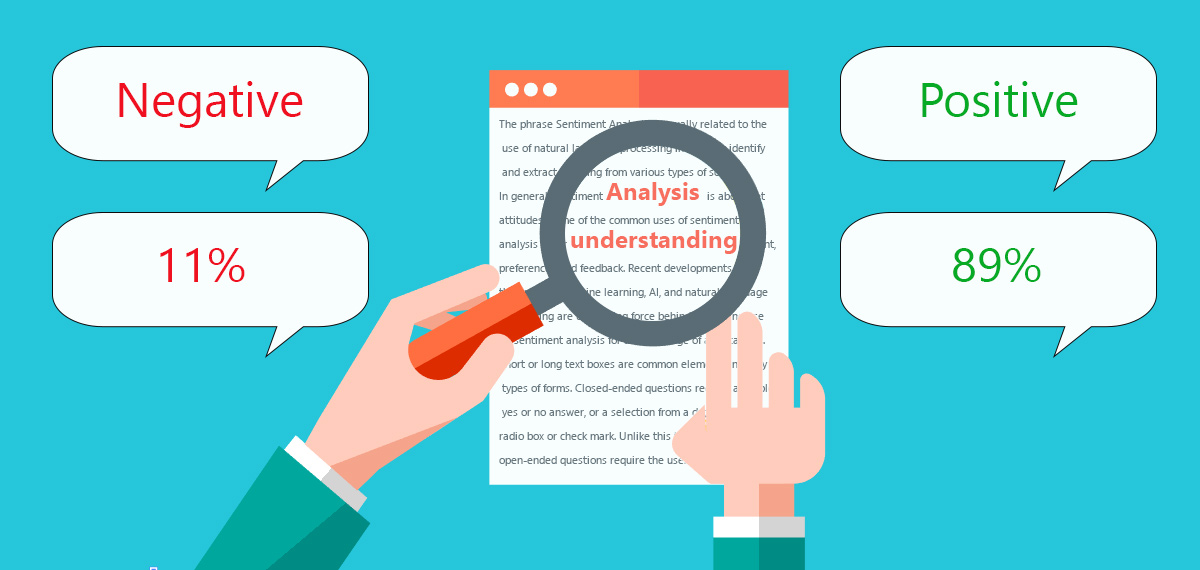
**Sentimental Analysis of Employee Review**

****

**Group 17: -**

Rakesh Jain

Sanket Bhaud

Ronak Seth

Anuj Rathod

Shivani Malandkar

Dipti Kalathiya

Contents

[**ABSTRACT** 3](#_Toc532682087)

[PROBLEM 3](#_Toc532682088)

[METHODS USED 3](#_Toc532682089)

[RESULTS OBTAINED 3](#_Toc532682090)

[**MATERIALS AND METHODS** 4](#_Toc532682091)

[DATASET DESCRIPTION 4](#_Toc532682092)

[**TOOLS AND TECHNIQUES** 5](#_Toc532682093)

[DATA PRE-PROCESSING 5](#_Toc532682094)

[EXPLORATORY ANALYSIS 6](#_Toc532682095)

[PREDICITVE MODELLING 11](#_Toc532682096)

[TOOLS USED 11](#_Toc532682097)

[**RESULTS & DISCUSSIONS** 12](#_Toc532682098)

[RESULTS 12](#_Toc532682099)

[DISCUSSION 12](#_Toc532682099)

[NLP 13](#_Toc532682099)

[**REFERENCES** 13](#_Toc532682104)

**ABSTRACT**

PROBLEM

The main reason for choosing this dataset was to get an insight on why employees from big companies like Google, Amazon, Microsoft are refraining from working there, is it their working culture or politics inside the company or the way they treat their employees. Through some detailed insights about the company from their recent or current employees, it can help the new joiner to decide whether is it the right place to start their career with.

METHODS USED

The methods used focus on cleaning the data followed by preliminary analysis, comprising heat maps and bar graphs. Also, the data contains many categorical variables, so multi class classification techniques are used for predictive modelling which predicts whether the employee is satisfied or not.

TARGET AUDIENCE

This dataset can be useful for both, The Organization and the User.

From Organizational Point of View: - Predicting what are the chances that the current employee will either stay/leave the organization. To improve the organization’s working culture or work-life balance, such reviews and the results will help them to spot a place for improvement. Employee Feedback is important for the organization.

From User Point of - It can help the new joiner to decide whether is it the right place to start their career with.

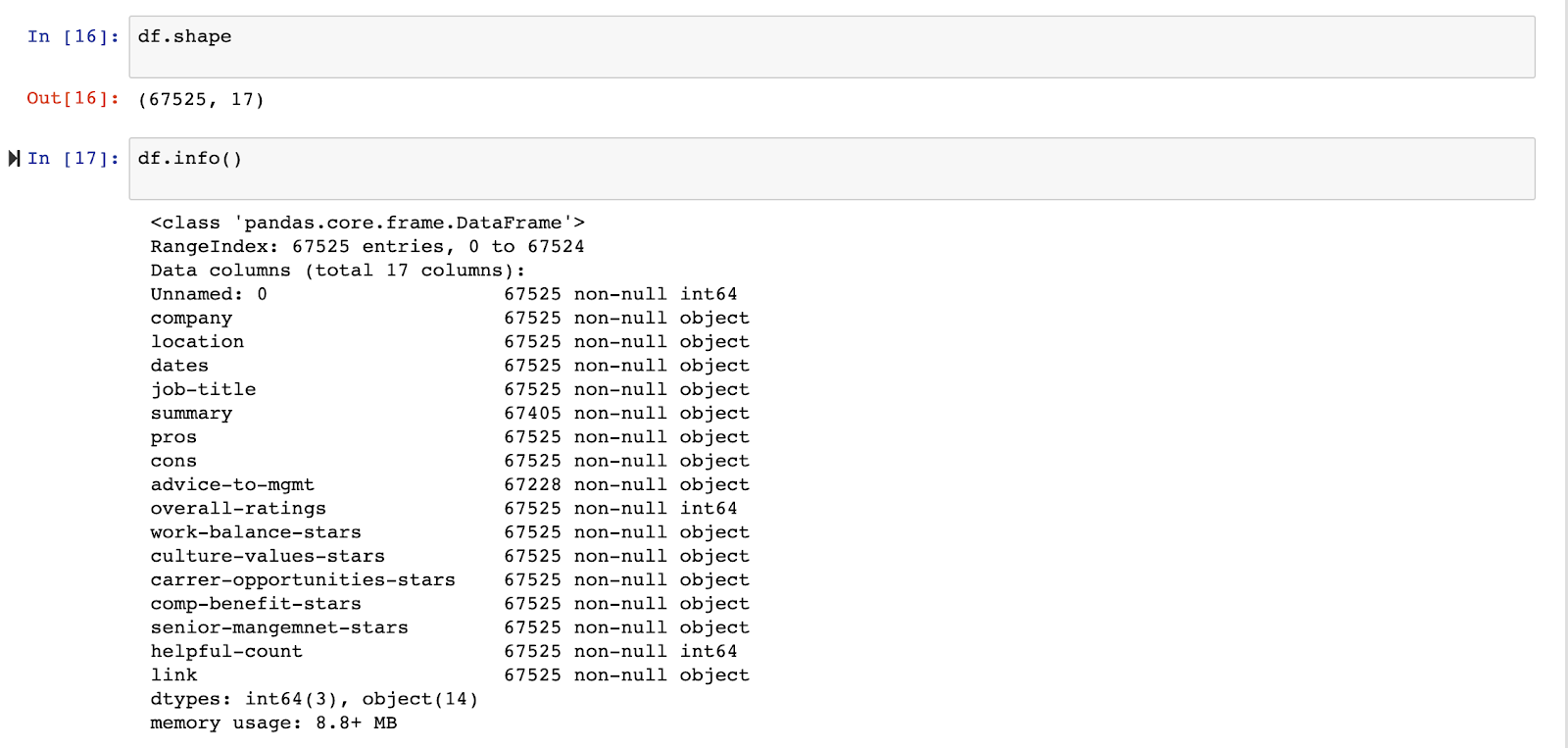
RESULTS OBTAINED

Our analysis mainly focuses on finding a suitable model to predict whether the employee is satisfied or not.

**MATERIALS AND METHODS**

DATASET DESCRIPTION

1. Index: index
2. Company: Company name
3. Location: This dataset is global, as such it may include the country's name in parenthesis [i.e. "Toronto, ON(Canada)"].
4. Date Posted: in the following format MM DD, YYYY
5. Job-Title: This string will also include whether the reviewer is a 'Current' or 'Former' Employee at the time of the review
6. Summary: Short summary of employee review
7. Pros: Pros
8. Cons: Cons
9. Overall Rating: 1-5
10. Work/Life Balance Rating: 1-5
11. Culture and Values Rating: 1-5
12. Career Opportunities Rating: 1-5
13. Comp & Benefits Rating: 1-5
14. Senior Management Rating: 1-5
15. Helpful Review Count: A count of how many people found the review to be helpful
16. Link to Review: This will provide you with a direct link to the page that contains the review. However, it is likely that this link will be outdated



**TOOLS AND TECHNIQUES**

The methods used to build the model are based on predictive modelling and exploratory analysis.

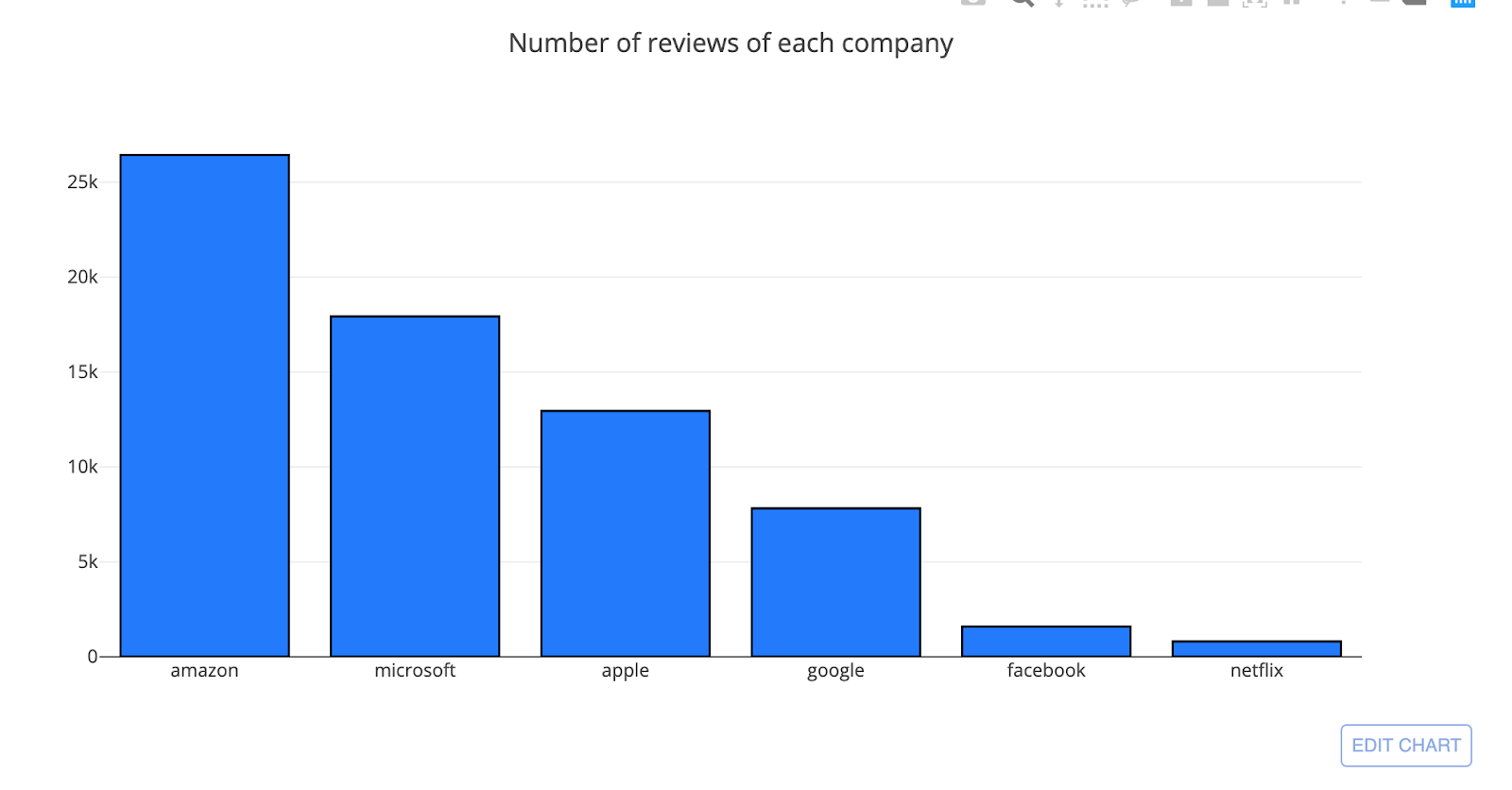
DATA PRE-PROCESSING

The Steps used for data pre-processing are as follows: -

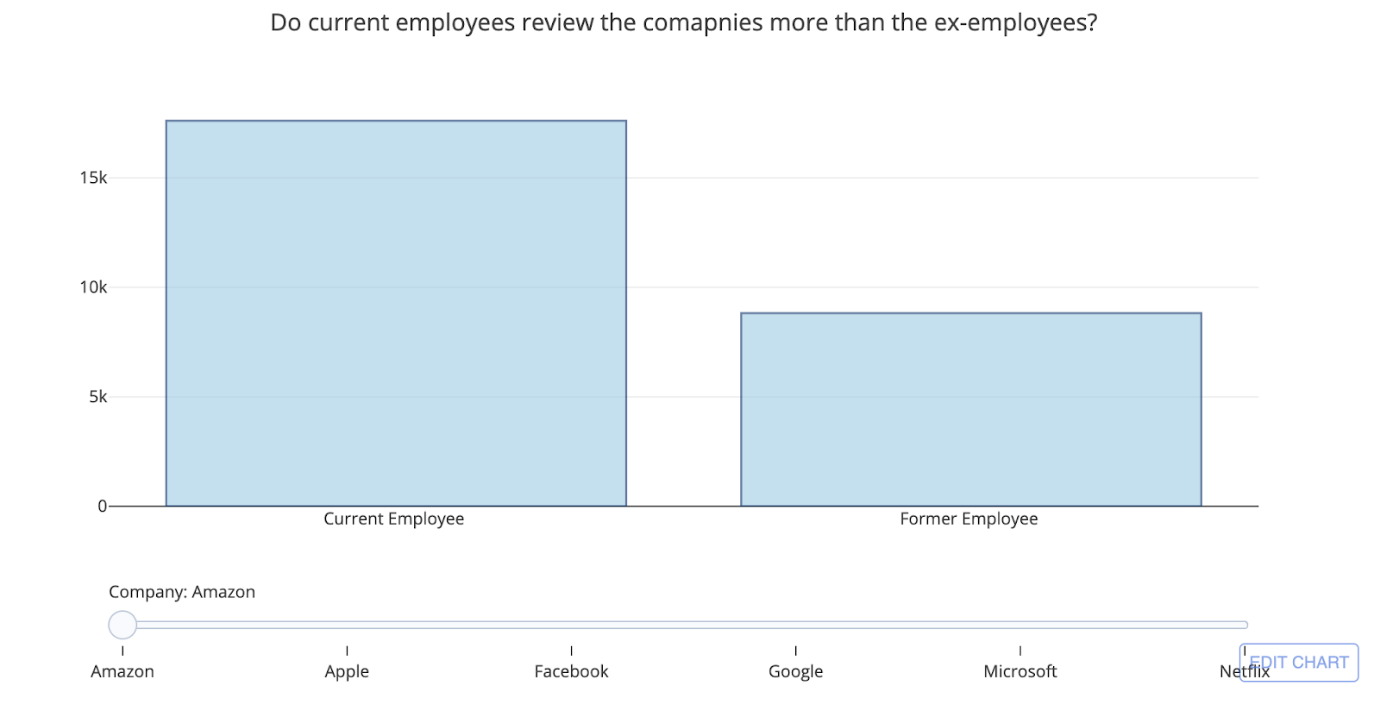
* The Date column is being formatted in YYYY-MM-DD format using DateTime method.
* The Job Title column is being split into Job Type (Current or Former employee) and Job Title (Manager, Software Engineer, etc)
* The Rating columns were first transformed into strings after which we applied round function and later converted into an integer for better predictions.
* The Text cleaning was done by removing null values, spaces, numbers and so on using regex function.
* Removed duplicate content and lemmatized the data using Lemmatization technique.

EXPLORATORY ANALYSIS

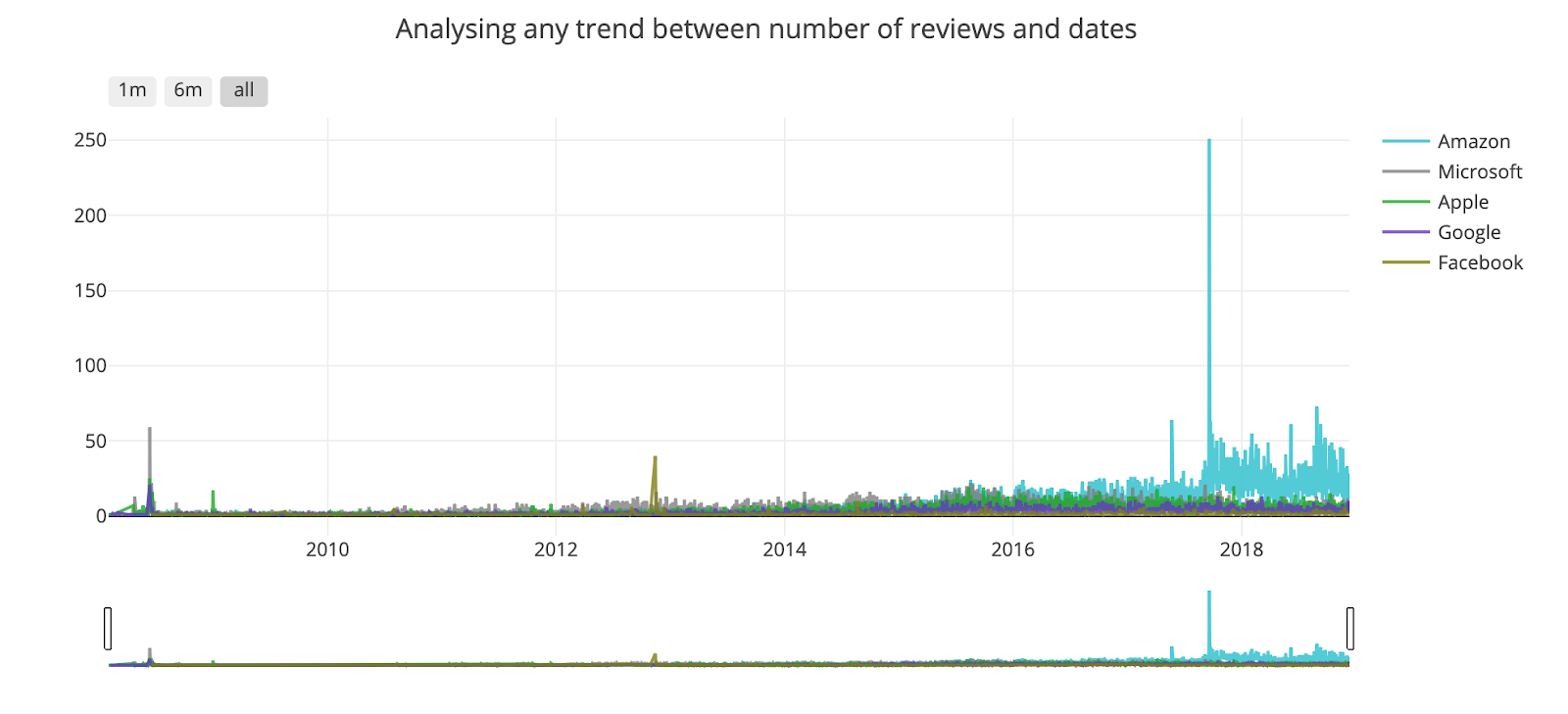
* **Number of reviews of each company -** In this graph, we have plotted the number of reviews each company has received and sorted them in ascending order



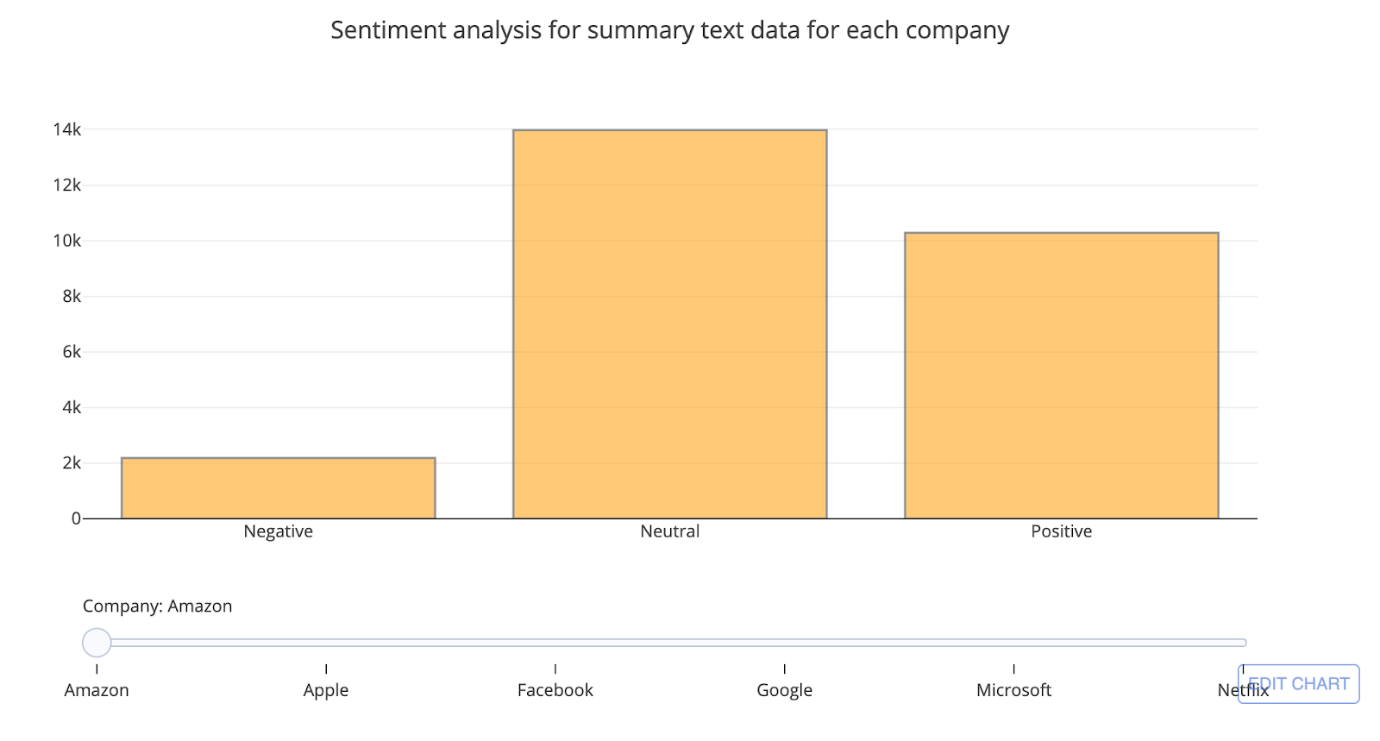
* **Which type of employees are commenting about the company most -** In this graph, we find out the which types of employees (whether current or former) are commenting about the company. We have used iploty’s seek bar feature to display the count of employee’s company wise.



**What is the trend in comments** - This time series graph shows at what time of the year most employees give their reviews. This is important from companies point to view to rectify whether the reviews are coming during the promotion period or after that.



**Sentiment analysis on the employee's comments -** In this graph, we find out whether the comments given by the employee in the summary section is Negative/Positive or Neutral comment. We have used iploty’s seek bar feature to display the count of employees who gave comments company wise.



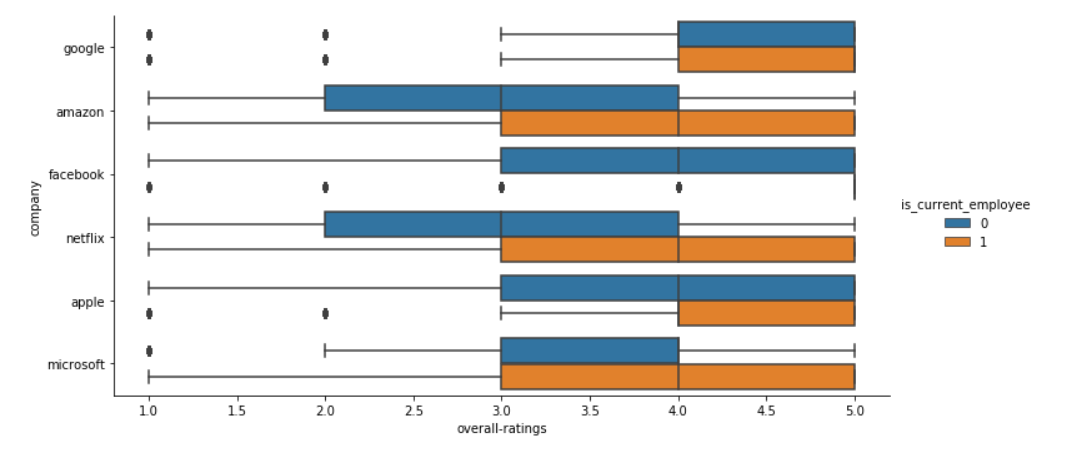
**Various ratings as per company -** This graph gives information about which company is best in work-life balance, culture values, career opportunities



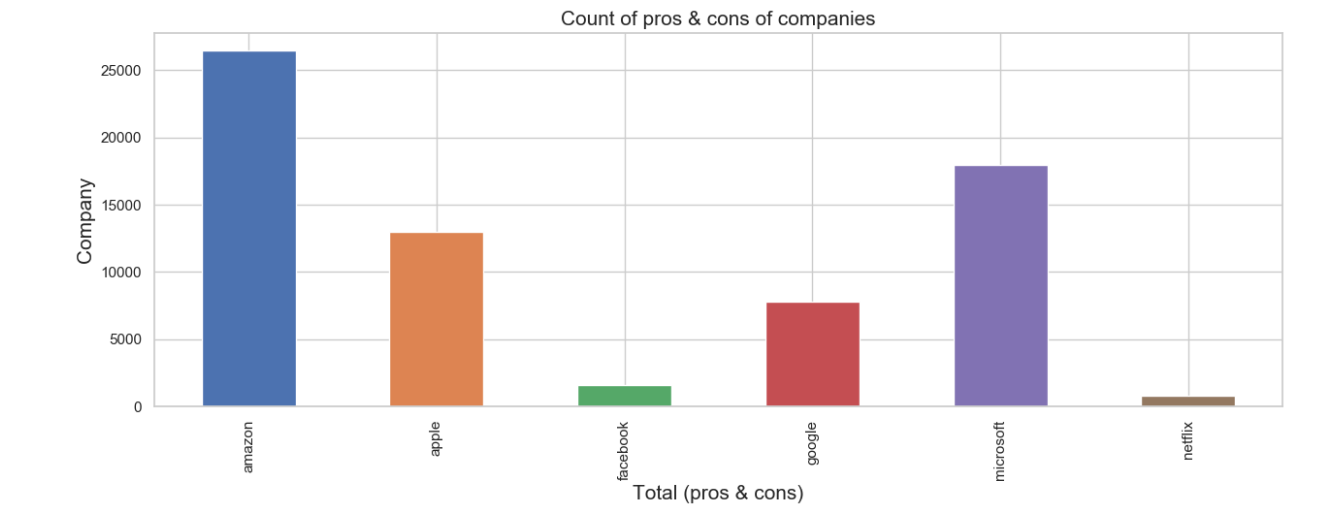
**Word Clouds -** This is the word cloud for Amazon which depicts what people are talking about the company the most.



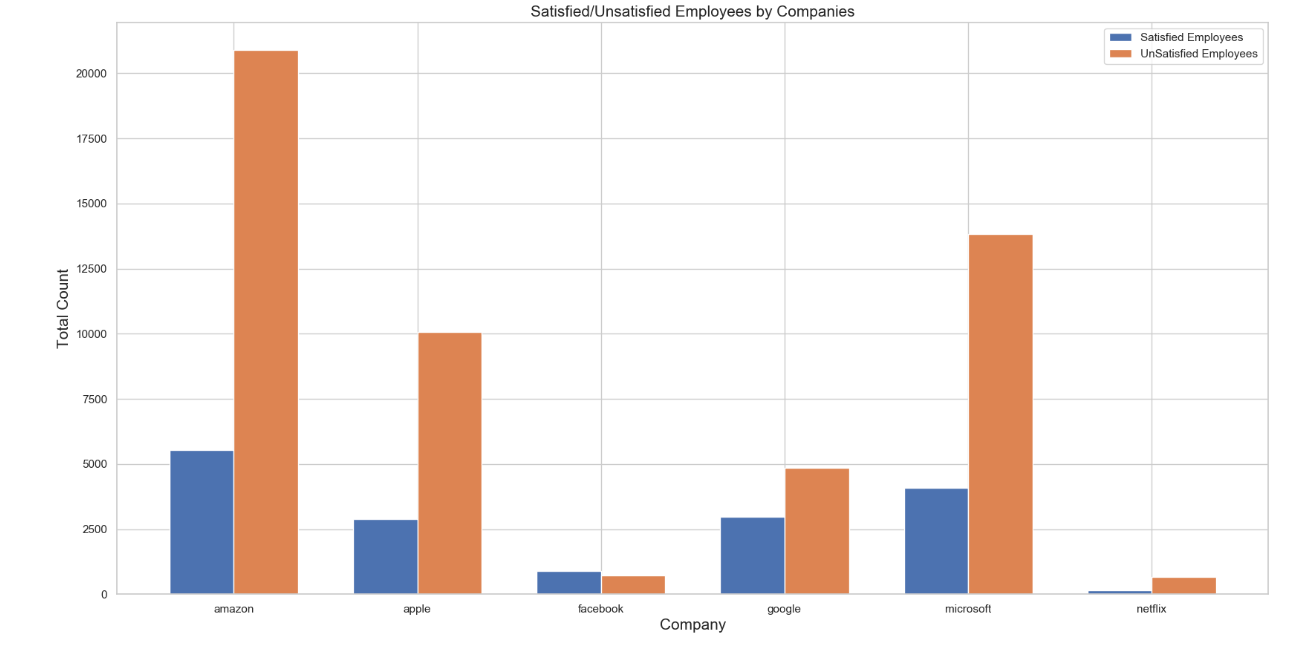
**Box-Plot for overall ratings for each company -** The graph shows that the overall rating is best for Google & Facebook and worst for Netflix. The ex-employees, as well as the former employees, have given good ratings for both these companies. If we consider the reviews given by the current employees of these two companies, almost all of them fall under 3 to 5 and maximum out of them fall under 4 to 5. Like the dots show, very few of the employees have given 1 or 2.



**Total pros and cons per company -** This graph shows the total number of pros and cons (reviews) that are being given by the employees and certainly, we can see that Amazon has the highest number of reviews followed by Microsoft. The reason can be anything for such high numbers which will be explained by heat maps.



**Total Satisfied/Unsatisfied employees by companies -** This graph shows the total number of satisfied Vs Unsatisfied employees in each company and we can clearly see that there are many unsatisfied employees in Amazon followed by Microsoft.



**Overall ratings using Heatmaps** - Assuming 4 and above to be very good rating, and less than 4 to be not so good, some observations from the above plot are:

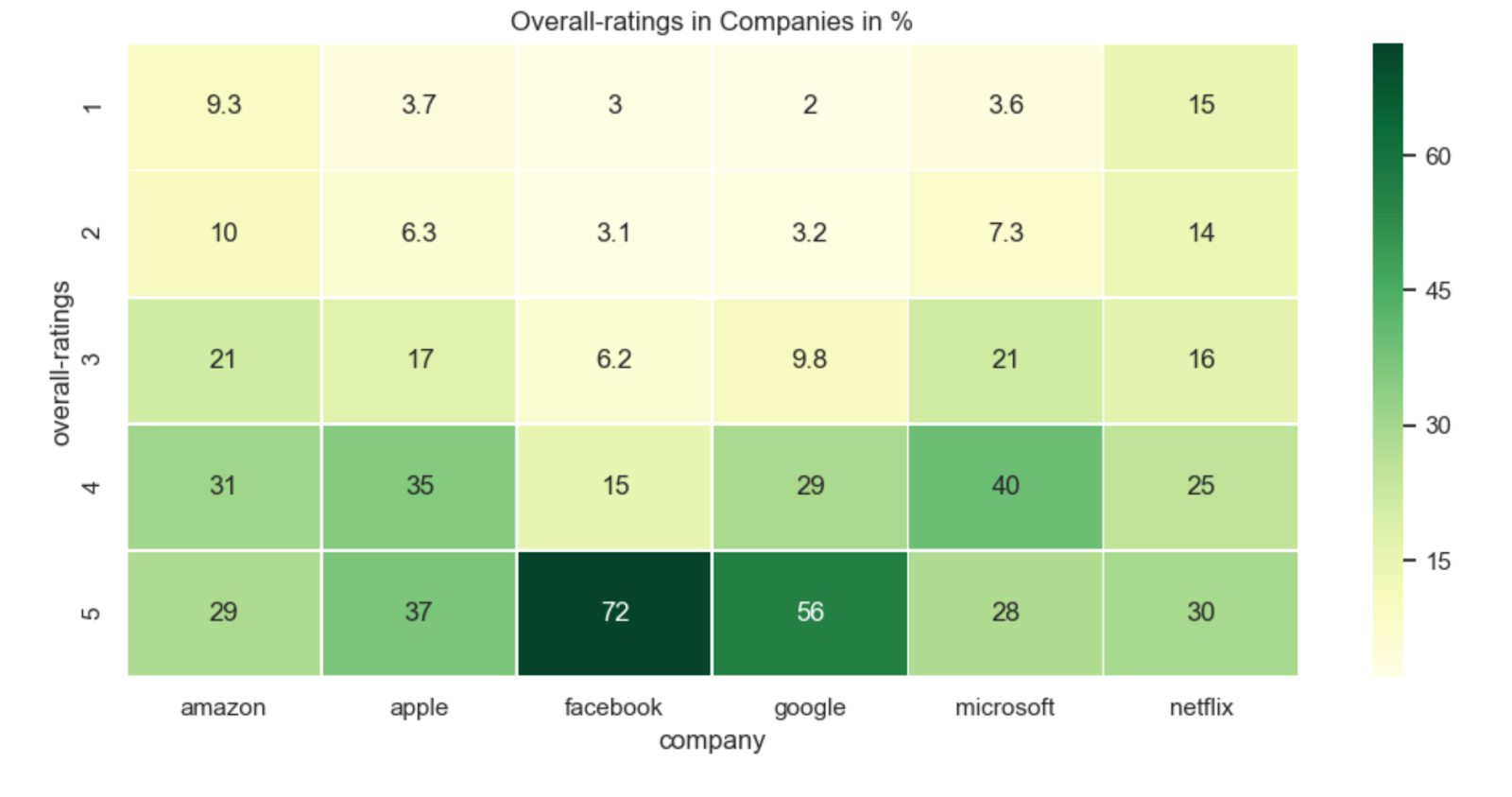
72% of people in FB have rated it 5 and overall 87% of people are very happy working in FB. This is followed by Google with 85%

NetFlix has got a 55-45% which means people have neutral say about the company

60% of people are happy with Amazon

72% of people are happy with Apple

68% of people are happy with Microsoft



PREDICITVE MODELLING

Based on the preliminary findings, the problem reduced to a classification problem to predict the primary crime type. The various methods we used for predictive modelling are SGD, Random Forest, KNN and Decision Tree classifier.

TOOLS USED

Python***.***

**RESULTS & DISCUSSIONS**

RESULTS

|  |  |  |
| --- | --- | --- |
| MODELS | TRAINING ACCURACY | TESTING ACCURACY |
| SGD | 75.66% | 75% |
| KNN | 81.28% | 76.76% |
| DECISION TREE | 83.26% | 76.87% |
| RANDOM FOREST | 82.74% | 76.24% |
| NAÏVE BAYES | - | 75.28% |
| LOGISTIC REGRESSION | - | 76.23% |

DISCUSSION

* The reason we’ve used **SGD** is that the output variable Y which is “Remarks” has two outcomes 1 = Satisfied and 0 = Unsatisfied thus classification problem works best for our output variable. As the updates in weight are done in batches on the training dataset the prediction is more accurate for the testing dataset.
* The reason we’ve used **KNN** is that the output variable Y which is “Remarks” has two outcomes 1 = Satisfied and 0 = Unsatisfied thus classification problem works best for our output variable. As we have data points separated into several classes knn works best to predict the classification of a new sample point in the output variable. Also, as there are no independent variables the algorithm works quite efficiently to give better predictions.
* We’ve used **Decision Tree** because the output variable Y which is “Remarks” has two outcomes 1 = Satisfied and 0 = Unsatisfied thus classification problem works best for our output variable. The data points are separated into several classes and Decision tree works best to predict when the data is numerical which our dataset has and thus helps in the classification of a new sample point in the output variable. To avoid overfitting, we have used limited variables and there is not much noise in the testing data.
* The reason we have used **Random Forest** is that as it does not require parameter tuning and it creates tree iteratively and summing them up to give the best output in terms of accuracy. As Random forest does overcome overfitting which could be the case in Decision tree this was the main reason to use Random forest to our dataset.

NLP

* After Performing text pre-processing on “Summary” column we applied Naive Bayes and Logistic Regression models to predict whether the employee is satisfied or unsatisfied.
* We had divided the data into training and testing in the ratio 80-20 respectively with satisfied/unsatisfied as our dependent variable.
* We transformed the Summary text into vectors by applying TF-IDF vectorizer and keeping the max\_features to 1500.
* We then fitted MultinomialNB algorithm on our training and testing data and got an accuracy of around 76 %.
* We then performed the same steps and fitted Logistic Regression model which gave us the accuracy of 76.40%.

**REFERENCES**

* <https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews>
* <https://plot.ly/#/>
* <https://scikit-learn.org/stable/>
* <https://www.kdnuggets.com/2018/03/text-data-preprocessing-walkthrough-python.html>
* <https://www.pythonforengineers.com/natural-language-processing-and-sentiment-analysis-with-python/>