

National College of Ireland

Project Submission Sheet – 2018/2019

Student Name: Pauline Stach
Student ID: 18177123
Programme: HDSDA January Start **Year:** 2018-2019
Module: Project
Lecturer: Enda Stafford
Submission Due Date: 03.12.2019
Project Title: Mining Employee Reviews: What Glassdoor.com Reveals About Company Culture
Word Count: 3657 (without references and appendix)

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Pauline Stach

Date: 03.12.2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Mining Employee Reviews: What Glassdoor.com Reveals About Company Culture

Pauline Stach
National College of Ireland
Dublin, Ireland
x18177123@student.ncirl.ie

Abstract—Company or Corporate Culture is a growing matter of interest in social science and the corporate world because it is found to be linked to two major aspects of organisational success: Employee Satisfaction and Corporate Performance. The question, however, is how corporate culture is defined or even more, who it is defined by. This project aims at analysing employee reviews from the homepage glassdoor.com in order to see how the aspect of company culture is defined by individuals, rather than by large corporations themselves. A statistical analysis will yield results showing that reviews of six tech companies are significantly diverging and that company culture is strongly correlated to overall job satisfaction. A closer linguistic and sentiment analysis will compare and match reviews to officially advertised corporate culture showing that culture values form a large part of review content and an alignment of the latter does at least partially reflect the general company ranking on glassdoor.com. Finally, an attempt is made to classify reviews and predict companies using classification algorithms such as Decision Trees, Random Forest and Naïve Bayes. A word2vec model will be implemented to allow for vector-based comparisons.

Keywords – Text mining, employee satisfaction, company culture, sentiment analysis, classifiers, word2vec

I. INTRODUCTION

The design of this project is based on the assumption that employee satisfaction is linked to company culture and corporate integrity [1]. Further, the idea that company culture is associated with company performance [2, 3, 4], first came up in the 80s and 90s of the 20th century [1, 5]. Interestingly, strong company culture regardless of its actual content is reported to have a strong influence on company success [6]. Since company culture, or also referred to as individual values, has strong ties to different levels within a corporation, it plays an important role in business strategy. For future references, company culture will be defined as “a set of norms and values that are widely shared and strongly held throughout the organization” [7]. This definition is especially useful from an analytical point of view, since it makes the concept of culture measurable.

Along with this knowledge comes the question of *what* company culture is for every single organisation and *who* defines it. Big data can thus be exploited in order to answer this question and help understand the concept of culture and how it is being perceived and evaluated by individual people in a more scientific way. Therefore, one of the best, since large and openly accessible, resources for anonymous employee review was chosen as a foundation for this project. Glassdoor.com is a homepage that enables current and former employees to write and view reviews, submit and view salaries or use it as a platform for job application and interview preparation. The dataset was originally sourced

from the homepage kaggle.com and consists of 67,000 rows of reviews on six major tech companies (*Amazon, Apple, Facebook, Google, Netflix, Microsoft*) with locations all over the world. Next to meta data like location, time, employee status and job title, there is written as well as rated feedback in six different categories, which allows for a linguistic as well as statistical analysis.

The following analysis will focus on the rating feedback categories and the *pros* review section, which was assumed to give more valuable insights than the *cons* and *summary* section. Keeping this in mind, the dataset has a much larger potential for various and more detailed analysis than performed here. The approach at hand is to establish that reviews print individual pictures of companies and that the difference between these is measurable and could ideally be used for classification and prediction tasks. The idea of predicting review ratings and deciding whether or not a review could be helpful to users could be valuable for companies who are, for example, recruiting potential employees through glassdoor.com [8]. Further, aligning reviews and official advertised company culture statements can be regarded as an indicator for strong and lived up corporate values and general employee satisfaction. In return, diverging notions should be taken as an issue to act upon, since weak company culture can be financially harmful. The project is based on the hypothesis that *Glassdoor reviews contain information that indicate strong or weak company culture*, which can be a valuable asset for business strategy.

II. BACKGROUND

A. Company Culture

Company *Climate and Culture* is one of five elements of an organisational architecture besides *Strategy, Structure, Core Processes/Systems* and *Skills* [9]. Among these, *Culture* is at the core of an organisational structure. *Culture, Climate* and *Moral* form the internal environment which is opposed to an external environment consisting of customer, suppliers, competitors etc. (Appendix I). These two environments shape an organization and whereas the external factors are rather easy to define, internal company culture is not. It is a collection of different norms, values and behaviours such as *Integrity, Teamwork* and *Innovation* [1]. Appendix II lists nine core values that can be further broken down into key words like *Ethics, Accountability, Trust, Honesty, Responsibility, Fairness* etc. (for the core value of *Integrity*). These will be used as categories for the final part of the analysis.

B. Mining Employee Reviews

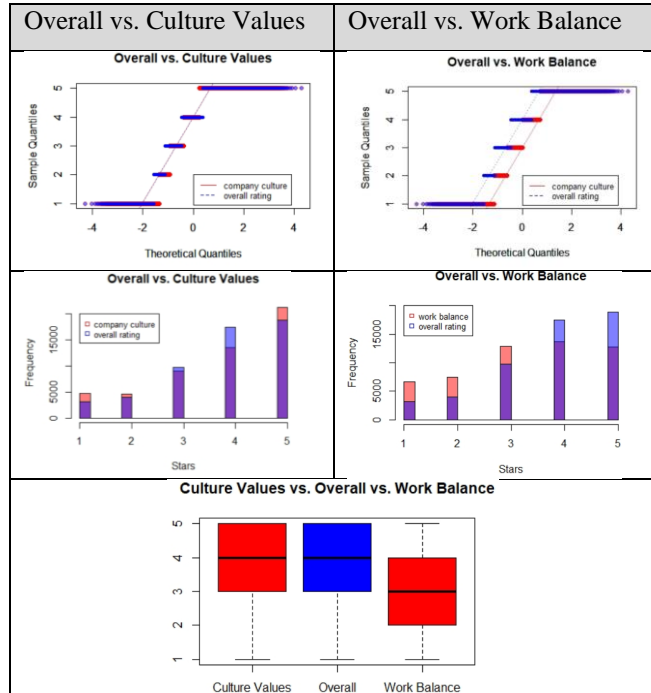
In order to find and define the right values for a given company, a promising bottom-up approach is to see what employees say instead of what leaders and CEOs wish to portrait [1, 10]. However, traditional methods of collecting feedback such as interviews or questionnaires are lacking in three major aspects. They are either not anonymised, not quantitative or not open enough to allow for more individual opinions. Thus, the glassdoor dataset presents itself as an ideal source of information in order to define core values and company culture. In recent social science research, this data has been successfully used to, for example, link employee satisfaction to corporate performance [5].

III. EXPLORATORY DATA ANALYSIS

A. Statistical Analysis

The statistical analysis of the numeric rating in six different categories (1. *overall*, 2. *work-balance*, 3. *culture-values*, 4. *career opportunities*, 5. *benefits* and 6. *senior management*) will provide first insights into characteristics of the data such as normality, disparity between groups and factors. First, graphical aids such as Q-Q plots, histograms and boxplots indicate that the data is not normally distributed which is confirmed by normality tests resulting in a p-value < 2.2e-16 for all groups. Interestingly, the conducted tests already show that the distribution of the *overall* rating is closer to the one of *culture-values* than to, for example, *work-balance*, as shown below, which indicates a close alignment between both groups (Appendix III).

TABLE I



More supporting evidence for a close alignment between overall and culture-values rating comes from a correlation matrix. It is clear to see (Fig. 1) that these two categories are strongly and higher correlated (75.86%), in contrast to work-balance (61.24%), opportunities (68.63%), benefits (53.79%) and senior-management (72.58%). This could indicate that for

employees, overall job satisfaction is dependent on or at least mostly influenced by company culture values.

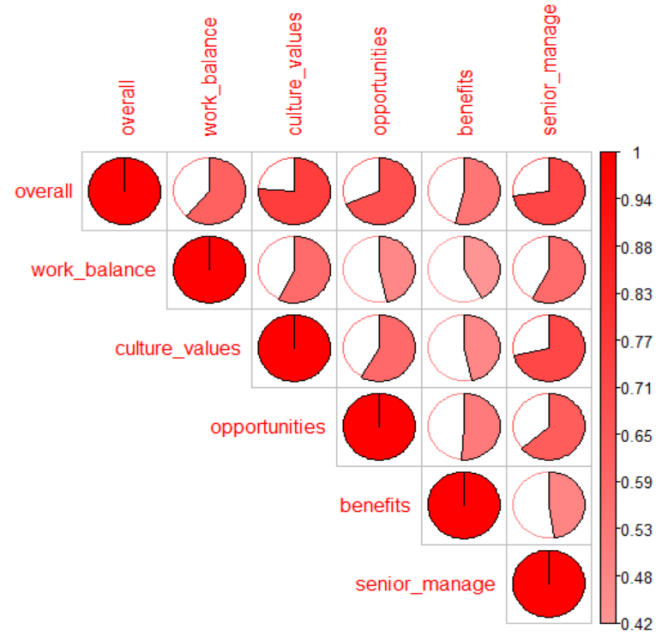


Fig. 1. Factor Correlation

Despite the fact that all distributions are non-parametric, a *two-way ANOVA* test was conducted in order to see if the two factors of *rating category* and *company* have similar/equal ratings (H1 and H2) and do not interact with each other (H3). The result shows that both factors as well as their interaction have a p-value of less than $\alpha=0.05$ which means that there is a statistically significant interaction effect (Appendix IV), meaning that both, rating category and company have a significant effect on the rating outcome and that both factors influence each other in terms of rating outcome.

Altogether, the statistical analysis shows that there is a relation between overall job satisfaction and company culture and sufficient evidence for diversity in the data in order to attempt to categorise and further predict companies based on their ratings.

B. Sentiment Analysis

In order to test the hypothesis at the center of this project, being that glassdoor reviews contain information on company culture, a linguistic approach is inevitable. A sentiment analysis was conducted to be able to see what employees say about their (previous) employers. The main focus here was on the *pros* review section for the following reason: this section was assumed to contain many and various positive aspects about a company. Even though much could be learned from negative feedback, the sentiment analysis did not capture word clusters including negation and qualifiers such as, ‘no opportunities’, ‘rarely communicating’ or ‘pretty unfriendly people’, which would have skewed the results. Instead, it only evaluated single word tokens, their frequency of occurrence and degree of positivity on its own. In addition to the first limitation mentioned above, other linguistic

feature such as sarcasm and irony were not considered in the analytical approach.

The analysis compared the six subsets of each company, sentences were tokenized, stop words, such as “and”, “the”, or “I” removed and content words were counted. At this stage, it needs to be mentioned that the number of reviews/rows and, therefore, tokens differs significantly between the six companies (Amazon 26430 rows, Apple 12950 rows, Facebook 1590 rows, Google 7819 rows, Netflix 810 rows and Microsoft 17930 rows). This difference will be considered in the sampling at a later stage. The first interesting finding is that the reviews varied significantly in their mean length of utterance (MLU), with Facebook yielding the longest reviews with a mean length of 39 (content) words per review and Google with only 19 (content) words. Secondly, the most positively evaluated words were in Amazon reviews (i.e. “benefits”, “fast”, “smart”), and the most negatively evaluated words in Google reviews (i.e. “challenging”, “hard”, “bad”). On average, Facebook has the highest average sentiment score (Table II). Similar to the MLU, the sentiment scores are significantly different.

TABLE II

	Amazon	Apple	Facebook	Google	Netflix	Microsoft
MLU	21.15	21.47	39.13	19.08	31.5	20.9
Max pos	47	29	41	44	15	22
Max neg	23	13	12	34	9	10
Avg	1.57	2	2.78	2.04	2.48	1.94

Further, the 15 most often occurring tokens across all companies are relatively similar which means that in general, employees seem to value certain aspects in similar ways and all companies seem to provide these values. As an example, the graphic below shows the word cloud of most frequent tokens for the company Netflix.



Fig. II Word Cloud Pros Review Company Netflix

Taking together all 10 most often used tokens and ranking them based on their occurrence and ranking, it turns out that *culture* is on position 4, occurring in 5/6 top 10 rankings with the average position of 6, after *environment* (pos 3), *benefits* (pos 2) and *people* (pos 1) (Appendix VII). This shows, how important the aspect of company culture is

to employees and secondly, that most of the other tokens represent values that occur within the vocabulary of company culture such as “people”, “environment”, “life”, “freedom”, or “employees”, which leads further to the next part of the analysis.

IV. COMPANY CULTURE IN REVIEWS

Having discovered that many of the most often used words in the glassdoor reviews represent culture values (with regard to the list in Appendix II), the aim was now to see if these values match with those publicly proclaimed by each individual company. Therefore, company websites and the homepage <http://panmore.com/> were scanned in order to find statements on culture, values and general company characteristics. On average, each company is communicating around 10-20 values that they use to advertise themselves and which form a great part of their corporate identity. These values were categorised based on the 9 categories of *Integrity*, *Teamwork*, *Innovation*, *Respect*, *Quality*, *Safety*, *Community*, *Communication*, *Reward* and *Other* (Appendix XI). At this stage, certain tendencies could be found, for example that Microsoft represents itself more with Integrity values, and Netflix with Respect values or Amazon with Innovation values.

Each company’s individual set of tokens, from the glassdoor dataset, was subset based on the list of words described above with the following results:

- With regard to all tokens, there is a high level of matches between officially advertised company culture values and glassdoor review values. All proclaimed values occur in Amazon’s, Apple’s, Facebook’s, Google’s and Microsoft’s reviews, whereas Netflix reviews match with 65%. The comparatively low result for Netflix could derive from the fact that Netflix has the fewest of all reviews.
- Focussing on the 100 most frequently used tokens, the percentage of matching values shift with Google being in first position (60%), followed by Amazon with more than 30% less matches, Apple (18%), Microsoft (16%), Facebook (15%) and Netflix (6%).

TABLE III

	Amazon	Apple	Facebook	Google	Netflix	Microsoft
All	100%	100%	100%	100%	65%	100%
Top 100	25%	18%	15%	60%	6%	16%

This shows us that in general, proclaimed company values are very well represented in employee reviews. They appear, in most cases, at least once in all reviews and many of them come up in the most frequently used words, especially for Google. Google employees’ most often given positive feedback is with 60% reflecting what the company is itself is advertising, namely that they value *people*, *innovation*, *flexibility*, *environment*, *fun* and *smartness* (Appendix XII, XIII). In this particular case, the argument that glassdoor reviews represent company values would be quite strong. However, even though the other five companies show a certain degree of overlap, it is a relatively small one.

Thus, five out of six companies lack in an accurate representation of company values when it comes to employee feedback.

Comparing this to the company rating deriving from the overall and culture-value rating categories, the results are as follows:

TABLE IV

Ranking	Most Value Matches	Culture-Value Rating	Overall Rating
1	Google	Facebook	Facebook
2	Amazon	Google	Google
3	Apple	Apple	Apple
4	Microsoft	Microsoft	Microsoft
5	Facebook	Amazon	Amazon
6	Netflix	Netflix	Netflix

Google, having the highest percentage of matching values is on position 2. on the culture-value and overall rating. Apple (3), Microsoft (4) and Netflix (6) have completely aligning positions across all categories, Amazon performs better in matching values compared to the two rating categories and Facebook performs relatively poor, dropping from 1st to 5th position. In four out of six cases, the degree of overlap between review values and official values is representative of the companies' overall ranking position. Hence, the initial hypothesis that *Glassdoor reviews contain information that indicate strong or weak company culture*, can be supported. For further analysis, a closer examination of the disagreement for the two companies of Facebook and Amazon would be of interest but will not be further focused on at this stage to keep the analysis manageable within the allocated timeframe and scope.

V. COMPANY CLASSIFICATION

Having established that glassdoor reviews significantly differ with regard to rating but also to the content of positive feedback, an attempt was made to classify and predict companies based on reviews. Both, numeric as well as linguistic data was used to run different classification algorithms: *Decision Trees (C5.0 Algorithm)* and *Random Forest*, *Naïve Bayes* and *kNN*. Giving a brief overview: all classifiers chosen are supervised algorithms. Decisions trees break down data into smaller and smaller subsets until the tree is fully grown or the node error rate has been reached. The trees generated here are fairly large with a size of 288 (Appendix XV, XVI). The Random Forest algorithm constructs a multitude of trees and classifies based on the mode of predicted classes. Naïve Bayes is a probabilistic classifier that assumes independence among predictors, which is rarely the case, since most classes are depending or correlating with each other (as outlined for the rating categories). In order to perform kNN, which classifies based on minimum distance between training and testing sample, the linguistic data was vectorised using the high-level neural networks library *Keras* in R [11]. After initial problems classifying the character data (tokens), the word embeddings were used for each of the algorithms outlined above. Steps used in building the word2vec model are:

- 1 Pre-processing: transforming reviews into sequence of 20000 integer tokens.

- 2 Skip-Gram Model: using each token as input to a log-linear classifier with a projection layer in order to predict

surrounding words. A generator function receiving a text vector, a tokeniser and different arguments such as the size of the window around each target word.

- 3 Keras Model: using Keras functional API, defining the embedding matrix, using cosine similarity as measure of similarity to compare target and context vector.

- 4 Model Training: using one instead of suggested five epochs, resulting in (only) 6 hours of training.

The result, a 100+ dimensional embedding matrix, can be found in the attached R code. The word embeddings showed clusters of semantically close words and those often used in the same context (Appendix XIV). Having two different types of data, single numeric, complex 100+ dimensional word vectors, classification was exercised in various attempts. Due to the difference in sample size for each company, subsets of different sizes were tested using random sampling techniques, and in some cases, Netflix as a category was excluded since it narrowed down the sample size significantly. F1 score was used as a measure of similarity, after initially testing different samples sizes and not sampling equal samples, also considering both, precision and recall. As a side note, classifying unequal samples yielded much *better*, but heavily skewed results towards the overrepresented classes and were not further considered. Unfortunately, all classification attempts achieved less than satisfying results as presented in TABLE V (also see Appendix XV). The highest similarity of .3727 was scored using Random Forest for numeric rating data.

TABLE V

	Vectors	Numeric ratings
Tree	.2425	.3587
Random Forest	.2627	.3727
Naïve Bayes	.1294	.3421
kNN	.2134	-

One explanation for the poor performance could be that the data was simply not diverse enough. With star rating categories only differentiating 5 levels, there is not much room for individual and decisive patterns. Appendix XVIII shows the attempt to draw clusters for each company, which results in an at least 90% overlap of clusters. As for the embedding classification, it appeared to be surprising that it performed even poorer than the star rating classification. However, based on the findings in the sentiment analysis, it can be argued that the reviews were often very similar. The 10 most commonly used words for each company formed a total of 33 tokens, meaning that all companies shared at least half of their top ten tokens with another company. With regard to decision trees, the tree size could be critically evaluated since large trees can easily be overfitted which prevents them from generalising to new data.

At this stage, the results of the statistical and sentiment analysis were significant enough to accept the fact that classification was not becoming the major focus of this project.

VI. CONCLUSION AND FUTURE DIRECTION

The aim of this project was to analyse the concept of company culture in relation to employee reviews. Company culture or rather its practical realisation, being at the core of organizational architecture and therefore at least partly indicative for its performance, is a growing matter of interest in today's corporate world. The underlying hypothesis was that glassdoor.com, a job application platform, captures reviews containing information that indicate strong or weak company culture. The approach of analysing employee feedback is contrasting a more top-down approach where companies and their leaders define cultural values which might often not be representative of the actual work life situation of their employees.

The analysis focused on two types of data, numeric ratings as well as written feedback. This allowed for different analytical approaches, providing various insights that could be combined to a coherent result. The statistical analysis revealed a strong correlation between culture-value and overall rating categories which could be indicative for the fact that employees consider cultural values as one of the most important aspects with regard to overall job satisfaction. The linguistic analysis showed a more in-depth picture of how company culture is represented in reviews. In general, positive feedback includes many words that relate to culture values such as *Ethics*, *Accountability*, *Trust*, *Honesty*, *Responsibility*, *Fairness* etc. Counting word frequencies and general number of occurrences across different company samples, the word *Culture* is on position 4. behind *Environment*, *Benefits* and *People*, which again, supports the importance of company culture for employees. In a final stage, values mentioned in reviews were matched to officially advertised company values. The amount of aligning values was then compared to the individual company's position in the ranking, resulting in a general agreement. Companies, whose reviews contain many/few of the advertised values were mostly ranked accordingly. All this information could be valuable from different perspectives. Customers who are using (glassdoor) reviews as a reference to decide whether or not they will apply for a job vacancy, will get a good overview of a company's culture and if this in accordance with their own values and work ethics. Secondly, companies might regard employee reviews as valuable feedback in order to either adjust their advertised values or introduce interventions in areas where there seem to be ambivalences.

The topic of review mining and company culture is, as mentioned initially, very important in today's job market, especially in tech, where companies try to distinguish themselves through their values and benefits. Employee reviews are therefore the ideal source of data to investigate company culture and to find and define areas for improvement. The dataset at hand has not been used to its full capacity due to the scope of this project. Many different analytical approaches could be taken, investigating, for example, negative and summary reviews, status of employment (current vs former employee), time and location of employment etc., only to mention a few. Also, the initial aim to tie company culture to corporate performance was not further investigated due to the limitations of this project. It

would, however, be very interesting to see, how the established ranking matches company performance rankings, which are accessible on, for example, forbes.com.

REFERENCES

- [1] L. Guiso, P. Sapienza, and L. Zingales, "The Value of Corporate Culture", *Journal of Financial Economics*, vol. 117, no. 1, pp. 60–76, July 2015.
- [2] D. R. Denison, "Corporate culture and organizational effectiveness", *The Academy of Management Reviews*, vol. 16, pp. 205–227, Jan. 1991.
- [3] C. Ostroff, "The relationship between satisfaction, attitudes, and performance: An organizational level analysis", *Journal of applied psychology*, vol. 77, no. 6, p.963, Dec. 1992.
- [4] J. K. Harter, F. L. Schmidt, and T. L. Hayes, "Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis." *Journal of applied psychology*, vol. 87, no. 2, pp. 268 –279, 2002.
- [5] N. Luo, Y. Zhou, and J. J. Shon, "Employee Satisfaction and Corporate Performance: Mining Employee Reviews on Glassdoor.com", *Thirty Seventh International Conference on Information Systems, Dublin, Ireland, 2016*, pp. 1–16. [Online]. Available: <https://pdfs.semanticscholar.org/b784/71ca2ac990e3ab2d70283e42e2bb5d3a8f7a.pdf> [Accessed on: Nov. 26, 2019].
- [6] G. G. Gordon and N. DiTomaso, "Predicting Corporate Performance from Organisational Culture", *Journal of Management Studies*, vol. 29, no. 6, pp. 783–798, Nov. 1992.
- [7] C. O'Reilly, and J. Chatman, "Culture as social control: corporations, cults, and commitment.", *Research in Organizational Behavior*, vol.18, pp.157–200, Jan. 1996.
- [8] Y. Berdugo, "Review Rating Prediction: A Combined Approach", 2019. [Online]. Available: <https://towardsdatascience.com/review-rating-prediction-a-combined-approach-538c617c495c> [Accessed on: Nov. 26, 2019].
- [9] M. D. Watkins, *Leadership Transitions: The Watkins Collection*, Harvard: Harvard Business Review Press, 2014.
- [10] J. D. Meier, "Brilliant Examples of Company Values (Amazon, Google, Microsoft and More)", 2017. [Online]. Available: <https://www.linkedin.com/pulse/brilliant-examples-company-values-amazon-google-microsoft-j-d-meier> [Accessed on: Nov. 26, 2019].
- [11] D. Falbel, "Word Embeddings with Keras", 2017. [Online]. Available: <https://blogs.rstudio.com/tensorflow/posts/2017-12-22-word-embeddings-with-keras/> [Accessed on: Nov. 26, 2019].

APPENDIX

I. Elements of Organisational Architecture [9]

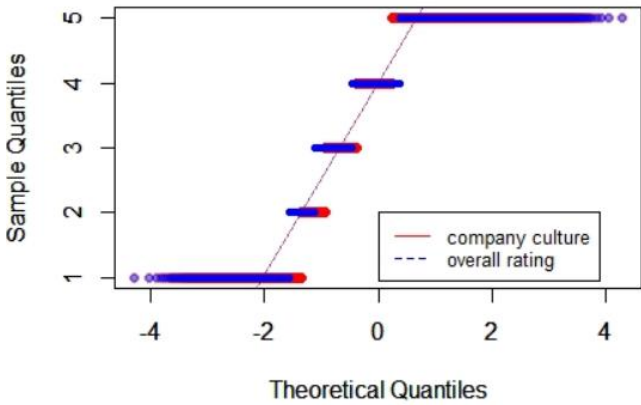
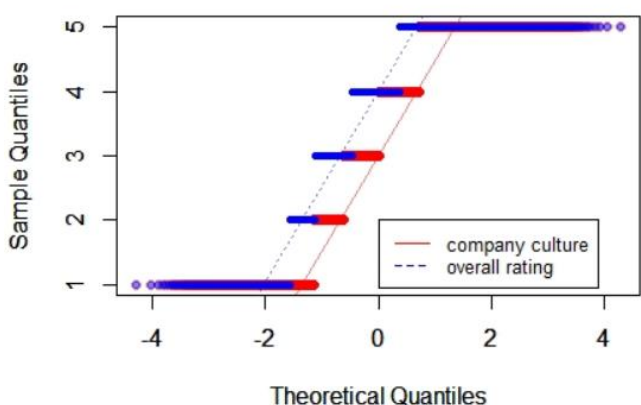
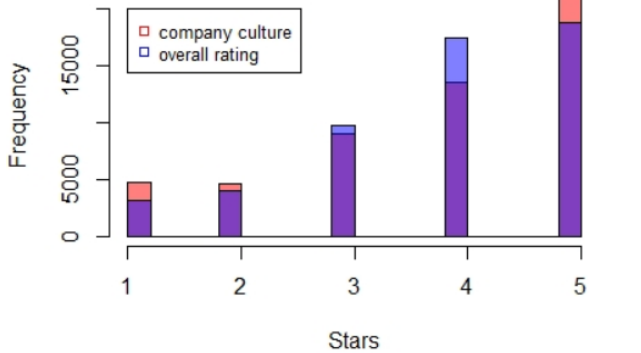
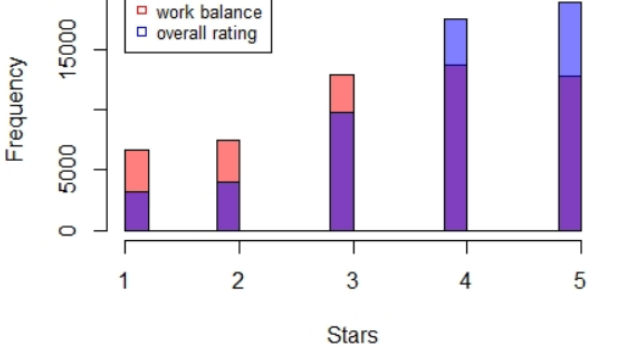
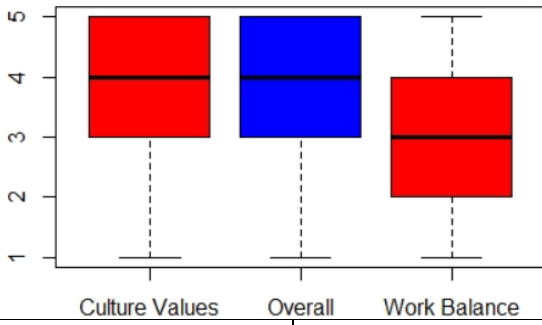
Elements of organizational architecture



II. Nine Categories of Advertised Corporate Value [1]

<i>Advertised Corporate Value</i>	<i>Keywords</i>
<i>Integrity</i>	<i>Integrity + Ethics + Accountability + Trust + Honesty + Responsibility + Fairness + Transparency + Ownership</i>
<i>Teamwork</i>	<i>Teamwork + Collaboration/Cooperation</i>
<i>Innovation</i>	<i>Innovation + Creativity + Excellence + Improvement + Passion + Pride + Leadership + Growth + Efficiency + Results</i>
<i>Respect</i>	<i>Respect + Diversity + Inclusion + Development + Talent + Dignity + Empowerment</i>
<i>Quality</i>	<i>Quality + Customer + Needs + Commitment + Dedication + Value</i>
<i>Safety</i>	<i>Safety + Health + Work/Life balance + Flexibility</i>
<i>Community</i>	<i>Community + Environment + Caring + Citizenship</i>
<i>Communication</i>	<i>Communication + Openness</i>
<i>Reward</i>	<i>Hard work + Reward + Fun + Energy</i>

III. Statistical Analysis: Normality Plots

Overall vs. Culture Values		Overall vs. Work Balance	
<p>Overall vs. Culture Values</p> 		<p>Overall vs. Work Balance</p> 	
<p>Overall vs. Culture Values</p> 		<p>Overall vs. Work Balance</p> 	
<p>Culture Values vs. Overall vs. Work Balance</p> 			
Oveall		ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST D = 0.23679, P-VALUE < 2.2E-16	
CULTURE VALUES		ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST D = 0.22484, P-VALUE < 2.2E-16	
WORK BALANCE		ONE-SAMPLE KOLMOGOROV-SMIRNOV TEST D = 0.18573, P-VALUE < 2.2E-16	

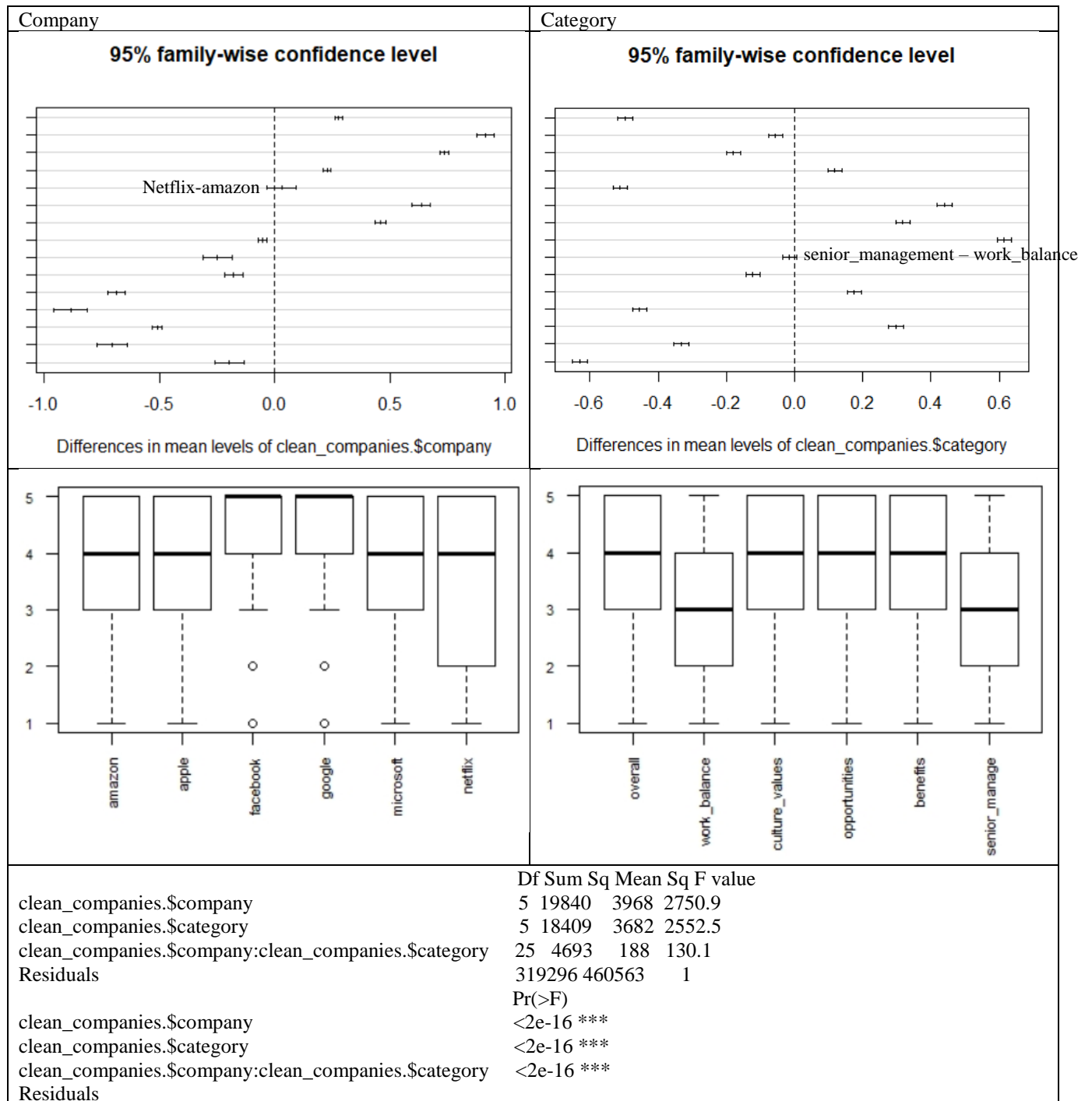
IV. Statistical Analysis: Two-Way ANOVA

H1: All Rating Categories have, on average, the same rating

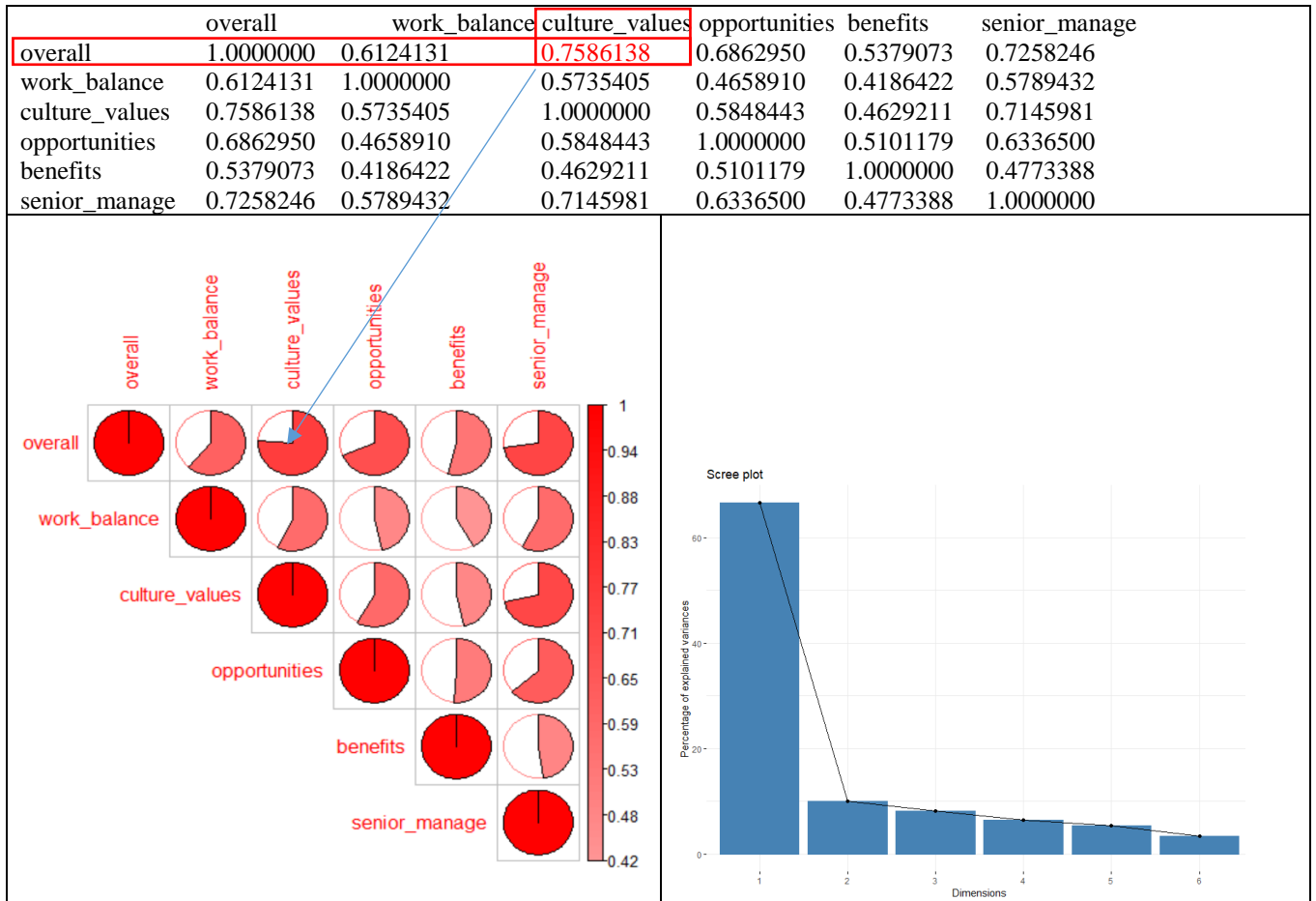
H2: All Companies have, on average, the same rating

H3: Both factors, Rating Categories and Companies, are independent or do not have an interaction effect

- Both Factors as well as their interaction are less than .05
- Sig difference everywhere except for: Netflix-amazon, senior_management-work_balance
- Work-balance and senior-management have lowest ratings
- Facebook and google have the highest ratings

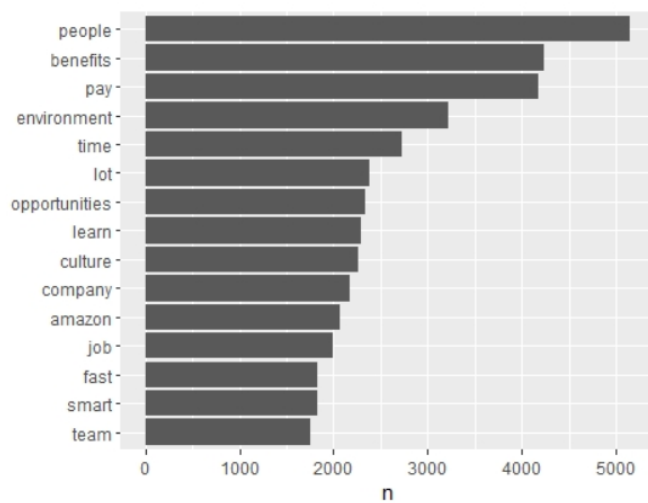


V. Statistical Analysis: PCA attempt

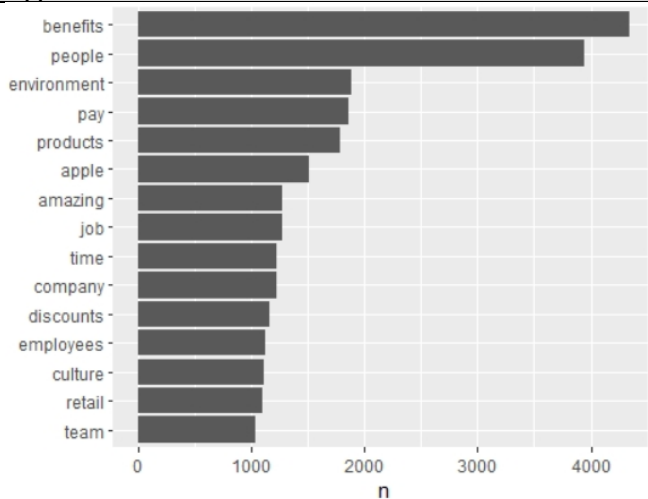


VI. Sentiment Analysis: Word Frequency Counts and Word Clouds

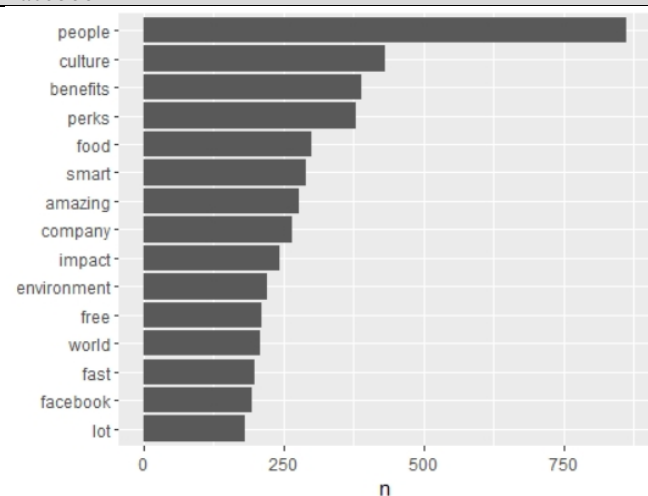
Amazon



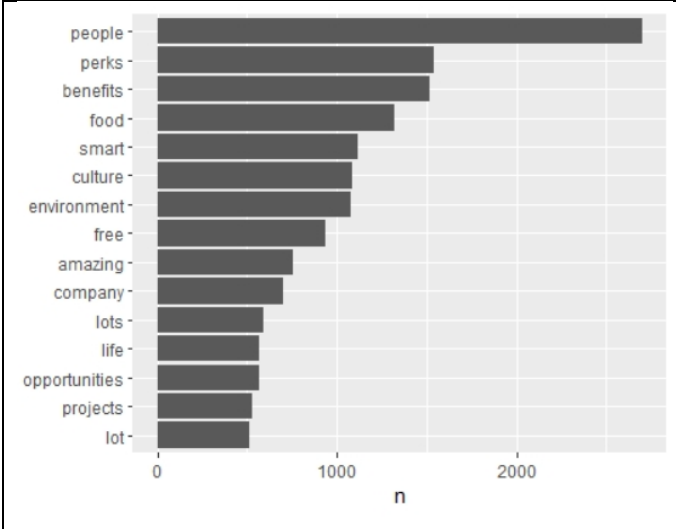
Apple



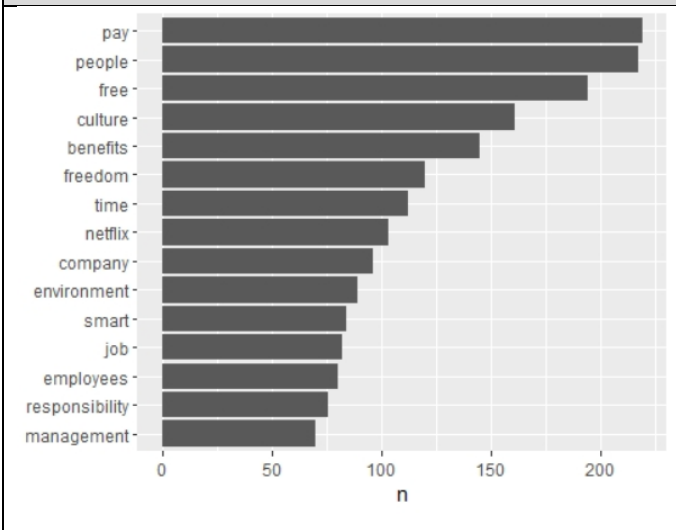
Facebook



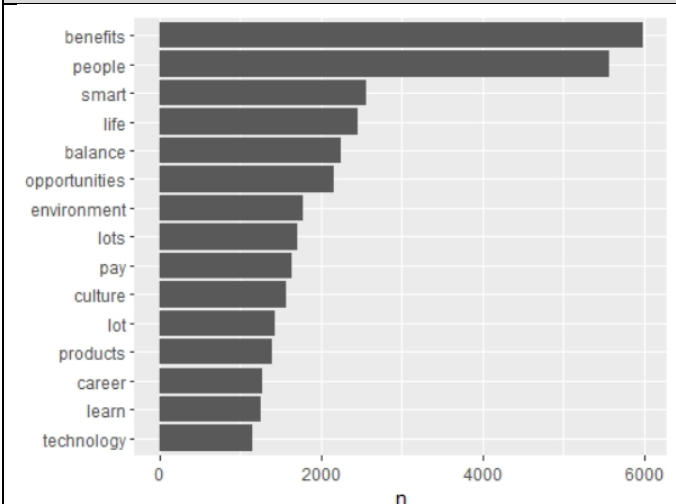
Google



Netflix



Microsoft	
-----------	--



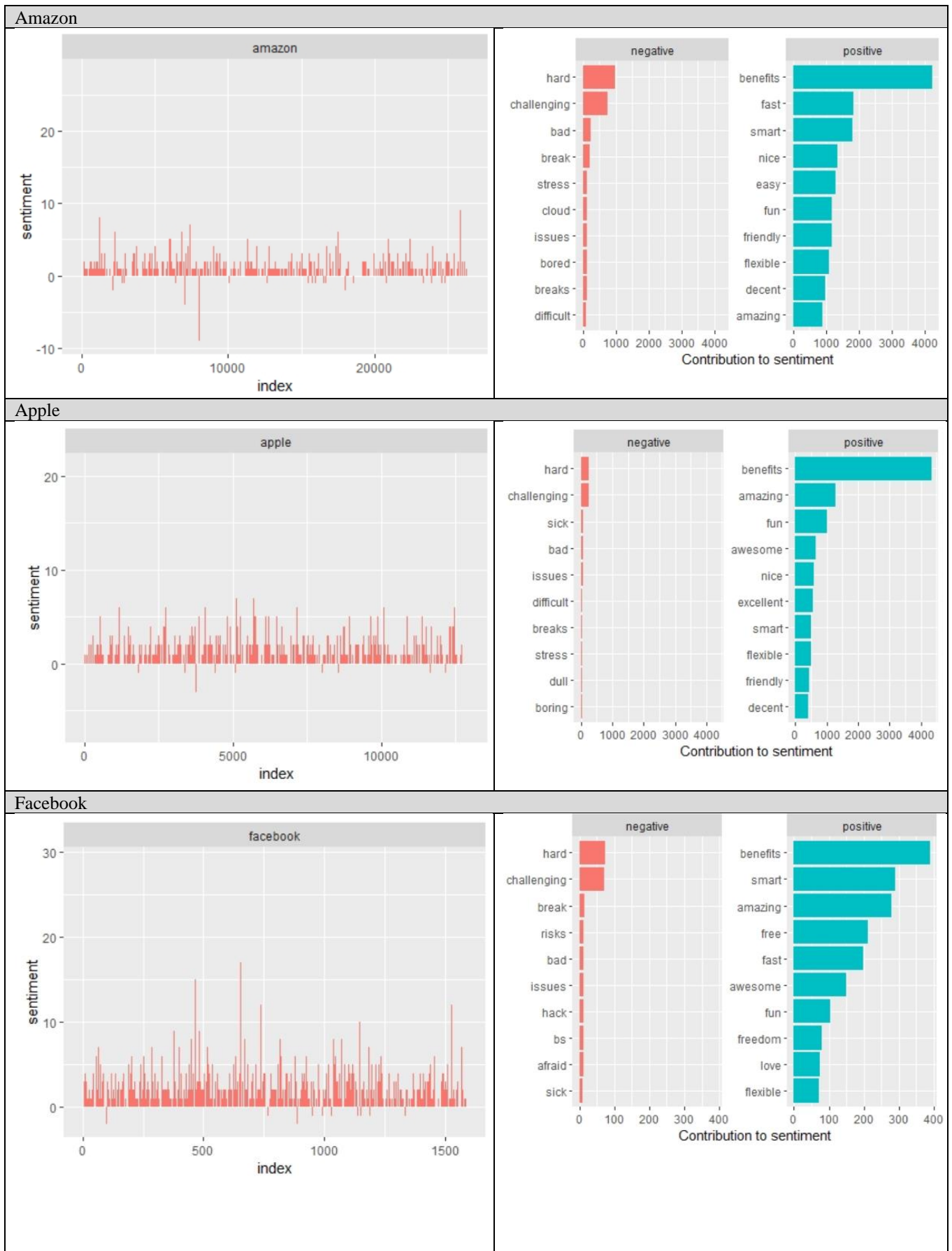
VII. Sentiment Analysis: Top 10 most commonly used words per company in the pros review section

Word	Position	Occurrence	Average Position	
People	1, 2, 1, 1, 2, 2	6/6	1.5	<3
Benefits	2, 1, 3, 3, 5, 1	6/6	2.5	<3
Environment	4, 3, 9, 7, 8, 7	6/6	6.3	<7
culture	9, 2, 6, 4, 10	5/6	6.2	<7
Pay	3, 4, 1, 9	4/6	4.25	<5
smart	6, 5, 9, 3	4/6	5.75	<7
Time	5, 8, 7	3/6	6.6	<7
free	10, 8, 3	3/6	7	
amazing	6, 7, 9	3/6	7.3	
perks	4, 2	2/6	3	<5
food	5, 4	2/6	4.5	<5
opportunities	7, 6	2/6	6.5	<7
products	5, 9	2/6	7	
job	10, 7	2/6	8.5	
lots	10, 8	2/6	9	
employees	10, 10	2/6	10	
life	4	1/6	4	
balance	5	1/6	5	
freedom	6	1/6	6	
Lot	6	1/6	6	
impact	8	1/6	8	
learn	8	1/6	8	
discounts	9	1/6	9	

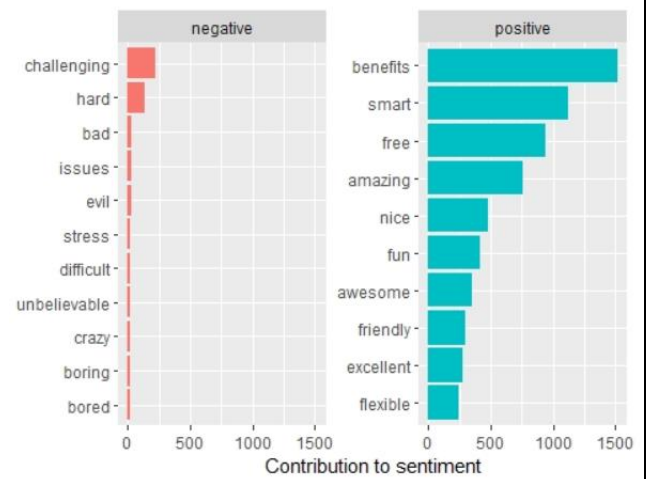
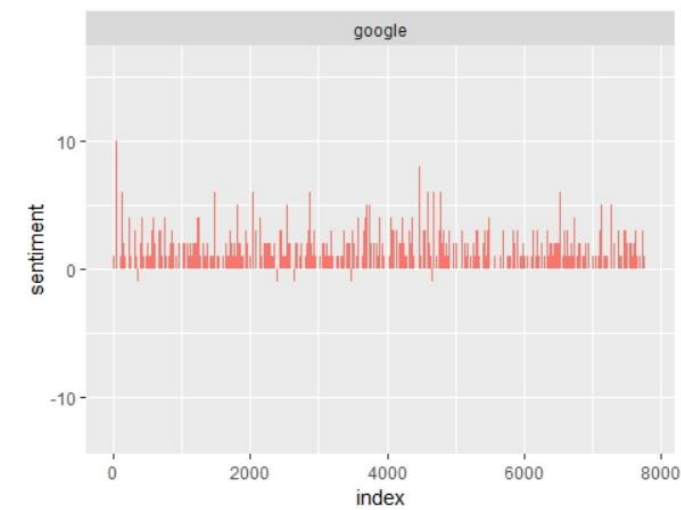
VIII. Sentiment Analysis: Sentiment scores per company in the pros review section: Mean Length of Utterance (MLU), max pos, max neg, min pos, min neg and average sentiment score

	Amazon	Apple	Facebook	Google	Netflix	Microsoft
MLU	21.15441	21.47336	39.1283	19.06906	31.50494	20.88957
Max pos	47	29	41	44	15	22
Max neg	23	13	12	34	9	10
Min pos	0	0	0	0	0	0
Mins neg	0	0	0	0	0	0
Average	1.574	2.003	2.778	2.041	2.478	1.939

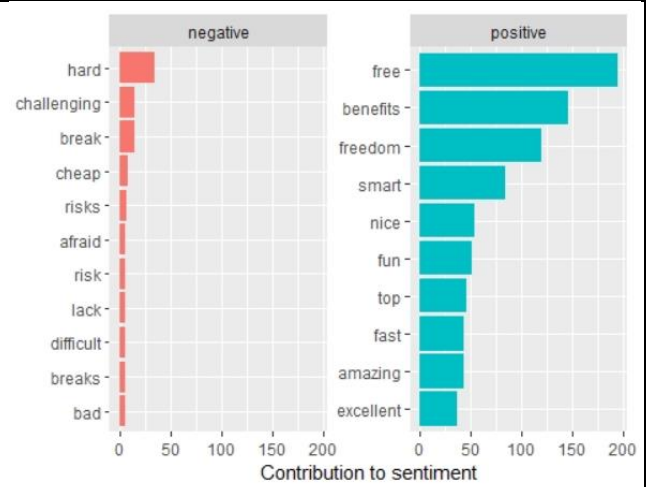
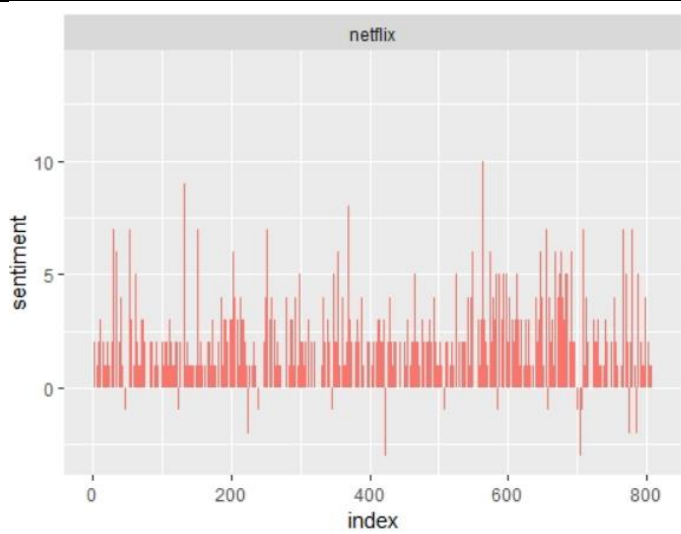
IX. Sentiment scores per company in the pros review section: top 10 most used positive and negative words (bar charts)



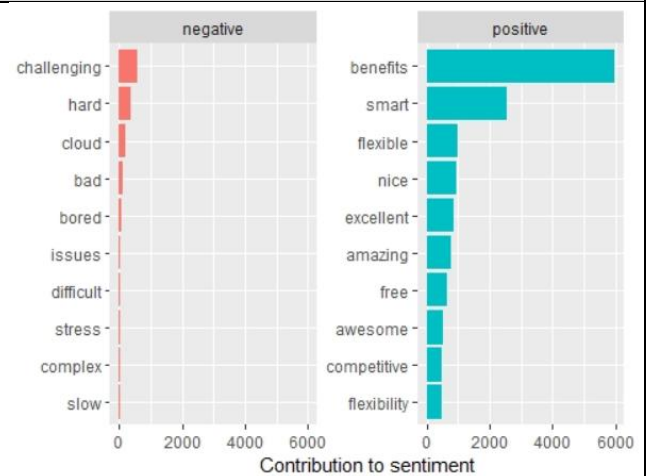
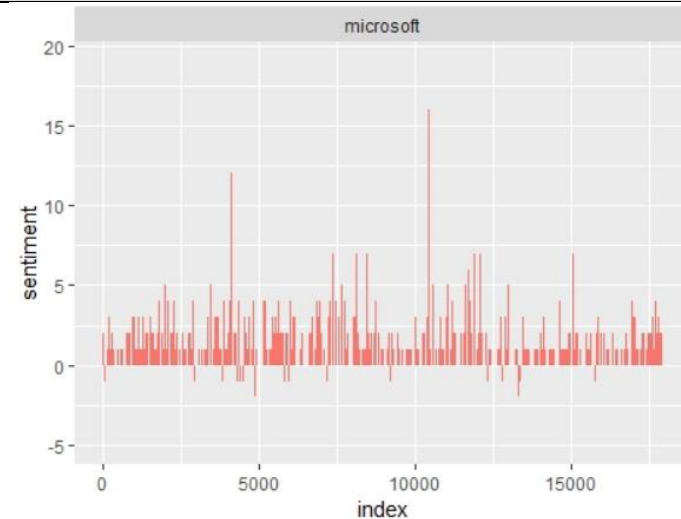
Google



Netflix

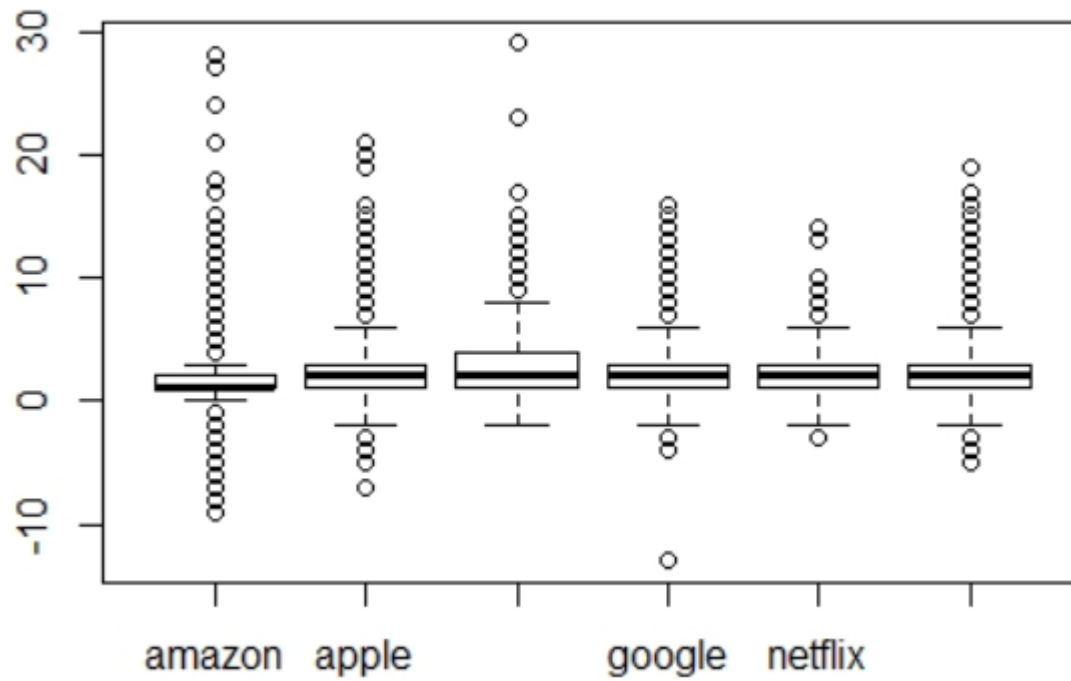


Microsoft



X. Statistically Significant Difference for Degree of Positivity and Negativity: Kruskal-Wallis rank sum test

Kruskal-Wallis chi-squared = 1326.9, df = 5, p-value < 2.2e-16



XI. Matching Advertised Company Values and Review Values: Advertised Company Values categorised according to II

Amazon	Apple	Facebook	Google	Netflix	Microsoft	Advertised Corporate Value
		open, openness	openness, open	open, openness		Communication
		interacting		communication		
innovators, innovation, innovating, inventing, invent	innovators, innovation, innovating, inventing, invent	improvement	innovators, innovation, innovating, inventing, invent	innovators, innovation, innovating, inventing, invent	innovators, innovation, innovating, inventing, invent	Innovation
passion, passionate	creativity, creative, create	creativity, creative, create	creativity, creative, create	passion	growth	
excellence	excellence	bold, boldness	excellence	excellence	AI	
think, thinking		fast, speed	smart, smartness	effective	think, thinking	
smart, smartness		problem-solving		curiosity	technology	
future		improvement			future	
results						
risks, risk						
	accessibility	community, communities, social	people, human	feelings	people	Community
	environment	connect	environment		social	
different, diversity, difference	different, diversity, difference, including, inclusion	different, diversity, difference, including, inclusion		Inclusion, including	different, diversity, difference, including, inclusion	Respect
learn, learning	education			respect	respect	
				selflessness		
				independent, independence		Reward
	combativeness		fun			
commit, commitment				integrity	integrity	Integrity
Sustainability, sustainable				courage, courageous	sustainability, sustainable	
trust	responsibility				responsibility	
					ethical	
					trust, trustworthy	
					accountable, accountability	
					honest, honesty	
customer		impact		impact	quality	Quality
bold, boldness		bold, boldness			customer	
				collaboration, collaborating		Teamwork
	privacy		flexibility			Safety
simple, simplify	secrecy		data	thoughts		Other
16	11	13	10	17	19	

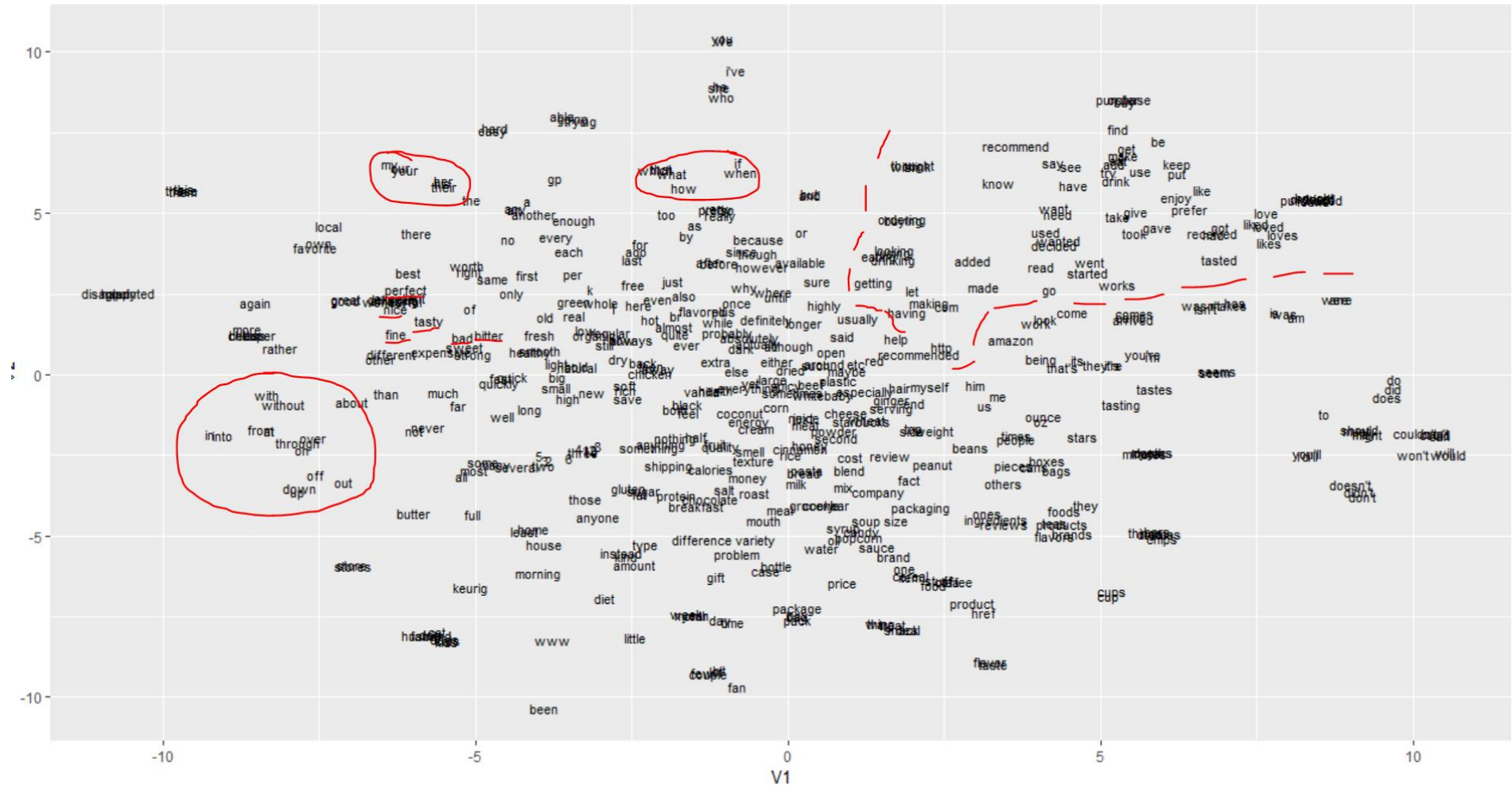
XII. Matching Advertised Company Values and Review Values: All tokens

Tokens from Reviews that Match Official Company Culture Values							
	n	amazon	apple	facebook	google	netflix	microsoft
8	1	bold	accessibility	bold	create	collaboration	people
12	2	commit	create	communities	creative	communication	accountability
26	3	commitment	creative	community	creativity	curiosity	accountable
27	4	customer	creativity	connect	data	effective	customer
99	5	difference	difference	create	environment	excellence	difference
258	6	diversity	diversity	creative	excellence	impact	diversity
291	7	excellence	education	creativity	flexibility	including	ethical
312	8	future	environment	difference	fun	inclusion	future
339	9	innovating	excellence	diversity	human	independence	growth
366	10	innovation	including	fast	innovating	innovating	honest
385	11	innovators	inclusion	impact	innovation	innovation	honesty
417	12	invent	innovating	improvement	innovators	inventing	including
561	13	inventing	innovation	including	invent	passion	inclusion
660	14	learn	innovators	inclusion	openness	respect	innovating
690	15	learning	invent	interacting	people		innovation
900	16	passion	privacy	openness	smart		innovators
1068	17	passionate	responsibility	social			integrity
1081	18	results	secrecy	speed			invent
1093	19	risk					inventing
1147	20	risks					quality
1525	21	simple					respect
1592	22	simplify					responsibility
1854	23	smart					social
2139	24	smartness					sustainability
3528	25	sustainability					sustainable
3820	26	sustainable					technology
11939	27	thinking					thinking
12224	28	trust					trust
1	29						trustworthy
		16/16	11/11	13/13	10/10	11/17	19/19
		100%	100%	100%	100%	65%	100%

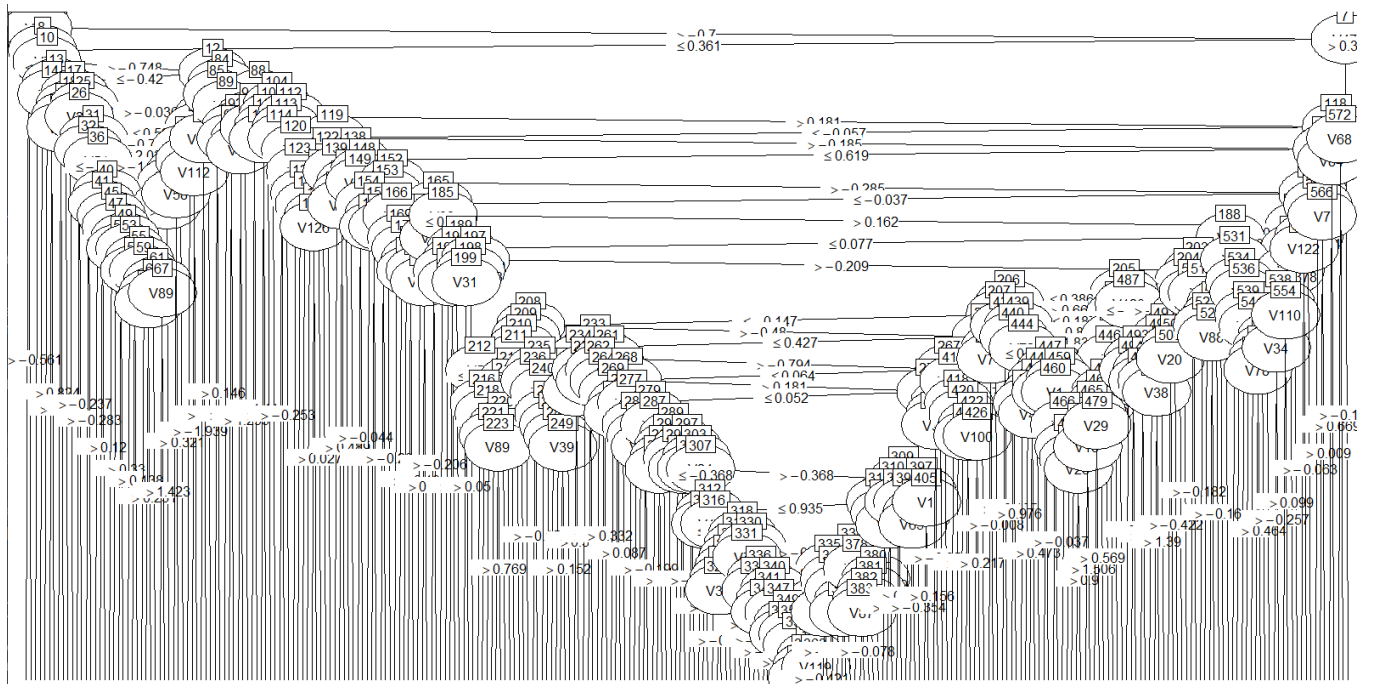
XIII. Matching Advertised Company Values and Review Values: 100 most often used tokens

100 Most Frequently Used Tokens from Reviews that Match Official Company Culture Values							
	n	amazon	apple	facebook	google	netflix	microsoft
8	1	customer	creative	impact	environment	communication	growth
12	2	innovation	environment	fast	flexibility		people
26	3	learn			fun		technology
27	4	learning			innovation		
99	5	smart			people		
1	6				smart		
		4/16	2/11	2/13	6/10	1/17	3/19
		25%	18%	15%	60%	6%	16%

XIV. Word Embedding in two dimensions

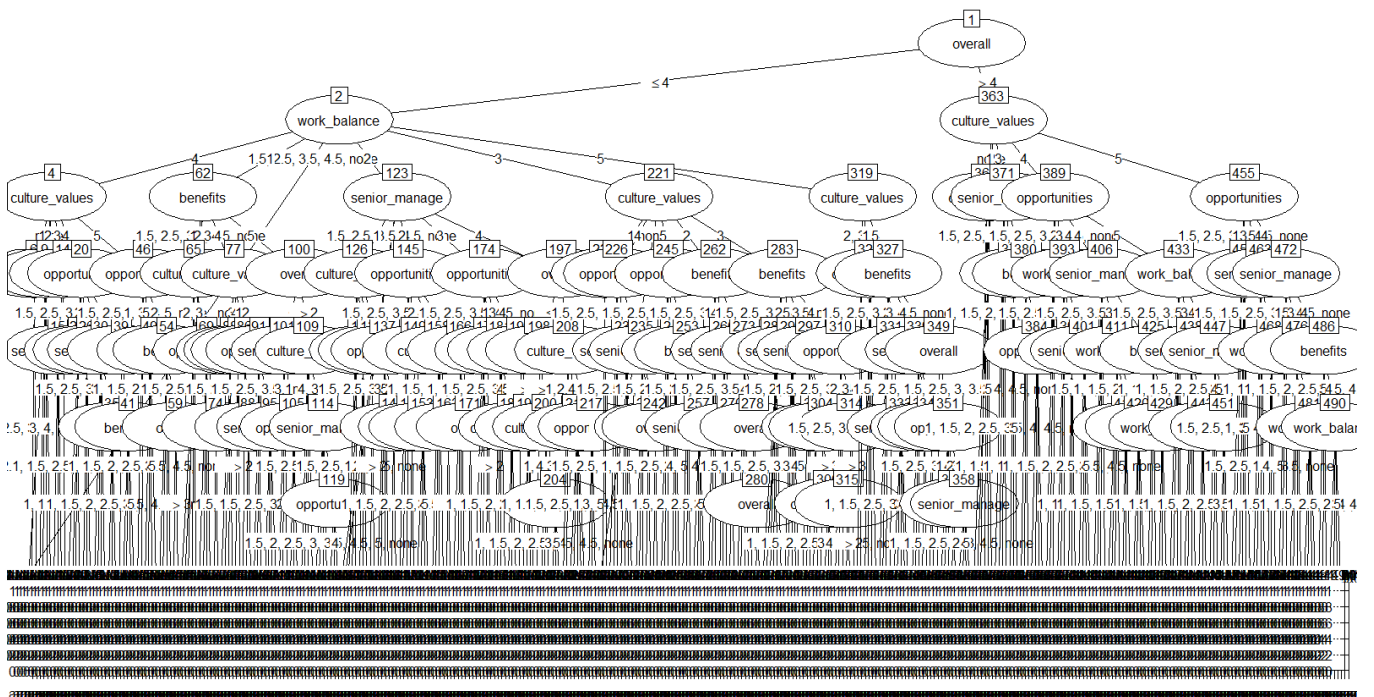


XV. Classification: Example Tree for Linguistic/Embedding Data



Tree size: 288
result
F1: 0.2424679

XVI. Classification: Example Tree for Numeric Data



Tree size: 288
result
F1: 0.3586744

XVII. Please find all other classification results in the R files attached

XVIII. Company Clusters based on star ratings

