provided against the reviews were reflecting the emotion or sentiment in the textual content of the review.

To achieve these objectives, we tried to first understand which library was working better by passing some reviews to NLTK and TextBlob. One key observation was that TextBlob was

failing when it was used on negative reviews. The figure below shows the sample reviews on which TextBlob and NLTK were applied and the resultant sentiment score. We can see that TextBlob returns high positive sentiment score for negative score whereas NLTK does a better job with the same reviews. So, we decided to use NLTK for our sentiment analysis.

```
In [157]: TextBlob('Lack of good "people managers", Troublesome process to recognize and award achievements, LAck of cross-group
              collaboration.').sentiment.polarity

Out[157]: 0.7

In [158]: analyzer.polarity_scores('Lack of good "people managers", Troublesome process to recognize and award achievements
                                   , LAck of cross-group collaboration.')['compound']

Out[158]: -0.128

In [159]: analyzer.polarity_scores('not as great salary as competitors')['compound']

Out[159]: -0.5096

In [160]: TextBlob('not as great salary as competitors').sentiment.polarity

Out[160]: 0.8
```

Figure 9: Using TextBlob and NLTK for sentiment score analysis on negative reviews

2. Identifying n-grams as Characteristics of a Company

We chose to do Frequency-based vectorization and are using CountVectorizer for extracting bi-grams and tri-grams from the text to identify the characteristics of the company which are most frequently used in the reviews. The reason we opted for bi-grams and tri-grams in place of single words was to preserve the meaning of the characteristic of a company as ‘Work life balance’ or ‘good pay’ has a completely different meaning to single words such as ‘work’, ‘life’, ‘balance’, good’ and ‘pay’.

Previous works have used single words and manually completed the characteristic based on the polarity of the word. For example, if the word ‘hours’ returned a negative sentiment score then the characteristic was manually considered as ‘long work hours’ for a company. To overcome this kind of manual estimation, we have extracted the bi-grams and tri-grams along with the sentiment score. The figure below shows the top 5 bi-grams and tri-grams for one of the companies- IBM. Note that some of the tri-grams and bi-grams are similar (used similarity analysis explained in next section to resolve this) and some actually do not have much meaning (for example- ‘not good’).

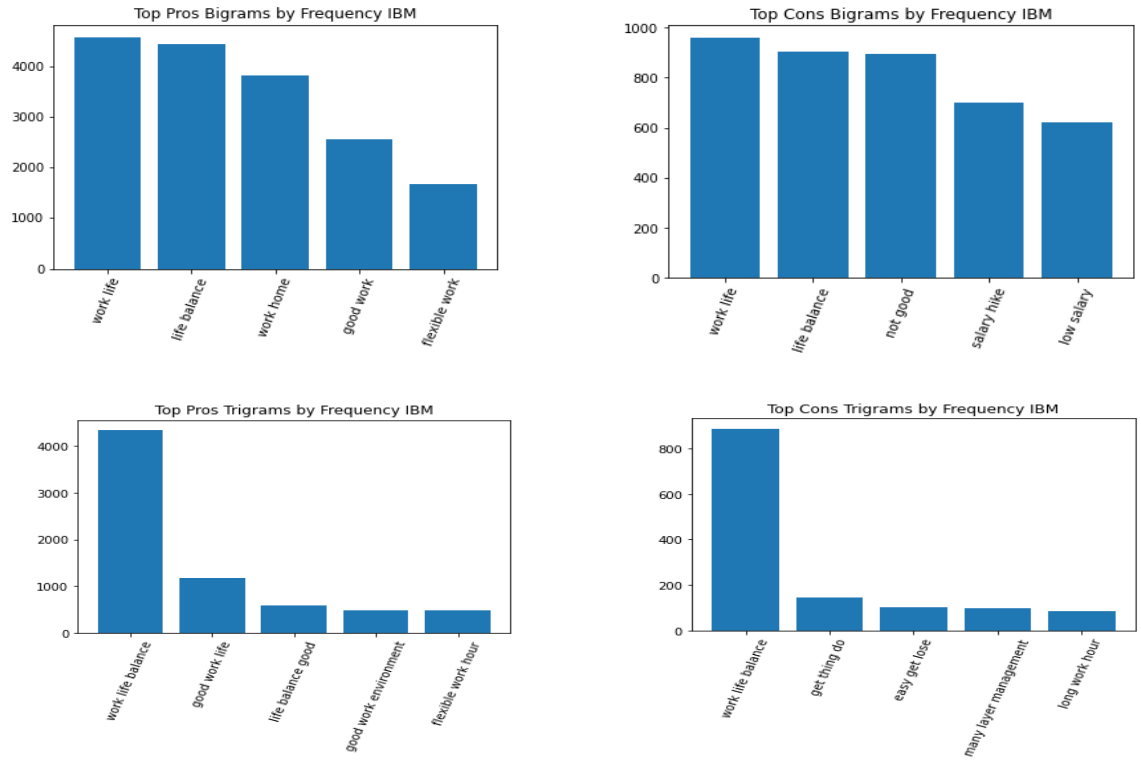


Figure 10: Top 5 trigrams and bigrams by frequency for IBM

3. Similarity Score Analysis

Once the tri-grams and bi-grams were extracted, we used the top 100 by frequency from both pros and cons to conduct similarity analysis and further zero-in on unique bi-grams or tri-grams which were having the most positive and negative effect on the employees satisfaction. For this purpose, we used cosine similarity in order to combine the tri-grams and bi-grams which were similar. Figure 11 shows the grouping of words based on cosine similarity for a company. We can see that words that mean the same are grouped into one parent word.

This analysis was an improvement on previous works that tried to extract keywords that had the most impact on employee satisfaction. Using this we were able to combine different forms of the same characteristic and find the ‘Employer Brand’ characteristic which had the most positive and negative effect on employees.

Bi Grams

'big company': ['big company', 'many people', 'not many', 'company not', 'may not', 'manager not', 'long time', 'become big', 'nothing bad', 'company work']

'low salary': ['salary hike', 'large company', 'less salary', 'low pay', 'salary increment',
'last year', 'not really', 'salary low', 'many layer', 'lay offs', 'salary less', 'less pay']

'life balance': ['fast pace', 'life balance', 'leadership principle', 'decent pay',
'flexible work', 'benefit great', 'best place', 'full time', 'place learn', 'feel like', 'talented people']

'free food': ['free food', 'work life', 'people work', 'good benefit', 'perk benefit', 'food great', 'great food', 'flexible work', 'good perk', 'good people', 'people good', 'food good',
'food gym', 'care employee', 'environment good', 'benefit work', 'perk good', 'perk free', 'perk food']

Tri-Grams

good work environment': ['good work environment', 'great work environment', 'learn new thing', 'work environment good', 'work culture good', 'work hard fun',
'lot opportunity learn', 'nice place work', 'good place learn', 'bring dog work', 'benefit good pay', 'good people work', 'opportunity learn grow',
'lot opportunity grow', 'customer centric company', 'learn new technology', 'good environment work', 'lot opportunity move', 'work home option',
'lot learn opportunity']

'long hour work': ['work long hour', 'hour short break', 'balance not good', 'not enough hour', 'hard work not', 'work hour per', 'not great work']

'life balance challenge': ['life balance challenge', 'great place work', 'mid level management',
'life balance difficult', 'long term career', 'many middle manager', 'good people leave', 'lack work life']

Figure 11: Grouping of similar bi-grams and tri-grams using Similarity score analysis

4. Results

We tried to aggregate all the findings into a dashboard that can be used as single source of truth by both, candidates exploring job opportunities looking for feedback on employers and employers evaluating their employer brand and employee satisfaction levels.

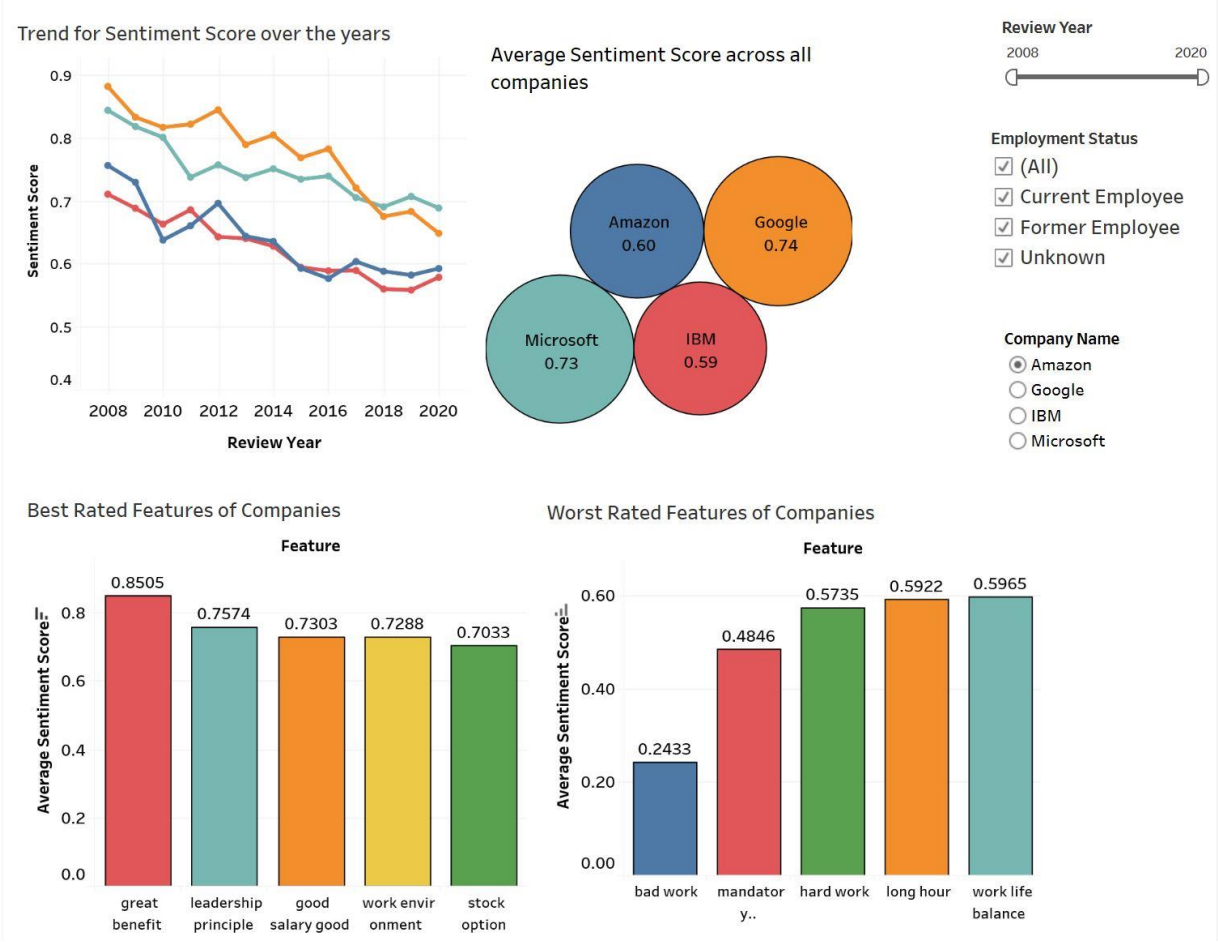


Figure 12: Dashboard showing the results

Link to the tableau dashboard:
https://public.tableau.com/profile/sagar.surendra.kulkarni#!/vizhome/Text_Analytics_Updated/Dashboard1

Key Insights

Some of the key insights drawn from this study are;

1. All the companies have seen a decline in employee satisfaction levels from 2008, although there were some years when the satisfaction levels improved compared to preceding years

2. Google and Microsoft have high average employee satisfaction followed by Amazon and IBM

3. Employer characteristic which had the most positive effect on employee satisfaction-

Amazon – Great Benefits

Google – Free Food

IBM – Great People

Microsoft – Great Pay Benefit

4. Employer characteristic which had the most negative effect on employee satisfaction-

Amazon – Bad Work

Google – Bad Work Life

IBM – Hike Not Good

Microsoft – Life Balance

Actions recommended;

1. Companies can look at some of their characteristics which are drawing highly negative reviews from employees and strive to make them better or create a positive attitude in people towards those characteristics. For example, employees in Microsoft have high negative impression of their 'Performance Review' process. So, Microsoft can look at the communications around this particular process and look inward to address any concerns from employees

2. Candidates can look at their preferences and see if their expectations will be met in a particular organization. For example, if a candidate is looking for 'flexible work hours' for taking care of personal responsibilities they would be better of look at positions in Microsoft as employees reviewed it as one of the best characteristics of the company

These are some of the ways both companies and candidates can make use of the findings from this study.

References

1. What Is Employer Branding and How Can It Grow Your Business? (https://business.linkedin.com/talent-solutions/blog/employer-brand/2018/employer-branding)

2. 10 Reasons Why Employer Branding Is Important (https://linkhumans.com/employer-branding-important/)

3. Employee Satisfaction and Corporate Performance: Mining Employee Reviews on Glassdoor.com (https://pdfs.semanticscholar.org/b784/71ca2ac990e3ab2d70283e42e2bb5d3a8f7a.pdf)

4. Improving Employee Satisfaction Through Text Analytics (https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-243_Final_PDF.pdf)

5. Text Mining Company Reviews (https://mguideng.github.io/2018-07-16-text-mining-glassdoor-big3/)

6. Influence of support leadership and teamwork cohesion on organizational learning, innovation and performance: an empirical examination (https://www.sciencedirect.com/science/article/pii/S0166497204000914)

7. Employer Branding for Dummies (http://resources.glassdoor.com/rs/glassdoor/images/employer-branding-for-dummies.pdf)

Appendix

1. Kaggle kernel link where the code can be executed: https://www.kaggle.com/rohan74/reviews-glassdoor

2. GitHub repository link where the web scraping code, raw corpus, cleaned corpus and tableau dashboard source files are stored: https://github.com/rohan74/Glassdoor-Review-Analysis

3. A python notebook is also attached with this document but request to use Kaggle kernel to execute the code as python notebook might take a long time to run the text preprocessing block