****

**School of Business**

**OPIM 5671 – Data Mining and Business Intelligence**

**Glassdoor Employee Reviews**  
**Sentiment Analysis**

**Group - 3**

**Aniket Jana**

**Pratiksha Nimesh**

**Kaman Dhami**

**Meghna Ashok**

**Haneesha Balireddy**

**Kushagra Goyal**

**Table of Contents**

[**Introduction** 3](#_Toc102251086)

[**Background and Problem Statement** 3](#_Toc102251087)

[**Data Overview** 3](#_Toc102251088)

[**Data Pre-processing** 4](#_Toc102251089)

[**Data Exploration:** 5](#_Toc102251090)

[**Overall rating by company:** 5](#_Toc102251091)

[**Review Count over the years** 7](#_Toc102251092)

[**Exploration – Amazon Reviews** 11](#_Toc102251093)

[**ANALYSIS** 17](#_Toc102251094)

[**KEY CONCEPT LINKS** 19](#_Toc102251095)

[**Modeling – Amazon Reviews** 20](#_Toc102251096)

[**Modeling – Microsoft Reviews (Supervised)** 25](#_Toc102251097)

[**Modeling – Microsoft Reviews** 30](#_Toc102251098)

[**Conclusion** 37](#_Toc102251099)

[**Recommendation** 37](#_Toc102251100)

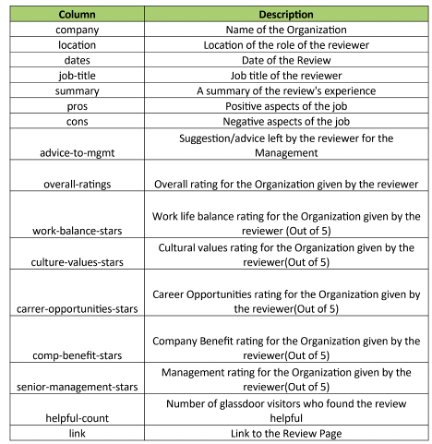
# **Introduction**

# **Background and Problem Statement**

* Companies today rely heavily on modeling techniques to predict employee retention
* This helps companies to be prepared for employee allocation, hiring, and future projects planning
* In our project we perform various text mining models to perform a sentiment analysis on employee reviews for Amazon and Microsoft employees on Glass door
* Our aim is to bucket the reviews into negative and positive reviews along with finding useful insights for the HR department, in order to devise strategies for customer retention as well as improve the company's public relations.

# **Data Overview**

* The data was vast and had a massive 67 thousand records in it.
* The dataset contains employees for companies such as Amazon, Google, Apple, Facebook, Microsoft, and Netflix, ranging between April 2009 and September 2018.
* However, we will be creating a subset of two companies for our project: Amazon and Microsoft.
* Below are the columns in the initial data set, and you can also find the data described below.



* With the data available, we plan to make recommendations to increase employee engagement and retention based on the positive and negative reviews.

# **Data Pre-processing**

* The data were preprocessed before we started modeling. The row was removed for the columns that had either a missing value or a "None" entry.
* If the rating were missing, we imputed them.
* The rescaling of the data has been done twice—one for supervised learning and the other time for unsupervised learning.
* For unsupervised learning, we used the pros, cons, and the rating to s to derive models
* For supervised learning, we rescaled the data for modeling purposes, where, if the rating was greater than or equal to 4, it is kept as a pro. Otherwise, if the rating is less than 4, we consider it a con.

After rescaling, we got the below number of rows:

**AMAZON:**

Pros (1): 8090 rows

Cons (0): 8025 rows

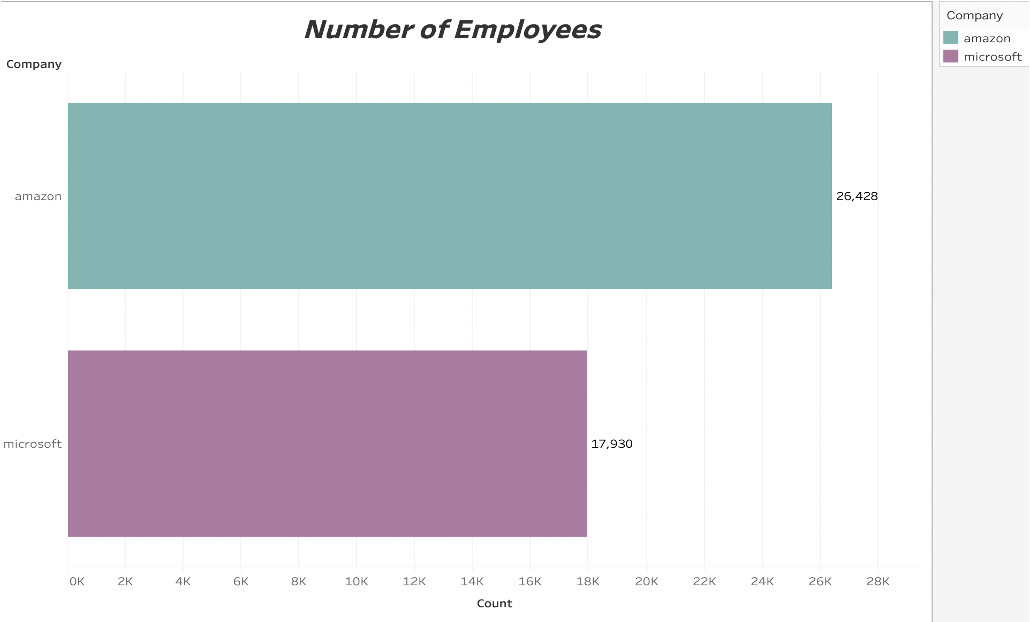
**MICROSOFT:**

Pros (1): 5063 rows

Cons (0): 5766 rows

# **Data Exploration:**

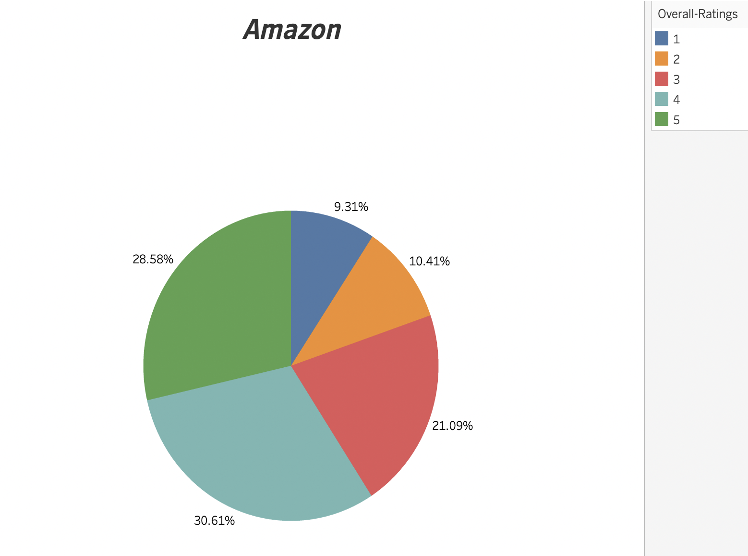
We first looked at the total employees. From the below bar chart, we can see that there are a total of26,428 employees whose reviews were recorded across the years, while the count of Microsoft reviews was 17930.



# **Overall rating by company:**

**Amazon:**

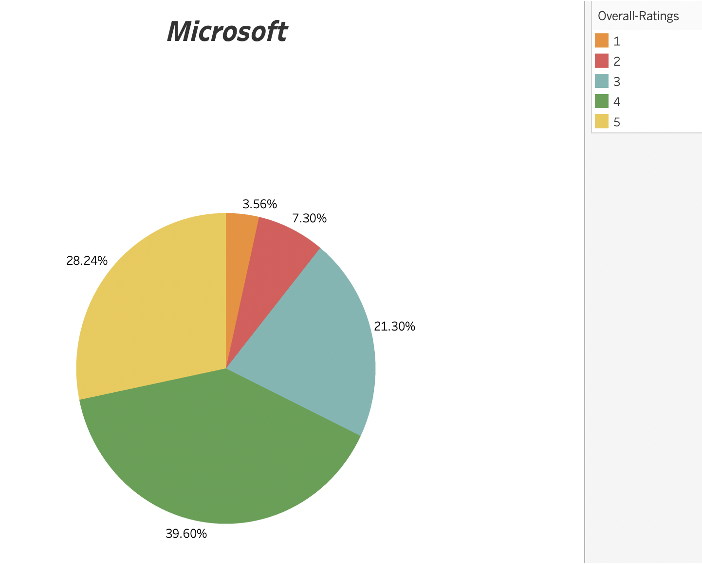
Out of the 26428 reviews, more than 50% the reviews lie in the pros bin suggesting that Amazon has more pros than cons



**Microsoft:**

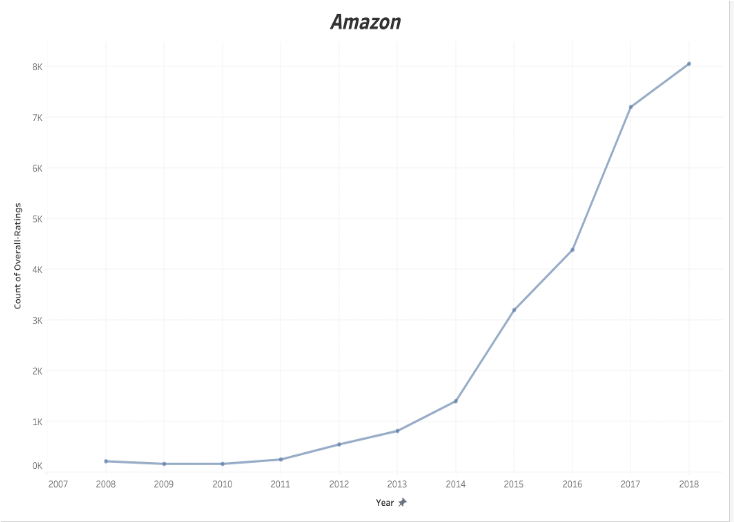
The same is the case with Microsoft. More than 50% of the reviews were equal to or greater than four ratings which tells that Microsoft has more than cons.

But when compared to Amazon, we can say that Microsoft has more pros and hance its employees might be more satisfied than their counterparts.

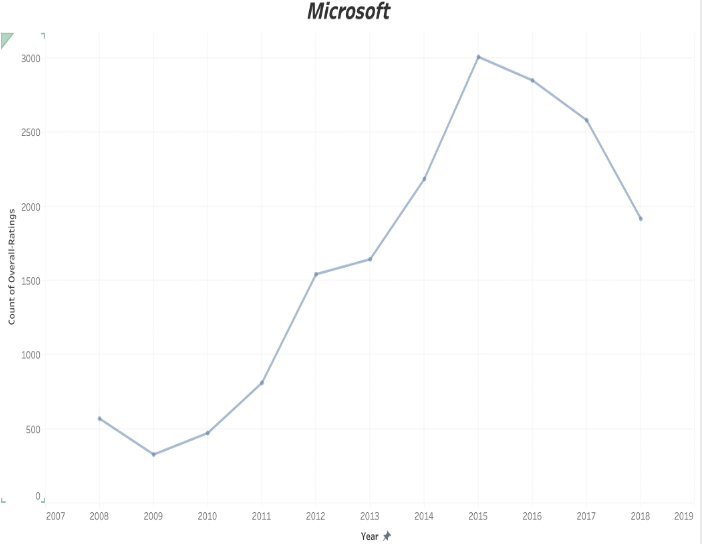


# **Review Count over the years**

Over the years, Amazon has increased the number of reviews received. We can see that the number of reviews was the least in the year 2008 with only around 300, which gradually increased to 8k in the year 2018



The same was not the case with Microsoft. There is no trend in the number of reviews received by Microsoft. We can see that the reviews were almost 600 in 2008 which fell the following year. The number of reviews gained a pace in 2010, but there was a considerable fall in 2015, with finally 1800 votes in 2018.



**Word Clouds:**

Word clouds are beneficial in grasping what the employees are saying about their company. There are numerous words, such as work, company, place, etc., that appear frequently and give no information. We removed such words from the word cloud. Also, we filtered out stop words, such as the, is, at, which, and, on, etc.

**Amazon- Word Cloud for pros**

From, the below word cloud, we can say that many employees emphasize on the words "good", "opportunities" etc. saying that their company provides them with various opportunities, a fun place to work, good salary etc.

From the word cloud, one can easily deduce how the employees felt about amazon and what its pros were.



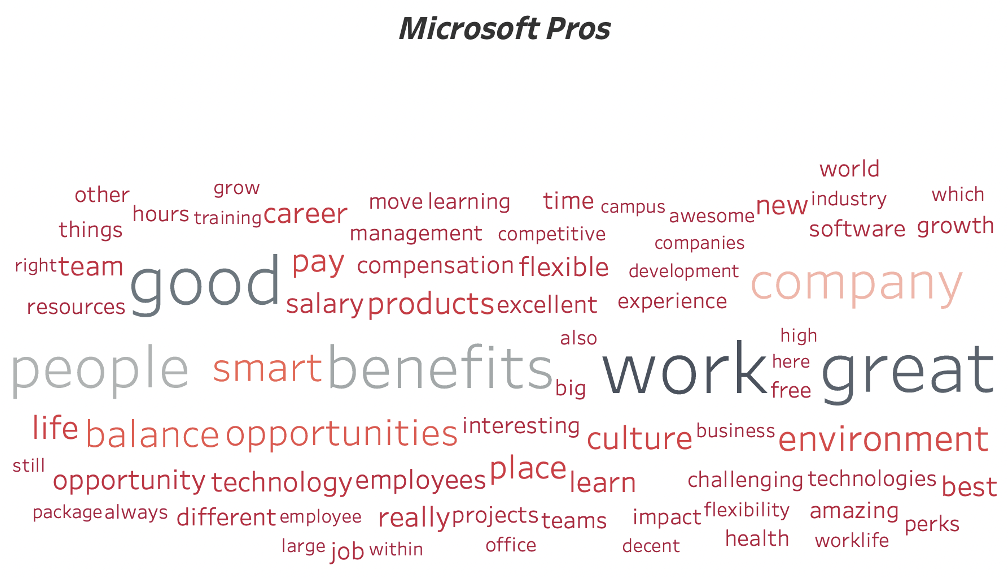
**Amazon- Word Cloud for Cons**

Many employees reported that its management was one of the biggest cons following by the number of hours they must work, suggested that they had no work life balance while some felt that they were not paid enough. Surprisingly the pay or salary was also one of the most said pros in amazon.



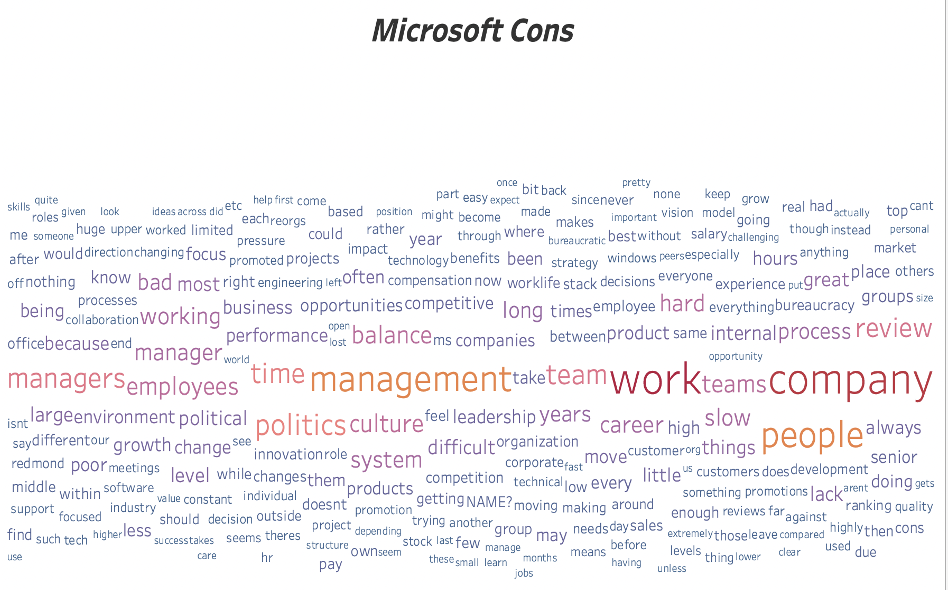
**Microsoft- Word Cloud for pros**

Microsoft employees has the same opinion as Amazon on their company's pros, emphasizing that they have great benefits, good opportunities, great pay etc.

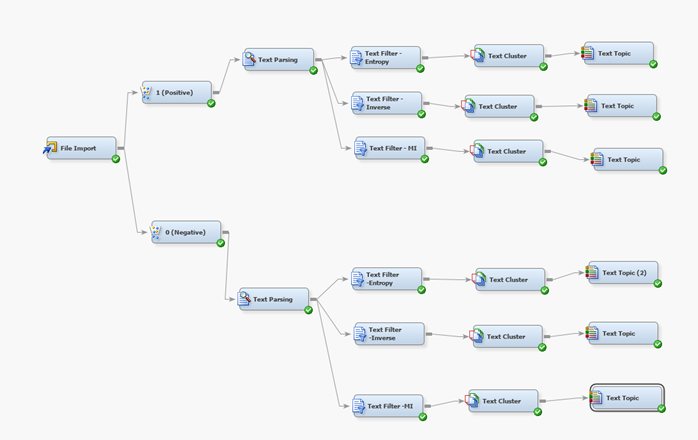


**Microsoft- Word Cloud for cons**

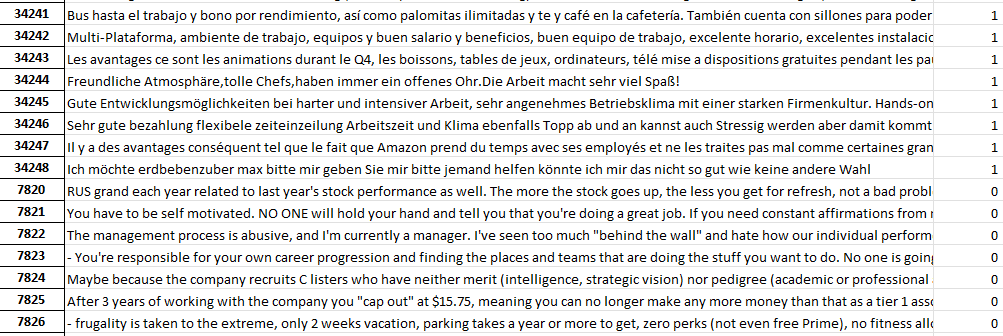
Surprisingly, the complaints about Microsoft were mostly similar to that of Amazon. The employees complained that their company has bad management, manager; there were lot of politics around them etc.



# **Exploration – Amazon Reviews**

Unsupervised learning Model built to analyze clusters and derive business Insights:  


For the sake of unsupervised learning, we altered and reshaped the dataset with equal number of 'pro' or positive values with the rating of 1, and 'cos' or negative values with a rating of 0, in order to differentiate the positives from the negative reviews for the sake of analysis.

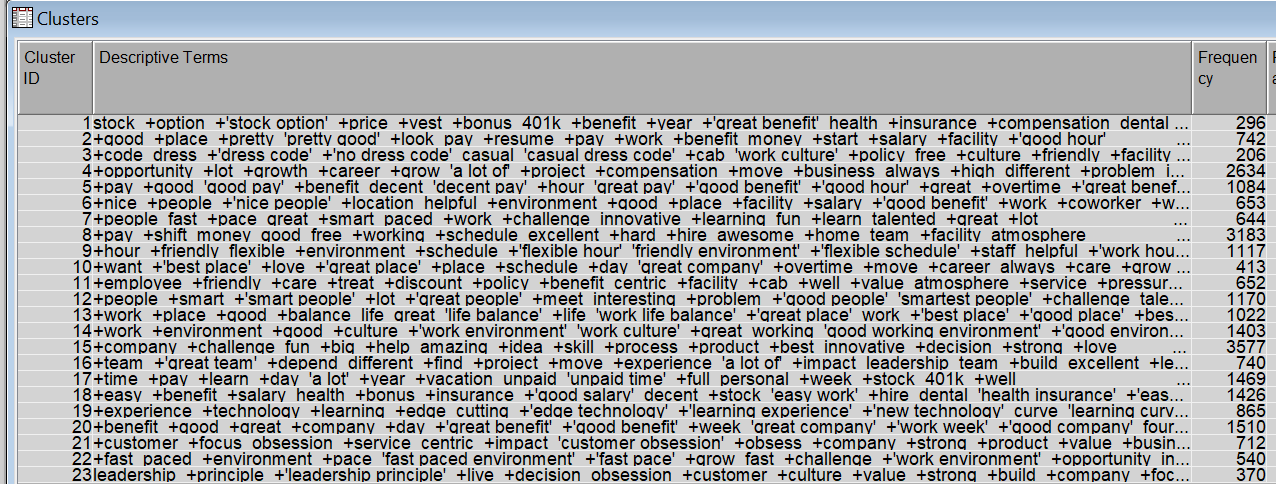


We then filtered the data into two subdivisions to separate the positive and negative reviews. And we then connected Text Filter, Text Cluster and Text Topic nodes for different Text Filter Term Weightings - Entropy, Mutual Information and Inverse document frequency in order to analyze different resulting clusters as per the above diagram.

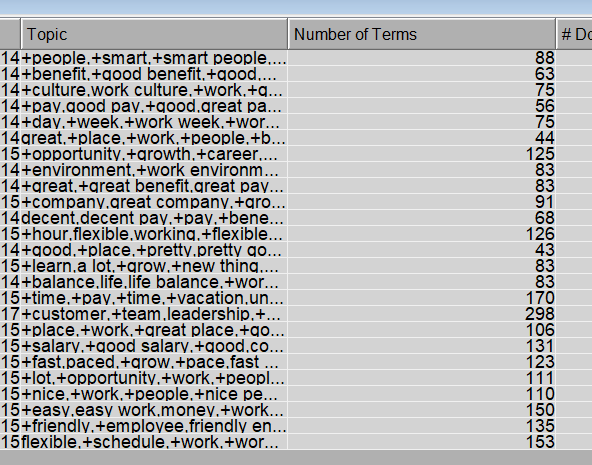
Text Cluster results and Text Topic results:

**Positive Cluster results:**

Text Cluster results:

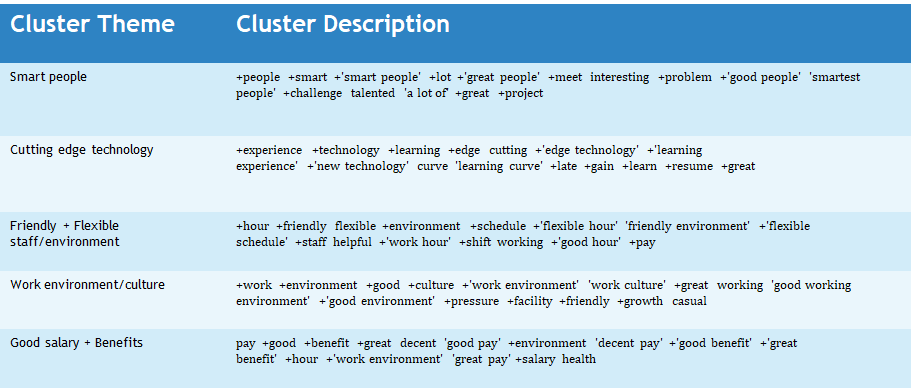
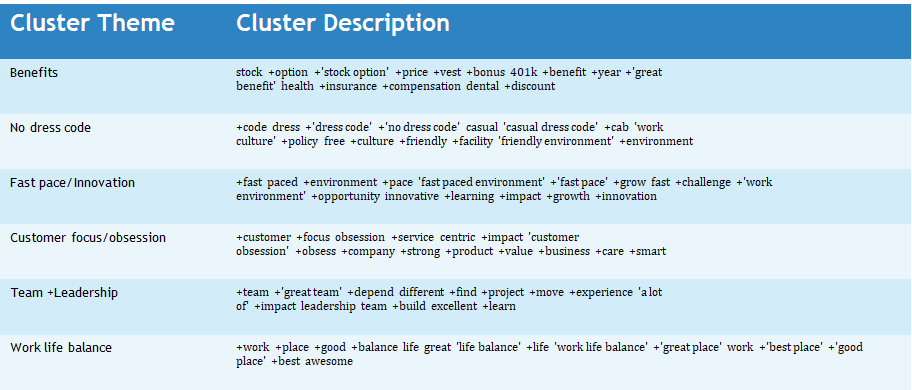


Text Topic results:

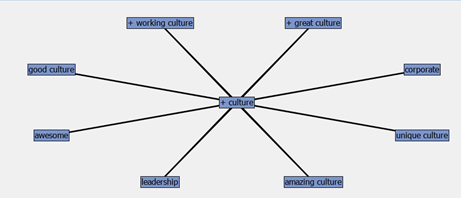


**ANALYSIS:**

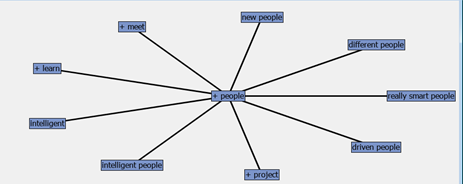
We analyzed our resulting clusters and identified a common theme for every cluster, from which we derived concept links that are representative of customer reviews and sentiment.

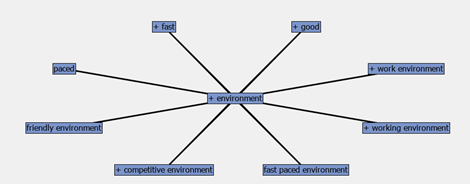


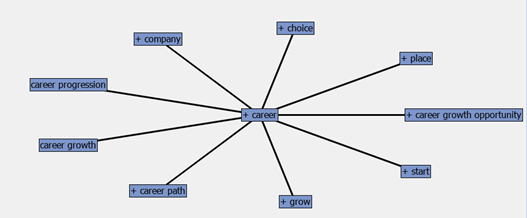
**Key Concept Links:**

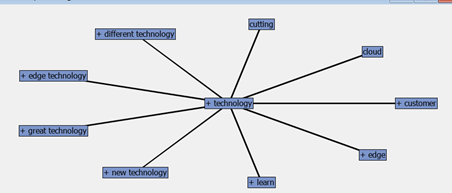


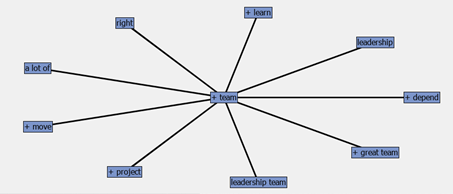






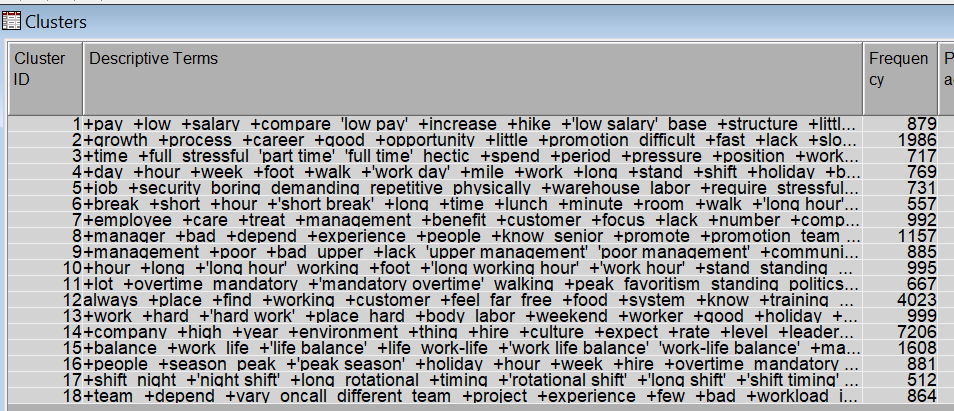




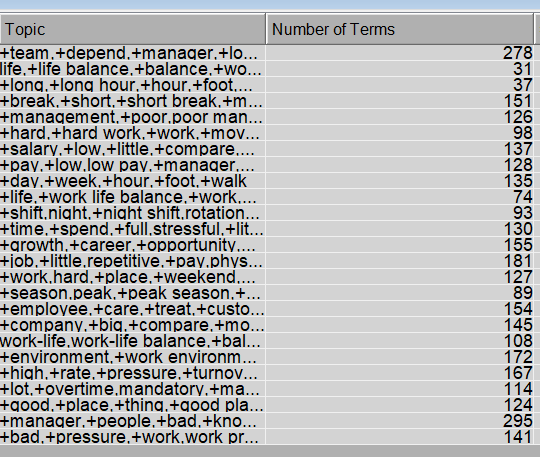


***Negative Cluster results:***

**Text Cluster:**

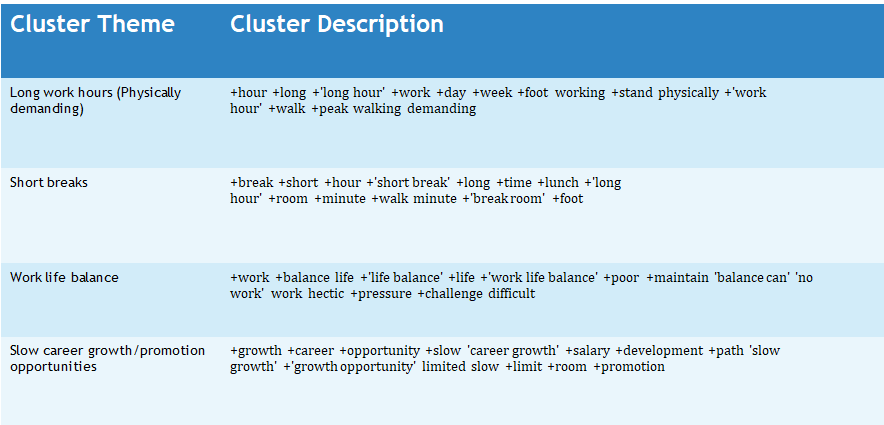
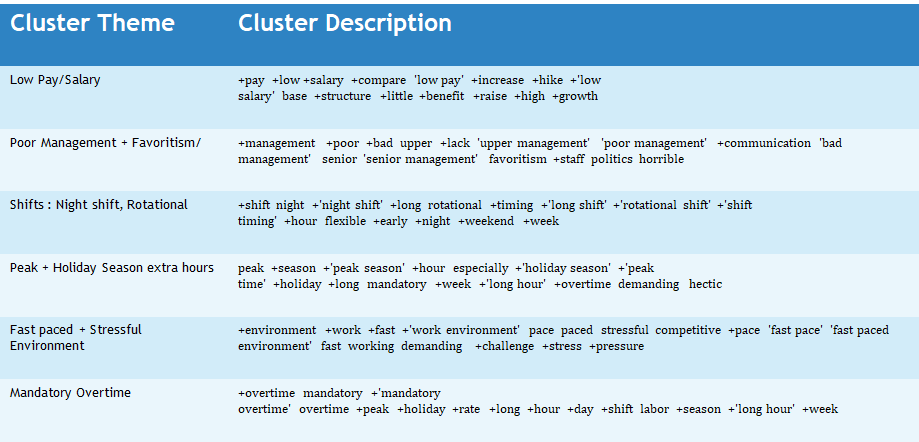


**Text Topic:**

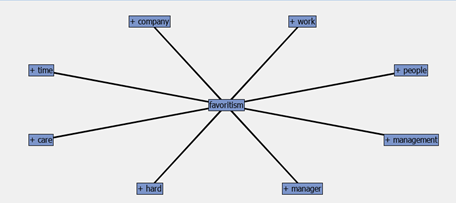


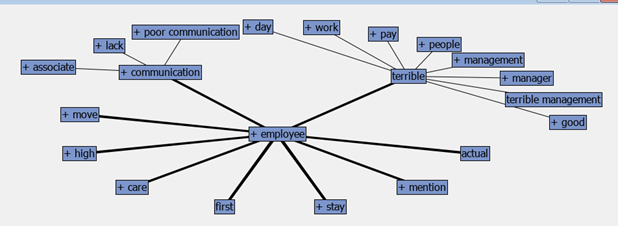
# **ANALYSIS**

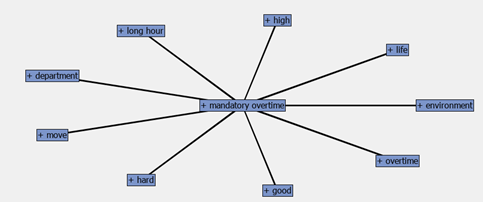
We analyzed our resulting clusters and identified a common theme for every cluster, from which we derived concept links that are representative of customer reviews and sentiment.



# **KEY CONCEPT LINKS**







**DISCREPENCIES**

There are some cluster themes that came up in both the pros and cons result. We further investigate this to clarify the true nature of sentiment.

●Work life balance in Pros: Frequency: 118 #documents: 117

Work life balance in Cons: Frequency: 364 #documents: 358

It is clear that most people feel there is a lack of work life balance at Amazon, in comparison to people that consider it a pro.

●Salary (High) in Pros: Frequency: 232 #documents: 232  
 Salary(Low) in Cons: Frequency: 30 #documents: 30  
A big percentage of employees feel the salary is high and are hence satisfied. Hence, Salary is a positive aspect at Amazon.

●Fast paced Environment in Pros: Frequency: 153 #documents: 153  
 Fast paced Environment in Cons: Frequency: 38 #documents: 38

The number of employees who consider the fast-paced environment at Amazon a positive, is almost 4 times more than the employees who consider it a Con that ends up being stressful.

# **Modeling – Amazon Reviews**

We first started creating models for **Amazon** dataset, below diagram is used with different variations in node values to find the optimal model.

Graphical user interface, text, application

Description automatically generated

Below attributes were used as the model inputs:

* Feedback column in the TEXT input variable, which holds values from 'pros' and 'cons' column.
* 'Overall ratings' is the Target variable, which is the binary.

Table

Description automatically generated

Now we ran different models with variations in text filter and text cluster values:

**Model1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Text Filter** | | **Text Cluster** | | |
| **Frequency Weight** | **Term Weight** | **SVD Resolution** | **Max SVD** | **Number of Clusters (Max)** |
| Log | Mutual Information | Low | 100 | 40 |

A picture containing application

Description automatically generated

**Model 2:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** |
| Log | | Entropy | Low | 100 | | 40 |

A screenshot of a computer

Description automatically generated with low confidence

**Model 3:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** |
| Log | | IDF | Low | 100 | | 40 |

A picture containing application

Description automatically generated

**Model 4:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | | **Number of Clusters (Max)** |
| Log | | Mutual Information | Medium | 100 | | 40 | |

A picture containing application

Description automatically generated

**Model 5:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | **Number of Clusters (Max)** |
| Log | Mutual Information | | High | 100 | 40 |

A picture containing application

Description automatically generated

**Model 6:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | | **Max SVD** | | **Number of Clusters (Max)** |
| Log | IDF | | Medium | 100 | | 40 | |

A picture containing table

Description automatically generated

**Model 6A:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** | |
| Log | | Entropy | Medium | 100 | | | 40 |

A picture containing diagram

Description automatically generated

**Model 7:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | | **Number of Clusters (Max)** |
| Log | IDF | | High | 100 | 40 | |

A screenshot of a computer

Description automatically generated with medium confidence

**Model 7A:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | | **Number of Clusters (Max)** |
| Log | Entropy | | High | 100 | 40 | |

A picture containing table

Description automatically generated

**Comparison between the above models:**

Below is the consolidated ROC index for the above models:

****

From above comparison, we found out that Model 5 performs best, and below is the ROC graph for the different models used in the diagram:

Graphical user interface, application

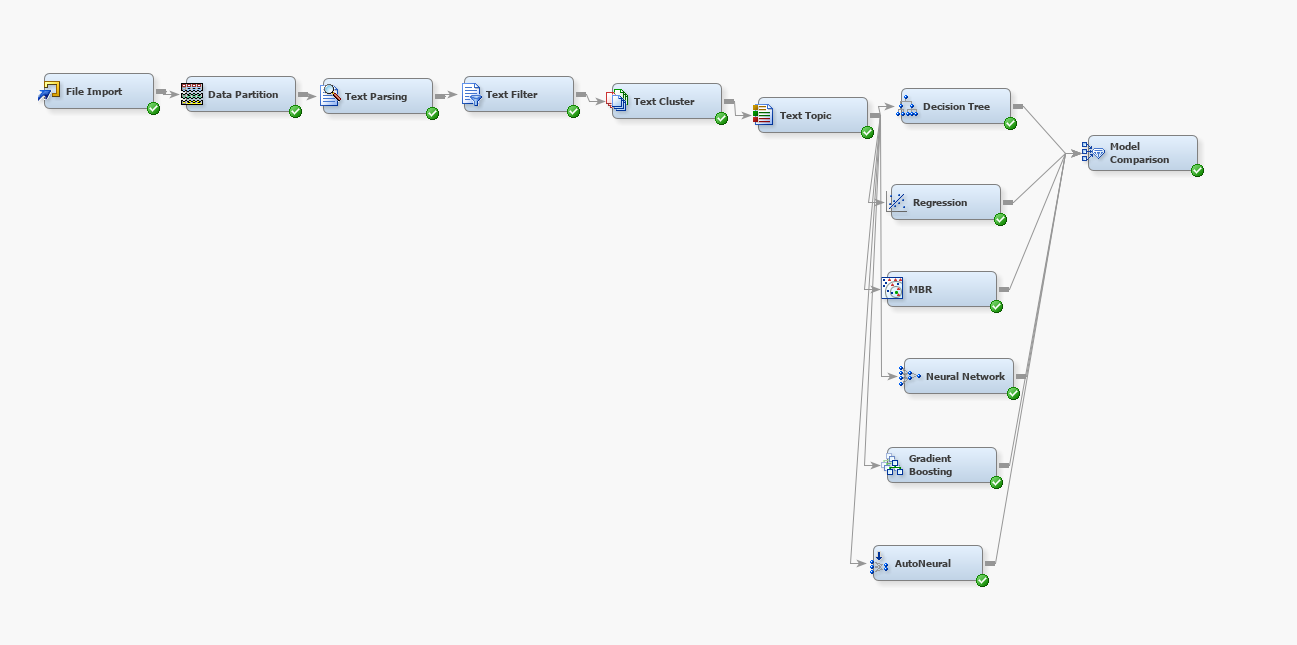
Description automatically generated

After comparing both the ROC index and misclassification rate, we concluded that Neural Network performs in terms of classifying the positive or negative sentiments in the reviews text.

****

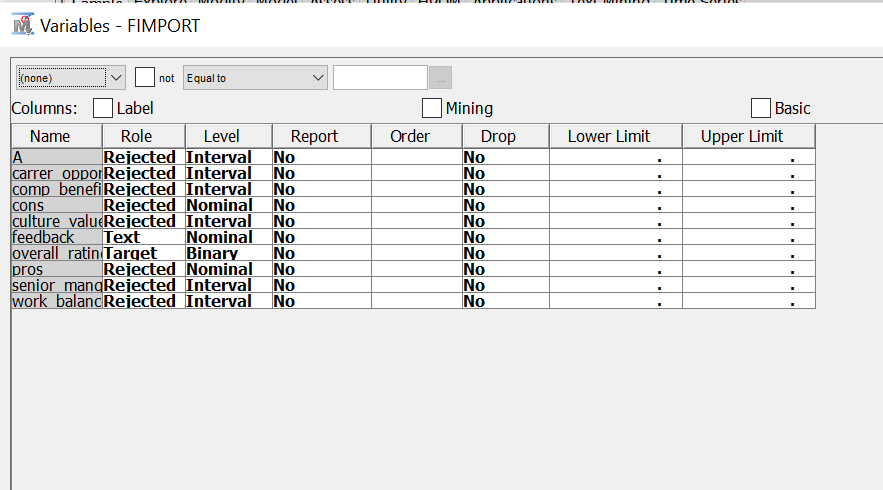
# **Modeling – Microsoft Reviews (Supervised)**

We first started creating models for **Microsoft** dataset, below diagram is used with different variations in node values to find the optimal model.



Below attributes were used as the model inputs:

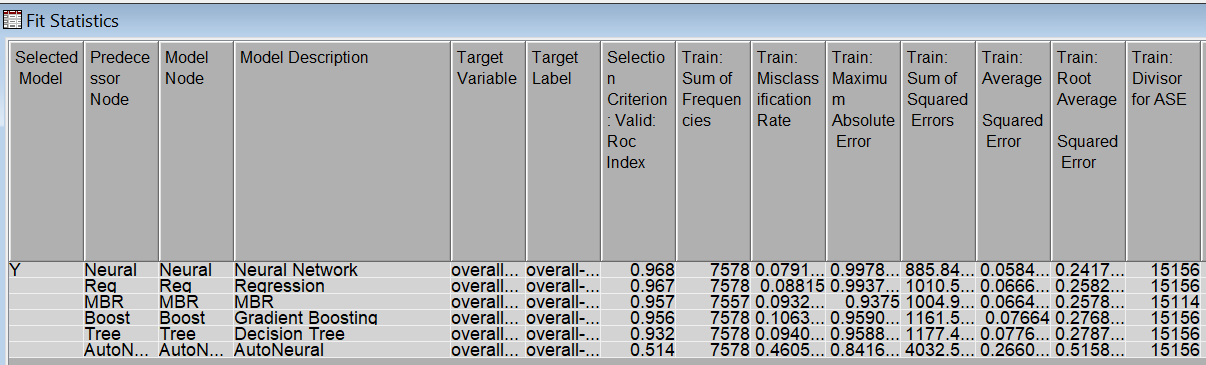
* Feedback column in the TEXT input variable, which holds values from 'pros' and 'cons' column.
* 'Overall ratings' is the Target variable, which is the binary.



Now we ran different models with variations in text filter and text cluster values:

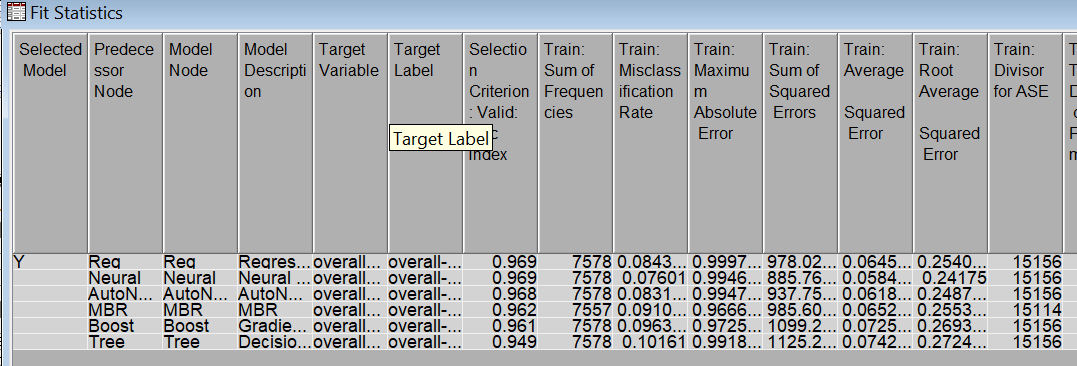
**Model1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Text Filter** | | **Text Cluster** | | |
| **Frequency Weight** | **Term Weight** | **SVD Resolution** | **Max SVD** | **Number of Clusters (Max)** |
| Log | Mutual Information | Low | 100 | 40 |



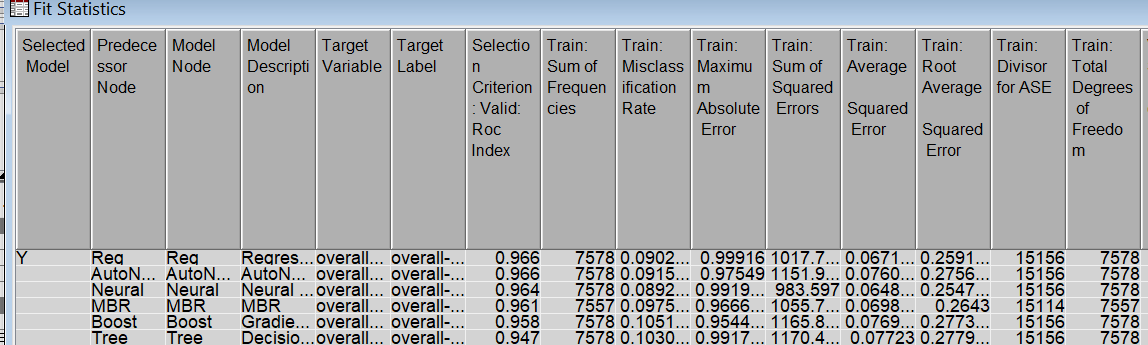
**Model 2:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** |
| Log | | Entropy | Low | 100 | | 40 |



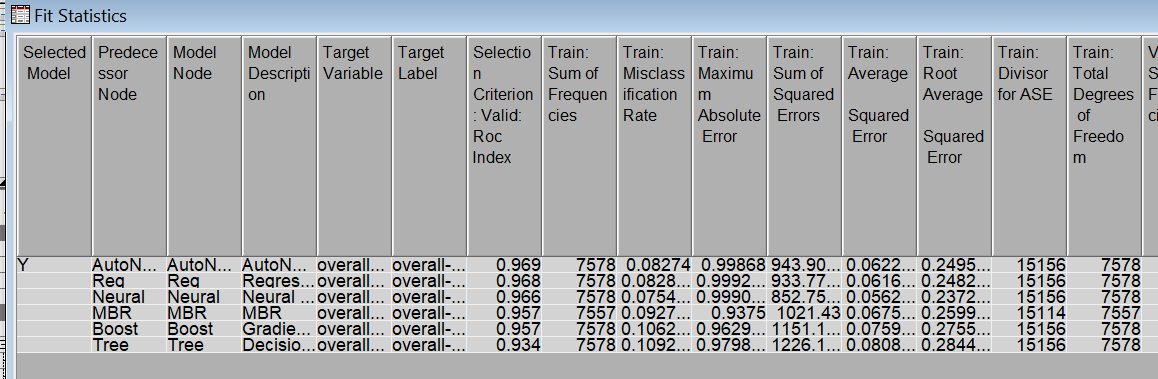
**Model 3:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** |
| Log | | IDF | Low | 100 | | 40 |



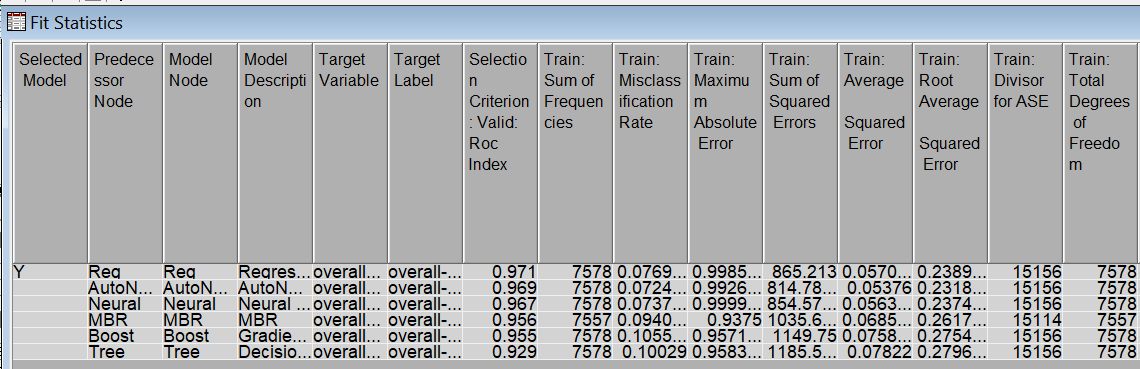
**Model 4:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | | **Number of Clusters (Max)** |
| Log | | Mutual Information | Medium | 100 | | 40 | |



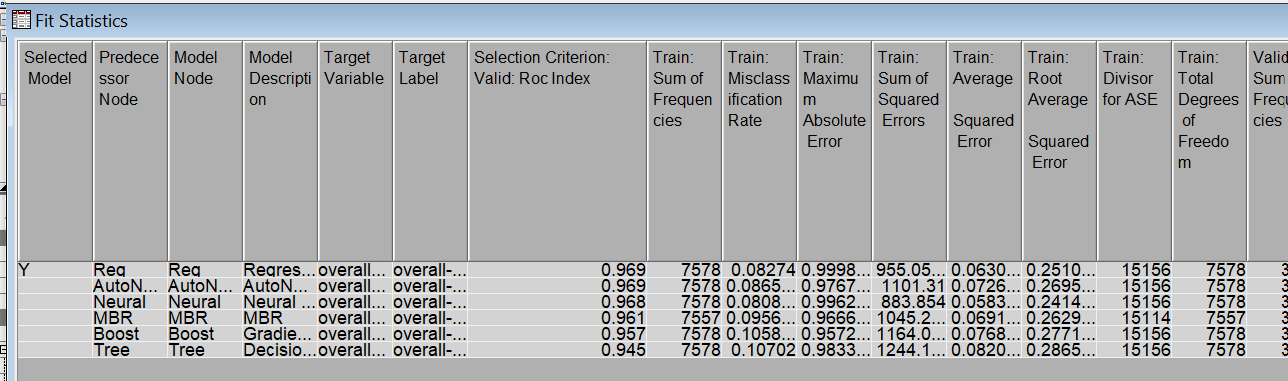
**Model 5:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | **Number of Clusters (Max)** |
| Log | Mutual Information | | High | 100 | 40 |



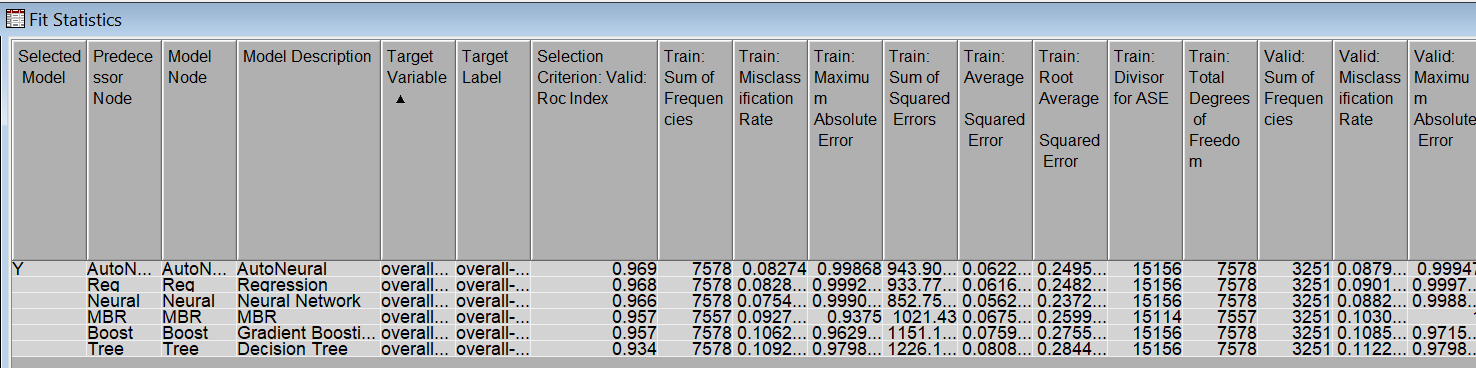
**Model 6:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | | **Max SVD** | | **Number of Clusters (Max)** |
| Log | IDF | | Medium | 100 | | 40 | |



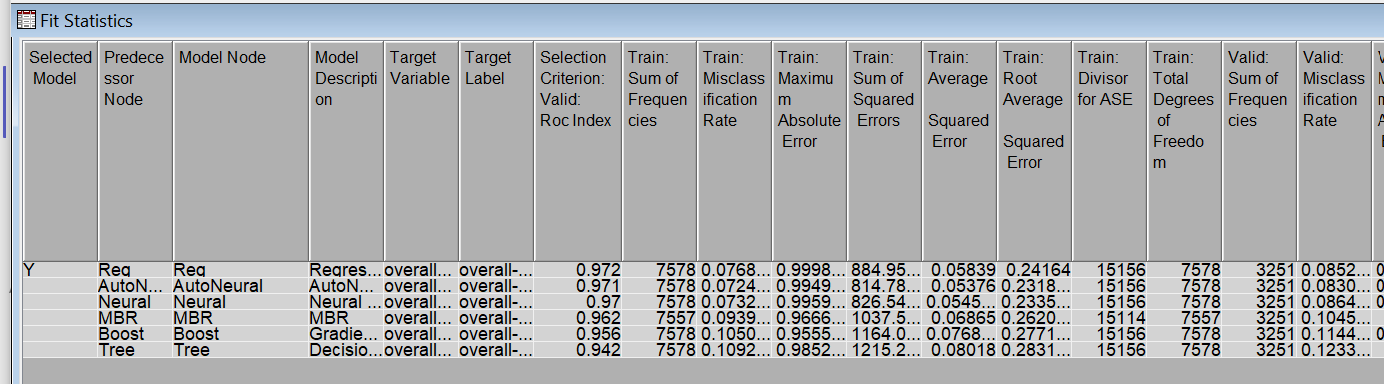
**Model 6A:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | | |
| **Frequency Weight** | **Term Weight** | | **SVD Resolution** | | **Max SVD** | **Number of Clusters (Max)** | |
| Log | | Entropy | Medium | 100 | | | 40 |



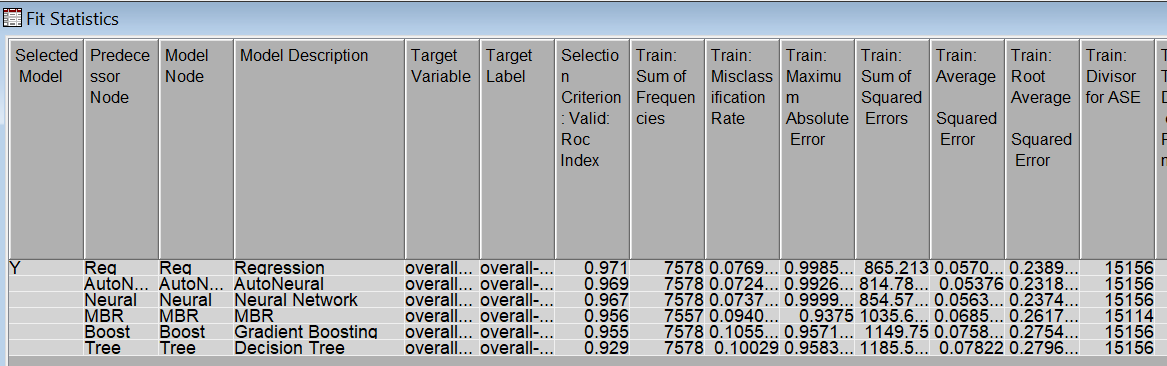
**Model 7:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | | **Number of Clusters (Max)** |
| Log | IDF | | High | 100 | 40 | |



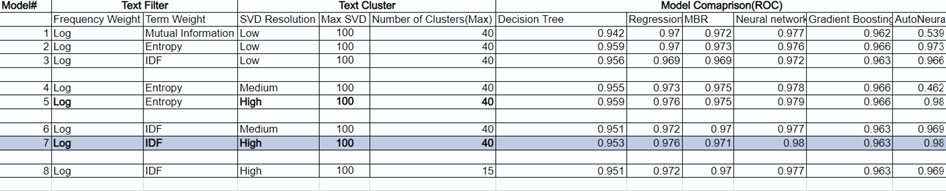
**Model 7A:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Text Filter** | | | **Text Cluster** | | | |
| **Frequency Weight** | | **Term Weight** | **SVD Resolution** | **Max SVD** | | **Number of Clusters (Max)** |
| Log | Entropy | | High | 100 | 40 | |

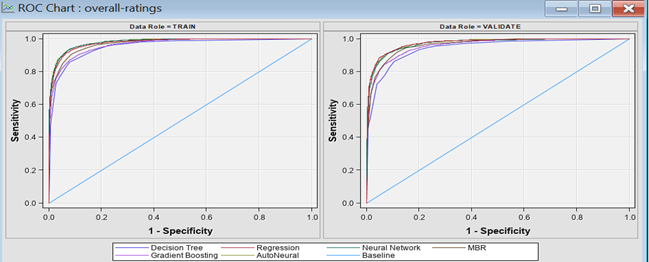


**Comparison between the above models:**

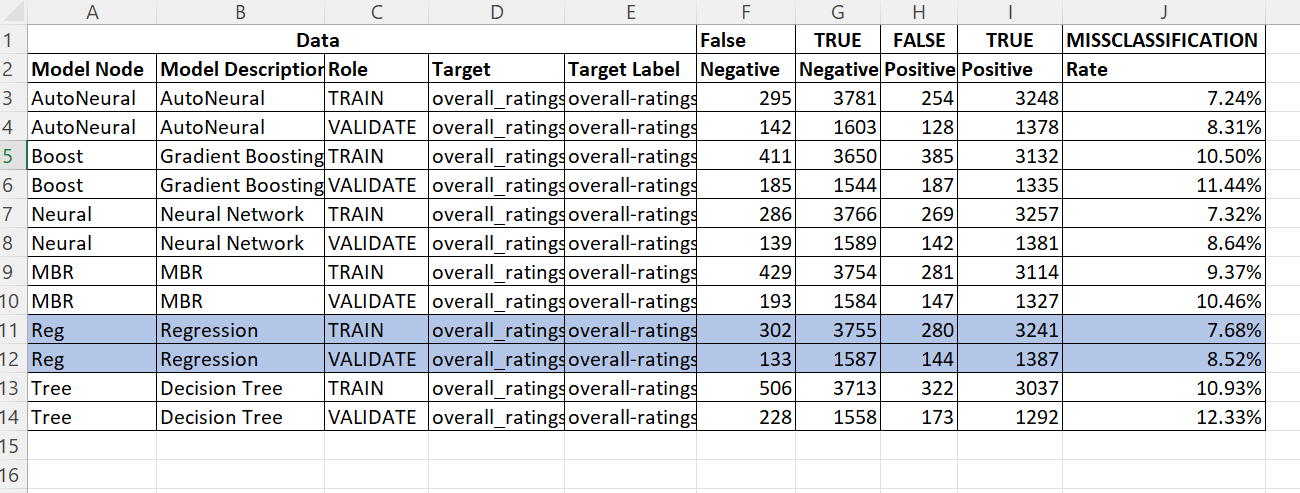
Below is the consolidated ROC index for the above models:



From above comparison, we found out that Model 7 performs best, and below is the ROC graph for the different models used in the diagram:



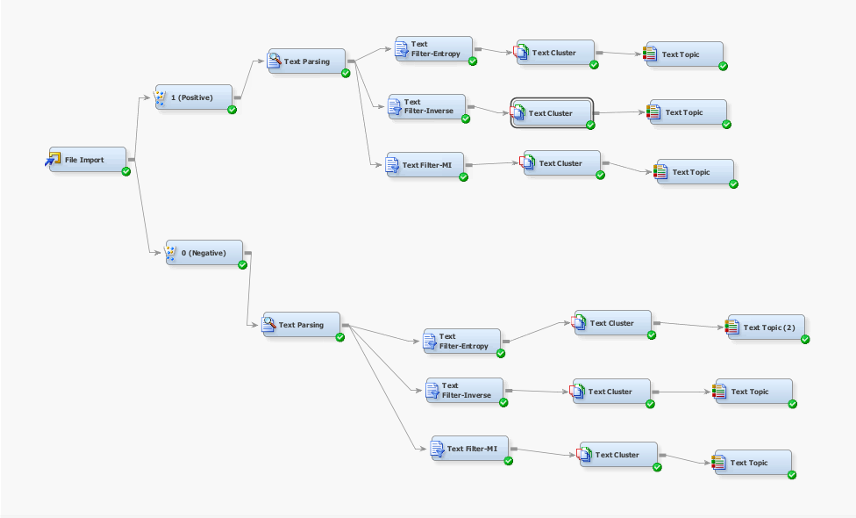
After comparing both the ROC index and misclassification rate, we concluded that Regression performs in terms of classifying the positive or negative sentiments in the reviews text.



# **Modeling – Microsoft Reviews**

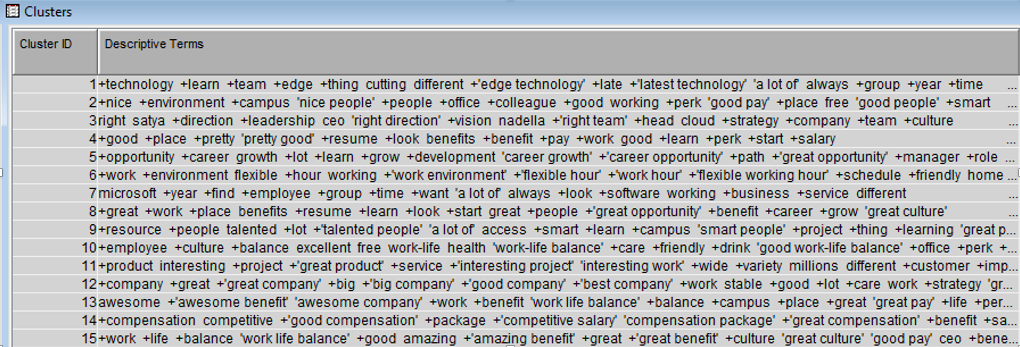
Similar to the Amazon Unsupervised learning Model, we built a model to analyze clusters and derive business Insights for Microsoft as well.

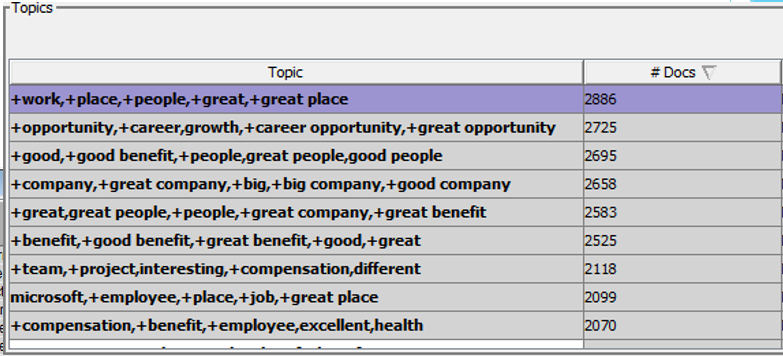
We started by filtering the data into two subdivisions to separate the positive and negative reviews. And we then connected Text Filter, Text Cluster and Text Topic nodes for different Text Filter Term Weightings - Entropy, Mutual Information, and Inverse document frequency in order to analyze different resulting clusters as per the below diagram.



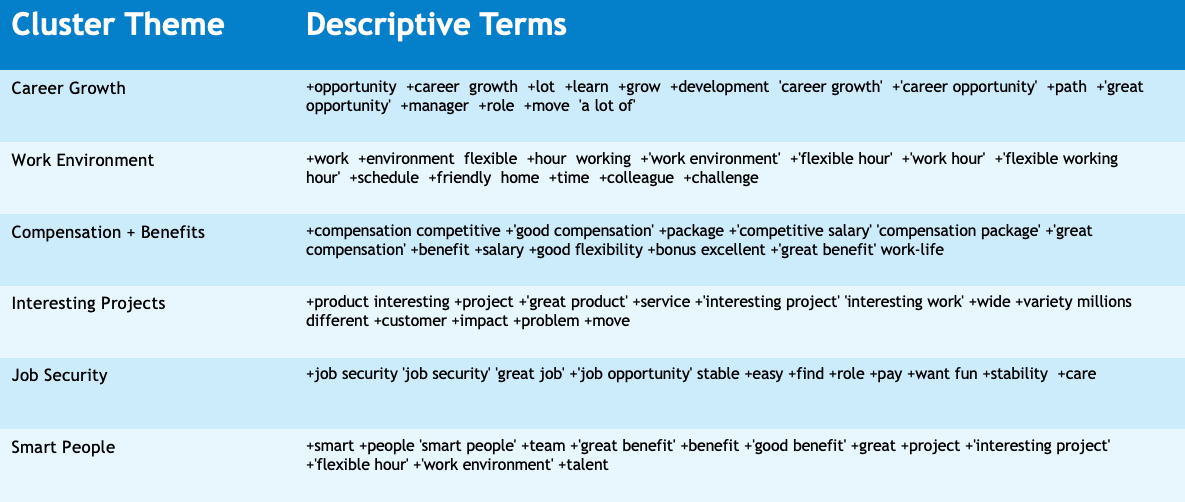
**Positive Cluster results:**

Text Cluster results and Text Topic results:

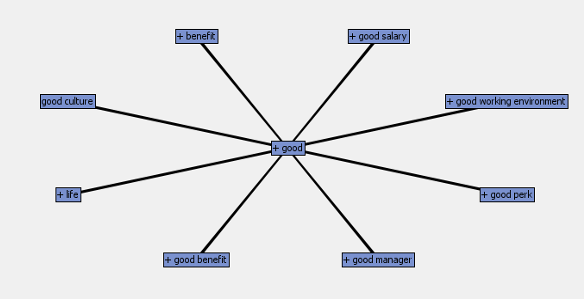


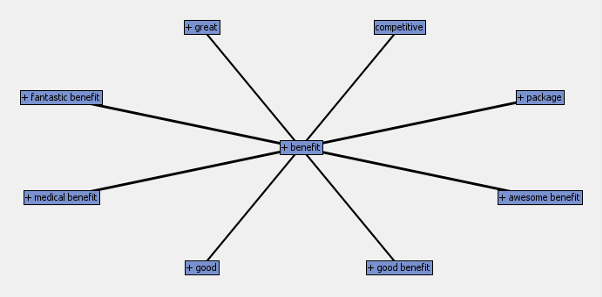


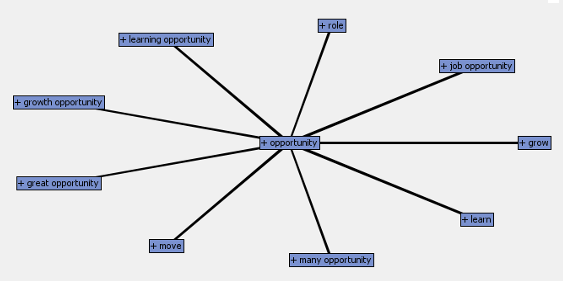
Text Cluster Analysis

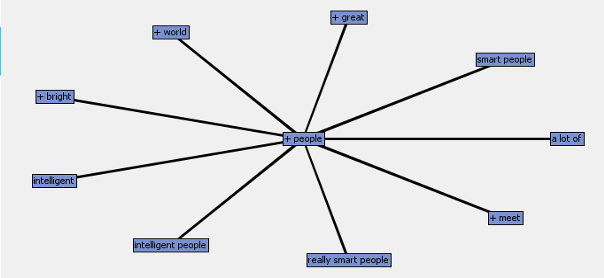


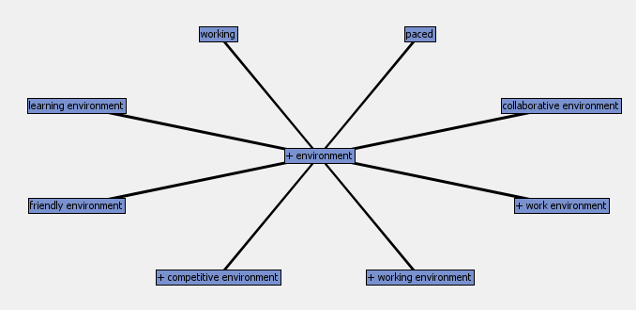
Key Concept Links

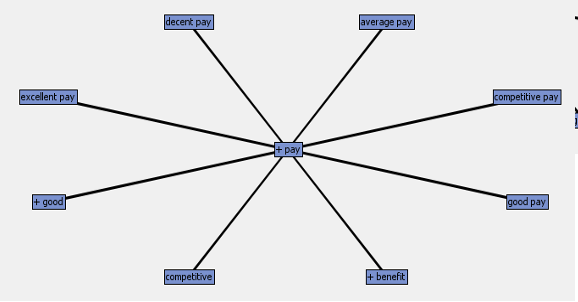


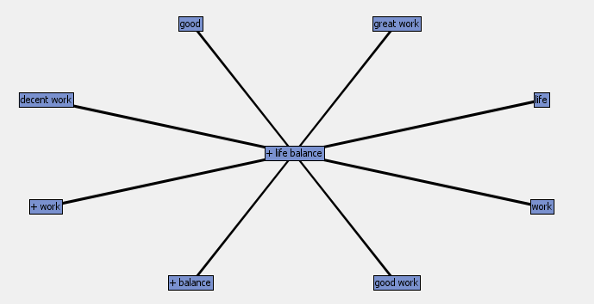






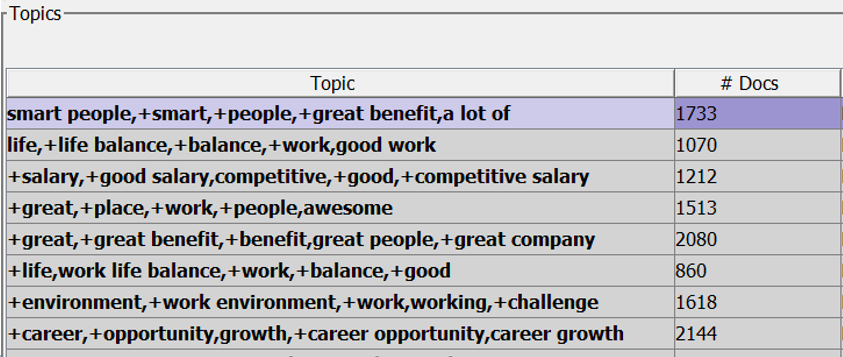






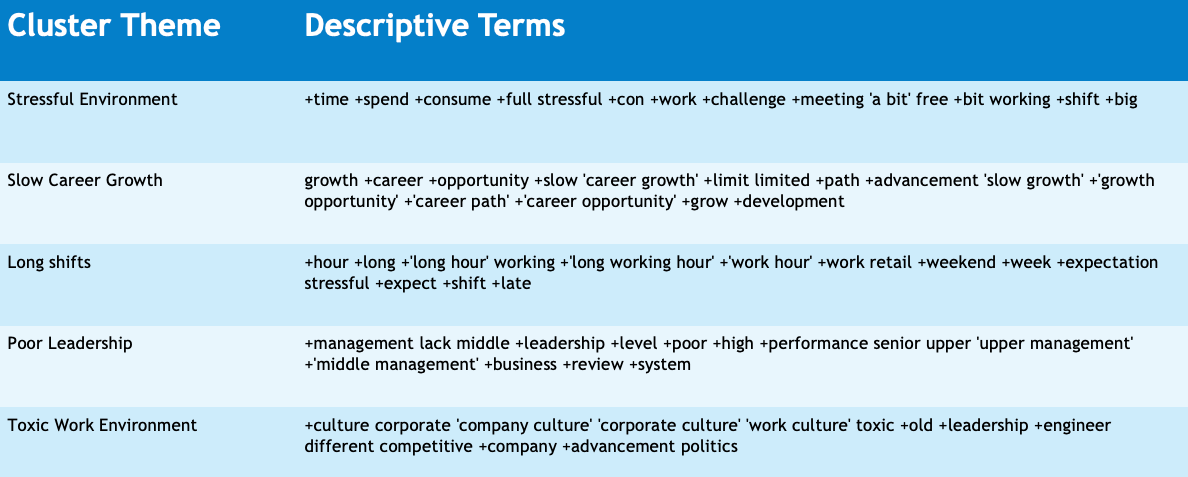
We analyzed our resulting clusters and identified a common theme for every cluster, from which we derived concept links that are representative of customer reviews and sentiment.

**Negative Cluster results:**

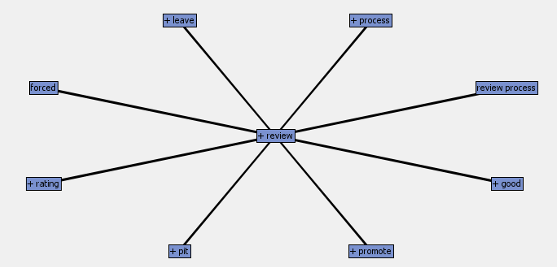


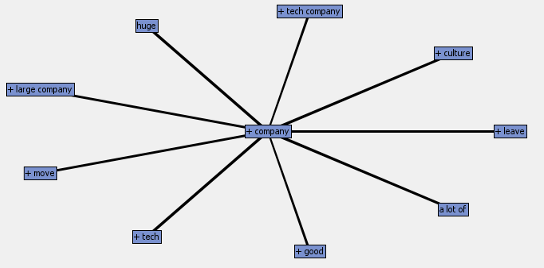


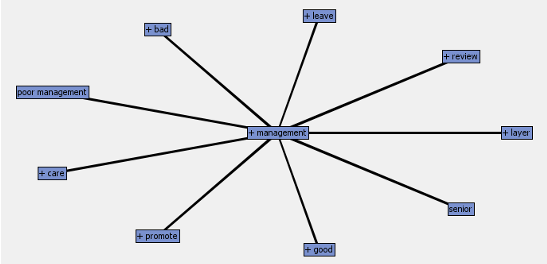
Text Cluster Analysis



Key Concept Links







We analyzed our resulting clusters and identified a common theme for every cluster, from which we derived concept links that are representative of customer reviews and sentiment.

**Discrepancies**

Similar to Amazon, in Microsoft analysis as well, there are some cluster themes that came up in both the pros and cons result. We further investigate this to clarify the true nature of sentiment.

* Career Growth in Pros: Frequency: 1013 #documents: 2725
* Career Growth in Cons: Frequency: 560 #documents: 2144

This shows that most of the people feel that there is a good career growth in Microsoft, in comparison to people that consider it a con.

* Compensation (High) in Pros: Frequency: 547 #documents: 2118
* Salary (Low) in Cons: Frequency: 726 #documents: 1212

This analysis has given quite close results however the number of employees is still higher for Compensation (low) than the employees who feel the compensation is good at Microsoft.

# **Conclusion**

**Amazon:**

Amazon Pros: Benefits (401K, Insurance, Stock options), No dress code, Fast pace/Innovation, Customer focus/obsession, Team +Leadership, Cutting edge technology, Smart people, Friendly + Flexible staff/environment, Work environment/culture emphasis, good salary + Benefits

Amazon Cons: Poor Management (Politics + Favoritism), Shifts: Night shift, Rotational, Peak + Holiday Season extra hours, Fast paced + Stressful Environment, Mandatory Overtime, Long work hours (Physically demanding) and short breaks, Bad Work life balance, Slow career growth/promotion opportunities.

**Microsoft:**

Microsoft Pros: Lot to Learn, Great opportunities, Flexible work environment, competitive compensation, good bonuses, interesting and impactful projects, Job Security, Smart and intelligent people

Microsoft Cons: full stressful, slow career growth, less career opportunity, long working hour, weekend, stressful expectation, lacks middle management, Corporate Culture, toxic and old leadership, politics, low compensations

# **Recommendation**

**Amazon:**

* **Run marketing internal and external campaigns to advertise the Pros of the companies to retain customers (i.e., the employees) about a free culture of no dress code, Flexible timings, etc.**
* **Work in improving in the Con's to improve the company's shortcomings and image.**
* **Take steps to improve work life balance and share it with the company and other social media (Like LinkedIn) to improve the judgement of the company**

**Microsoft:**

* **Similarly, run internal and external marketing campaigns to advertise the pros. The pros would include flexible work hours, great work opportunities, etc.**
* **Same as above, work in improving the Con's to better the company's image**
* **Take steps to improve the image about the company on long working hours and improve stress levels among the employees**