ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
MSc Data Science and Marketing Analytics

The effect of industry's idiosyncrasies in crowdsourced employer branding: an LDA and regression approach

Author: Adriana Cañas (597943)

Supervisor: Prof. dr. Bas ACD Donkers
Second Assessor: dr. Anastasija Tetereva

Date final version: August 12th, 2022

ABSTRACT

The Great Resignation has become the latest buzzword in the employment market. With employees looking for better opportunities that increase their workplace wellbeing, organizations are in fierce competition to not only attract but also retain talented individuals. In this battle, employer reputation and attractiveness – employer *branding* – play a huge role. Employer branding is the package of diverse benefits that employees identify with a firm's employment. With the extended use of the internet, employer reputation has become easily accessible through crowdsourced platforms like Glassdoor. Using Latent Dirichlet Allocation, this paper unveils the benefits/practices that matter the most to candidates from scrapped online reviews, which include (1) salary, (2) management, (3) work/life balance, (4) growth opportunities, (5) other non-monetary benefits, (6) colleagues, (7) working hours, (8) suitability to start a career, (9) work environment, (10) if firms care about their employees and (11) general feeling. Differences across industries are also clarified, which is demonstrated using regression techniques.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Contents

Introduction

With resignation rates peaking by the end of 2020, companies all around the world are experiencing the biggest challenges in their employment practices, especially when retaining and recruiting employees. As workers quit at unprecedented rates and current employees demand more and better benefits, organizations are competing to not only attract but also retain talent. Now, the power is in the hands of the workforce, as a rush of job openings is saturating the labour market, giving prospective employees myriads of choices and possibilities. Given the long-lasting job dissatisfaction situation, particularly aggravated by companies not being able to match their wages to the current inflation levels, workers are more likely to leave their current jobs and search for more satisfactory roles at companies with better offerings. In these times, it is crucial that organizations understand what employees are looking for in a job and in a company, and adapt their employment practices to embrace and accept these changes.

Yet, how can organisations unveil what – potential – employees are seeking in such a changing period of time? In an era in which individuals interact more than ever online, it is crucial for firms to understand and manage their reputation and attractiveness as employers not only offline, but also on the web. Writing reviews on online platforms has become the new word-of-mouth: user-generated content now affects hotel online bookings (Ye et al., 2011), product sales (Zhu & Zhang, 2010) and employer branding (Dabirian et al., 2017).

Employer branding narratives – or how *attractive* an employer is for current and potential employees –, were naturally developed behind closed doors with conversations shared between colleagues. In some cases, employees could speak up to their managers about issues that were causing job dissatisfaction but the fear of potential negative consequences withheld workers from making this move. In today's world, several platforms have emerged in which employees can share their experiences anonymously without any negative repercussions affecting their careers. Sites like Glassdoor, offer companies direct and easy access to these newly developed narratives. Nevertheless, most companies are not yet exploiting the data and insights behind these reviews.

Natural language processing (NLP) techniques can help companies to analyze extensive amounts of textual data from crowdsourced employer branding platforms like

Glassdoor. For organizations, ignoring their crowdsourced employer reputation can be detrimental, as job seekers might rely on this information to separate good employers from the bad. Not only firms could become unattractive to potential new joiners, but also they could lose their best recruits to better firms.

Moreover, it is not yet known whether employees value the same employment offerings across different industries. One could think that perks like free food are highly valued in the Technology industry, compared to the Investment Banking industry which might be more salary-driven. These industries' differences are also worth exploring, as companies should adapt their practices to their specific area of expertise to maximize the success rate of their human resources department's efforts.

Because employer branding is changing, it is important to understand previous narratives around this topic, which is explored in the Literature Review section of this paper. After that, an explanation of the Methodology used to unveil the dimensions that employees use when talking about their employers is given. The Data section goes through the data collected for this study, which has been scrapped from Glassdoor. The motivation of the industries selected for the data collection as well as an exploratory data analysis of the novel dataset is also described in this section. Then, both topic modelling and regression results are both presented and discussed. Finally, the whole research is summed up in the Conclusions section together with a discussion of the limitations of the study and a proposal for further lines of research.

Literature Review

Human capital as a source of competitive advantage

Back in 1991, Barney (1991) discussed the idea that firms are said to have a sustained competitive advantage when they are implementing a value-creating strategy that is not being executed by other competitors and that is difficult to replicate by other current or potential firms. This sustained competitive advantage can be potentially achieved if firms possess a set of attributes or characteristics that are valuable, rare, imperfectly imitable and non-substitutable (Barney, 1991). Contrary to prior theories in the field of strategic management, the resource-based view theory popularised by Barney (1991), shifted the emphasis away from the external resources (i.e., opportunities and threats) toward a more internal-focused perspective based on internal strengths and weaknesses (Hoskisson et al., 1999). This expanding embrace of internal attributes as sources of competitive advantage supported the claim that human capital is a strategic source of value creation for a firm (Wright et al., 2001).

Employer branding and employer attractiveness

In this context, many authors have argued how human capital and human resource management can be viewed as a resource of sustained competitive advantage (Boxall, 1996; Lado & Wilson, 1994; Wright & McMahan, 1992). With this goal in mind, firms strive to appear as attractive employers in order to recruit talented job seekers, who initially have to decide which jobs should be taken into consideration (Sivertzen et al., 2013; Cable & Turban, 2003). To make this decision, several factors are taken into account. Cable and Turban (2003) state that *organizational reputation* is one of the main determinants of a firm's ability to recruit new talent. Similar to product branding, organizational reputation can become a source of competitive advantage, making the candidate decide on one employer or another depending on the firm's name and reputation, or what Akhjhmbler and Barrow (1996) describe as *employer brand.* This concept is defined as "*the package of functional, economic and psychological benefits provided by employment, and identified with the employing company*" (Ambler & Barrow, 1996, p. 197). Many definitions of this term have been given in the last decades. Lloyd (2002) describes employer branding as the "sum of a company's efforts to communicate to existing and prospective staff that it is a desirable place to work".

In one of the most popular researches, Backhaus and Tikoo (2004) summarize several proposed definitions into one unique proposition: employer branding involves advocating a clear view of what differentiates a firm and makes it desirable as an employer; both within and outside the firm. This places importance on both *attracting* external talent and *retaining* internal employees. *Employer attractiveness*, defined as "the envisioned benefits that a potential employee sees in working for a specific organization" (Berthon et al., 2005), becomes particularly important in this discussion. Understanding the contributors of what makes an employer attractive to outsiders is crucial for the success of the recruiting strategy of a firm.

In light of the above points, it can be argued that the employer's image (i.e. how the company is perceived) directly affects the attractiveness of the firm to potential job candidates. Having a unique value proposition as an employer can help the firm to differentiate its employment offering from those that other companies offer (Edwards, 2010). This value proposition and brand image arise from a firm's employer branding, which is – to some extent – controlled by the employer. However, potential candidates also develop an image of the employer-based on information sources that are not under the control of the hiring firm. With the spread of technology and accessibility to an internet connection, new communication channels have emerged in all departments of a business. In human resources management, attraction, recruitment and selection are now being managed mostly digitally (Chhabra & Sharma, 2012). Job seekers are one click away from not only finding new job postings but also researching and comparing organisations that are appealing to them. While some of these channels might be employer-controlled (i.e.: the firm's social networks, website or job boards), some others are crowdsourced and out of the firm's control (i.e.: review platforms).

Employer branding under crowdsourced platforms

Crowdsourced platforms pose a challenge for employer branding. Former and current employees now have the opportunity of sharing their job experiences allowing other job candidates to gather more information about their potential employers (Dabirian et al., 2017). This new form of employer branding falls outside the control of the firm and its human resources team. Regulating what a firm's employees share online would obliterate the transparency that online reviews offer against the beautified version that firms themselves communicate about their workplaces. Thus, companies need to (1) understand this new source of crowdsourced employer branding, (2) comprehend the factors that contribute to employer attractiveness and (3) implement strategic initiatives aimed at implementing these factors into the employer branding. As Dabirian et al., 2017 puts it: "*Great work environments do not emerge by happenstance, but rather result from deliberate and strategic initiatives aimed at attracting, engaging, and retaining employees*" (p. 198). This paper seeks to contribute to the existing literature by identifying the dimensions of employer attractiveness.

Dimensions of employer attractiveness

Previous work has been done on the matter of unveiling what dimensions drive employer attractiveness. Ambler and Barrow (1996) proposed three dimensions: functional, psychological and economic benefits. Berthon et al. (2005) refined and extended these dimensions by accounting for five factors obtained through Principal Components Analysis: (1) interest value which measures *"the extent to which an individual is attracted to an employer that provides an exciting work environment, novel work practices and that makes use of its employee's creativity to produce high-quality, innovative products and services"* (p. 159); (2) social value which captures *"the extent to which an individual is attracted to an employer that provides a working environment that is fun, happy, provides good collegial relationships and a team atmosphere"* (p. 159); (3), economic value *"assesses the extent to which an individual is attracted to an employer that provides above-average salary, compensation package, job security and promotional opportunities"* (p.159 and 162); (4) development value, measures the attractiveness of an employer based on if it provides *"recognition, self-worth and confidence, coupled with a career-enhancing experience and a springboard to future employment"* (p. 162); and lastly the (5) application value *"assesses the extent to which an individual is attracted to an employer that provides an opportunity for the employee to apply what they have learned and to teach others, in an environment that is both customer orientated and humanitarian"* (p. 162).

While the work of Berthon et al. (2005) is promising, their work was based on questions led by a moderator on subjects related to "ideal" employers as well as the factors that the subjects regarded as important when considering potential employers. Yet, the amount of data that is available to firms today could not be exploited back then.

Crowdsourced employer branding platforms like Glassdoor unleash a great potential to obtain a more real and detailed look at the drivers of employer attractiveness. Glassdoor.com is a platform that provides millions of ratings, reviews, salaries, and job interview insights in an attempt to increase workplace transparency (Glassdoor, 2022). Founded back in 2008, Glassdoor relies on anonymous user-generated content to form a more truthful version of a firm's employment practices. Users can give their opinions via two main mechanisms: first, a 5-point Likert scale gives participants the chance to evaluate their employers on 6 different dimensions: work/life balance, culture & values, diversity & inclusion, career opportunities, compensation & benefits and senior management. Then, participants have a dedicated space to share both "pros" and "cons" of their employment experience. Although the dimensions used by Glassdoor are easy to understand for users (who can also get tired of giving a review if a more extensive list was provided), it may be ignoring the complexity of employer attractiveness and the numerous factors that might be affecting an employee's overall rating. We see that this is the case in some reviews, in which the overall rating given to a firm is higher than any of the individual dimensions' scores or than the average itself. This shows that there might be additional attributes that employees care about when giving an evaluation of their employer and that are not being considered in the platform, while also illustrating that some dimensions might have a higher weight in determining the overall score.

Dabirian et al. (2017) go a step further in order to determine the most important dimensions of employer attractiveness. They use 38,000 reviews from Glassdoor to investigate what employees care about when giving an evaluation of their employer, focusing on Glassdoor's 10 best and 10 worst places to work in 2016. The authors confirm the dimensions proposed by both Ambler and Barrow (1996) and the propositions by Berthon et al. (2005) and unveil two new dimensions which they name "management value" and "work/life balance". While Dabirian et al. (2017) extend the existing propositions with the use of Natural Language Processing (NLP), their research uses companies across all industries, without accounting for the potential differences across industries in terms of company values and employer branding.

The complexity of employer attractiveness dimensions raises a number of important questions that are relevant not only for employer branding but for overall human resources management: (1) What dimensions better capture the attributes that employees find attractive about their employers?; (2) What dimensions matter the most for the rating of a company?; and (3) How do these dimensions differ across industries?”

Methodology

The following section examines the methods and analyses applied to the extracted reviews to obtain further insights and results. The research aim is to understand the employees' assessment and evaluation of their employer's practices and to identify differences across four industries: tech, investment banking, consumer goods (FMCG) and accounting. To understand the evaluation of employees, topic modelling is employed as the means of unveiling relevant abstract topics that occur in the collection of reviews. Moreover, to quantify the impact of the topics in each of the four industries, regression analysis will follow.

Latent Dirichlet Allocation

Topic models are statistical Bayesian models for discovering the underlying semantic structure of a collection of documents to unveil hidden topical patterns in a text body. Many methods have been developed in the past decades, including Latent Dirichlet Allocation (LDA), which has grown to be one of the most used topic modelling methods (Zhao et al., 2015). Presented by Blei et. al (2003), LDA is an unsupervised generative probabilistic model of a corpus. It builds on the idea that documents can be represented as a probabilistic distribution over latent topics, with each topic being characterized by a distribution over words. Given a corpus consisting of a collection of D documents, these being a sequence of N words, LDA process can be represented as:

Figure 1: Plate representation of LDA (Blei & Lafferty, 2009).

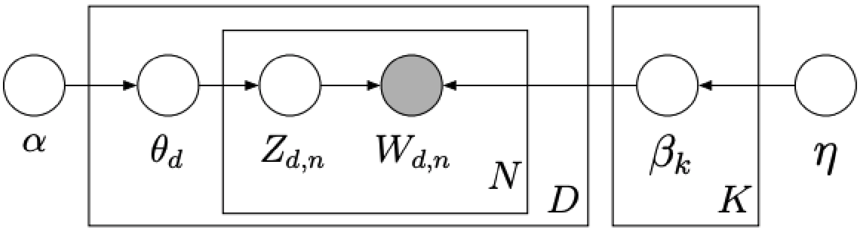
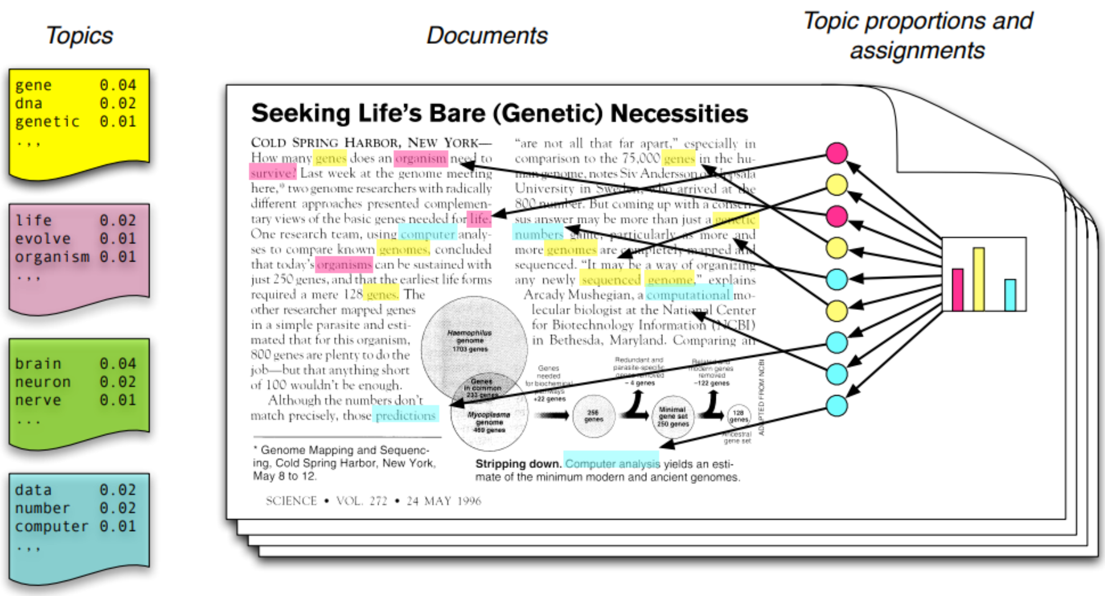


Figure 1 illustrates the dependencies among the model variables and parameters: *α* represents the Dirichlet prior parameter of per document-topic distribution (i.e., the parameter governing the prior distribution of θ_d); while β_k is the Dirichlet prior parameter of per topic-word distribution (i.e., the term distribution for each topic *K,* where *K* is the number of

predefined topics). Similarly, θ_d represents the per-document d topic proportions; $Z_{d,n}$ is the per-word topic assignment which is drawn from a distribution with parameter θ_d; and $W_{d,n}$ is the n^{th} word in the d document and it is the only observed variable in the model. The intuition behind the LDA model and its plate representation can be better understood using the following illustration:

Figure 2: Intuitive representation of LDA (Blei, 2012).



On the far left part of **Figure 2**, the author assumes a number of topics K, which are distributions over terms (β_k). For each document, the generative process of LDA is as follows: a distribution over the topics is chosen (histogram on the far right), θ_d; a topic assignment is picked (coloured coins), $Z_{d,n}$; and lastly, a term from the corresponding topic is selected from the corpus, $W_{d,n}$ (Blei, 2012).

Determining the number of topics

So far, how to determine the number of topics, K, has been disregarded in this paper. Because topic modelling is an unsupervised learning method, the set of possible topics is unknown prior to running the model. While this task is, most of the time, an educated guess or a trial-and-error evaluation, some authors have used coherence score as a resource to determine the optimal number of K (Islam, 2019). Topic coherence measures help discern what topics are semantically interpretable (i.e., human-understandable) from those that are just "artifacts of statistical inference" (Stevens et al., 2012, p. 954). While many coherence measures exist in the literature (Röder et al., 2014; Rosner et al., 2015), this paper relies on the algorithm developed by Tommy Jones in the R package *textmineR* (Jones, 2021). The author recommends fitting several topic models across a sequence of topics, calculating the probabilistic coherence for each topic in each model and averaging the probabilistic coherence across all models for each topic (Jones, 2021). Following this idea, a list of models (each with its own different number of *K* topics) is evaluated to find the final number of *K*. The optimal number of topics to be chosen can be selected by evaluating the coherence measure for each topic and specifying *K* with the highest – average – topic coherence (Islam, 2019).

Additionally, perplexity can also be used to evaluate language models. It is a statistical measure commonly used to capture how well a probability model predicts a sample. Lower perplexity values normally indicate better generalization performance (Blei et al., 2003). For a test set of *M* documents, perplexity can be defined as:

$$perplexity(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

However, it should always be kept in mind that a higher number of topics will offer a high held-out sample fi t(generally, better perplexity and coherence) but it will make the model highly complex and unintelligible. It is then required for the researcher to balance model fit and interpretability to obtain the best results. For that reason, this research forces the number of topics to exist between four and sixteen. A topic model with less than four topics will result in very generalized topics, with little nuance across them. On the other hand, a topic with more than sixteen topics will become too complex as well as time-consuming not only to interpret the results but previously to run the models in the selected software. Thus, in this research, seven LDA models have been built with 4, 6, 8, 10, 12, 14 and 16 topics respectively and the final number of *K* has been selected after averaging the topic coherence across all topics for each of the seven models.

Topic modelling in short-text

Unveiling latent topics in long documents like books, articles or websites has proven to be successful (Albalawi et al., 2020). However, with the increase in popularity of sites like Twitter or instant messaging platforms, topic models might not perform as expected. Because of the short extension of these documents (i.e., tweets, Q&As, Glassdoor reviews, etc.), inferring topics can become a challenge, as short text is often noisy and sparse. Albawali et al. (2020) applied several topic modelling methods to two (short-text) textual datasets and compared their performance across different metrics like topic coherence, precision, recall and *F*-score. Their findings suggest that both LDA and Non-Negative Matrix Factorization (NMF) deliver meaningful topics and good overall results (Albawali et al., 2020).

Yet, it is still likely that these methods will not perform as desired on short-text. To minimize the effect of short-text in the topics, some authors have suggested the use of biterms as an addition to the unigrams used in a conventional LDA model (Yan et al., 2013; Park et al., 2015). Park et al. (2015) found that using a mixture of bigrams and unigrams yields better accuracy than traditional LDA with only unigrams and that an LDA model with only bigrams. However, Yan et al., 2013 found that the LDA model with bigrams performed better than both the mixture model and the traditional model with word co-occurrences. In both cases, the traditional LDA model with unigrams is the worst-performing model for short-text data. Given these conflictive findings, this research tested both an LDA model with only bigrams and an LDA model with both bigrams and unigrams. For all the set number of topics, the LDA model with a mixture of terms and bigrams yielded a higher coherence score than the model that only included bigrams. However, the topics were harder to interpret as the most probable words were a mix of what it felt two different topics. Thus, the next sections will solely present and discuss the findings of the biterm LDA model.

Regression analysis

Multiple Linear Regression

Compared to other more advanced machine learning techniques, multiple linear regression is simpler and more interpretable. Determining how the topics from the LDA model differ across different industries, might benefit from not applying a black box model in which comparisons become harder to make. Thus, the goal of this analysis is to build a linear regression with *review rating* as the dependent variable and use the topics from the LDA model as regressors. The model is built on 80% of the total data, which forms the training set.

The equation for this regression model takes the form

$$Rating = \beta_0 + \beta_i X_i + \beta_j D_j + \beta_{ij} X_i D_j + \epsilon$$

where X_i represents the *i*th topic, D_j represents the *j*th industry and X_iD_j represents the interaction of the *i*th topic with the *j*th industry.

Because probabilities across all topics for each review add up to 1, a wise choice needs to be made with regards to what topic is left out as a reference topic. Failing to do so, will cause the model to not be adequately defined, as there will be a strong correlation between the independent variables. The choice of what topic to use as a baseline is arbitrary, meaning that interpretation of the rest of the topic coefficients will depend on the baseline topic that it's chosen. Following this thought, it is useful to select an extreme topic (i.e., whose interpretation is clearly positive or clearly negative) or a neutral topic (i.e., whose interpretation is neither positive nor negative). If a very positive topic is chosen, it can be expected that all the rest of the coefficients will be negative compared to the left-out topic. Although selecting a neutral topic would make the interpretation of the rest of the coefficients more intuitive, it is not always the case that such a clear neutral topic is available in the model. Choosing a neutral topic is a somewhat subjective task; thus, the interpretation of the results will greatly vary depending on the left-out topic. In this case, a neutral topic has been selected as a reference, to sort of display both positive and negative effects on ratings.

Furthermore, checks have to be made when building a linear model as to whether the assumptions of the model hold or not. These assumptions are: 1) there is a linear relationship between the dependent variable and the regressors; 2) observations are independent of each other; 3) the variance of the error terms is constant (homoscedasticity), and 4) the errors follow a normal distribution. Because of the discreteness of this research's dependent variable (*review rating*, values = 1-5), the linearity, homoscedasticity and normality assumptions are violated by nature. While a violation of the assumptions might lead to inefficiencies in the insights extracted from the regression model, this paper tries to overcome these violations by introducing more advanced regression techniques which will be discussed next.

Beyond linearity: Random Forest

The multiple linear regression model specified above assumed that all of the linear regression assumptions are met. In the perfect scenario, linear regression is expected to produce the best results when the data presents a linear relationship between the dependent and independent variables. Yet, this is a rare find and not the case with the topic probabilities data.

To better capture non-linear features, other algorithms have emerged to improve the accuracy of both prediction and classification tasks. Random Forest is one of the most widely used algorithms in past literature, not only for the Data Science community but also beyond this discipline. It is an ensemble supervised machine learning method that combines the prediction outcomes from numerous decision trees, where each tree is built from the values of an independent set of random vectors (Tan et al., 2016). This algorithm can be summarized in the following steps:

1. From the data set, B random samples are drawn with replacement.
2. A random subset of features for each of the B bootstrapped samples is selected.
3. A regression tree is built on each bootstrapped sample B.
4. The final outcome is obtained by computing the average from all the individual predictions generated by the regression trees.

In this paper, random forest is applied to the review and topic probabilities data to capture the non-linearity presented in the case.

Variable importance

Visualizing variable importance plots is a widely used tool to interpret black-box models. In a random forest, the importance given to each variable comes from the improvement in the split-criterion at each split in each tree, which is then accumulated for each variable across all trees (Hastie et al., 2009). Random forest also constructs a different measure for variable importance, built from the permuted out-of-sample (OOB) samples. For each bth tree, the prediction accuracy (in the case of regression, MSE) is recorded on the OOB portion of the data. Then, each predictor variable is randomly permuted in the OOB samples and the prediction error is recorded again. Because of this permitting, a decrease in accuracy can be expected. These values are averaged over the whole forest and it is used as a measure of the importance of each variable in the model. Thus, the higher the value for this important measure, the more important the variable is.

Partial dependence plots

So far, interest has only lied down in the interactions between the industries and the topic probabilities. Yet, visualizing functional relationships between two topic probabilities can be a powerful interpretational tool. Partial dependence plots are a visual way of displaying the marginal effect of one or two features on the predicted variable of the regression. To visualize these relationships, two-variable partial dependence plots are used for some of the most influential features.

Data

Data collection

Reviews for sixteen companies belonging to four different industries have been scraped[1] from the Glassdoor.com website. The selected industries include high-tech, accounting, asset management and consumer goods. The motivation for this choice lies in the fact that these industries often hold strong stereotypes when discussing employment experiences.

High-tech conglomerates such as Silicon Valley are globally known for their amazing perks and work-life policies. Until 2017, Google had been awarded the top position in Fortune's Top 100 Best Companies to Work For ranking, for six years in a row (Fortune, 2017). Among the reasons explaining this phenomenon, we find luxurious perks like gourmet food, haircuts, a gym subscription or great parental-leave policies. Other industries have also strong occupational stereotypes associated with them. A quick search on the Internet unveils that investment banking employees can be expected to work from 60 to up to 100 hours per week. In the accounting industry, the Big 4 are also known for their long hours and vertical hierarchical structure, which can be often perceived as undesirable by some individuals.

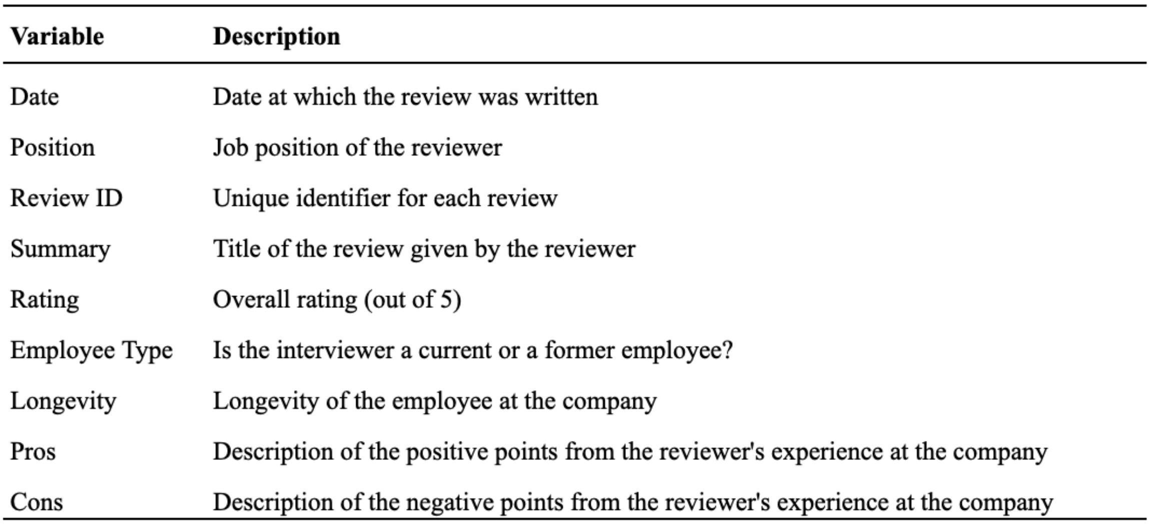
To select four companies within each industry it has been useful to look at those with the highest market capitalizations. For the high-tech industry, Microsoft, Alphabet (Google), Amazon and Meta (Facebook) have been selected. Despite having the highest market capitalisation, Apple has been left out to avoid having reviews of employees working in retail rather than in tech- and management-driven positions. JP Morgan, Goldman Sachs, Morgan Stanley and Citigroup have been selected as a representation of the Investment Banking/Financial industry. The worldwide known Big 4 in accounting (Deloitte, KPMG, PwC, EY) will be representing the accounting industry. Lastly, the firms representing the FMCG industry are Procter and Gamble, Nestlé, Coca-Cola and Pepsico. It must be noted that market capitalization was used as a *guide*. In some cases, results might differ as it can become challenging to draw the line between different industries for a given company (i.e.:

[1] Note: the data has been scrapped on May 10th 2022.

Morgan Stanley can fall under asset management, banking, investment, financial services, etc.).

A web scraper has been built to construct the dataset that will be used for the methods explained in the previous section. The final dataset consisted of 129,680 reviews. **Table 1** summarises the items that have been scrapped:

Table 1: Description of the variables in the data.



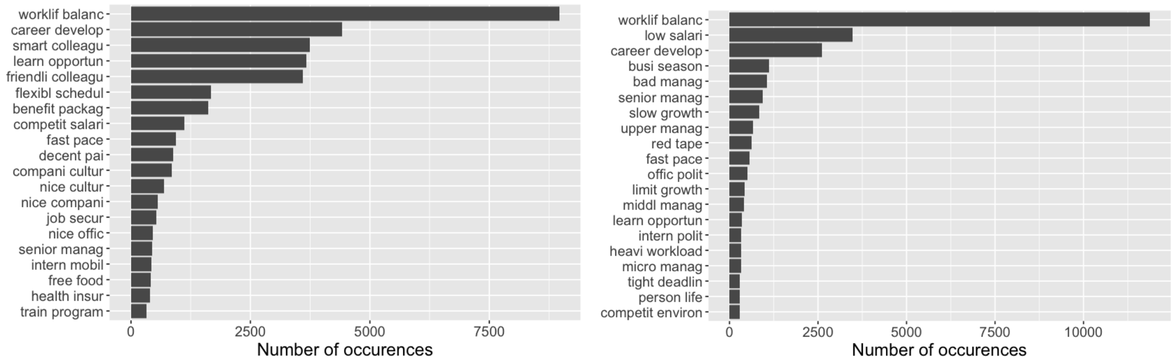
Variable	Description
Date	Date at which the review was written
Position	Job position of the reviewer
Review ID	Unique identifier for each review
Summary	Title of the review given by the reviewer
Rating	Overall rating (out of 5)
Employee Type	Is the interviewer a current or a former employee?
Longevity	Longevity of the employee at the company
Pros	Description of the positive points from the reviewer's experience at the company
Cons	Description of the negative points from the reviewer's experience at the company

On top of scraping data at the review level, general data on the company level has also been mined from the overview firm site on Glassdoor. The variables include company name, size of the company, type (private, public or NGO), estimated revenue, headquarters location, year in which it was founded and the industry it belongs to.

Exploratory Data Analysis

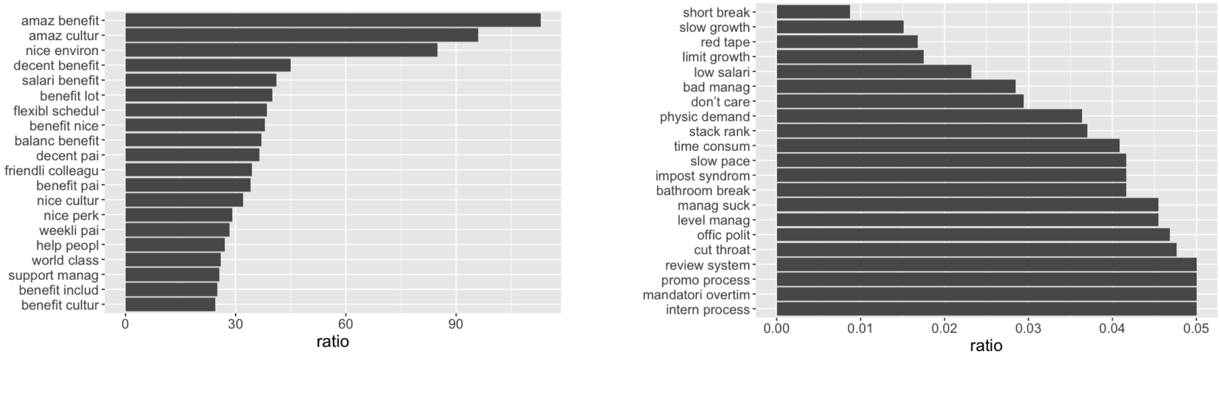
Before moving to a more advanced analysis, this section will now present some key exploratory facts and figures about the data just described above. Word frequency histograms are a useful tool to visualize and represent qualitative data. In this paper's case, it is extremely valuable to represent which bigrams occur more often in a set of reviews. For instance, taking a general perspective, **Figure 3** depicts the most frequent bigrams in the set of positive aspects (*pros*) and negative aspects (*cons*) of the reviewer's experience:

Figure 3: Bigram frequency histogram for pros (left) and cons (right).



As it can be observed, bigrams that appear frequently in positive reviews include *work-life balance, career development, smart colleagues, learning opportunities, flexible schedule, competitive salary, nice office* or *free food.* On the other hand, negative reviews have some distinct bigrams for this group (i.e.: *low salary, red tape, slow growth)* but also share similarities with the positive reviews (i.e.: *work-life balance, career development)*. Although one can extract some insights already from these graphs, it is still not very clear what reviewers talk more about in positive versus negative reviews. To overcome this issue and gain more precision in the direction of the insights, a ratio of the count of bigrams appearing in positive reviews versus the count of bigrams appearing in negative reviews is considered for further analysis. More specifically, this ratio provides some nuance into the relativity of the frequency in which bigrams appear in either positive or negative reviews. A high ratio of bigram count in positive descriptions to bigram count in negative descriptions indicates that the bigram appears more frequently relative to its appearance in negative descriptions. Moreover, as the ultimate interest is to also investigate how employees perceive their employer's attractiveness across industries, this ratio analysis of positive to the negative word count has been applied in each of the four industries selected in this paper. Figure 4 plots the relatively most frequent bigrams in positive descriptions and negative descriptions for the Tech industry:

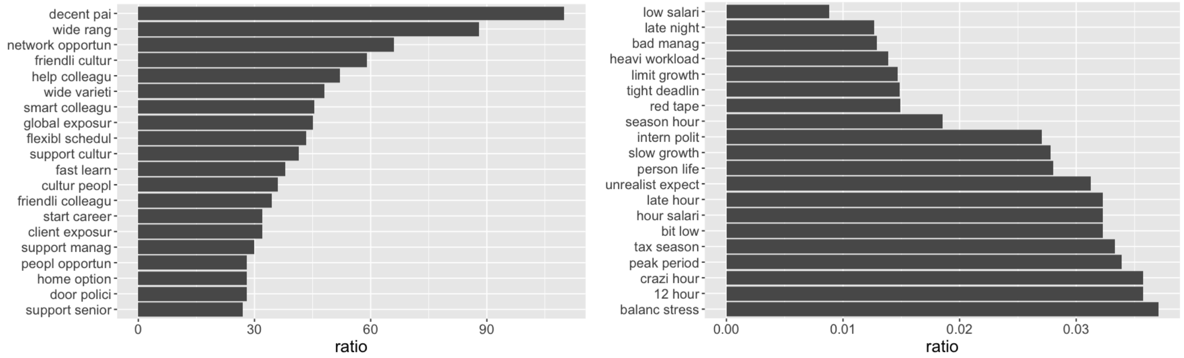
Figure 4: Relatively most frequent bigrams in the Tech industry's pros (left) and cons (right).



From the above plots, it can be argued that employees working for a tech company, think of its *amazing benefits* and *nice perks, flexible schedules* and *salary* when expressing the positive aspects of their working experience. On the other hand, negative comments mention *short breaks*, *slow/limited growth, red tape,* its *review system* and its *promotion process*. These first insights support this paper's idea of how the Tech industry is very well known for the perks that they offer as part of their employment contracts, and this is apparently very valued by employees of companies in this industry.

Interestingly, in the accounting industry employees often make reference to its *networking opportunities, global* and *client exposure,* working from *home options*, and *fast learning.* Employees also mention quite often how *friendly, helpful* and *smart* their colleagues are, creating a *friendly* and *supportive* business culture. On the other hand, working-hours-related bigrams appear often in negative comments: *late nights, crazy hours, 12-hour* working days and *peak periods* sustain the common stereotype of working at a Big 4. Further insights can be extracted from **Figure 5**:

Figure 5: Relatively most frequent bigrams in the Accounting industry's pros (left) and cons (right)



Relative frequencies for bigrams in the Consumer Goods and Investment Banking industries can be found in **Appendix A** and **Appendix B**, respectively. Looking at all four industries, there are some groups of words that could potentially fall under one big group. For instance, there seems to be a recurrent theme about salary, with bigrams both in pros and cons (*decent pay, low salary, salary benefits*). One can also argue the existence of a general benefits theme (*flexible schedule, home option, unlimited sick leave, paid vacations, nice office, free food),* a theme capturing the culture and environment at the firms (*friendly colleagues, supportive culture/environment, brand recognition, team-oriented),* as well as some other topics related to working hours *(12-hour* work day, *late nights, 50-hour* work week), internal processes (*red tape, political environment, slow pace, legacy system)* and growth/development (*limit/slow growth, career development, opportunities)*.

Although this analysis brings already clearer insights, more nuance is needed to determine actual differences across industries. To further gather more concrete insights, the following section will discuss the findings from performing *topic modelling* on the dataset, to gather common topics under one unique umbrella.

Topic Modelling

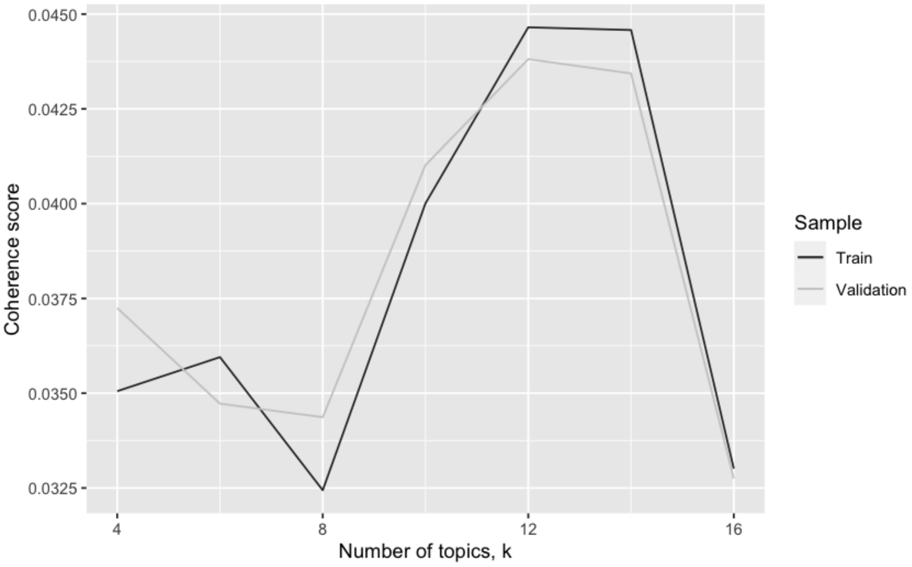
Results

Prior to building the final LDA model, selection of the optimal number of K topics had to be done. With hyperparameters α and β set to 0.1 and 0.05 respectively, seven models with topics $K = 4$ to $K = 16$ (by 2-unit increases) have been built. To assess the optimal number of topics, the coherence and perplexity measures of these seven topics were evaluated in both training and validation sets. As can be inferred from **Table 2** and **Figure 6**, $K = 12$ generated the highest coherence across both sample sets, with a coherence score of 0.045 in the training set and 0.044 in the validation sample.

Table 2: Coherence and perplexity measures for optimal K selection.

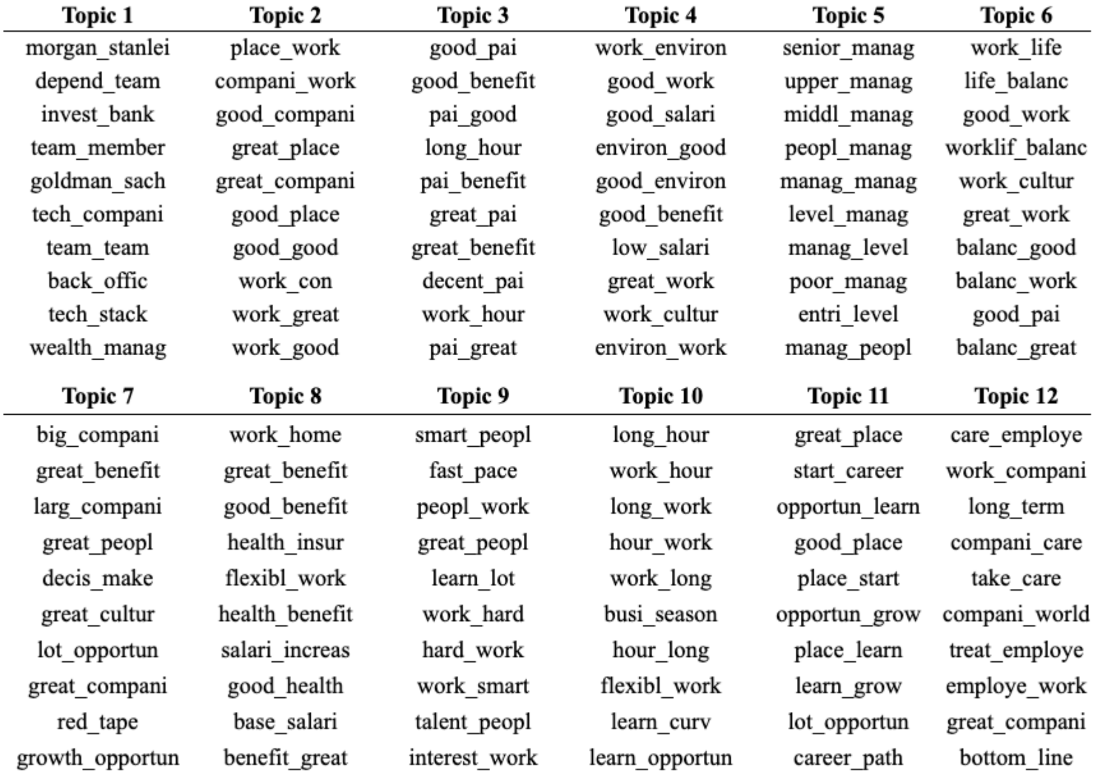
Number of topics, *K*	Coherence training set	Coherence validation set	Perplexity training set	Perplexity validation set
4	0.035	0.037	-113.141	-118.893
6	0.036	0.035	-111.051	-116.690
8	0.032	0.034	-109.822	-115.346
10	0.040	0.041	-108.873	-114.256
12	0.0447	0.044	-108.210	-113.538
14	0.0446	0.043	-107.723	-112.901
16	0.033	0.033	-107.403	-112.454

Figure 6: Coherence measure across training and validation sets.



Moreover, building a model with K = 12 yields a good balance between coherence and perplexity, which are normally key metrics for the good performance of a model. Thus, with these results, an LDA model with K = 12 is then built. **Table 3** summarizes each topic with the most frequent bigrams.

Table 3: Top 10 bigrams for each topic in the LDA model with K = 12.



Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
morgan_stanlei	place_work	good_pai	work_environ	senior_manag	work_life
depend_team	compani_work	good_benefit	good_work	upper_manag	life_balanc
invest_bank	good_compani	pai_good	good_salari	middl_manag	good_work
team_member	great_place	long_hour	environ_good	peopl_manag	worklif_balanc
goldman_sach	great_compani	pai_benefit	good_environ	manag_manag	work_cultur
tech_compani	good_place	great_pai	good_benefit	level_manag	great_work
team_team	good_good	great_benefit	low_salari	manag_level	balanc_good
back_offic	work_con	decent_pai	great_work	poor_manag	balanc_work
tech_stack	work_great	work_hour	work_cultur	entri_level	good_pai
wealth_manag	work_good	pai_great	environ_work	manag_peopl	balanc_great
Topic 7	**Topic 8**	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**
big_compani	work_home	smart_peopl	long_hour	great_place	care_employe
great_benefit	great_benefit	fast_pace	work_hour	start_career	work_compani
larg_compani	good_benefit	peopl_work	long_work	opportun_learn	long_term
great_peopl	health_insur	great_peopl	hour_work	good_place	compani_care
decis_make	flexibl_work	learn_lot	work_long	place_start	take_care
great_cultur	health_benefit	work_hard	busi_season	opportun_grow	compani_world
lot_opportun	salari_increas	hard_work	hour_long	place_learn	treat_employe
great_compani	good_health	work_smart	flexibl_work	learn_grow	employe_work
red_tape	base_salari	talent_peopl	learn_curv	lot_opportun	great_compani
growth_opportun	benefit_great	interest_work	learn_opportun	career_path	bottom_line

Discussion

This model consisted of twelve different topics, which are referred to as employer attractiveness dimensions. Topic 1 includes bigrams related to working in teams: *team member* and *dependant team* but also some industry-specific bigrams (e.g.: *Morgan Stanley, Investment Banking, Goldman Sachs* and *wealth management*). While this topic might not be strong in the rest of the industries, it might be hinting a strong relationship between teamwork and companies in the investment banking industry. This topic will not be considered a dimension of employer attractiveness, as it is not a clear and generalizable indicator of employer branding across industries.

The second topic includes mostly positive bigrams related to the overall employment experience at the company: *good company, great place* and *working great.* Thus, the second topic represents a "General Feeling": do employees generally have a good/positive feeling about their employers?

Topic number 3 gathers several bigrams that revolve around the theme of "Salary": *good pay, good benefit, decent pay,* etc. This topic captures if a job is fairly rewarded through salaries but potentially also other benefits, as the term benefit appears quite often. They partly constitute what Berthon et al. (2005) called economic value.

The fourth topic includes bigrams like *work environment*, *work culture* and *good environment*. Topic 4 can then be named "Work environment" and represents the idea that employees seek good and positive work environments to thrive and where there is a company culture in place.

Topic number 5 is mainly related to "Management": *upper management, senior management* and *people management* are some of the bigrams included in this topic. Aligned with Dabirian et al. (2017), this topic accounts for the idea that managers should be genuine leaders that care about their employees and possess good managing skills.

The sixth topic includes bigrams like *work life, life balance* and *balance good*, this topic represents the "Work/life balance" dimension, which was also discovered by Dabirian et al. (2017). As they unveiled, a company should provide opportunities for all employees to have a state of equilibrium in which a worker's private and professional lives are equally weighted as they interfere with each other.

Often, employees mention working for *large companies*, which goes together with *decision-making and growth opportunities.* Together with a *great culture* and *great people*, the environment of a large company creates the perfect set up for employee development by providing power in the decision making process, leading to growth opportunities within the company. However, big corporations also suffer from excessive regulations that hinder decisions and actions (i.e., *red tape*) and limit growth opportunities. Overall, topic 7 can then be described as "Growth opportunities": to what extent does the company environment provide opportunities for growth and decision making?”. Berthon et al. (2005) found this dimension to be *development value.*

The eighth topic contains bigrams like *work* [from] *home, health insurance* or *great benefits*. These could all be understood as benefits that fall outside the monetary compensation and that employees value even more after COVID-19. The newer generations of employees value companies that will encourage taking care of one’s mental health, which could be aligned with the offering of other benefits and policies like working from home. Topic number two is then described as “Benefits (excl. salary)”: does the company provide other perks and benefits outside of salary? Together with topic number 3, they complete what Berthon et al. (2005) called economic value.

The ninth topic covers bigrams like *smart, great* and *talented* people but also *fast-paced, work hard* and *learn a lot*. It seems like a challenging environment fosters collaboration between talented employees, which results in a learning experience for them. This topic could be understood as “Colleagues”: are there nice and talented people in the workplace? Back in 2005, Berthon et al. (2005) named this dimension as social value.

Topic number 10 can be named “Working hours”. Bigrams like *long hours, working hours, busy season* and *flexible hours* summarize the bigrams included in this topic. This is a new dimension that has not been unveiled by previous literature and would suggest that employees care about the number of hours they spend at work as well as the option for flexible hours.

The eleventh topic is led by bigrams like *great place* [to] *start* [a] *career, opportunities* [to] *learn* and to *grow.* Topic 11 captures the idea that employees often suggest their companies as a great place to start one’s career in online reviews, often appealing to the learning and growth opportunities. Thus, topic 11 can then be named “Starting a career” and represents the idea that some employees recommend their employers in their reviews for fresh graduates looking to start and build their professional paths.

Lastly, topic twelve is all about how companies *treat* their employees. *Care* [about] *employees,* [the] *company cares, take care* or *treat employees* summarize this topic. At the end of the day, employees expect more from their employers than just a monetary reward at the end of each month. Taking care of the employees throughout their whole journey at the company will make them less likely to be dissatisfied and leave for another firm. As important as attracting new talent is, so is retaining extraordinary employees. This topic can be named "Caring about employees": does the organization take care of its employees and their needs?

Through this LDA model, eleven employer attractiveness dimensions have been unveiled: *general feeling, salary, work environment, management, work/life balance, growth opportunities, benefits (excl. salary), colleagues, working hours, starting a career* and, *caring about employees.*

The results of this paper confirm three out of the five original dimensions discussed by Berthon et al. (2005): social (*colleagues*), development (*growth opportunities*) and economic value (*salary*). It also confirms the two recently uncovered value propositions unveiled by Dabirian et al. (2017): management value and work/life balance. Additionally, this paper discovers six additional topics that employees often mention in their reviews about their employers: benefits (excl. salary), starting a career, work environment, working hours, caring about employees and general feeling. In other words, when employees were writing a review, they considered a total of eleven different employer attractiveness dimensions.

While all eleven dimensions are relevant to not only current, former and potential employees but also to all employers, their weight will most likely differ across the four different industries at hand: technology, accounting, investment banking and consumer goods. Industry's idiosyncrasies can play an important role in determining to what extent each of the twelve dimensions matters the most for employees. The next section will present and discuss the regression results to gain further insights into the weight of each of the employer attractiveness dimensions.

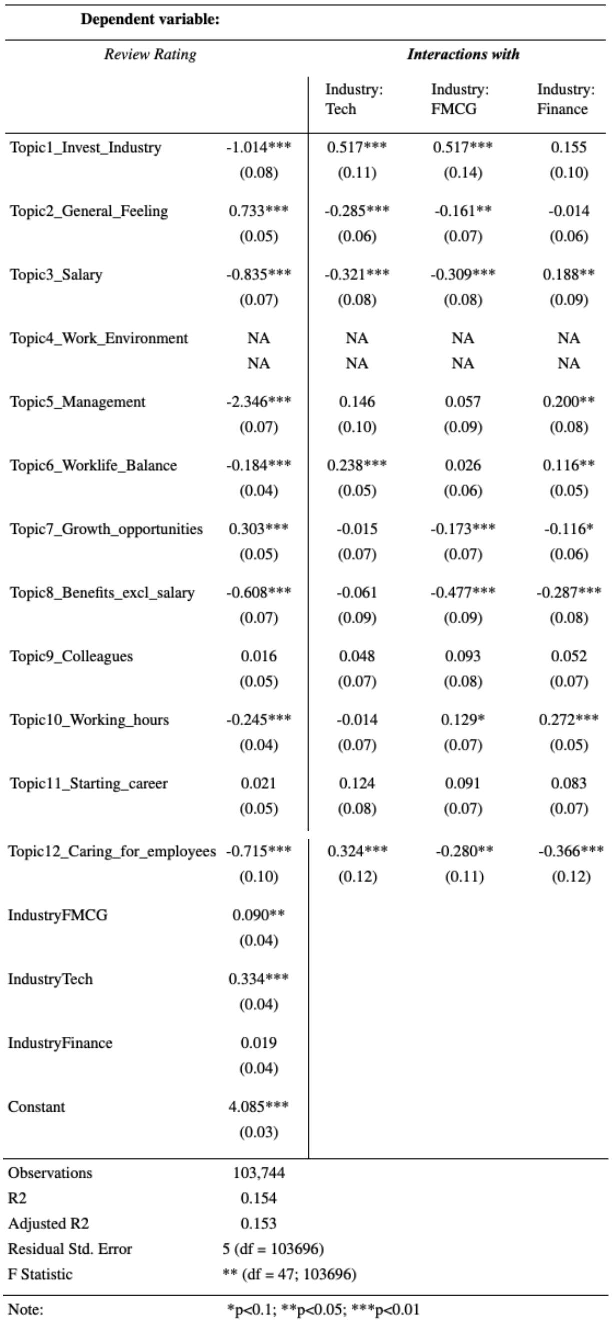
Regression Analysis

Results

In the main analysis, twelve topics were unveiled that related to employer branding and attractiveness. Further interest lies in how these topic probabilities affect review rating in the online platform Glassdoor, as well as the mediating effect of the four different industries.

Table 4 depicts the regression outcome of the model:

Table 4: Summary of the regression model.



Dependent variable:				
Review Rating		***Interactions with***		
		Industry: Tech	Industry: FMCG	Industry: Finance
Topic1_Invest_Industry	-1.014***	0.517***	0.517***	0.155
	(0.08)	(0.11)	(0.14)	(0.10)
Topic2_General_Feeling	0.733***	-0.285***	-0.161**	-0.014
	(0.05)	(0.06)	(0.07)	(0.06)
Topic3_Salary	-0.835***	-0.321***	-0.309***	0.188**
	(0.07)	(0.08)	(0.08)	(0.09)
Topic4_Work_Environment	NA	NA	NA	NA
	NA	NA	NA	NA
Topic5_Management	-2.346***	0.146	0.057	0.200**
	(0.07)	(0.10)	(0.09)	(0.08)
Topic6_Worklife_Balance	-0.184***	0.238***	0.026	0.116**
	(0.04)	(0.05)	(0.06)	(0.05)
Topic7_Growth_opportunities	0.303***	-0.015	-0.173***	-0.116*
	(0.05)	(0.07)	(0.07)	(0.06)
Topic8_Benefits_excl_salary	-0.608***	-0.061	-0.477***	-0.287***
	(0.07)	(0.09)	(0.09)	(0.08)
Topic9_Colleagues	0.016	0.048	0.093	0.052
	(0.05)	(0.07)	(0.08)	(0.07)
Topic10_Working_hours	-0.245***	-0.014	0.129*	0.272***
	(0.04)	(0.07)	(0.07)	(0.05)
Topic11_Starting_career	0.021	0.124	0.091	0.083
	(0.05)	(0.08)	(0.07)	(0.07)
Topic12_Caring_for_employees	-0.715***	0.324***	-0.280**	-0.366***
	(0.10)	(0.12)	(0.11)	(0.12)
IndustryFMCG	0.090**			
	(0.04)			
IndustryTech	0.334***			
	(0.04)			
IndustryFinance	0.019			
	(0.04)			
Constant	4.085***			
	(0.03)			
Observations	103,744			
R2	0.154			
Adjusted R2	0.153			
Residual Std. Error	5 (df = 103696)			
F Statistic	** (df = 47; 103696)			
Note:	*p<0.1; **p<0.05; ***p<0.01			

As it can be seen, most of the coefficients are significant at least at the 0.05 level, with some of them being also significant at the 99% confidence interval. However, a very low adjusted R-squared is reported (15.32%) which can be explained by the fact that most of the assumptions of linear regression are not met in this model. Yet, the model allows making comparisons between different industries and the effects of the topic probabilities on the *review ratings*.

Taking as a reference topic *Work environment* – considering it a neutral/slightly positive topic – the highest decrease in *rating* happens when a review for an Accounting firm is related to the *Management* topic ($\hat{\beta}$ = -2.35, *p-value < 0.001*). The next strongest effect comes from *topic 1,* which was specific to the Investment Banking industry and is not considered a employer branding dimension *per se*. Yet, mentioning this topic in a review had a strong negative effect on *rating* ($\hat{\beta}$ = -1.01, *p-value* < 0.001) for the accounting industry. *Salary* has also a crucial effect on *rating*: reviews in the accounting industry revolving around the topic of salary have lower ratings ($\hat{\beta}$ = -0.83, *p-value < 0.001*) than those reviews referring to the work environment. Other negative effects specific for the accounting industry come from reviews containing topics like *Caring about employees (*$\hat{\beta}$ *= -0.71, p-value < 0.001*), *Benefits – excl. salary – (*$\hat{\beta}$ *= -0.61, p-value < 0.001*)*, Working hours (*$\hat{\beta}$ *= -0.25, p-value < 0.001*) and *Work life balance (*$\hat{\beta}$ *= -0.18, p-value < 0.001*).

On the other hand, there are some other topics that have a positive effect on *rating*, compared to the baseline topic *Work environment*. While *Starting a career* and *Colleagues* have a slight positive effect, this effect is not significant at the 90% confidence level. The strongest positive effect for the accounting firms comes from *General Feeling (*$\hat{\beta}$ *= 0.73, p-value < 0.001*), a topic that included bigrams like *good company* or *great place*. *Growth opportunities* also have a positive significant impact in the *rating* of accounting firms' reviews *(*$\hat{\beta}$ *= 0.30, p-value < 0.001*).

The regression model also incorporates the interactions of all topic probabilities with the rest of the industries. Eleven of these interactions are highly significant with a *p-value < 0.001.* Reviews related to the topic of *Salary* had a greater decrease in *review rating* for the Tech industry compared to the Accounting industry, the reference group ($\hat{\beta}$ = -0.32*, p-value* <

0.001). This also happened for the FMCG industry, which experienced a decrease of 0.31 in *review rating* for a one unit increase in the topic probability of *Salary* (*p-value < 0.001).* The interaction between *Salary* and the Investment & Asset Management industry was found to be not significant (*p-value > 0.05).* Thus, we fail to reject the null hypothesis that the estimated coefficient of the interaction between *Salary* and the Investment & Asset Management industry is zero at significance level $\alpha = 0.05$. This means that no differences in *review rating* are reported between the Accounting and Investment & Asset Management industries.

The presence of the topic *Management* did not differ greatly from one industry to another. The effect for both the Tech and FMCG industries were not significant (all *p-values > 0.1)* when compared to the Accounting industry. However, there is a slightly significant positive effect for reviews mentioning *Management* in the Finance group, with an estimated coefficient of $\widehat{\beta} = 0.20$*, p-value < 0.05*.

For *Worklife Balance* it was found that for the Tech industry, a review mentioning this topic had the most significant positive effect on *review rating* ($\widehat{\beta} = 0.24$*, p-value < 0.001*), compared to the Accounting industry. A similar effect was found for the Investment & Asset Management industry causing an increase of 0.12 in *review rating* (*p-value < 0.05)* for a 1 unit increase in the topic probabilities for *Worklife Balance* (compared with the accounting industry). The null hypothesis that the interaction coefficient between *Worklife Balance* and the FMCG industry equals 0 failed to be rejected at significance level $\alpha = 0.1$, concluding that no significant differences were found between the FMCG and the Accounting industry when the review contains *Worklife Balance* theme.

The presence of the *Growth Opportunities* theme in reviews decrease the *rating* for all industries with respect to the Accounting industry. The most significant effect was found for the FMCG industry, whose *review rating* decreased by 0.17 (*p-value < 0.01)* in the presence of this topic, compared to the reference group. For the Investment industry the effect is similar, as reviews containing this topic experienced a lower review rating ($\widehat{\beta} = -0.12$*, p-value < 0.1).* No significant effect was found for the Tech industry, meaning that we fail to reject the null that the interaction coefficient of *Growth Opportunities* with the Tech industry at the significance level $\alpha = 0.1$.

For *Benefits (excl. Salary)*, significant differences were reported between the FMCG industry and the reference group (the Big 4). Writing about additional benefits had a negative impact for the Consumer Goods industry ($\hat{\beta} = -0.48$, *p-value < 0.001*) and for the Financial industry ($\hat{\beta} = -0.29$, *p-value < 0.001*). For the Tech companies however, no significant differences were found compared to the Accounting industry; the null hypothesis that this coefficient is equal to zero cannot be rejected at significance level $\alpha = 0.1$ (*p-value = 0.50)*.

The interactions between the presence of the topic *Colleagues* with the industries were all not significant (all *p-values > 0.1)*. As we fail to reject the null hypotheses that these coefficients are zero at significance level $\alpha = 0.1$, no differences are reported in the way a review on the topic of *Colleagues* affects *review rating* for any of the industries with respect to the Accounting industry.

Reviews revolving around the topic of *Working Hours* had a positive and significant effect for the Finance industry ($\hat{\beta} = 0.27$, *p-value < 0.001*) but not as significant for the FMCG industry ($\hat{\beta} = 0.13$, *p-value < 0.1*). Yet, for the Tech industry, no significant differences were found when compared to the Accounting industry.

The effect of the topic *Starting a career* on *Rating* did not differ from one industry to another. The effect for all industries were not significant (all *p-values > 0.1)* when compared to the Accounting industry.

Caring about employees presented signficant differences across the industries. For the Tech industry, this effect was positive and significant with *p-value* < 0.01 ($\hat{\beta} = 0.32$), compared to the accounting industry. Contrary to this, both the FMCG and Finance industries displayed a negative effect in *review rating* when *Caring about employees* was present. The strongest effect came from the Financial industry ($\hat{\beta} = -0.37$, *p-value < 0.01*) which experienced ratings 0.37 units lower than the Accounting industry in the higher the presence of this topic was. Similarly, the FMCG perceived ratings 0.28 units lower than the Accounting industry for higher probabilities of this topic ($\hat{\beta} = -0.28$, *p-value < 0.05*).

Good and positive words related to the experience of the employees (topic: *General Feeling*) had a not so strong positive impact in the *rating* when it was in an interaction with the other industires. In particular, reviews for companies in the Tech industry saw a decrease of 0.29 rating units when the topic *General Feeling* was present, compared to the Accounting Industry ($\widehat{\beta}$ = -0.29*, p-value < 0.001*). A similar effect was found for the FMCG industry ($\widehat{\beta}$ = -0.16*, p-value < 0.05*) but no significant differences were reported between the Accounting and the Finance industry at the significance level α = 0.1.

Random Forest

Because of the issues presented by the non-linearity of the data, a random forest with 1000 trees had been built on a train set of 80% of the data. Surprisingly, this model reported a 13.04% of total variance explained, which is slightly lower than the multiple R-squared of the multiple regression model. Using a variable importance approach, a plot can be generated to display to *rank* the importance of the independent variables in predicting the *rating*.

Figure 7: Variable importance plots for the Random Forest model

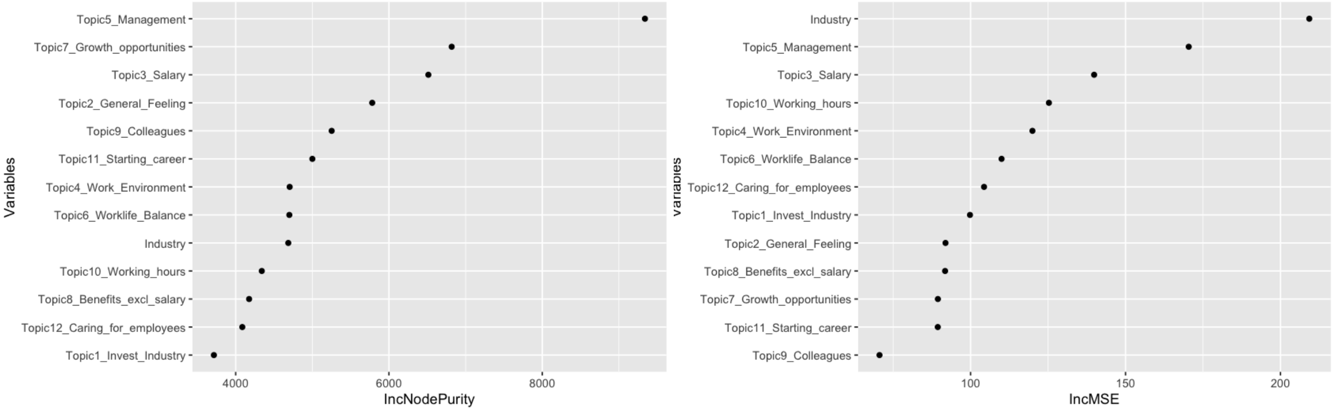
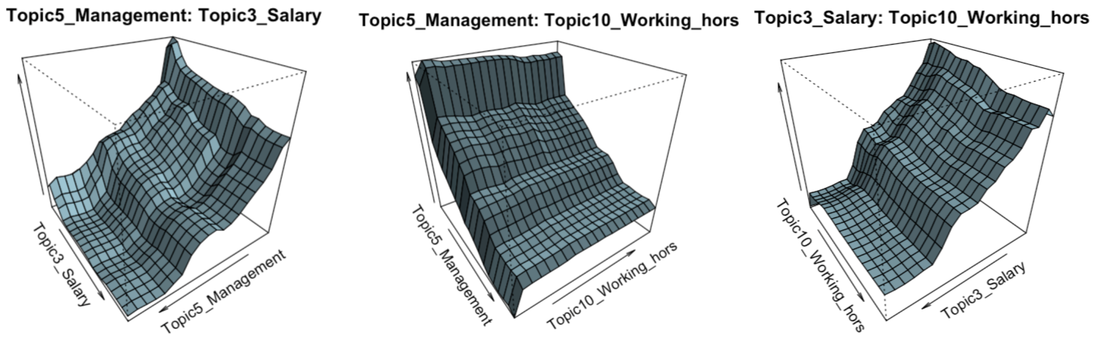


Figure 7 illustrates the relative variable importances using both importance measures described in the Methodology section. Both rankings differ slightly with each other: with the first measure (left graph), the topic of *Management* is the most relevant predictor, followed by *Growth opportunities* and *Salary.* Other influential predictors follow: *General Feeling*, *Colleagues* and *Starting a career,* but they are significantly less influential the top three. Using the OOB permutation measure (right graph), *Industry* has the strongest influence in predicting *rating*, followed by the *Management* and *Salary* topics, which are also shared by

the first measure of importance. *Working hours, work environment* and *work-life balance* are also important predictors for the rating in an employment review.

Management, salary and *working hours* are highly influential in the outcome of the random forest model. **Figure 8** illustrates the partial dependence plot for the interaction of these features:

Figure 8: Bivariate partial dependence plots for *management, salary* and *working hours.*



The three graphs display strong interaction effects. For high values on a review of *management* topic probabilities, review *rating* is nearly independent of the topic probabilities for the *salary* topic, whereas for lower values there is a strong dependence on the salary *topic probabilities*. The same reasoning follows for the middle and right graphs. Despite these strong interactions, they are not noteworthy, as if the topic probability for a specific topic in a review approaches 1 (i.e.: the review is mostly/only about that unique topic), there is no room for other topics to have a probability that approaches 1 in that same review as well. When both topic probabilities are far from 1 (i.e.: the review contains a mix of both topics), then there is a strong dependence of both topic probabilities on *rating*, as they coexist in the same review.

Discussion

The results presented for the linear regression model unveil crucial findings for the management of talent in firms across the four industries studied. Firstly, reviews for the accounting industry that were related to the *Management* topic had a very strong negative effect on the overall rating that the employee gives to these firms. This hints that a good employee-management relationship is strongly demanded by employees in this industry who require better practices from their leaders and overall better management skills. However, these negative effect improves slightly – but still is negative – when this topic is paired with the Finance industry in which management is still a topic to work on but is not as urgent as in the accounting industry.

On a similar note, employees value that the company and their managers care about them, as suggested by the *Caring about employees* topic which, if present in a review, had a negative impact on the general rating of the accounting employers. The only industry which presents slightly better results is the Tech industry. While still impacting negatively the rating, the tech industry seems to have better policies in place to support the employees and care about them and their wellbeing. The financial industry is the one displaying the worst data for this topic which requires immediate action in order to improve their attractiveness as an employer.

In addition, reviews in which the topic of *Salary* or *Benefits (excl. Salary)* was predominant, had – on average – lower ratings as well, indicating that accounting companies should reconsider the monetary compensation that they offer to stay competitive in the market and match the expectations and value of their employees. The negative effect of Salary is even stronger for high-tech companies and the FMCG industry, when compared to the accounting industry. While the negative effect on high-tech companies was not expected, it can be explained by the fact that some firms have non-tech job positions (i.e.: Amazon warehouse or delivery workers) which might account for this negative feeling. The Finance industry presented slightly better results compared to the accounting firms, but the effect on *rating* remains negative. For other non-monetary benefits, the accounting industry presented a higher rating than those reviews belonging to the FMCG or Financial industry. It is specially concerning the case of the Consumer Goods industry, who shows the strongest negative effect when a review is highly about this topic.

Another sharp decrease in the *rating* of the accounting reviews is caused by both *Working hours* and *Work-life balance*. These topics being present in a review indicate that employees are not satisfied with the number of hours that they have to work and thus, the little work-life balance that the company offers them. Although *Work life balance* had a negative effect on the accounting reviews, the Tech industry presents a positive effect on *rating*, meaning that employees are less dissatisfied with the work-life balance offered by their employers. For the Finance industry the effect is still negative but not as much as for the accounting industry, who seems to be the worst-performant in this category. Similarly, the effect of *Working hours* in a review is quite positive on the rating for the finance industry, hinting that the companies in the accounting industry are the ones who should take care of the number of working hours and overwork that their employees have to take in.

On the other hand, *growth opportunities* are positively valued and important for employees. Companies that have the right policies in place to create an environment that allows for personal and professional growth development will be attractive to outsiders but will also give reasons for current workers to not leave. All companies present a positive effect in this topic but the accounting industry has the strongest effect, followed by the finance and the consumer goods industry. The tech industry performs as well as the accounting industry in this category.

Random Forest

Results from the random forest model are aligned with the multiple regression outcomes just described. A review about the topic of *Management* has a very influential impact on the rating of such review, meaning that employees demand and require genuine leaders with the right skills to perform well as managers. Not only the management style should be correct but also the hierarchies in place, which might be the cause of extreme bureaucracy and micromanagement. As the results show, *Salary* is also a main cause of changes in rating. Companies need to align their monetary compensations with the market's requirements, as well as the levels of expertise, know-how and experience of the candidates. *Working hours* have also a great influence on the rating of a review; working overtime can be the cause of long-term dissatisfaction in a company, which can end in resignation by the employees. In relation to this, *Work-life balance* shows to be quite influential as well, again showing that a good balance between working hours and personal life is key for retaining and attracting employees. Lastly, good company culture and *work environment* are also key determinants of the rating of a company and hence, its attractiveness as an employer. Working on the factors that drive a high rating and thus, high attractiveness, will improve the online reputation of a company as an employer, signalling themselves as a good place to work at.

Conclusion

Firms have been competing to recruit the most talented individuals for the last two decades. Yet, the competition has never been more intense. With a pandemic that made a lot of employees rethink their employment situation, a wave of workers resigning from their current jobs started, and has not stopped yet. Employees have been leveraging this situation as they now have the power to look elsewhere for a better offer with conditions that match their expectations. Yet, a lot of employers do not know what these expectations are and how their offerings should change to avoid employee attrition. Ignorance over these expectations can be detrimental for a company as the lack of change in their employment practices can lead to a worsening of their employer's brand and attractiveness. In this paper, employer attractiveness was represented by the rating of a company in Glassdoor reviews. This research has unveiled the set of value propositions that employees value the most in an employer and has also made a distinction on how these vary across industries. These results provide new guidelines for improving a firm’s employer branding and minimizing attrition cases as well as maximizing the recruiting rates of new talent.

At the beginning of this paper, three research questions were raised which have been answered through the analyses and results presented and discussed in this paper. Firstly, “What dimensions better capture the attributes that employees find attractive about their employers?”. Through LDA topic modelling, twelve different topics have been unveiled: colleagues, growth opportunities, salary, management value, work-life balance, other benefits (excl. salary), starting a career, work environment, working hours, caring about employees, general feeling and one last topic that did not stand for a dimension *per se*. These eleven topics (excluding the last one) were mentioned recurringly in the set of more than 129 thousand reviews when employees were sharing a review about their employment experiences.

Secondly, the impact of the different dimensions on the overall rating for a company had to be quantified, i.e.: “what dimensions matter the most for the rating of a company?”. For that, a linear regression approach was used and then complemented with a random forest model to capture non-linear effects. The linear regression findings support that reviews on the topics of management, salary and benefits (excl. salary) had the strongest negative effect on the rating of a company. Moreover, the random forest outcome also identifies management and salary as key drivers of review rating, but the model also unveiled the importance of working hours, working environment (culture) and work-life balance in the employer branding of an enterprise. This suggests that firms need to re-evaluate their policies and prioritize change in management, salary & benefits, working hours/work-life balance and company culture to see an improvement in their online reputation.

The last question was aimed at discovering differences across industries: “How do these dimensions differ across industries?”. Using hypothesis testing in the linear model, the following recommendations can be drawn: (1) a change in monetary compensation for the Accounting, Tech and FMCG industries is required as reviews on this topic have lower ratings; (2) management values within a company are required to change across all industries as they all present very strong negative effects on the review rating; (3) companies from all industries can follow the example of the Tech industry in regards to their work-life balance policies; (4) the Tech, FMCG and Finance industries need to offer better development and growth opportunities within the firm to retain more of their employees; (5) issuing other non-monetary benefits needs to be made a priority for companies in all industries; and (6) future work in the area of caring for the employee is required from all industries but specially for the Finance industry, as their ratings get drained by the negativity surrounding this topic.

Limitations and further research

While this research gives companies the knowledge of the propositions that matter the most to employees, it does not provide any guidelines or recommendations on how to improve in each individual area as a deep understanding of human resources techniques and practices is required. Yet, the pairing of the results presented in this paper with a human resources qualitative approach can be a starting point for valuable future research. Moreover, this paper only focused on very big corporations in which polarization might be stronger. Small and medium-sized companies might provide more homogeneity in terms of the topics presented in the data, giving more accurate results and recommendations. Lastly, this paper did not account for the diversity in roles and positions within a company (i.e.: not all reviews within a tech company might be from a tech-driven position). For further research, making this distinction is recommended to really capture the distinctive characteristics and requirements of each industry.

References

Ambler, T., & Barrow, S. (1996). The employer brand. *Journal of Brand Management*, *4*(3), 185–206. https://doi.org/10.1057/BM.1996.42

Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using topic modelling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, *3*, 42.

Backhaus, K., & Tikoo, S. (2004). Conceptualizing and researching employer branding. Career development international.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management*, 17(1), 99-120.

Berthon, P., Ewing, M., & Hah, L. L. (2015). Captivating company: dimensions of attractiveness in employer branding. *Http://Dx.Doi.Org/10.1080/02650487.2005.11072912*, *24*(2), 151–172. https://doi.org/10.1080/02650487.2005.11072912

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. In *Text mining* (pp. 101-124). Chapman and Hall/CRC.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Boxall, P. (1996). The strategic HRM debate and the resource‐based view of the firm. *Human resource management journal*, 6(3), 59-75

Cable, D. M., & Turban, D. B. (2003). The Value of Organizational Reputation in the Recruitment Context: A Brand-Equity Perspective. *Journal of Applied Social Psychology*, *33*, 2244–2266.

Dabirian, A., Kietzmann, J., & Diba, H. (2017). A great place to work!? Understanding crowdsourced employer branding. Business horizons, 60(2), 197-205.

Edwards, M.R. (2010), "An integrative review of employer branding and OB theory", *Personnel Review*, Vol. 39 No. 1, pp. 5-23. https://doi.org/10.1108/00483481011012809

Fortune (2017). The Top 100 Best Companies to Work For. https://fortune.com/best-companies/2017/google/

Glassdoor (2022). About Us. https://www.glassdoor.com/about-us/

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Hoskisson, R. (1999). Theory and research in strategic management: swings of a pendulum. *Journal of Management*, *25*(3), 417–456. https://doi.org/10.1016/S0149-2063(99)00008-2
Lado, A. A., & Wilson, M. C. (1994). Human resource systems and sustained competitive advantage: A competency-based perspective. *Academy of management review*, *19*(4), 699-727.

Islam, T. (2019). Yoga-veganism: Correlation mining of twitter health data. *arXiv preprint arXiv:1906.07668*.

Jones, T. (2021). textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.5. https://CRAN.R-project.org/package=textmineR

Lloyd, S. (2002). Branding from the Inside Out, Business Review Weekly, 24(10), pp.64-66.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *arXiv preprint arXiv:1403.6397*.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

Sivertzen, A. M., Nilsen, E. R., & Olafsen, A. H. (2013). Employer branding: Employer attractiveness and the use of social media. *Journal of Product and Brand Management*, *22*(7), 473–483. https://doi.org/10.1108/JPBM-09-2013-0393/FULL/PDF

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 952-961).

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Park, Y., Alam, M. H., Ryu, W. J., & Lee, S. (2015). BL-LDA: Bringing bigram to supervised topic model. In *2015 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 83-88). IEEE.

Wright, P. M., & McMahan, G. C. (1992). Theoretical perspectives for strategic human resource management. *Journal of management*, *18*(2), 295-320

Wright, P. M., Dunford, B. B., & Snell, S. A. (2001). Human resources and the resource based view of the firm. *Journal of management*, 27(6), 701-721.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456).

Ye, Q., Law, R., Gu, B., & Chen, W. (2011). The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human behavior*, *27*(2), 634-639.

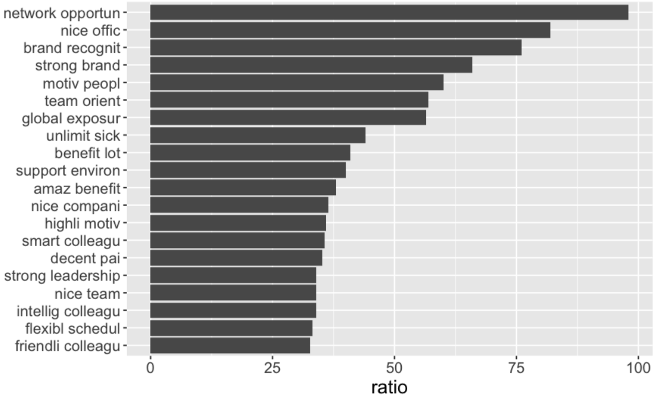
Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. In *BMC bioinformatics* (Vol. 16, No. 13, pp. 1-10). BioMed Central.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, *74*(2), 133-148.

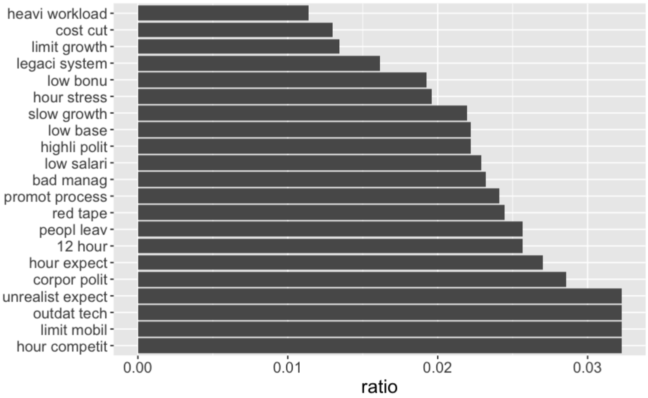
APPENDIX A

Relative frequency of bigrams in the Consumer Goods industry reviews

Bigrams appearing relatively more often in positive reviews:



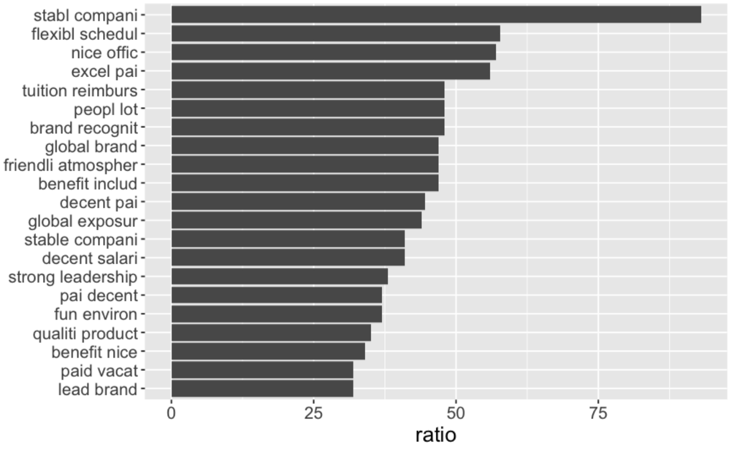
Bigrams appearing relatively more often in negative reviews:



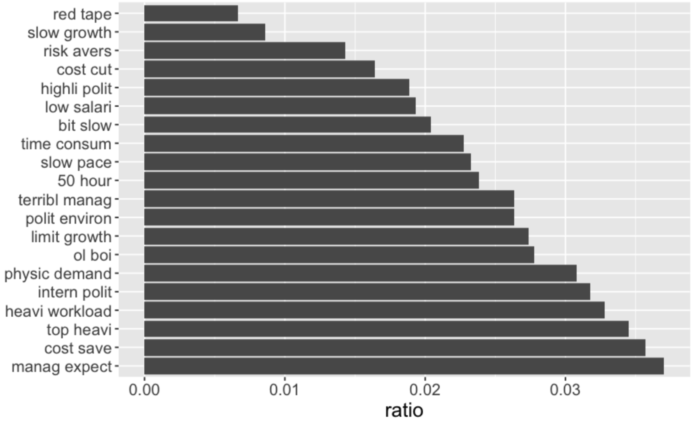
APPENDIX B

Relative frequency of bigrams in the Asset Management industry reviews

Bigrams appearing relatively more often in positive reviews:



Bigrams appearing relatively more often in negative reviews:


red tape
slow growth
risk avers
cost cut
highli polit
low salari
bit slow
time consum
slow pace
50 hour
terribl manag
polit environ
limit growth
ol boi
physic demand
intern polit
heavi workload
top heavi
cost save
manag expect
ratio