SESUG Paper 243-2019

# Improving Employee Satisfaction Through Text Analytics

Luis David Martinez, Oklahoma State University

## ABSTRACT

The success of any business is heavily influenced by the productivity of its employees, making employees one of the most important assets that a company holds. One factor that is linked with employee productivity is work satisfaction. In other words, how well an employee works is partly determined by how happy he or she is with their job. Therefore, knowing the positive and negative feelings that employees have toward their employer can be useful for making better human resource decisions. One source of data on employee's opinions about their employer and work environment is online reviews. In this project, we take a collection of online reviews from a popular job search website and analyze it to gain insights into how employees feel about their company. The dataset includes a pros and cons comment sections as well as a score from 1-5 on the overall opinion of the company. Thousands of reviews were collected on six popular technology companies. Online reviews have been analyzed descriptively in other research to determine what employees most commonly mention when reviewing their employer. Our own preliminary descriptive analysis showed a statistically significant difference among the six technology companies in relation to overall scores. This research paper, however, takes the analysis of these scores further by creating a predictive rule-based model that uses the text portions of the review as variables to classify comments into positive or negative. By interpreting the model, we can determine what factors may be most significant for causing high or low employee satisfaction. These factors can then be used by companies in taking appropriate steps to improve their employee sentiment and in turn, their overall productivity. The tools used to perform this analysis include SAS® Enterprise Miner™ and SAS® Studio. This research can also be easily expanded to include other companies and improved upon by adding more training data from other job sites.

## INTRODUCTION

In present day, analytics has been innovatively used in most company departments. From marketing and sales to research and development, it is easy to see how gleaning useful information from large datasets can be beneficial to a company. One such area that has leveraged data to improve its efficiencies is human resources. The applications of data analysis in this area have achieved many goals including hiring the people most likely to succeed (Van Vulpen) to predicting when employees are likely to leave the company (Alkuwaiti et al, 2016).

The reason for focus on human resources is simple; the success of a company is largely determined by the success of its employees. One specific area that may be improved is employee satisfaction, which is of major interest to companies, given that the more satisfied employees are, the higher their productivity (Sorensen, 2013). Keeping employees happy and engaged helps to increase company performance by improving quality of work done and decreasing turnover and dishonest behavior (Sorenson, 2013). A satisfied and engaged workforce also lends to a better overall work environment which is more conducive to creativity (S. Yekanialibeiglou, H. Demirkan, 2018).

The benefits of having employees that are satisfied with their employer are clear, but companies must also keep employees happy in the most cost-conscious way possible. To do this, companies need to focus on the things that employees value the most and which will have the greatest impact with the least amount of input. This information, however, is not always easily accessible.

Surveying employees is one possible way to get their opinion, but this is time consuming, expensive and may not lead to the best results, as people who are required to take surveys do not typically provide the best feedback.

A way to overcome this problem is to mine available data found on-line. Job sites such as *Glassdoor* and *Indeed* provide employees a place to submit written reviews of companies as well as give them a score on several metrics. These websites then compile these scores and rank companies based on their overall employee review score. By analyzing this freely available data, we can learn about the most common topics when people are complaining or praising a company without running our own surveys or reading through thousands of reviews.

This paper presents a descriptive analysis of the most significant variables in the dataset and also presents models that can be used to classify comments written by employees into positive and negative. The model results are then interpreted to provide insights into the common drivers of employee satisfaction.

## DATA DESCRIPTION AND PREPARATION

The original dataset from which all of the other subsets used in this paper are derived is a collection of over 67,000 online reviews of companies written by employees. The companies in the dataset are Google, Microsoft, Facebook, Netflix, Amazon, and Apple. The dates on the reviews range from 2008 to 2018 with more reviews in the recent dates. While the dataset included 17 total variables, the following subset of variables were selected for further preparation and analysis in SAS® Studio:

- ID - The unique identifier of the review. (Nominal)
- Company - Company being reviewed by the employee. (Nominal)
- Job_Title – Employee's job title and whether they are current or former. (Nominal)
- Pros- Positive comments about the company left by the employee. (Text)
- Cons - Negative comments about the company left by the employee. (Text)
- Overall_Ratings - Overall rating given by the employee to the company, on a scale of 1 to 5. (Ordinal)

From the cons portion of the reviews, many of the comments in the cons column indicates that there are no negatives to report, according to the employee writing that review. For example, there are many instances where the comment in the cons row is "None" or "No cons". By looking through the top few pages of the data and noting the most common occurrences of these instances, the cons column is cleansed of the majority of these types of comments.

From the job title column, the actual job title of the employee is removed from the column because approximately 40% of the actual employee titles in the original dataset are entered as "Anonymous Employee". Whether an employee was current or former is present in all of the observations, therefore, this portion of the column was kept for all datasets.

This set of variables and all observations are used in the descriptive portion of the project, however, for the modeling portions of the project the data is sampled and manipulated to create five additional datasets for text analysis and modeling.

The process followed in creating these five subsets of data is the same and is outlined below:

1. The original data is divided into two datasets, one which has the pros text column removed and the other which has the cons text column removed. We are left with a pros dataset and a cons dataset.
2. The text column's name in both datasets is changed to "comment".
3. Both datasets are given a new variable called "sentiment", the dataset with pros is given a value of "positive" for all observations and the cons dataset is given a value of "negative" for all observations.
4. When creating the new datasets, these two datasets are sampled and the two samples of data are concatenated to create a dataset with both positive and negative comments.

The five sampled datasets created using this process are the following:

**Pros_and_Cons –** Includes 1,500 randomly sampled positive comments and 1,500 randomly sampled negative comments.

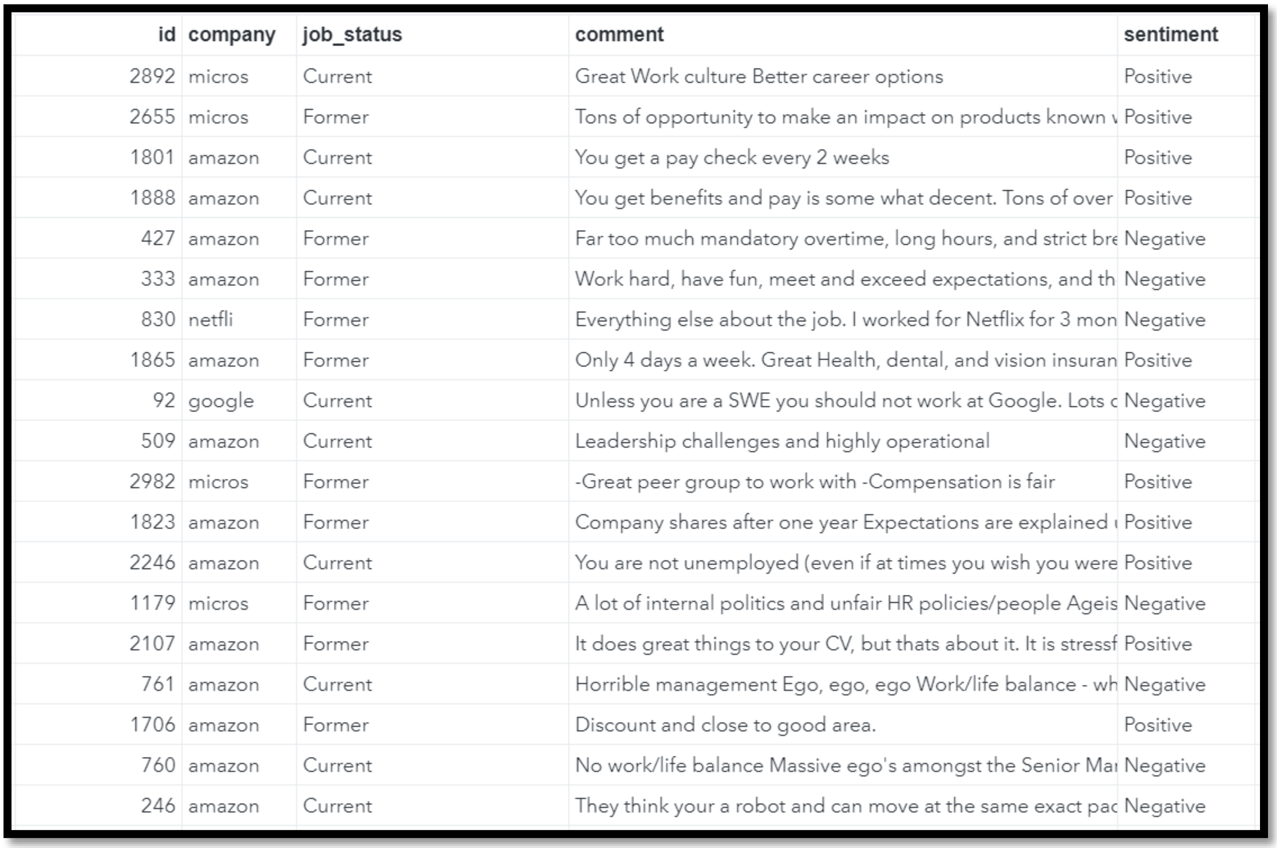
**Current-** Includes 1,500 randomly sampled positive comments and 1,500 randomly sampled negative comments, with the condition that all observations are from current employees. (3,000 total observations)

**Former-** Includes 1,500 randomly sampled positive comments and 1,500 randomly sampled negative comments, with the condition that all observations are from former employees. (3,000 total observations)

**Facebook-** Includes 1,500 randomly sampled positive comments and 839 negative comments, with the condition that all observations are from Facebook employees. This dataset used all of the available negative reviews of Facebook. (2,339 total observations)

**Amazon-** Includes 1,500 randomly sampled positive comments and 1,500 randomly sampled negative comments, with the condition that all observations are from Amazon employees. (3,000 total observations)

An example of the structure of the final datasets is shown below.



| id | company | job_status | comment | sentiment |
|---|---|---|---|---|
| 2892 | micros | Current | Great Work culture Better career options | Positive |
| 2655 | micros | Former | Tons of opportunity to make an impact on products known | Positive |
| 1801 | amazon | Current | You get a pay check every 2 weeks | Positive |
| 1888 | amazon | Current | You get benefits and pay is some what decent. Tons of over | Positive |
| 427 | amazon | Former | Far too much mandatory overtime, long hours, and strict br | Negative |
| 333 | amazon | Former | Work hard, have fun, meet and exceed expectations, and th | Negative |
| 830 | netfli | Former | Everything else about the job. I worked for Netflix for 3 mon | Negative |
| 1865 | amazon | Former | Only 4 days a week. Great Health, dental, and vision insuran | Positive |
| 92 | google | Current | Unless you are a SWE you should not work at Google. Lots | Negative |
| 509 | amazon | Current | Leadership challenges and highly operational | Negative |
| 2982 | micros | Former | -Great peer group to work with -Compensation is fair | Positive |
| 1823 | amazon | Former | Company shares after one year Expectations are explained | Positive |
| 2246 | amazon | Current | You are not unemployed (even if at times you wish you were | Positive |
| 1179 | micros | Former | A lot of internal politics and unfair HR policies/people Ageis | Negative |
| 2107 | amazon | Former | It does great things to your CV, but thats about it. It is stressf | Positive |
| 761 | amazon | Current | Horrible management Ego, ego, ego Work/life balance - wh | Negative |
| 1706 | amazon | Former | Discount and close to good area. | Positive |
| 760 | amazon | Current | No work/life balance Massive ego's amongst the Senior Ma | Negative |
| 246 | amazon | Current | They think your a robot and can move at the same exact pa | Negative |

**Figure 1. Sample of Final Dataset**

## METHODS

Following the data preparation in SAS studio, the dataset is analyzed descriptively. Having a descriptive understanding of the data helps the formulation of questions and ideas for the text analytics portion of the project. The descriptive analysis of the data is performed and presented in SAS Studio. Once the descriptive analysis is finished the following sequence of methods are used on the modeling datasets to analyze the text portions of the data in SAS® Enterprise Miner™:

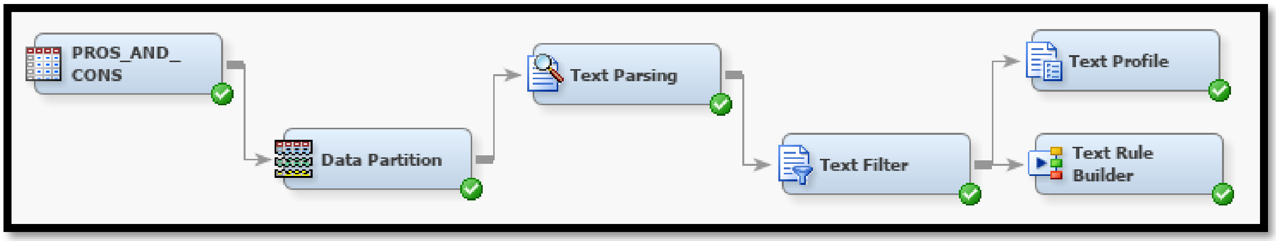


**Figure 2. Sequence of Text Analysis in SAS Enterprise Miner**

A brief explanation of each of the nodes, their purpose, and the important settings is listed below:

**Data Node (Pros_and_Cons in this figure)-** the data node represents the dataset being used in the current sequence. The data can be changed to run the same sequence of nodes on different datasets.

**Data Partition Node-** separates the dataset into 70% training, 30% validation. Using validation data helps to ensure that we don't overfit our models in the training dataset when we use the Rule Builder Node and gives us a way to assess the accuracy of the models.

**Text Parsing Node-** parses dataset and creates a term by document matrix which shows you the most frequently occurring terms within the dataset. This node is also able to be configured to tag the part of speech of the words in our dataset and to ignore certain parts of speech and punctuation marks. The "Different Parts of Speech" and "Noun Groups" are set to "Yes".

**Text Filter Node-** is used to remove words from the data that is passed on to subsequent nodes for analysis. There are many reoccurring words within the data (such as "the", "of", "in", "and", etc.) that provide no value to the models and are therefore filtered out. This node also allows for the combination of terms into synonyms. For example, in our data the terms "politics" and "political" were combined to make one term given that the observed reviews were referring to the same thing but were being treated as different terms after the original parse. The "check spelling" option is changed to "Yes", this allows the node to create synonyms for misspelled words.

**Text Profile Node-** is used to determine the terms that have the highest likelihood of describing the different levels of the target variable. In the case of this dataset, this node gives terms that describe the positive and negative comments. This node not only gives a count of terms that are most frequently present but uses a hierarchical Bayesian model to give lower weights to terms that are likely to describe multiple target levels.

**Text Rule Builder Node-** creates rules that are combined into a model that can predict target value based on the content of the text. Rules are a simple test of whether the observation of text includes a certain term or (groups of terms) or not. This node outputs the same results as other modeling tools in SAS Enterprise Miner and can be interpreted for descriptive purposes as well as used to score new data. The settings for generalization error, purity of rules, and exhaustiveness are left at medium for all of the results in this paper.

## RESULTS

### DESCRIPTIVE ANALYSIS

The overall_ratings column is present for all of the observations and is the easiest variable to analyze to give us an idea of the satisfaction of employees regarding their company.

Overall the ratings are skewed toward the more positive ratings as shown in figure 3 below:

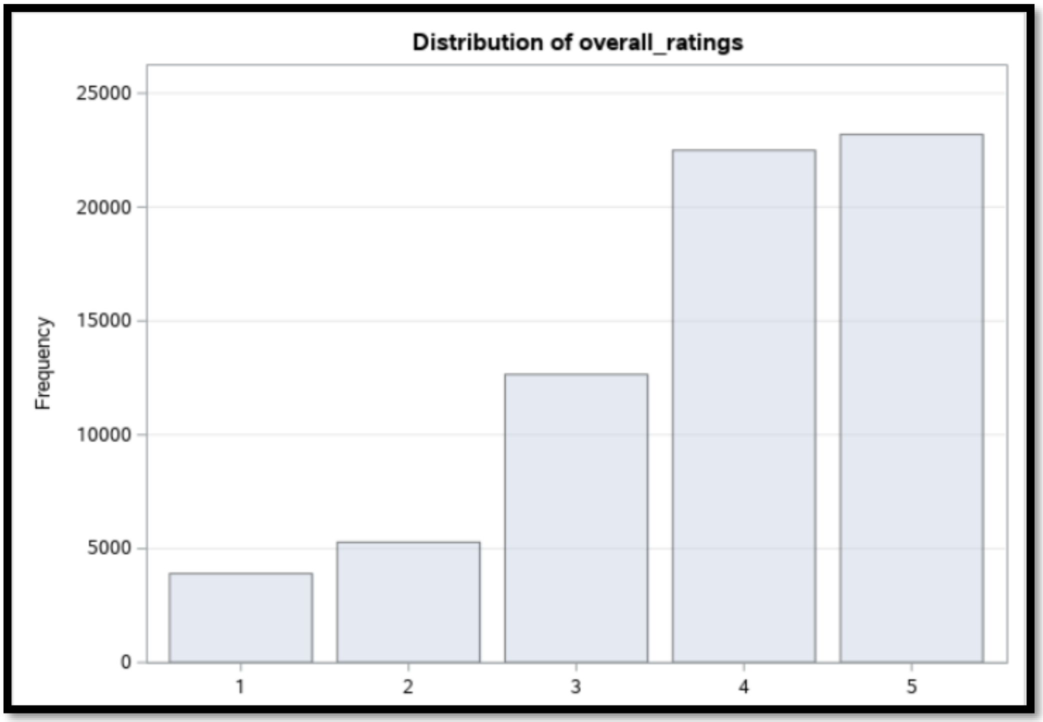


**Figure 3. Count of Observations by overall_ratings**

For additional insight, we take the overall rating variable and analyze based on company to determine how the companies rank. The companies are ranked in order of their average overall score in the figure below:

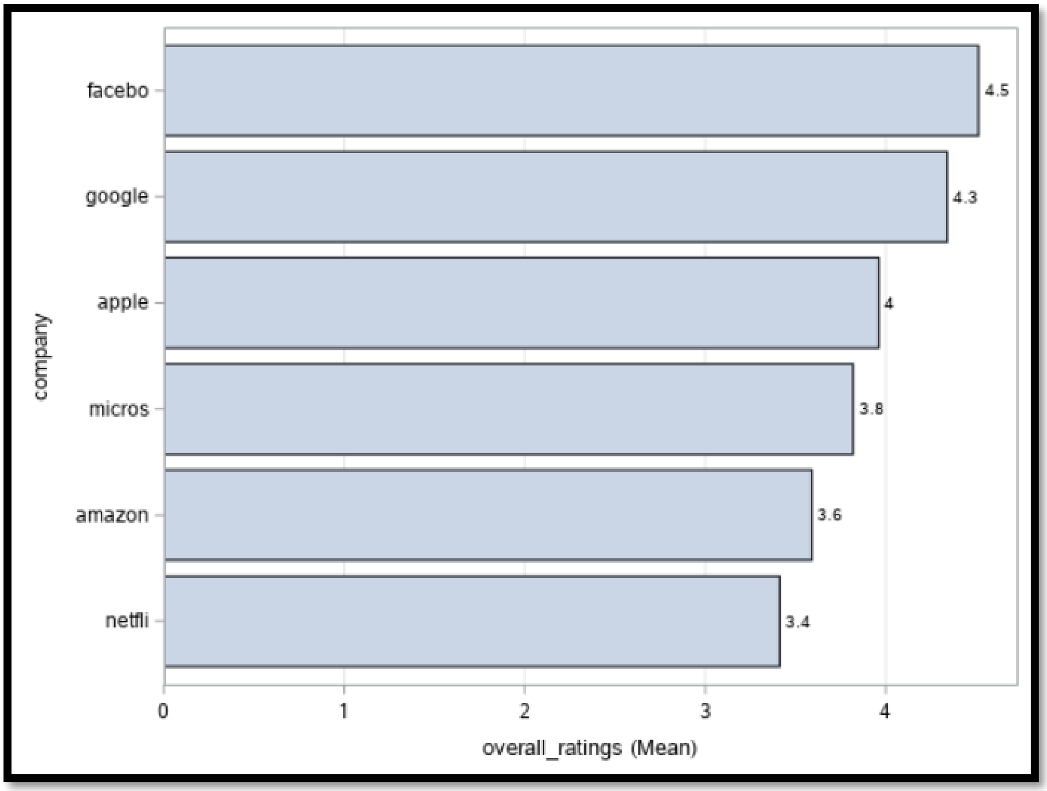
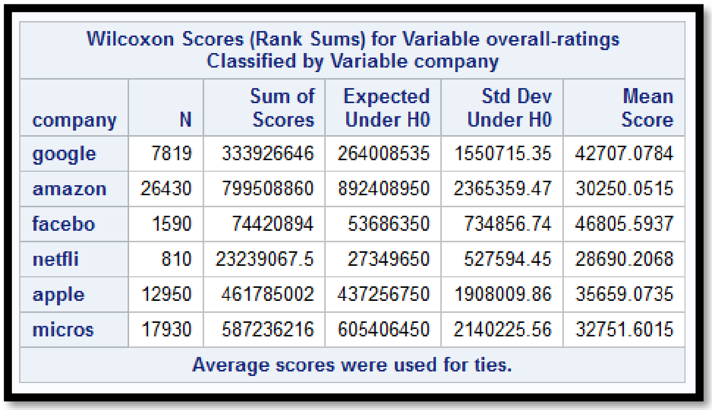


**Figure 4. Ranking of Company by Average overall_score**

These scores are all relatively good as the average score for all companies on *Glassdoor* is 3.3 (Glassdoor, 2017). Meaning that even the lowest rated company in our dataset is above average, however the higher rated companies are much better than average and can be studied in detail to determine the reason for their high scores. To determine whether the differences in the scores of the companies, we run a Kruskal-Wallis test for significance. This test was chosen given that our independent variable (company) has more than two levels and our dependent variable (overall_ratings) is ordinal. The figure below shows the results of the Kruskal-Wallis test:



Wilcoxon Scores (Rank Sums) for Variable overall-ratings Classified by Variable company

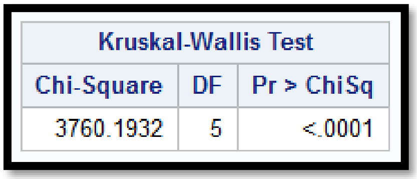
| company | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| google | 7819 | 333926646 | 264008535 | 1550715.35 | 42707.0784 |
| amazon | 26430 | 799508860 | 892408950 | 2365359.47 | 30250.0515 |
| facebo | 1590 | 74420894 | 53686350 | 734856.74 | 46805.5937 |
| netfli | 810 | 23239067.5 | 27349650 | 527594.45 | 28690.2068 |
| apple | 12950 | 461785002 | 437256750 | 1908009.86 | 35659.0735 |
| micros | 17930 | 587236216 | 605406450 | 2140225.56 | 32751.6015 |

Average scores were used for ties.

Null Hypothesis: There is no difference in the median values of overall_score for the companies in the dataset.

Alternative Hypothesis: Null hypothesis is not true.

(Prior to test we select a 95% confidence level)



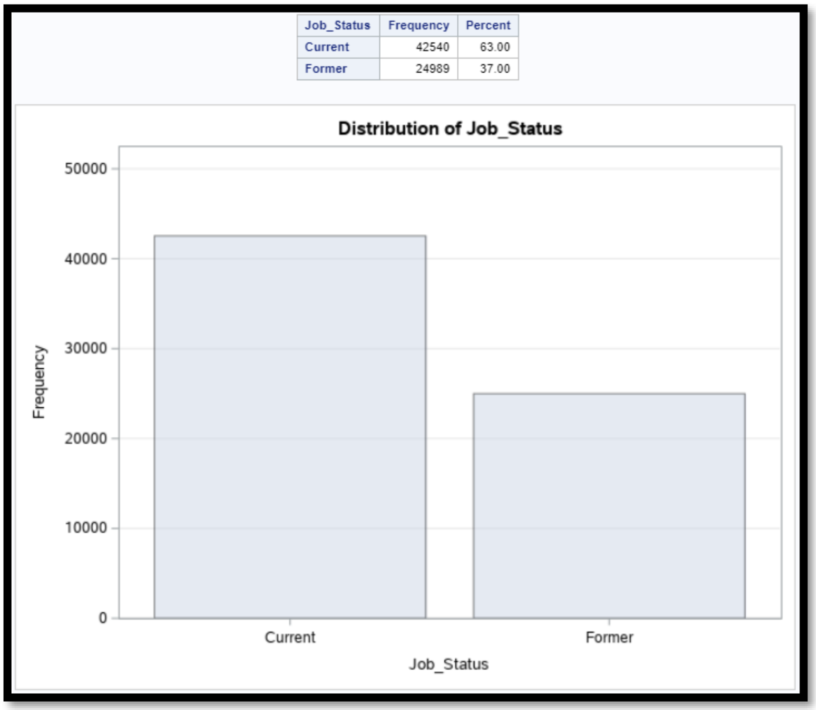
Kruskal-Wallis Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 3760.1932 | 5 | <.0001 |

**Figure 5. Results of Kruskal-Wallis Significance Test**

Given the results of the test, we reject the null hypothesis that there is no difference in the median overall scores given to companies. This signifies that the companies are indeed being rated differently by their employees and the observed difference are not due to chance. Once again, the highest scoring company was Facebook while the lowest was Netflix.

We then look at the average overall_score based on the job_status of the employee to determine if the scores of current and former employees are different. First, we look at the number of former and current employees as a whole in the figure below:



| Job_Status | Frequency | Percent |
|---|---|---|
| Current | 42540 | 63.00 |
| Former | 24989 | 37.00 |

**Figure 6. Frequency of Current vs. Former Employees in original Dataset**

The data set has a much higher number of current employees reviewing the companies than former. However, there are enough former employees that we can segment the data based on this variable and get meaningful insights. The figure below shows the average ranking given to companies based on the status of the reviewer:

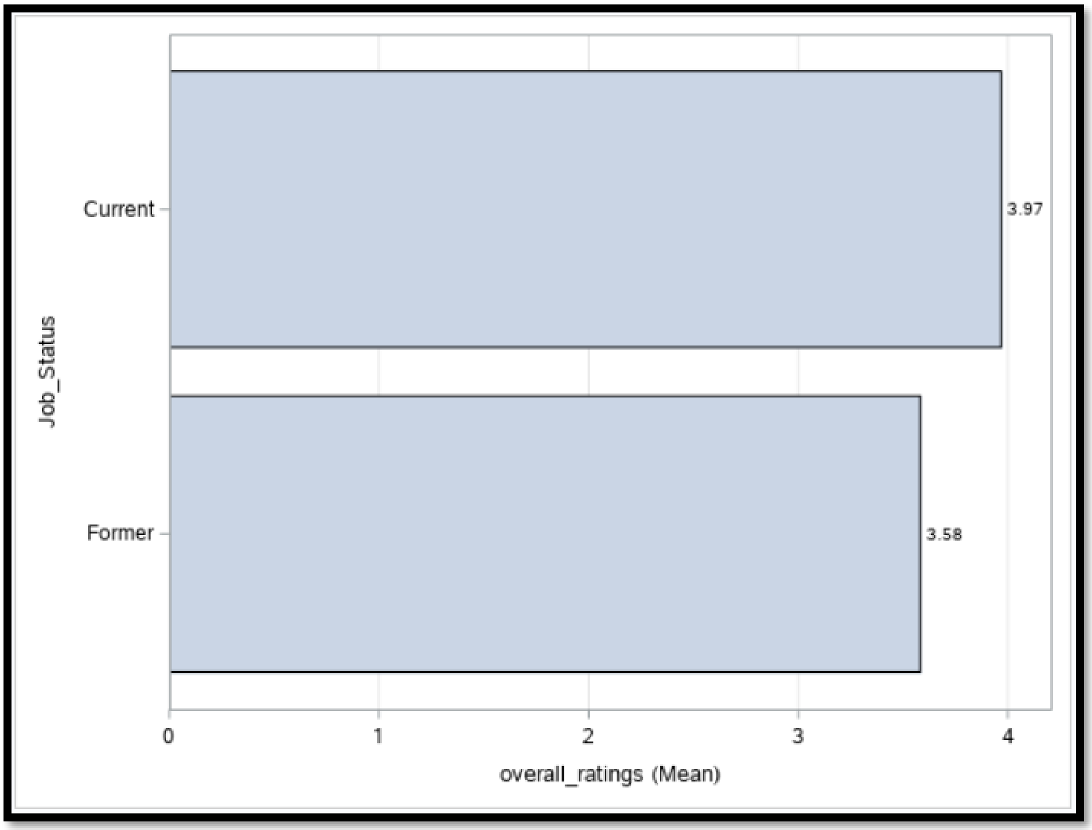


**Figure 7. Overall Score by Former vs. Current Employees**

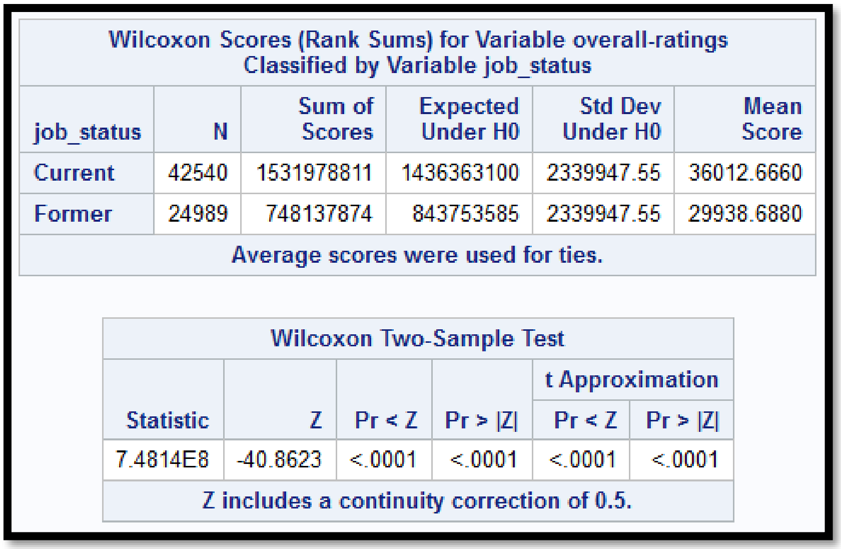
We again run the appropriate test to determine whether this difference is significant or due to chance. In this case, our independent variable is categorical with two levels (job_status) and our dependent variable is ordinal (overall_ratings), so the appropriate test is a Wilcoxon-Man Whitney test.

The results of this test are shown in figures 6 and 7 below:

Null Hypothesis: There is no difference in the median value of overall_ratings for current and former employees.

Alternative Hypothesis: The null hypothesis is not true.

(Prior to test we select a 95% confidence level)



Wilcoxon Scores (Rank Sums) for Variable overall-ratings
Classified by Variable job_status

| job_status | N | Sum of Scores | Expected Under H0 | Std Dev Under H0 | Mean Score |
|---|---|---|---|---|---|
| Current | 42540 | 1531978811 | 1436363100 | 2339947.55 | 36012.6660 |
| Former | 24989 | 748137874 | 843753585 | 2339947.55 | 29938.6880 |

Average scores were used for ties.

Wilcoxon Two-Sample Test

| Statistic | Z | Pr < Z | Pr > \|Z\| | t Approximation Pr < Z | t Approximation Pr > \|Z\| |
|---|---|---|---|---|---|
| 7.4814E8 | -40.8623 | <.0001 | <.0001 | <.0001 | <.0001 |

Z includes a continuity correction of 0.5.

**Figure 8. Results of Wilcoxon Test for Significance in overall_ratings**

The results of the Wilcoxon two sample test allow us to reject the null hypothesis and determine that there is a statistically significant difference between the median scores of the two groups. In this case, the current employees are more likely to rate their company higher than former employees.

Through the brief descriptive analysis, we find two statistically significant differences in scores within segments of the data. The text analysis portion focuses on these differences to discover if these differences can also be seen in the text variables. More specifically, we attempt to find whether we can observe in the comments section any differences in what current and former employees are saying as well as in what employees from different companies may be saying.

## TEXT ANALYSIS AND MODELING

We begin the text analysis by looking at the Pros_and_Cons dataset, which is not segmented in any way and includes random observations from all companies and all job statuses. This analysis gives us an overall idea of how employees feel at these companies in general. We should note that given that some companies have a much higher number of observations in the original dataset, this sampled dataset includes more observations from some companies than for others, since it was derived by a simple random sample. After the appropriate text preparation nodes are run (see methods section) the text profile node is used to explore the dataset.

**Text Profile-** As previously mentioned, the text profile node is used to extract terms that are most likely to describe the two target variable values (positive or negative). These terms are also likely to be significant rules when developing a text rule builder model.

Figure 4 shows the results of the text profile node. It visualizes the terms that are most likely to be found in the positive and negative comments, the brighter red color signifies a higher likelihood of the term appearing in that target value while the brighter blue shows a lesser likelihood of that term appearing:

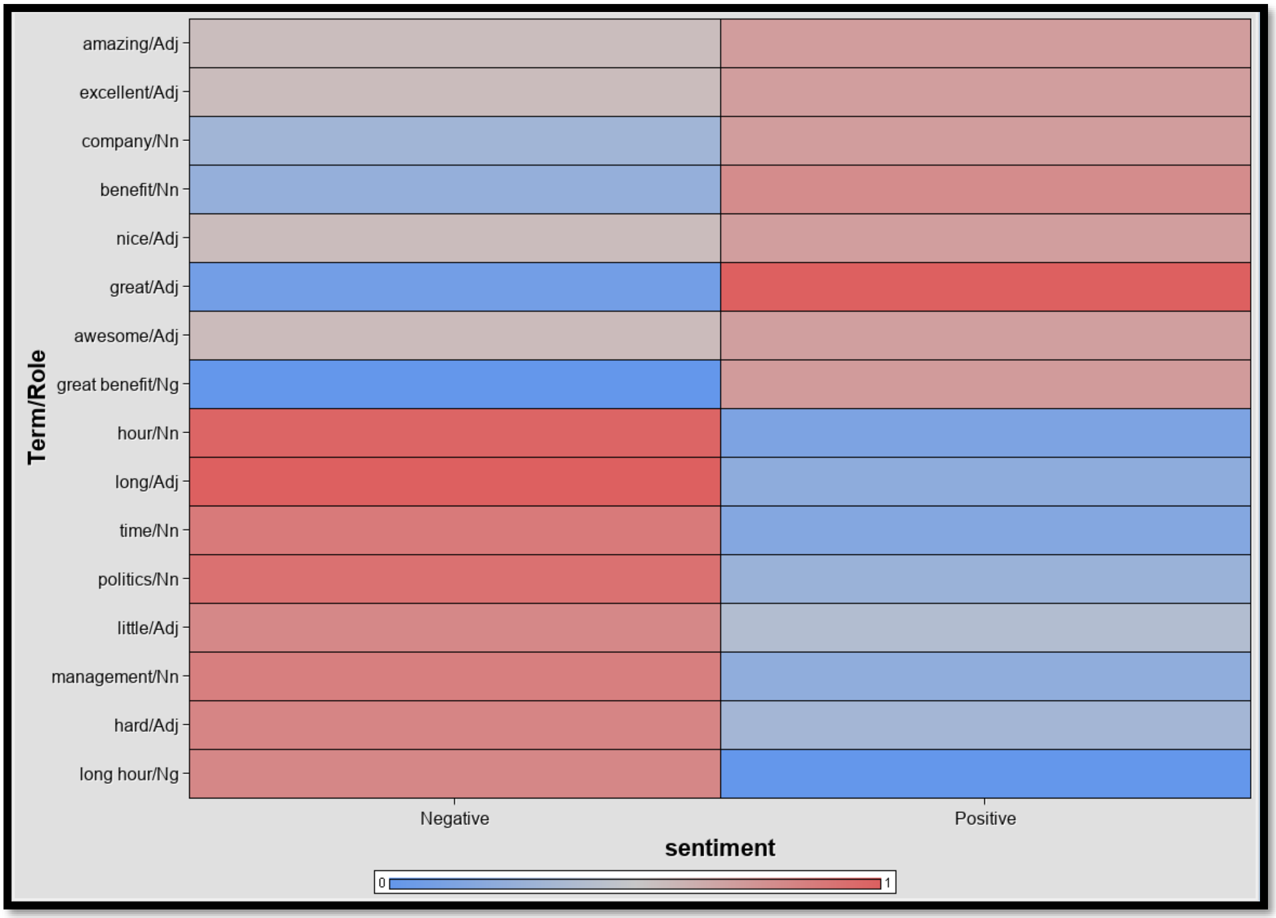


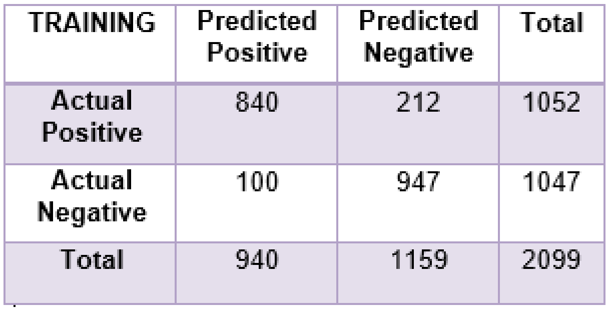
**Figure 9. Target Variable Value Term Profile**

The profile results show that the positive comments are most likely to include the term “great benefit” as well as various descriptive positive adjectives like “amazing”, “excellent”, and “great”. The negative comments are mostly centered around the hours or time that employees are working, with the common terms being “hour”, “long”, “time”. Other noteworthy terms are “management” and “politics” which are commonly used when making negative comments but not when making positive ones. From profiling the positive and negative comments, it appears that the single most important thing for driving positive feelings among employees is good benefits. For negative comments, the single most important driving factor may be working long hours.

**Text Rule Builder-** The text rule builder creates a model that gives us a more in depth look at the terms that are important for classifying negative and positive comments. While it is true that there are other modeling techniques that may do a better job of classifying the comments, the ease of interpretation of the rule builder model makes it an excellent choice for this research.

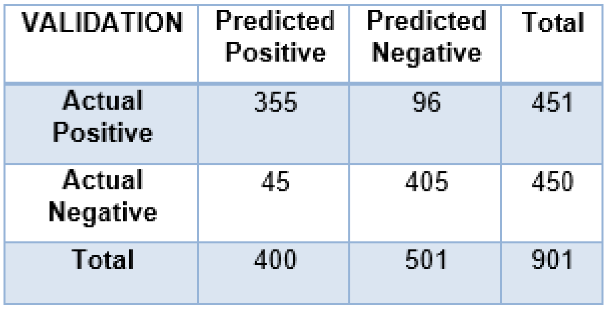
Once again, we show the results of the Pros_and_Cons dataset which includes observations from all companies and from both current and former employees. By interpreting the rules that the model creates to classify the comments, we can see what things employees are most commonly talking about when giving pros and cons about a company. The rules are simply a condition which classifies the comments based on the presence of a term. For example, if a positive rule is “happy”, then all comments which include this term will be classified as positive.

Before interpreting the rules created by the model however, we must ensure that the model is successfully classifying the observations. First, we look at the confusion matrices of the training and validation datasets:



| TRAINING | Predicted Positive | Predicted Negative | Total |
|---|---|---|---|
| Actual Positive | 840 | 212 | 1052 |
| Actual Negative | 100 | 947 | 1047 |
| Total | 940 | 1159 | 2099 |

**Table 1. Confusion Matrix for Training Data**



| VALIDATION | Predicted Positive | Predicted Negative | Total |
|---|---|---|---|
| Actual Positive | 355 | 96 | 451 |
| Actual Negative | 45 | 405 | 450 |
| Total | 400 | 501 | 901 |

**Table 2. Confusion Matrix for Validation Data**

To further asses the accuracy of the model we use the confusion matrices to derive the following performance metrics commonly used in evaluating text classification models:

**Recall (Sensitivity)-** Of all of the positive comments, how many were correctly classified positive?

**Positive Precision-** Of all of the positive predictions, how many of them were correct?

**Negative Precision-** Of all of the negative predictions, how many of them were correct?

**Overall Precision-** Of all of the observations in the dataset, how many were classified correctly?

**Misclassification Rate-** Of all of the observations classified, how many were incorrect?

Table 3 below shows these metrics for both training and validation data predictions:

| | Training | Validation |
|---|---|---|
| **Recall** | 79.8% | 78.7% |
| **Positive Precision** | 89.36% | 88.8% |
| **Negative Precision** | 81.7% | 80.8% |
| **Overall Precision** | 85.1% | 84.4% |
| **Misclassification Rate** | 14.9% | 15.6% |

**Table 3. Model Performance Assessment Metrics**

The performance assessment metrics signify a relatively successful model from which we can make valid interpretations. For the purpose of this paper, the primary metrics to look at are the overall precision and misclassification rate; this is because we are not interested in classifying positive or negative comments more or less that the other one.

Another thing to consider is the difference in how the model is predicting training and validation data. In this case, it is a good sign that the assessment metrics are stable, with only slight decay from training to validation. This suggests that the model is properly fit and is likely to translate well into additional data. Now that we have determined that the model is accomplishing its purpose well, we can look at the rules that it uses to classify the observations to learn the opinions of employees.

The rules show similar results as the text profile node. Some of the most important and useful rules for determining positive comments are shown in table 4 (not all rules are selected, only those which are determined to have managerial significance):

| Rule | Training True Positive/Total | Validation True Positive/Total |
|---|---|---|
| Long Hour | 45/45 | 14/14 |
| Difficult | 40/43 | 14/17 |
| Stressful | 24/26 | 10/11 |
| Politics | 59/68 | 23/26 |
| Time & Not: Benefits, Great, People, Pay | 104/120 | 36/49 |
| Shift | 32/38 | 6/7 |
| Pressure | 26/30 | 8/12 |

**Table 4. Selected Rules for Predicting Negative Comments**

The selected rules show that the environment and conditions of an employee are very important for his/her satisfaction. It will come as no surprise that negative comments often mention “long hours”, as well as “difficult”, “stressful”, and “high-pressure” work environments. Office politics are also found to be a common critique. Although management was a significant term in the topic profile node results it did not show up as a rule in the rule builder node. In a similar way, the most significant and potentially useful rules for determining positive comments are shown in table 5:

| Rule | Training True Positive/Total | Validation True Positive/Total |
|---|---|---|
| Great Benefits | 120/120 | 50/51 |
| Great Environment | 47/47 | 25/25 |
| Nice | 36/37 | 23/23 |
| Smart | 101/108 | 41/43 |
| Benefit | 220/236 | 97/101 |
| Fun | 48/53 | 21/25 |
| Friendly | 36/40 | 15/18 |
| Flexible | 55/60 | 26/28 |

**Table 5. Selected Rules for Predicting Positive Comments**

As shown in the results of the text topic node, the most common praise given to companies by employees are the benefits. Other important rules involve the nature of other people in the company such as “nice”, “smart”, “fun”, “friendly” and “flexible”. From these results, it is clear that the culture of a workplace is important for creating positive feelings in employees.

It is interesting to note that pay was not mentioned in either positive or negative comments. Further research must be done, but it appears from these results that increases in salaries may not be the best avenue for keeping employees satisfied and motivated.

## Facebook vs Amazon

Now that we have analyzed a sample of the data in the aggregate, we compare the text portions of two companies in the data. Facebook and Amazon are chosen for this comparison because Facebook (4.5/5) had the highest overall average score and Amazon (3.6/5) had the second lowest (we did not select

Netflix even though it had the lowest average review score because it only had 800 total reviews). The purpose of this comparison is to determine what may be causing the differences in these companies' overall scores.

Two rule-based classification models are built: one for each company. These are then assessed in a similar way as the model in the previous section. The rules are then compared for the two models to determine what negative rules are present for Amazon that may be driving its score down. The negative rules of the two companies are compared in table 6. Because we are interested in the differences of the two companies, we picked out the rules that were found in one companies' model but not in the other companies' model:

| **Facebook** | **Amazon** |
|---|---|
| Work-life Balance | Manager & Not: Good, Friendly, Great |
| Commute | Monotonous |
| Bus | Stressful |
| Busy | Repetitive |
| Cost | Frugality |
| Expectation | Favoritism |
| Require | Walking |
| Limit | Tire |
| Hard & NOT Work | Operational |

**Table 6. Negative Rules Unique to Facebook and Amazon.**

While some rules chosen by the models are common to both companies, by looking at the rules which are unique to each company, we can determine which things may be driving down the overall satisfaction of employees in the lesser rated company. For Facebook, it appears that employees are complaining about the cost of living as well as the commute given the rules "bus", "commute", and "cost". For Amazon, the negative rules that are mentioned look to be more regarding the actual conditions of the job such as "monotonous", "stressful", "repetitive", and "walking". These negative comments are very likely coming from employees who work warehouse related jobs given the nature of the work done by Amazon. It is very possible then that the reason for the much lower scores is that the Facebook employees are mostly working in offices doing technology related tasks while Amazon employees are coming from a broader range of occupations, which may have lower work satisfaction.

These results point to the importance of providing work which employees deem meaningful and engaging. Having employees perform the same repetitive and even boring tasks is likely to lead to lower work satisfaction.

## Current vs Former

The same technique was used to analyze the differences between former and current employees. This segmentation was chosen based on the difference in overall scores of current versus former employees. Current employees gave higher average scores than former employees and discovering why this is so may give us insights into how we can lower employee turnover. The rules were compared to determine the differences between them and gain insights regarding the importance placed on different factors by employees. The comparison of the negative rules for current vs former employees is found in table 8:

| Former | Current |
|---|---|
| Political | Favoritism |
| Management & NOT great, company | Promote |
| Time & Manager & NOT Great | Deadline |
| Bureaucracy | On call |
| Ranking | Dependent |

**Table 7. Unique Negative Rules for Former vs. Current Employees.**

The main differences between current and former employees appear to be regarding the management of the company. Terms like “political”, “management”, “bureaucracy”, and “ranking” suggest that former employees are more likely to perceive the company they reviewed as having poor management. Current employees on the other hand mention things which are not related to a single topic.

We also note that in both of the groups which we analyzed (Facebook vs. Amazon and Current versus Former) the group with the lower overall ratings had a management related rule while the higher scoring group did not.

The results of this comparison point heavily to the importance of having management that is perceived to be good by employees. It would make sense that an employee’s main reason for wanting to leave a company would be the stress caused by working for a difficult and unprofessional manager.

## CONCLUSION

### SUMMARY

The analysis of the dataset uncovered several differences in the data that can be utilized in managerial decision making. Through descriptive analytics and statistical tests this paper determined that there was a substantial difference in the employee satisfaction at each of the companies being reviewed. It also determined that former employees in our dataset gave lower ratings than current employees. From going deeper into the textual portions of the dataset, the following insights were uncovered:

- In general, one the most positive factors mentioned are benefits, good pay is seldom mentioned enough to be used as a predictive rule.
- In general, the most negative factor mentioned are long hours and time of work. Once again pay is not commonly mentioned when critiquing an employer.
- When comparing Facebook which was highly rated, to Amazon, which was lowly rated, the most commonly cited negative rules for Amazon are related to warehouse type work.
- When comparing current employees to former employees, the main negative issue mentioned by former which is absent from current employees is the management and political work environment of a company.
- For both groups that were compared, “management” appeared as a significant rule for the lower rated group but does not appear as a significant rule for the higher rating group. This suggests that employees are able to deal with other downsides of a company but are more likely to leave if they do not approve of the management.

### RECOMMENDATIONS

Based on the findings of this research there are several key drivers that can be addressed to improve the satisfaction of employees. From a compensation perspective, if the goal of a company is to increase the employee’s perception of their job and their happiness with it, it may be better to invest in added benefits or perks before focusing on increasing salary. This is because benefits and perks were mentioned very frequently as a positive rule while pay was not a common topic. This could be a source of savings to the company as increasing an added benefit could cost less than increasing salary**,** but provide higher impact to employees' satisfaction.

From a work-environment perspective, it is clear that having appropriate hours and good management is extremely important to employees. Overall, decreasing the stress levels and the amount of pressure put on employees is going to increase their satisfaction. Although these things will be impossible to completely eliminate from a work environment, having management that can motivate employees without added stress is one key to a productive workforce. It is also important for employees to feel like the people they work with are positive and nice.

While this research focused on a limited group of data, it shows a process which can be followed and repeated to derive insights from data that is freely available online. If more specific results and recommendations were desired by leaders of a company, it would be of value to get the data from their own company and run similar research. This would provide tailored and specific insights which would likely be even more useful and impactful to that company.

## LIMITATIONS AND FUTURE RESEARCH

One of the main limitations of this research is that the reviews are voluntary and may not be representative of the population of the company. There may be many other opinions that are not present in the dataset simply because those employees chose not to review the company. There are also many websites which can be used to review companies and we studied only one. Other websites may have differing opinions on the same company based on the different employees that chose to review the companies.

Another limitation was the sampling that was done to analyze the text data. Due to the high amount of processing power necessary to run models on text data, a small sample was chosen for each of the text classification models. It is likely that more accurate results could have been possible if we had the computing capacity and time to analyze all of the reviews instead of a sample of them.

As future research, we hope to expand on this research by adding other sources of data that may be useful in assessing employee satisfaction. For example, adding data on companies' turnover rate, compensation packages, average salaries, number of employees, and others, would allow us to compare companies regarding their reviews to see how the reviews align with the actual company data. Using financial data could also help us to determine if there exist any clear correlations between employee review content and financial success of a company. Additionally, we can also include more companies from different industries and regions and create profiles of employee opinions based on industry and area of the country.

## REFERENCES

T Alkuwaiti, Ahmed and Raman, Vinoth and Arun, Vijay and Rm, Palanivel and Prabaharan, Sivasankar. 2016. “Predicting The Exit Time of Employees in an Organization Using Statistical Models”. International Journal of Scientific & Technology Research. 5:213-217.

Chakraborty, Goutam and Liu, Jiawen and Sarkar, Kumar, Mantosh. 2016. “Feature-Based Sentiment Analysis on Android App Reviews Using SAS Text Miner and SAS Sentiment Analysis Studio.” Proceedings of the SAS Global Forum, Las Vegas.

Chakraborty, Goutam and Pagolu, Murali and Garla, Satish. 2013. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS. Cary, North Carolina: SAS Institute Inc.

Glassdoor, Inc. “50 HR and Recruiting Statistics for 2017”. Accessed Aug 2, 2019. https://resources.glassdoor.com/rs/899-LOT-464/images/50hr-recruiting-and-statistics-2017.pdf

SAS Institute Inc. 2017. “SAS Text Miner 14.3: Reference Help”. Accessed on July 17, 2019. https://documentation.sas.com/?docsetId=tmref&docsetTarget=p07ldkae1pq2jmn1p58qiqez004s.htm&docsetVersion=14.3&locale=en

S. Yekanialibeiglou and H. Demirkan. 2018. “Enhancing Creative Performance in Work Environments”. The Fifth International Conference on Design Creativity, Bath, UK: Demirkan Department of Interior Architecture and Environmental Design.
Available at: https://pdfs.semanticscholar.org/3f8a/8189f7ba1965674fb53f86b809e95a1736b9.pdf

UCLA Institute for Digital Research and Education. “Choosing the Correct Statistical Test in SAS, STATA, SPSS and R”. Accessed August 19, 2019. https://stats.idre.ucla.edu/other/mult-pkg/whatstat/.

Van Vulpen, Eric. “Predictive Analytics in Human Resources: Tutorial and 7 case studies”. AIHR Academy. Accessed on April 12, 2019. Available at https://www.analyticsinhr.com/blog/predictive-analytics-human-resources/.

## ACKNOWLEDGMENTS

I would like to thank Dr. Miriam McGaugh and Dr. Goutam Chakraborty for their guidance and mentoring as well as all of the members of the 2020 Ok. State MSBAN cohort from whom I am always learning.

## RECOMMENDED READING

- *Text Mining and Analysis. Practical Methods, Examples, and Case Studies Using SAS®*

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Luis David Martinez
Oklahoma State University
(505) 236-9324
Luis.Martinez11@okstate.edu
https://www.linkedin.com/in/ldavidmartinez/