The Informational Value of Employee Online Reviews

Efthymia Symitsi[1], Panagiotis Stamolampros[1], George Daskalakis[2], Nikolaos Korfiatis[2]

European Journal of Operational Research (Forthcoming)

[1]Leeds Business School, University of Leeds. 2.13 Maurice Keyworth Building,
Leeds, LS2 9JT United Kingdom

[2]Norwich Business School, University of East Anglia, Norwich Research Part,
Norwich NR50WE, United Kingdom

Abstract

This paper investigates the informational value of online reviews posted by employees for their employer, a rather untapped source of online information from employees, using a sample of 349,550 reviews from 40,915 UK firms. We explore this novel form of electronic Word-of-Mouth (e-WOM) from different perspectives, namely: (i) its information content as a tool to identify the drivers of job satisfaction/dissatisfaction, (ii) its predictive ability on firm financial performance and (iii) its operational and managerial value. Our approach considers both the rating score as well as the review text through a probabilistic topic modeling method, providing also a roadmap to quantify and exploit employee big data analytics. The novelty of this study lies in the coupling of structured and unstructured data for deriving managerial insights through a battery of econometric, financial and operational research methodologies. Our empirical analyses reveal that employee online reviews have informational value and incremental predictability gains for a firm's internal and external stakeholders. The results indicate that when models integrate structured and unstructured big data there are leveraged opportunities for firms and managers to enhance the informativeness of decision support systems and in turn, gain competitive advantage.

Keywords: *Analytics, Employee online reviews, Topic modeling, Big data, Decision processes*

1. Introduction

The proliferation of online platforms that extend the digital discussion beyond consumption experiences offers new opportunities for the study of organizational interventions and has the potential to provide key insights to managers. Employee online reviews form a special case of e-WOM that contains internal information from all-level employees that managers and researchers alike can use to extract valuable insights. Nonetheless, this is a rather untapped source of information as it has received only limited attention in the literature (e.g., Huang, Li, Meschke, & Guthrie, 2015; Symitsi, Stamolampros, & Daskalakis, 2018). This study aims to explore the informational value of employee online reviews for firms and their internal and external stakeholders from various perspectives.

Initially, we investigate the drivers of employee satisfaction using both structured and unstructured data derived from online employee reviews (i.e., numerical rating and review text respectively). Then, we demonstrate the informational gain of using this information in operational and financial applications. More specifically, we explore the information content of these reviews in predicting employee turnover and firm profitability. Finally, we estimate the economic significance of this information for investors and the gains it produces if incorporated in an investment strategy. Our analysis is based on a battery of methodologies that among others combine recent advances in the probabilistic topic analysis, allowing us to capture the information embodied in the review text of a novel dataset of 349,550 employee review ratings and texts for 40,915 UK firms across all sectors of the economy, sourced from *Glassdoor*.[1]

Our study contributes to several streams in the extant literature. First, we add to the literature that investigates the value of structured and unstructured data for corporate operations and

[1] We are grateful to Glassdoor's data science team for providing us the data used in this study.

managerial decisions. Firms invest in big data initiatives for collecting and exploiting massive amounts of data to derive value, but still, managers question the benefits that this information generates (LaValle, Lesser, Shockley, Hopkins & Kruschwitz, 2011). In principle, big data consist of raw data that firms should extract, analyze and convert into insights that will form the basis for competitive advantage, improve business decisions and eventually performance (Kunc & O'brien, 2019; Xu, Frankwick & Ramirez., 2016). However, academic guidance for practical application is still limited. The intuition behind this is that a successful strategy not only requires proper infrastructure but also changes in business practices (McAfee & Brynjolfsson, 2012). Therefore, we demonstrate how information from employees can be of value for key managerial issues, such as employee satisfaction, firm profitability, employee turnover and investment decisions.

Second, this study presents the benefits of coupling structured and unstructured data through various empirical applications. Specifically, we showcase how managers can practically employ unsupervised textual methods and harvest valuable information from employee big data in order to feed-in decision support systems. To this end, we respond to recent calls for the incorporation of unstructured data in Operational Research and Management Science (OR/MS) models (Mortenson, Doherty & Robinson, 2015) providing not only empirical evidence for the benefits of integrating rich big data sources in various models but also a roadmap for managers that will allow the sustainable enhancement of business practices and result in competitive advantage.

Third, we contribute to the literature that studies the forecasting ability of information from online platforms, such as online review aggregators, forums, social media and search engines. Several scholars find that through the exploitation of such data, accurate predictions can be made for sales, stock prices, and electoral performance (Chen, De, Hu & Hwang, 2014; DiGrazia, Mckelvey, Bollen & Rojas, 2013; Kulkarni, Kannan & Moe, 2012), among others.

Our study extends this literature by testing the predicting ability of an unexplored informational cue derived from employee review platforms.

Finally, we add to the voluminous literature that explores the factors that explain employee satisfaction and subsequent outcomes, such as employee turnover and firm performance (Hancock, Allen, Bosco, McDaniel & Pierce, 2013; Rubenstein, Eberly, Lee & Mitchell, 2018). Our approach addresses several of the disadvantages of the established academic and managerial practices. At the same time, we address some of the gaps identified in the extant literature (Hom, Lee, Shaw & Hausknecht, 2017) by employing a more content-specific approach that aligns our empirical applications with what is proposed as effective evidence-based management practices and by integrating multiple sources of data within a particular context (Briner, Denyer & Rousseau, 2009).

This study is structured as follows: Section 2 presents the research background. Section 3 describes the data. Section 4 provides the framework for topic modeling of the review text. The research design comprises of three empirical applications: Section 5 investigates the drivers of employee satisfaction considering both structured and unstructured information from online reviews; Section 6 demonstrates the predictive power of structured and unstructured data in determining employee turnover, a critical factor of operational performance; Section 7 studies employee satisfaction as a determinant of firm profitability and assesses the informational value of both structured and unstructured data for managers and investors. Section 8 discusses the implications and limitations of this research, presenting also a machine learning pipeline that can be used by firms and managers for combining structured and unstructured data to harness the *"wisdom of employees"*. The study concludes in Section 9.

2. Related literature

Managers and academics alike, are nowadays overwhelmed with a large scale of online user-generated information. Online reviews, in particular, have received considerable attention in the consumer decision-making literature (Zhu & Zhang, 2010). Several studies examine the effect of review valence and/or volume on sales and how this information can benefit the production process through more accurate demand forecasting (Chintagunta, Gopinath & Venkataraman, 2010; Schneider & Gupta, 2016). In addition to customers, review platforms offer access to valuable information for researchers and managers by eliminating considerably the cost and time resources spent on surveys or focus groups. Moreover, firms gain access not only to reviews related to their products/services but also to reviews about their competitors and the market.

Hitherto, academic research is focused mainly on online consumer evaluations in a post-transactional context, while other forms of online reviews are neglected. Online platforms that cover issues beyond consumption-based experiences offer new opportunities for the study of other topics of interest for researchers and managers. Employee online reviews is a special case of e-WOM that gains popularity in the literature as a proxy of employee satisfaction (Huang et al., 2015; Symitsi et al., 2018), and a source of valuable internal information from employees across all levels of hierarchy; from senior managers to rank-and-file employees. However, the possibilities that online reviews offer as an internal source of information from all-level employees, and a managerial tool for HR and OR practices, are unexplored.

This study fills the knowledge gap of the informational value of online employee reviews by offering an extensive analysis of how both structured and unstructured data can be beneficial for managers and other stakeholders. As argued by Zhan and Tan (2018), "*the great value of big data generates from the use of integrated data sources*" *(p.2)* . Through operational and financial frameworks, we demonstrate that the enriched content of online reviews may provide valuable assistance in predicting and explaining several critical managerial issues such as job satisfaction, employee turnover and firm performance.

Numerous studies in the literature focus on factors that drive employee satisfaction, employee turnover and their relationship with various aspects of financial and operational performance (Bernhardt, Donthu & Kennett, 2000; Edmans, 2011; Huang et al., 2015; Symitsi et al., 2018). Employee turnover is also a popular topic in the management and operational research strands (Bordoloi & Matsuo, 2001; Corominas, Lusa & Olivella, 2012; Darmon, 2004; De Bruecker, Van den Bergh, Beliën & Demeulemeester, 2015; Griffeth, Hom & Gaertner, 2000; Hom & Kinicki, 2001; Maertz Jr & Griffeth, 2004; Song & Huang, 2008). High turnover has been found to interfere with performance, and even delay or disrupt firm operations (Hausknecht, Rodda & Howard, 2009; Mohr, Young & Burgess, 2012; Ton & Huckman, 2008). Moreover, it has been associated with elevating HR costs associated with hiring, training and productivity loss (Cascio, 1991). Thus, investigating labor turnover has been a main focus of academic literature in operational research with considerable managerial implications. Such implications are magnified in cases of human-centered operations (Bordoloi & Matsuo, 2001) that workforce planning plays a core role (e.g., tourism and services, in general).

Employee online reviews offer an alternative source of information that may address several limitations found in existing methods in this field of research. First, the collection of survey data for employee satisfaction, motivation and engagement is a rather static task in the extant research and management practice as it commonly takes place only once per year (Lee, Hom, Eberly, Li & Mitchell, 2017). In contrast, online reviews arrive at a higher frequency offering dynamic information from employees. Importantly, the representation of former employees in web platforms is similar to that of current employees, providing significant advantage given the difficulty of firms and researchers in approaching and collecting data from previous employees, thus offering a unique opportunity to investigate employee turnover factors. Second, employee satisfaction surveys are based on measurement scales with predefined constructs (e.g., *Minnesota Satisfaction Questionnaire* and the *Job Descriptive Index*). However, the variables of interest are characterized as multidimensional and therefore, predefined constructs are unlikely to capture latent factors (Jung et al., 2009). Online reviews, through the review text, provide employees the opportunity to discuss the critical drivers of their experience with an employer (either positive or negative) beyond the established measures, allowing the investigation of further unexplored determinants. The distinct narratives of employees regarding the advantages and disadvantages of working with an employer offer the opportunity to investigate potential asymmetries in the factors that drive satisfaction and dissatisfaction. Third, researchers and practitioners can very easily access a huge pool of employee reviews that arrive from all sectors of the economy; this would be extremely difficult, time-consuming and costly to achieve if having to rely on collecting primary data. Fourth, online reviews that arrive voluntarily and anonymously from employees across all levels of hierarchy address issues arising due to the potential reluctance of employees to provide accurate and critical feedback subject to managerial pressures and intolerance (Holland, Cooper & Hecker, 2016). Finally, employee online reviews offer advantages over other samples beyond survey data that are used as a proxy of employee satisfaction such as the "*Best Places to Work*" list (e.g., Edmans, 2011). Analyses based on that list may suffer from potential self-selection bias as it is the firm's decision whether to participate in this survey or not, with a higher probability of firms that know (or believe) that have satisfied employees to participate. On the contrary, in the case of online reviews, employees are free to decide whether they provide a review for their employer or not.

3. Data

Our dataset consists of online employee reviews for UK firms provided to the authors by *Glassdoor*. *Glassdoor* is one of the most popular jobs listing websites, with a strong presence in the US and the UK markets, that allow current and former employees to anonymously review the working experience with an employer. In addition to an overall rating, employees evaluate several job elements, such as career opportunities, compensations and benefits, senior leadership, work/life balance, and culture and values. The majority of online reviews are accompanied with demographics, such as gender, age, and education and firm-specific information, such as the sector the company operates in, whether the company is public or private, and economic variables. Table 1 provides a description of the dataset.

Table 1: *Employee online review sample characteristics*

Reviewer Characteristics	
Total number of reviews	349,550
- *Former employees*	165,441
- *Current employees*	184,109
Gender: Female	88,670
Gender: Male	132,404
Education: High School graduate	11,259
Education: University graduate (Bachelor)	70,119
Education: University postgraduate (MSc/MBA/PhD)	22,766
Average reviewer age	34.5
Employer Characteristics	
Total number of employers	40,915
Average number of employees per employer	3,587
Average annual profitability (million £)	31,892

The *Glassdoor* platform allows employees to accompany their numerical text with open-ended narratives discussing positive and negative aspects and providing feedback to management. Considering the availability of this unstructured information in our dataset, we are interested in extracting qualitative dimensions in order to shed light on latent determinants of job satisfaction. The distinction of the review narratives to discuss advantages and disadvantages in different sections offers an insightful analysis of potential differences/asymmetries on the factors that make employees satisfied or dissatisfied (Herzberg, Mausner & Snyderman, 2011).

To this end, we use topic modeling techniques. Topic models are unsupervised text mining techniques that identify and organize a textual corpus of documents or words in specific topics (deriving a document-topic or word-topic distribution) based on their co-occurrence likelihood. Recent developments in topic modeling methods gain popularity in academic research as textual analysis methods that allow scholars to harness the information content of large corpora (e.g., Korfiatis, Stamolampros, Kourouthanassis, & Sagiadinos, 2019; Tirunillai & Tellis, 2014). Another advantage of these methods is that they offer reproducibility, since human coders are not used, and fast processing of unstructured big data for deriving meaningful information. In particular, this study employs the structural topic model (STM) (Roberts, Steward & Airoldi, 2016) which allows the inclusion of additional covariates (e.g., document metadata) into the estimation of document-topic and word-topic distributions. This approach relaxes the restrictive assumption which considers an equal probability of all authors to write a document, allowing the modeling of the probability of topic prevalence to be observed across a range of covariates.

4. Extracting qualitative dimensions from reviews through topic modeling

The topic modeling analysis involves three steps: (a) the pre-processing of the text, (b) the identification of the number of topics that explain most of the variability of the corpus, and (c) the inclusion of the overall rating as a covariate to estimate how the topics change for more satisfied and dissatisfied employees.

4.1 Text preparation for analysis

We prepared the text for the analysis following the steps described in previous literature (Stamolampros, Korfiatis, Kourouthanassis & Symitsi, 2019; Tirunillai & Tellis, 2014). These include: *(i)* word text tokenization, *(ii)* removal of numbers and punctuation marks, *(iii)* removal of English stop words (using the SMART stop-word list), and (iv) removal of context-specific stop words (e.g., company names, job roles). We then used Part-Of-Speech (POS) tagging to extract adjectives, adverbs and nouns from the tokenized text as these parts of speech contain the highest level of information entropy in the baseline document-term matrix, due to the origin of English from Indo-European languages (Baeza-Yates & Ribeiro-Neto, 1999). The remaining words were lemmatized to form groups of words with the same root using the Stanford NLP (Natural Language Processing) parser. Finally, we performed a frequency filtering of the terms to maintain those that appear in at least 1% of the total reviews in the initial corpus.

4.2 Estimating the topic solution

The topic solution was estimated using the STM package in R (Roberts, Stewart, and Tingley, 2017). The number of topics was calculated through an iterative process based on three criteria: (i) the held-out likelihood, (ii) the semantic coherence of the topic structure and (iii) the exclusivity of topic words to the topic. Semantic coherence is a criterion developed by Mimno, Wallach, Talley, Leenders & McCallum (2011) which increases with the frequency of co-occurrence of the most probable words in each topic of the estimated solution. Exclusivity considers the mutual appearance of the most probable words in more than one topic and evaluates the overall topic quality for each candidate model. We employed the spectral decomposition algorithm of Lee and Mimno (2014) in order to evaluate the range of the topic solutions in each of the corpora and constructed a seed vector of the candidate number of topics (K). This ranged from $K_{min}=6$ topics to K_{max}=14 topics with a two-step increment. The intuition behind setting the minimum seed value is associated with the number of *Glassdoor*'s individual rating aspects. STM considers the assignment of an employee review to a finite set of topics by estimating a review-topic proportion and review-word distribution against a vector of covariates. Considering the nature of the problem and the context of the employee review data from Glassdoor, we evaluated both distributions against the overall rating, whether the employee is a current or former employee and also the sector of the company. A detailed description of the STM process is provided in Appendix A.

The optimal number of topics for our topic solution, *K*, was estimated based on the highest held-out likelihood against the ratio of their semantic coherence and exclusivity, which describe the topic-word distribution. These criteria (see Appendix B) indicate that a *K*=12 describes better the variability of each corpus subject to the overall rating and employee status (former or current). For the identification of the most important words in a given topic, Roberts et al. (2016), proposed the FREX criterion which combines these measures using a weighted harmonic mean of a word's rank in terms of exclusivity and frequency in a *k*-topic solution:

$$FREX_{k,w} = \left(\frac{\omega}{ECDF\left(\beta_{k,w}/\sum_{j=1}^{k}\beta_{j,w}\right)} + \frac{1-\omega}{ECDF\left(\beta_{k,w}\right)}\right)^{-1}, \tag{2}$$

where $k \in K$ is the k^{th} topic, *w* is the word under consideration, β is the topic-word distribution, and ω is a prior set equal to 0.5 that imposes equal weight on the influence of exclusivity and frequency. The top loading reviews along with representative reviews from the topic solution for each corpus were estimated. In addition to the topic-word and topic-document distributions, we computed the proportion of each topic in the overall corpus. Then, topic labels were assigned using input from experts in Human Resource Management through a manual labeling task. In particular, for each topic, a selection of the top 10 loading reviews (based on maximum values of the θ loadings) was provided along with the top loading words. The experts had to mutually agree a label comprising of up to two words and assign it to the given topic. The process was replicated in the same way for both positive and negative feedback corpora found on Glassdoor employee online reviews. Table 2 presents the topic solutions and the corresponding topic labels in Panel A and B, respectively.[2]

There are particular topic solutions such as *Compensation/Benefits*, *Company Reputation*, *Career Opportunities*, *Task Variety* and *Management* that dominate in the positive feedback corpora, while issues raised by employees in the negative feedback narratives are mostly concentrated on *Management*, *Office/Premises*, *Career Opportunities*, *Job Roles*, and *Compensation* topics. As expected, the individual rating aspects are indeed significant determinants of employee satisfaction/dissatisfaction. However, through analyzing the content of review feedback, we uncover additional dimensions that are not captured through the pre-selected criteria.

Table 2: *Topic Solution for Positive/Negative Feedback*

Topic Label	Prop. (%)	Top 7 FREX Words
Panel A: Positive Feedback		
Compensation/Benefits	13.25	pay, salary, benefit, good, pension, scheme, decent
Company Reputation	12.46	great, place, product, smart, really, brand, amazing
Career Progression	12.04	career, opportunity, progression, development, high, excellent, market
Task Variety	11.94	year, best, new, thing, way, just, better
Management	8.74	management, team, senior, service, member, support, manager
Work Environment	8.56	friendly, environment, atmosphere, fun, colleague, helpful, relaxed
Employee Perks	7.39	free, discount, food, staff, store, nice, pro
Work Life Balance	6.26	life, balance, work, interesting, project, hard, variety
Office Location	5.99	office, location, people, london, event, social, area
Working Hours	5.74	hour, job, easy, time, working, student, home
Flexibility	4.62	flexible, long, available, day, flexibility, plenty, different
Company Culture	3.01	culture, strong, value, worklife, employee, leadership, focus
Panel B: Negative Feedback		
No Negatives	15.40	really, con, people, place, great, good, many
Management/Leadership	13.70	poor, senior, leadership, culture, management, direction, employee
Office/Premises	9.78	office, bit, big, quite, head, need, small
Career Progression	8.96	progression, career, opportunity, development, little, process, limited
Job Role	8.04	life, lot, sometimes, difficult, project, change, fast
Benefits	7.31	salary, low, pay, benefit, market, bonus, industry
Compensation	7.17	month, never, money, even, wage, minimum, ever

[2] This analysis is based on the "positive" and "negative" review text. This information explicitly identifies positive and negative factors that affect employees' experience with an employer. We do not use information contained to "feedback to management" column as its content is limited and reflects actions that should be taken by the management team upon the comments mentioned in "positive" or "negative" columns."

Working Hours	6.66	long, shift, busy, late, break, holiday, enough
Staff pressure	6.06	high, staff, turnover, member, pressure, target, support
Recruitment	5.88	hour, short, time, term, period, recruitment, full
Work/life Balance	5.56	work, hard, balance, much, home, amount, quality
Customer Facing	5.46	customer, store, sale, service, rude, shop, colleague

Note: This table presents in Panel A and Panel B the topic solutions from Structural Topic Model (STM) using the positive and negative feedback corpora from Glassdoor online reviews, respectively. The second column presents the proportion of reviews that are assigned to each topic solution. The last column presents the top 7 FREX words, i.e., the most probable words in each topic.

5. Predicting the determinants of employee satisfaction

5.1 Evidence from numerical and textual features

In this section, we investigate what drives employee satisfaction. Our basic model regresses the overall satisfaction rating on the rating of five job characteristics namely career opportunities, compensation and benefits, senior leadership, work/life balance and cultural values, in all samples of reviews (Model 1). Since we are interested in investigating the informational benefits of structured and unstructured data and in order to allow comparability, we accommodate five additional models in a complete sample across all the variables used in this analysis, i.e., ratings, topic loadings and control variables. More specifically, model 2 replicates the regression with only the numerical ratings for the complete sample. Model 3 controls for employee and employer characteristics. Model 4-6 augment Model 3 by adding positive, negative and both positive and negative topic loadings, respectively. With the dependent variable being ordinal, we perform ordered logistic regressions in line with previous research as follows:

$$\Pr(S_{ij} = \lambda) = \Pr(\mu_{\lambda-1} < S_{ij}^{*} \leq \mu_{\lambda})$$

$$S_{ij}^{*} = \beta X_{ij} + \gamma Z_i + \delta W_j + \epsilon_{ij,} \qquad (1)$$

where S_{ij}^{*} is the latent variable of reviewer's i evaluation for firm j, S_{ij} is the observed rating score with $\lambda \in [1,5]$, μ_2 to μ_5 are the cutoffs, X_{ij} is a vector with the rating scores for the five job-specific rating aspects, Z_i is a vector of reviewer demographics, and W_j is a vector with firm-specific variables as presented in the previous section. ϵ_{ij} is the error term, assumed to be independent and identically distributed with the logistic distribution. Table 3 presents the results of the regression analysis. To conserve space, we suppress coefficients for topics and industry controls.[3] All the individual rating dimensions are statistically significant determinants of the overall job satisfaction. Surprisingly, *Compensation and Benefits* is one of the least influential factors compared to the other job characteristics with coefficients ranging from 0.326 to 0.433. We find that *Culture and Values* and *Senior Leadership* matter most for employee satisfaction. All the topics are also statistically significant with the positive (PROS) topics (coefficients range from 3.377 to 3.734) having a stronger effect on employee the overall rating than the negative (CONS) topics (coefficients range from 0.547 to 0.900). *Employee Perks, Career Progression* and *Working hours* are the most influential positive topics and *Management/Leadership, Recruitment,* and *Job Role* are the most influential negative topics. Reviewer characteristics, such as the level of education and the age, are systematically associated with job satisfaction. In particular, we document that the *Overall Rating* reduces significantly with the level of education but increases with the age of the reviewer. The results remain qualitatively similar when we employ subsamples of current and former employees. From ANOVA analysis we find a statistically significant improvement in Model 3 when this is incremented with textual features indicating that the accommodation of both structured and unstructured data offers benefits. This is also illustrated by an increase in R^2 (Mc Fadden) by 0.31.

Table 3: *Results of the Ordered Logistic Regressions: Factors affecting the Overall Rating*

	(1)	(2)	(3)	(4)	(5)	(6)
Rating Dimensions						
Career Opportunities	0.750*** (0.005)	0.726*** (0.010)	0.735*** (0.010)	0.392*** (0.013)	0.587*** (0.011)	0.344*** (0.015)
Compensation and Benefits	0.433*** (0.004)	0.417*** (0.009)	0.420*** (0.009)	0.407*** (0.012)	0.326*** (0.010)	0.342*** (0.014)
Senior Leadership	0.873***	0.808***	0.811***	0.465***	0.483***	0.278***

[3] The full output of the regression analysis is available upon request by the authors.

	(0.005)	(0.011)	(0.011)	(0.014)	(0.012)	(0.016)
Work/life Balance	0.397***	0.376***	0.385***	0.340***	0.378***	0.412***
	(0.004)	(0.008)	(0.009)	(0.012)	(0.010)	(0.013)
Culture and Values	0.827***	0.797***	0.794***	0.419***	0.604***	0.378***
	(0.005)	(0.011)	(0.011)	(0.014)	(0.012)	(0.015)
Employee Controls						
Gender: Female			-0.007	0.096***	-0.007	0.073***
			(0.018)	(0.024)	(0.020)	(0.026)
Age			0.003***	-0.001	0.001	-0.002
			(0.001)	(0.001)	(0.001)	(0.001)
Education: University graduate			-0.090***	-0.072*	-0.145***	-0.103**
			(0.029)	(0.038)	(0.032)	(0.042)
Education: University postgraduate			-0.111***	-0.176***	-0.250***	-0.245***
			(0.033)	(0.044)	(0.037)	(0.049)
Employer Controls						
Public			-0.015	-0.048	-0.027	-0.046
			(0.022)	(0.029)	(0.025)	(0.032)
Log (Employees)			-0.005	0.008	-0.007	0.009
			(0.006)	(0.008)	(0.007)	(0.009)
Log (Revenues)			-0.007	-0.021***	-0.015***	-0.026***
			(0.005)	(0.006)	(0.005)	(0.007)
Supressed Coefficients						
Industry	Yes	Yes	Yes	Yes	Yes	Yes
Textual: PROS	No	No	No	Yes	No	Yes
Textual: CONS	No	No	No	No	Yes	Yes
Pseudo R^2	0.57	0.45	0.46	0.72	0.59	0.77
ANOVA: Pr(Chi) vs Model 3	-	-	-	0	0	0
Log Lik.	-228,642	-50,370	-50,244	-25,673	-37,669	-21,071
AIC	457,302	100,758	100,568	51,448	75,441	42,266
Observations	292,370	58,823	58,823	58,823	58,823	58,823

Note: Model 1 employs five individual aspects of employee satisfaction as captured in the full dataset of Glassdoor. Models 2-6 employ a complete subsample across all the tested variables (numerical and textual features) and the control variables for employee and employer characteristics. ANOVA analysis compares the performance of Model 3 (structured data) vs. Models 4-6 that are incremented with textual features (unstructured data). $^{*}p<0.10$; $^{**}p<0.05$; $^{***}p<0.01$.

5.2 The effect of employee satisfaction on topic prevalence

This section presents further insights on how critical factors determined by the structural textual analysis of positive and negative narratives vary among more satisfied and dissatisfied employees by accommodating additional covariates. A particular topic may dominate in each review corpus. However, it is possible that employees consider two or more topics when providing feedback. The structural topic methodology estimates the marginal effects on the topic distribution in the continuum between low (dissatisfied) and high (satisfied) overall ratings.

Figure 1 illustrates these effects. The dotted vertical line represents no effect, with the topics on the right-hand side being discussed more when the overall rating satisfaction increases and the topics on the left-hand side being discussed more by dissatisfied employees. The horizontal axis indicates the marginal effect on the topics of a unit increase in overall satisfaction. For instance, a unit increase in the overall rating signifies a decrease of almost 2% on the reviews that discuss mainly *Compensation* issues (positive feedback). For satisfied employees, the dominance of the topics related to *Career Progression*, *Management*, *Culture*, *and Working Environment* increases, while for less satisfied employees, topics associated with *Employee Perks* and *Compensation* tend to become more prevailing when the positive feedback topics are considered.

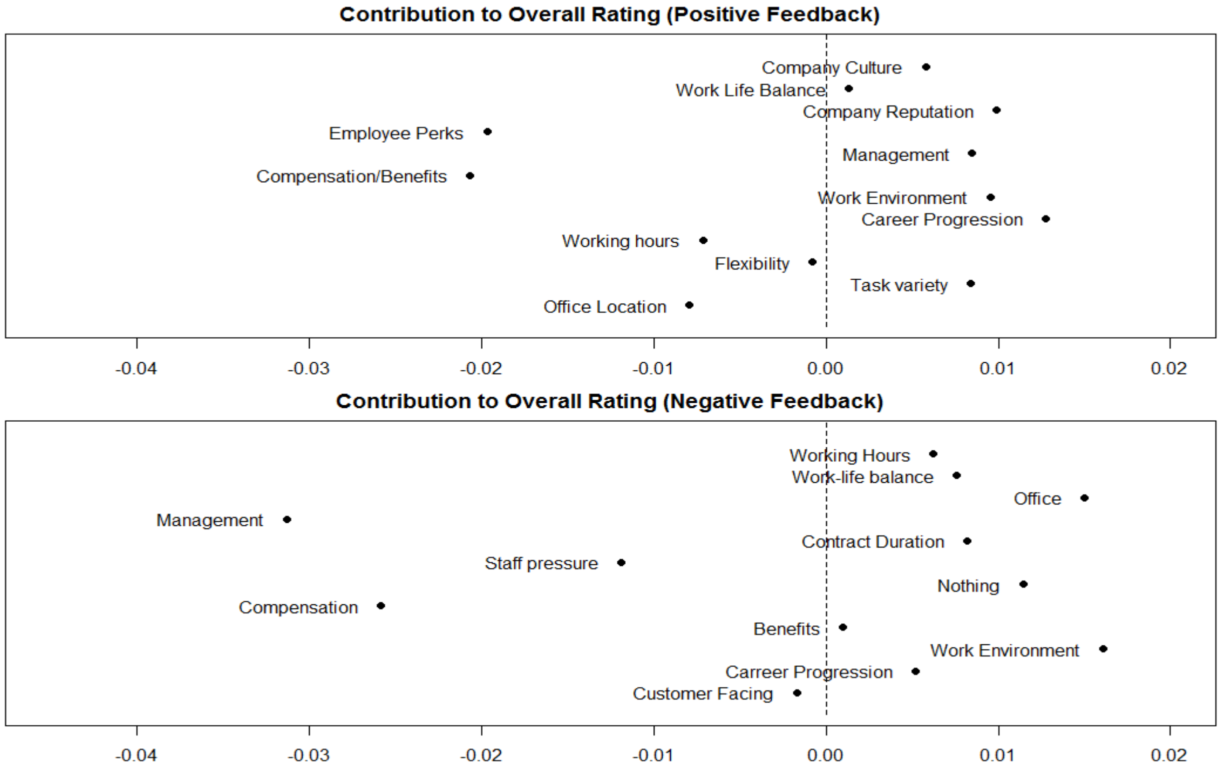


Figure 1*: Marginal effects of overall rating (low to high) for the topic distribution of positive (upper) and negative (lower) aspects of the review text. The dotted line represents no effect.*

6. Predicting employee turnover

Job satisfaction is a critical factor in recruiting and maintaining workforce and employee dissatisfaction has been greatly associated with employees' voluntary quitting (Hom & Kinicki, 2001). However, accessing former employees is a great challenge for researchers that often constrains studies in investigating behaviors, such as the intention to quit a company than actually quitting. Previous studies have shown that actions and intentions of quitting are different concepts and are thus, predicted by different sets of variables (Cohen, Blake & Goodman, 2016).

Other limitations of prior research derive from the use of survey data (Mitchell, Holtom, Lee, Sablynski & Erez, 2001; Steel, 2002). As a result, the empirical findings of several studies are based on a limited number of questionnaires or concern particular job roles and/or industries. Moreover, even the most exhaustive turnover models fail to incorporate important constructs and unveil latent variables with the explained variation in employee turnover being low (Maertz Jr & Griffeth, 2004). This study overcomes these limitations and offers a new framework to investigate employee turnover determinants utilizing a massive dataset of opinions expressed by former employees who have actually departed from a post (action of quitting) than intending to depart (scenario). Our purpose is to investigate not only particular constructs derived from *Glassdoor* platform in predicting employee turnover, but also the informational value from extracting topics through textual analysis providing new academic and managerial insights.

6.1 Contrasting the information value of structured and unstructured job satisfaction data

6.1.1 Structured job satisfaction determinants of employee turnover

Having considered the drivers of employee ratings in both numerical and textual aspects, we first evaluated their usefulness in predicting the likelihood of employee turnover considering the current/former status supplied by each employee at the time of the review. In so doing, we demonstrate the informational value deriving from online employee reviews from (i) structured data – comprising the overall rating and the rating aspects in numerical scales (Models 1-2) and (ii) unstructured data – comprising the textual content of positive and negative aspects (Models 3-5). To examine the association between various satisfaction aspects and employee turnover, we employed a logit generalized linear model specified as follows:

$$P(IsFormer = 1|\, x_{ij}) = \frac{\exp\,(a + \ \beta X_{ij}\ + \gamma Z_i + \delta W_j)}{1 + \ \exp\,(a + \ \beta X_{ij}\ + \gamma Z_i + \delta W_j)}, \qquad (3)$$

where the dependent variable is 1 when an employee *i* has already left the company *j* at the time of the review and 0 otherwise, X_{ij} is a vector including the scores employee *i* from company *j* provides for one or more of satisfaction aspects including the overall score, Z_i is a vector of reviewer demographics, and W_j is a vector with firm-specific variables. The estimated response probabilities are strictly between zero and one. The logit model assumes a non-linear association between the probability of an output and the covariates. Table 4 reports the estimation output of logit regressions.

Overall, the job satisfaction variables are statistically significant and carry the expected sign. The higher the satisfaction, the less likely the employee is to have left the company. Converting the coefficients to odds ratio estimates by e^{β} makes the coefficients easier to interpret and gives us an indication of which variables have the largest effect on the probability of an employee to have left the company. The percent change in the odds ratio for a unit increase in the overall rating is estimated as $100 \times (e^{\beta} - 1)$. Thus, an increase in the overall rating leads to a 31.82% decrease in the odds ratio indicating that job satisfaction is also an economically significant driver of the probability of an employee to have left the company (Model 1). Among the various job satisfaction aspects, a unit increase in each aspect leads to a decrease in the odds ratio from 22.66% to 29.39%.[4] ANOVA analysis finds that adding individual ratings improves significantly the information content of the model (Model 2 vs. Model 1). Importantly when we include textual loadings, we find that all topics significantly improve the fit of the models increasing the R^2 by 0.22 for PROS, 0.18 for CONS, and 0.31 altogether. Moreover, as expected the turnover odds ratio decreases more for positive than negative topics. The odds of turnover decrease by 0.49 times for employees in reputable companies and by 0.47 times for companies that offer flexibility, while the odds of turnover are higher when there is staff pressure. Other important findings exhibit that gender and age are the most economically significant determinants of turnover after job satisfaction. For female employees the odds of turnover increase by 1.13 times versus male employees. For more educated staff the odds of turnover increase by 1.17 times for those holding an undergraduate degree (Bachelor's) and a postgraduate degree compared to high-school diploma holders.

Table 4: *Logit model of the likelihood of an employee to have left a company*

	(1)	(2)	(3)	(4)	(5)
Rating Dimensions					
OverallRating	-0.383*** (0.007)				
Career Opportunities		-0.185*** (0.009)	0.039*** (0.011)	-0.084*** (0.011)	0.035*** (0.012)
Compensation and Benefits		0.022** (0.009)	0.120*** (0.01)	0.042*** (0.01)	0.095*** (0.011)
Senior Leadership		-0.106*** (0.01)	0.085*** (0.012)	0.067*** (0.012)	0.113*** (0.013)
Work/life Balance		-0.047*** (0.008)	0.063*** (0.01)	0.062*** (0.009)	0.132*** (0.01)
Culture and Values		-0.115*** (0.01)	0.072*** (0.012)	0.048*** (0.011)	0.103*** (0.013)
Employee Controls					
Gender: Female	0.112*** (0.018)	0.114*** (0.018)	0.063*** (0.02)	0.103*** (0.019)	0.062*** (0.021)
Age	-0.005*** (0.001)	-0.006*** (0.001)	-0.005*** (0.001)	-0.004*** (0.001)	-0.003*** (0.001)
Education: University graduate	0.149*** (0.028)	0.159*** (0.028)	0.187*** (0.031)	0.182*** (0.031)	0.197*** (0.033)
	0.140***	0.151***	0.184***	0.249***	0.248***

[4] Due to correlated covariates in Model 2, we report the coefficient range from models that run each numerical rating aspect individually.

Education: University postgraduate	(0.033)	(0.033)	(0.036)	(0.035)	(0.038)
Employer Controls					
Public	0.027 (0.022)	0.023 (0.022)	0.019 (0.024)	0.028 (0.024)	0.019 (0.026)
Log (Employees)	-0.035*** (0.006)	-0.035*** (0.006)	-0.022*** (0.007)	-0.052*** (0.006)	-0.037*** (0.007)
Log (Revenues)	-0.001 (0.005)	-0.001 (0.005)	0.005 (0.005)	0.007 (0.005)	0.012** (0.006)
Suppressed Coefficients					
Industry	Yes	Yes	Yes	Yes	Yes
Textual: PROS	No	No	No	Yes	No
Textual: CONS	No	No	No	No	Yes
Pseudo R^2	0.09	0.1	0.31	0.27	0.4
ANOVA: Pr(Chi) vs Model 1	-	0	0	0	0
Log Lik.	-38,559	-38,455	-32,868	-33,986	-30,189
AIC	77,184	76,984	65,831	68,067	60,496
Observations	58,823	58,823	58,823	58,823	58,823

Note: Models (1)-(5) present the logit coefficients along with standard errors in parentheses of the probability of an employee that reviews a company to be former and how this is associated with various numerical and textual job satisfaction factors, controlling for reviewer and firm characteristics. $^{*}p<0.10$; $^{**}p<0.05$; $^{***}p<0.01$

6.1.2 The information gain of unstructured job satisfaction data in employee turnover

In this section, we investigate whether the information contained in employee feedback narratives can predict employee turnover. This problem is of high economic significance from an operational planners' perspective as it provides an indication of capacity constraints that may arise from the departure of employees and, in turn, may result in high recruitment and training costs for the replacement staff.

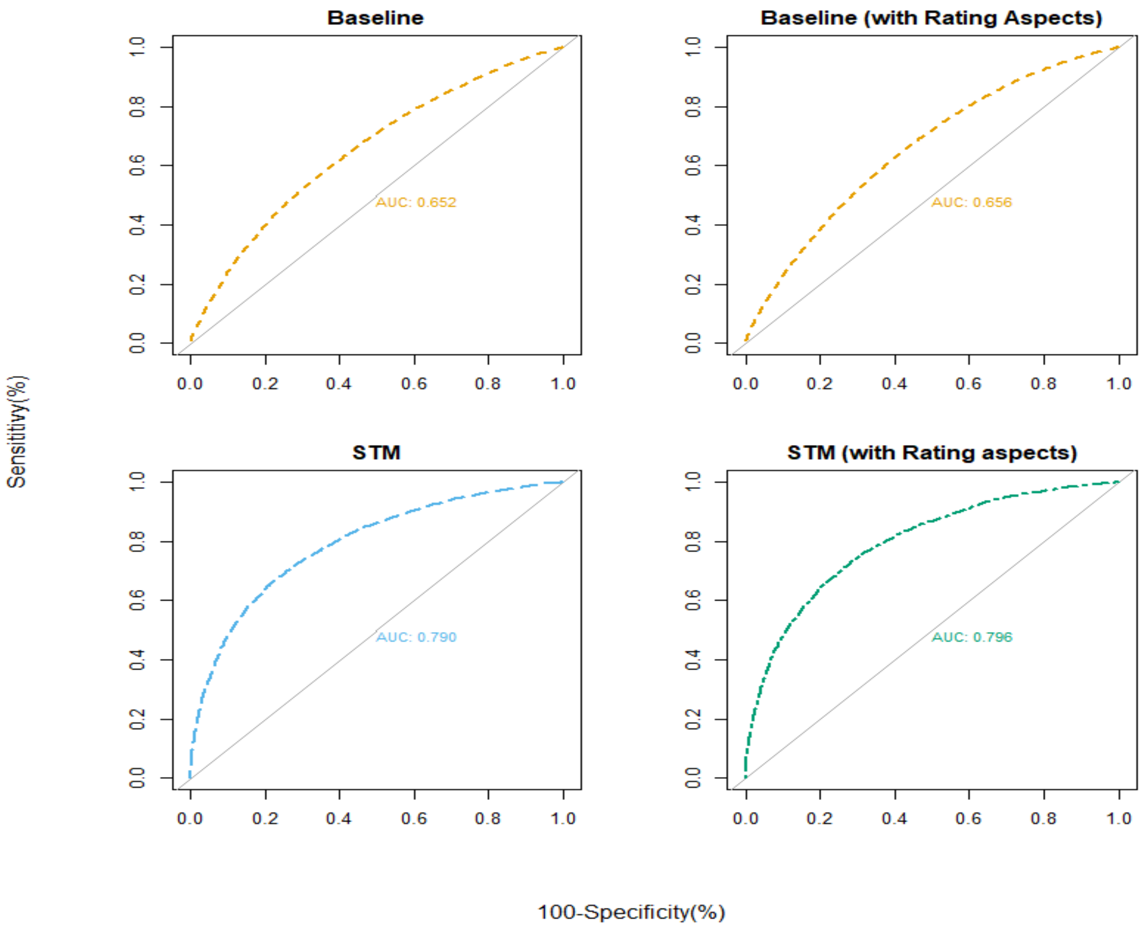


Figure 2: *ROC curves for comparing the accuracy of the different specifications of the classification task (See Table 5). The 45-degree line represents the classification threshold.*

From a measurement viewpoint, this analysis represents a classification task that assigns the probability of an employee to be current or former based on the numerical ratings and the narrative of the positive and negative aspects. Under the assumption that former and current employees are similar the factors that explain the departure of former employees can be used also to predict the likelihood of current employees to leave the company. While various models and methods have been proposed in the literature in both parametric and non-parametric variants, following the same line of argumentation as in Bertels, Jacques, Neuberg and Gatot (1999), we employed Linear Discriminant Analysis as a method of choice for this classification task when including the topic parameters. Hence, topic loadings for the positive and negative text are considered qualitative criteria that can be linearly combined to capture multiple dimensions of employee satisfaction; dimensions that are not necessarily reflected on numerical ratings. We used an 80/20 train and test split ratio and evaluated the predictive accuracy of the numerical and textual features using maximum likelihood for each model in terms of sensitivity, specificity, and the area under the curve (AUC). Table 5 reports these findings and Figure 2 provides the ROC curves for the reported models.

Using the textual features extracted from the topic models for positive and negative topics, the accuracy of the classification model increases by 21.2% compared to a baseline model (overall rating plus controls, i.e., gender, age, education, industry, private/public, number of employees, annual revenues). This result clearly shows that the qualitative aspects contained in the narrative of an employees' experience increase significantly the accuracy of the model. On the other hand, the use of individual rating aspects in comparison with the overall rating does not improve the classification accuracy significantly.

Table 5: *Comparison of employee turnover using numerical and textual features.*

Model	Sensitivity	Specificity	AUC (%)	% Improvement
Baseline (Only Rating)	-	-	65.2	
Baseline (With Rating aspects)	0.51	0.70	65.6	0.6%
STM (topic loadings)	0.67	0.74	79.0	21.2%
STM (topic loadings) and overall rating	0.68	0.76	79.6	22.1%

Note: The baseline model examines the overall rating controlling for gender, age, education, industry, private or public, number of employees, annual revenues. Number of Observations: 58,823.

7. Predicting the economic value of employee satisfaction

7.1 Employee ratings and firm performance

In previous sections we examined structured and unstructured data in predicting employee satisfaction and turnover. In this section, the economic value of this data is presented by investigating the informational content in predicting corporate performance. An increasing number of scholars argue that the traditional view of the firm needs to give its place to a human-centered one due to the changing nature of modern western economies.[5] As eloquently put by Zingales (2000): *'Employees are not merely automata in charge of operating valuable assets but valuable assets themselves, operating with commodity-like physical assets (p.1641).'* Extant scholarly findings reflect this thought and substantiate a positive relationship among employee satisfaction, customer satisfaction and corporate performance (Bernhardt et al., 2000; Edmans, 2011; Huang et al., 2015; Symitsi et al., 2018). The service-profit chain model (Heskett, Jones, Loverman, Sasser & Schlesinger, 1994) describes the mechanism that governs this relationship where more satisfied employees offer a better value of service to customers leading to higher customer satisfaction and loyalty that further stimulates firm profitability and growth.

To empirically extend this literature, by investigating the information content of structured and unstructured data in predicting financial performance, we focused on the overall satisfaction rating reported on a 5-point Likert scale by current employees. This is to ensure that opinions from disgruntled former employees do not skew our results. Our sample includes employees from both private and public companies allowing a more accurate generalization of our findings. We gather annual financial data from the Bureau van Dijk's FAME database for all firms with at least 100 reviews. In particular, we collect data on profitability (return on assets - ROA), leverage, total assets, sales growth, capital intensity and expenditure, R&D expenses, number of employees and year of incorporation. The final sample includes 35,231

[5] The traditional view posits that investments in employees should not be undertaken as employees are considered to perform unskilled tasks and, therefore, management should focus on extracting the maximum output at the least possible cost. With maximizing the shareholder wealth as the central goal of firms, according to standard corporate finance theory, managers proceeding to employee satisfaction investments do not perform their duties appropriately as they transfer value from shareholders to employees. In contrast, the human-centred view of the firm stipulates that due to the service and knowledge-based nature of modern developed economies, firms need creative employees that have the skills to innovate and translate their innovations into services and products. Thus, to create value for their companies and shareholders managers need to invest in their employees in order to attract and retain the required creative and highly skilled workforce. Relevant literature, therefore, focuses on understanding whether firms with high employee satisfaction create more value than those with low levels of employee satisfaction.

reviews for 161 firms (55 public with primary listing in the UK, 41 public but not listed in the UK and 65 private) for the period spanning from 2014 to 2017, yielding 623 firm-year observations. Table 6 presents the description of variables and key statistics.

Table 6: *Descriptive statistics of the firm-year variables.*

Variable	Obs.	Mean	St.Dev.	Skew.	Kurt.
Overall Employee Rating	623	3.19	0.62	-0.15	0.41
Career Opportunities Rating	623	3.05	0.56	0.04	0.59
Compensation and Benefits Rating	623	3.07	0.65	-0.16	0.26
Senior Leadership Rating	623	2.81	0.61	0.26	0.96
Work/life Balance Rating	623	3.13	0.62	-0.25	0.23
Culture and Values Rating	623	3.17	0.68	-0.15	0.11
ROA	517	8.53	11.74	-0.79	16.60
Leverage	519	0.44	0.29	5.43	58.19
Sales Growth	501	128.39	2871.82	22.25	494.01
Total Assets ('000000s)	519	8.62	28.31	6.87	56.24
Capital Intensity	515	11.15	224.95	22.56	507.96
Capital Expenditure Ratio	623	-0.01	0.06	-13.09	187.71
R&D dummy	623	0.11	0.31	2.47	4.13
R&D index	623	0.00	0.03	9.34	110.06
Employees ('000s)	508	33.96	79.80	5.26	30.54
Firm Age	623	30.33	25.72	1.73	3.00

Note: Obs. represents firm-year observations.

We examine the relationship between employee satisfaction and corporate performance using the following baseline regression model (Model 1):

$$ROA_{jt} = \alpha + \beta Employee\ rating_{jt-1} + \gamma' x_{jt-1} + \varepsilon_{jt}\ , \tag{4}$$

where j and t correspond to firm and year, respectively. $Employee\ rating_{jt-1}$ is the annual employee satisfaction rating computed by averaging all available ratings for each firm each year, while the vector x_{jt-1} contains firm-specific characteristics. We also control for time fixed-effects (Model 2), time and industry fixed-effects (Model 3), industry/time fixed-effects and lagged profitability (Model 4). ε_{jt} is the error term clustered at firm level which assumes that standard errors are robust to heteroskedasticity and within-firm serial correlation. The results presented in Table 7 indicate a statistically significant positive relationship between employee satisfaction rating and firm profitability. More specifically, our analysis shows that UK firms rated highly by their current employees in terms of satisfaction, achieve superior profitability (ROA) compared to those rated poorly. We document that a one unit increase in the overall employee satisfaction rating increases the percentage of profit a company earns in relation to its assets from 1.668% to 2.239%. The adjusted R^2 ranges from 19% to 20.1% indicating that the augmented model with industry and time fixed effects has a better fit (Model 3 vs Model 1 and Model 2). Not surprisingly the adjusted R^2 doubles when a lagged dependent variable is added in the model. These findings, which are robust to controlling for firm characteristics, industry and time fixed-effects, suggest that the human-centred view of the firm is confirmed not only in the US, but also in the UK. Importantly, the significant positive relationship between employee satisfaction rating and profitability indicates that online employee reviews can be used to forecast the financial results of UK firms. This highlights the value-relevance of online employee reviews for UK investors supporting the view that non-financial indicators are of key importance for security valuation as they address the inability of standard accounting measures to capture investments in intangibles (Amir & Lev, 1996).

To increase the robustness of our results, in Table 7 we also present two additional models. Model 5 accounts for heterogeneity across firms by adding firm fixed effects. This model yields the weakest coefficient for the tested variable, though, the p-value is marginally insignificant (0.13). The insignificant coefficient though may be explained by the significant reduction in the degrees of freedom by adding firm fixed effects in combination with the small number of years in our panel data set. Model 6 accounts for endogeneity concerns by employing a dynamic GMM estimation method in line with Huang et al. (2015) who mitigate such issues by using lagged instruments at the context of family firms. This approach is widely applied in the empirical literature of corporate finance, when a model controls for several covariates that may be endogenously associated with each other (Arellano & Bond, 1991; Flannery & Hankins, 2013; Wintoki, Linck & Netter, 2012). The coefficient for average employee rating is statistically significant and positive. The p-value of the Arellano-Bond AR(2) test is 0.522, implying that we cannot reject the null hypothesis of no second-order serial correlation of the first-differenced errors and justifying the lagged levels of the dependent variable as instruments for the first-differenced model. The Hansen test reports p-value of 0.857, thus, the difference-in-Hansen test, yielding a p-value of 0.781, is interpreted as a test for the validity and exogeneity of the instruments. Hence, we cannot reject the null hypothesis that our lagged instruments are valid.

Table 7: *Employee satisfaction and firm profitability*

Dependent Variable: ROA	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) FE	(6) D-GMM
Overall Employee	2.298**	2.339**	2.049**	1.638*	1.388	1.644*
Rating	(0.916)	(0.906)	(0.942)	(0.864)	(0.921)	(0.968)
Leverage	-6.926***	-6.955***	-6.373**	-1.511***	-7.801	-4.532
	(2.428)	(2.431)	(2.546)	(2.622)	(6.195)	(3.848)
log(Total assets)	-1.938***	-1.949***	-1.841***	-0.883**	0.171	-1.395
	(0.408)	(0.411)	(0.491)	(0.421)	(4.038)	(1.332)
Sales growth	-0.000***	-0.000***	-0.000***	-0.000*	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Capital intensity	-1.134**	-1.128**	-1.017*	-0.287	-7.907*	-1.127
	(0.523)	(0.525)	(0.524)	(0.535)	(4.334)	(0.944)
Capital expenditure	-2.498	-3.035	-1.692	0.650	-8.406**	-4.323
Ratio	(4.103)	(4.212)	(3.929)	(3.001)	(2.685)	(3.796)
R&D intensity	2.100	1.833	2.250	-0.698	-57.903**	-59.639*
	(14.510)	(14.600)	(13.852)	(9.805)	(24.985)	(32.933)
R&D dummy	0.124	0.090	0.106	0.055	3.515*	1.853
	(1.456)	(1.456)	(1.170)	(0.895)	(1.977)	(1.987)
log(Employees)	-1.206**	-1.193**	-1.473**	-1.205***	-5.263	-1.679
	(0.477)	(0.478)	(0.572)	(0.446)	(3.500)	(1.989)
log(Firm age)	-0.992	-0.974	-1.005	-0.733	3.505	-0.305
	(0.714)	(0.724)	(0.733)	(0.600)	(8.776)	(1.036)
Public(UK listed)	4.794***	4.883***	4.974***	2.804**		5.633**
	(1.263)	(1.284)	(1.358)	(1.240)		(2.481)
Public(Private)	-0.348	-0.288	-0.970	-0.320		0.429
	(1.301)	(1.315)	(1.458)	(1.293)		(1.992)
ROA_{t-1}				4.974***	-0.057	0.076
				(0.104)	(0.162)	(0.137)
Constant	45.288***	45.343***	48.163***	25.185***		41.147***
	(5.610)	(5.506)	(6.824)	(6.856)		(12.466)

Observations	481	481	481	481	481	481
Industry fixed-effects	No	No	Yes	Yes	No	Yes
Time fixed-effects	No	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.195	0.190	0.201	0.408	0.185	-
AR(2) test p-value	-	-	-	-	-	0.522
Hansen test p-value	-	-	-	-	-	0.857
Diff-in-Hansen test p-value	-	-	-	-	-	0.781

Note: Model 1 performs OLS regression of firm profitability (ROA) on the overall employee rating and control variables. Models 2, 3, and 4 control for time fixed effects, time and industry fixed effects, and time/industry fixed effects and lagged ROA, respectively. Model 5 performs a panel data firm fixed effect regression and Model 6 performs a dynamic GMM regression including lagged ROA GMM-type instrumental variables. Robust standard errors are reported in parentheses clustered at firm level. $^{*}p<0.10$; $^{**}p<0.05$; $^{***}p<0.01$.

7.1.1 The information gain from structured and unstructured employee data

We investigate the information content of online employee information in predicting firm performance measured with ROA. In particular, we study both structured and unstructured data and compare the informational value added in the model compared to a baseline model (Model 0) that does not include any information from employees.[6] In so doing, we present in Table 8 the changes in the adjusted R^2 versus the baseline model as well as the p-values of an ANOVA analysis.[7] In addition to the average overall employee satisfaction rating (Model 1), the structured data analysis examines individual aspects of employee satisfaction, namely career opportunities (Model 2), compensation and benefits (Model 3), senior leadership (Model 4), work/life balance (Model 5), and culture values ratings (Model 6). Then, Model 7 accommodates all the individual aspects simultaneously.

In line with the results of Table 7, Table 8 shows the benefit of the basic model when this is augmented by the overall employee satisfaction rating. We also find that the information content of additional employee satisfaction aspects can offer further advantages. All but the compensation and benefits category increase the adjusted R^2 from 0.4% to 1.4% and improve

[6] The baseline model (0) is described by: $ROA_{it} = \alpha + \gamma' x_{it-1} + \varepsilon_{it}$. In addition to this specification (1) and in line with the previous analysis, specification (2) adds time fixed effects, specification (3) adds time and industry fixed effects and specification (4) adds time and industry fixed effects and the lagged dependent.

[7] To conserve space, we do not tabulate the results (coefficients from each model along with standard errors), but all outputs are available upon request by the authors.

significantly the fits of the model. In line with the previous analysis, the effects of the individual aspect rating vary, indicating that they are not of equal importance for employees.

Table 8: *Structured and unstructured data and firm profitability*

Model	Dependent Variable	(1)	(2)	(3)	(4)
1	Overall Employee Rating	0.012	0.012	0.008	0.005
	ANOVA: 1 vs 0 (p-value)	0.005	0.005	0.017	0.027
2	Career Opportunities Rating	0.008	0.008	0.005	0.005
	ANOVA: 2 vs 0 (p-value)	0.024	0.024	0.045	0.034
3	Compensation and Benefits Rating	0.000	0.001	-0.001	0.003
	ANOVA: 3 vs 0 (p-value)	0.316	0.303	0.504	0.074
4	Senior Leadership Rating	0.014	0.014	0.009	0.010
	ANOVA: 4 vs 0 (p-value)	0.003	0.003	0.011	0.003
5	Work/Life Balance Rating	0.010	0.010	0.011	0.010
	ANOVA: 5 vs 0 (p-value)	0.009	0.009	0.007	0.003
6	Culture Values Rating	0.011	0.011	0.006	0.004
	ANOVA: 6 vs 0 (p-value)	0.007	0.007	0.032	0.041
7	All individual aspects	0.013	0.013	0.012	0.009
	ANOVA: 7 vs 0 (p-value)	0.031	0.033	0.038	0.037
	ANOVA: 7 vs 1 (p-value)	0.343	0.371	0.192	0.140
8	Textual: PROS	0.030	0.030	0.030	0.013
	ANOVA: 8 vs 0 (p-value)	0.003	0.003	0.002	0.033
	ANOVA: 8 vs 1 (p-value)	0.024	0.025	0.011	0.000
	ANOVA: 8 vs 7 (p-value)	0.014	0.013	0.010	0.000
9	Textual: CONS	0.004	0.004	0.000	-0.003
	ANOVA: 9 vs 0 (p-value)	0.292	0.285	0.408	0.643
10	Textual: PROS and CONS	0.025	0.025	0.025	0.007
	ANOVA: 10 vs 0 (p-value)	0.026	0.027	0.027	0.192
	ANOVA: 10 vs 7 (p-value)	0.116	0.115	0.106	0.000
	Industry fixed-effects	No	No	Yes	Yes
	Time fixed-effects	No	Yes	Yes	Yes
	Lagged ROA	No	No	No	Yes

Note: This table presents the change in the adjusted R^2 of models that regress return on assets (ROA) on employee satisfaction variables (Models1-10) versus the baseline model (Model 0) controlling for leverage, log(Total assets), sales growth, capital intensity, capital expenditure ratio, R&D index, R&D dummy, log(employees), log(firm age) and a dummy of Public vs Private firms. Specification (1) is an OLS regression with robust standard errors clustered at firm level, specification (2) includes time fixed effects, specification (3) includes both time and industry fixed effects and specification (4) includes time and industry fixed effects and controls for the lagged ROA. Models are also compared via ANOVA analysis where the p-values are provided.

Our findings suggest that senior leadership (coefficients range from 2.194 to 2.491) and work/life balance (coefficients range from 2.283 to 2.433) are the most prevailing factors that could drive firm profitability with changes in adjusted R^2 ranging from 0.9% to 1.4% and significant improvements over the baseline model. While the extended Model 7 generates a better fit than the baseline model, we find that the overall fit does not improve significantly against Model 1, that includes the overall employee satisfaction rating. Our analysis further studies the information gain from extending models with unstructured data employing the STM methodology. More specifically, we study whether the topics extracted from employees' review text featuring the positive (PROS) and negative (CONS) aspects of working with a particular employer can predict firm profitability. Models 8, 9, and 10 of Table 8 augment the baseline Model 0 by the PROS, CONS and both PROS and CONS textual ratings, respectively, for all the topics extracted from the textual analysis in 4.2.2. The results demonstrate that the information extracted by the positive review text offers significant advantages in predicting firm profitability. Importantly, company reputation (coefficients range from 1.064 to 1.074), career progression (coefficients range from 1.132 to 1.143), management (coefficients range from 1.140 to 2.257), work environment (coefficients range from 0.968 to 0.975), work/life balance (coefficients range from 1.094 to 1.095), and office location (coefficients range from 1.838 to 2.342) are the topics that yield a statistically significant effect on predicting ROA. The adjusted R^2 of Model 8 improves Model 0 by 1.3% to 3%. Moreover, Model 8 offers significant benefits compared to Models 0, 1 and 7. On the contrary, CONS do not carry information that is helpful in predicting firm profitability.

7.2 The value relevance of employee online reviews for investors

As satisfied employees lead to higher profitability, this relationship should also be reflected on equity prices. Relevant empirical research takes this stance and investigates if this information is incorporated in stock prices. For example, Filbeck & Preece (2003) examine the shareholder value effects of a firm's inclusion in the *'100 Best Places to Work for in America'* and reveal a significant positive market reaction on the day of the announcement. Edmans (2011) evaluates the performance of a stock portfolio for firms and finds that employee satisfaction positively impacts firm value, though, this is not fully reflected in stock prices.

We assess the economic value of online employee reviews in portfolios that invest in firms with the highest levels of employee satisfaction. This allows us to examine whether this intangible is fully priced in the stock market. Explicitly, we construct an equally-weighted and a value-weighted portfolio by including the stocks of the top 25% of the public firms (with primary listing in the UK) in our sample in terms of the monthly overall employee satisfaction rating, that is, the average of all available reviews for a firm in each month.[8] Constructing portfolios that sort stocks by a particular characteristic (here the overall employee satisfaction) and looking for abnormal alphas at the extremes is a standard practice in the finance literature. We use monthly re-balancing of the portfolios and account for risk by assessing portfolio performance on the basis of the following popular asset pricing models:

Capital Asset Pricing Model (CAPM)

$$R_{pt} = \alpha + \beta_{MKT}\, MKT_t + \varepsilon_{pt} \tag{5}$$

Fama-French's three-factor model (FF3)

$$R_{pt} = \alpha + \beta_{MKT}\, MKT_t + \beta_{HML}\, HML_t + \beta_{SMB}\, SMB_t + \varepsilon_{pt} \tag{6}$$

Carhart's four-factor model (C4)

$$R_{pt} = \alpha + \beta_{MKT}\, MKT_t + \beta_{HML}\, HML_t + \beta_{SMB}\, SMB_t + \beta_{MOM}\, MOM_t + \varepsilon_{pt} \tag{7}$$

Fama-French's five-factor model (FF5)

$$R_{pt} = \alpha + \beta_{MKT}\, MKT_t + \beta_{HML}\, HML_t + \beta_{SMB}\, SMB_t + \beta_{RMW}\, RMW_t + \beta_{CMA}\, CMA_t + \varepsilon_{pt}\,, \tag{8}$$

where R_{pt} is the monthly return on portfolio p in excess of the risk-free rate, obtained from Ibbotson Associates. Daily stock prices for all listed companies in our sample and the FTSE 100 stock market index are taken from Thomson Reuters Datastream. MKT_t, HML_t SMB_t,

[8] For robustness reasons, we also constructed portfolios based on the top 30% and 50% of firms with the most satisfied employees and our results remain qualitatively unchanged.

MOM_t, RMW_t and CMA_t are the excess market returns and the value, size, momentum, profitability and investment pattern of European factors, respectively, taken from Kenneth R French's website.[9] ε_{pt} is the error term which allows standard errors to be heteroskedastic and serially correlated.

The CAPM (Eq. 5) is an equilibrium asset pricing model that relates the expected returns of an asset with the market returns and models how investors make asset allocation decisions based on a series of assumptions such as competitive and frictionless markets, rational and risk-averse investors, investors that have similar expectations in terms of expected returns and return variances of all assets (for a detailed description of the assumptions of asset pricing models see (Fabozzi, Neave, & Zhou, 2011, p.288). The FF3 (Eq. 6), C4 (Eq.7) and FF5 (Eq. 8) identify additional linearly associated variables for returns (risk factors), extending the CAPM in line with the Arbitrage Pricing Theory (APT). It has been shown that these models have better explanatory performance to the CAPM, although in practice, the CAPM is commonly used due to difficulties in estimating the additional factors of the FF3, C4, and FF5 factor models. The inference of the aforementioned models is based on standard OLS regressions with heteroskedasticity and autocorrelation (HAC) robust standard errors reported in parentheses.

The intercept α (alpha) captures the abnormal risk-adjusted return. The null hypothesis tested is that the alphas of these portfolios do not differ from zero, with a positive and statistically significant alpha indicating outperformance. The portfolio analysis results are presented in Table 9 in Panel A. These indicate statistically and economically significant abnormal returns in the case of an equally-weighted (value-weighted) portfolio when using the CAPM, C4 and FF5 (FF3 and C4) to account for risk. In panel B of Table 9, we extend the

[9] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_libraryhtml

analysis to investigate the informational value of various employee satisfaction aspects including career opportunities, compensation and benefits, senior leadership, work/life balance, and culture and values.

Table 9: *Employee satisfaction and portfolio performance*

	Equal-weighted portfolio				Value-weighted portfolio			
	CAPM	FF3	C4	FF5	CAPM	FF3	C4	FF5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Overall Employee Rating								
alpha	0.965***	0.630	1.281**	1.074*	0.141	0.711	0.866*	0.097
	(0.350)	(0.568)	(0.489)	(0.589)	(0.413)	(0.422)	(0.442)	(0.427)
MKT	0.383*	0.519**	0.339*	0.362*	1.215***	0.979***	0.936***	1.012***
	(0.112)	(0.204)	(0.172)	(0.197)	(0.143)	(0.152)	(0.156)	(0.142)
SMB	0.508	0.452	-0.105		-0.883***	-0.897***	-0.427	
	(0.389)	(0.319)	(0.436)		(0.289)	(0.288)	(0.316)	
HML	-0.080	-0.672***	0.564		0.096	-0.045	0.812**	
	(0.246)	(0.244)	(0.537)		(0.182)	(0.221)	(0.389)	
MOM			-0.924***				-0.220	
			(0.216)				(0.196)	
RMW				0.187				1.317***
				(0.649)				(0.470)
CMA				-1.553***				0.491
				(0.544)				(0.394)
Obs.	39	39	39	39	39	39	39	39
Adjusted R^2	0.090	0.084	0.387	0.220	0.651	0.709	0.711	0.764
Panel B: Job Satisfaction Aspects' alpha								
Career	1.393***	0.529	1.035*	1.183*	-0.098	-0.185	0.294	-0.287
Opportunities	(0.505)	(0.593)	(0.569)	(0.597)	(0.472)	(0.541)	(0.513)	(0.596)
Compensation	1.232***	0.904*	1.237**	1.358**	0.039	0.524	0.559	0.020
& Benefits	(0.398)	(0.515)	(0.519)	(0.565)	(0.402)	(0.388)	(0.414)	(0.406)
Senior	1.070***	0.631	1.200**	1.343**	-0.095	0.173	0.356	-0.152
Leadership	(0.433)	(0.614)	(0.576)	(0.617)	(0.414)	(0.46)	(0.481)	(0.513)
Work/life	0.837**	0.714	1.232***	1.019**	-0.079	0.163	0.387	-0.003
Balance	(0.348)	(0.462)	(0.403)	(0.496)	(0.36)	(0.403)	(0.412)	(0.451)
Culture	1.326***	0.877*	1.252**	1.050*	0.311	0.715*	0.771*	-0.101
& Values	(0.274)	(0.487)	(0.477)	(0.526)	(0.375)	(0.401)	(0.427)	(0.345)

Note: Panel A reports the coefficients along with heteroskedasticity and autocorrelation robust standard errors in parentheses for abnormal returns (alphas), controlling for standard asset pricing factors. Specifications (1), (2), (3) and (4) present the results from a Capital Asset Pricing Model (CAPM), Fama-French 3 factor model (FF3), Carhart 4 factor model (C4) and Fama-French 5 factor model (FF5), respectively. Results are presented for both equal-weighted and value-weighted portfolios that invest in the top 25% of firms and short the bottom 25% firms, sorted on the overall employee satisfaction rating. Panel B presents the abnormal alphas when portfolios are sorted on various job satisfaction aspects. $^{*}p<0.10$; $^{**}p<0.05$; $^{***}p<0.01$.

To conserve space, we only report the abnormal alphas along with HAC robust standard errors in parentheses. Our findings exhibit that this information offers statistically significant abnormal profits after controlling for standard risk factors, which are particularly elevated compared to the abnormal profits generated when portfolios are constructed based on the overall employee satisfaction. The abnormal returns are more pronounced for compensation and benefits and culture and values, indicating that various aspects from firms’ policies that affect employee satisfaction seem not to be priced in the market.

8. Discussion, implications and limitations

The results of our analysis extend the extant scholarly thought in several directions. While several studies investigate employee satisfaction drivers, we extend the findings by harnessing the value of both structured (numerical rating) and unstructured (review text) data. Moreover, we differ from previous studies that focus on overall employee satisfaction by decomposing the analysis to additional factors based on both numerical and narrative content. In so doing, we present methodological advances derived from the combination of textual and numerical features and enriched information content by unveiling latent job satisfaction dimensions not captured by established measurement scales. Our findings indicate that the information content varies significantly among different employee satisfaction aspects justifying the purpose of this analysis.

The findings of this research require careful interpretation. For example, while compensation and benefits and work-life balance seem to be less influential factors compared to career opportunities, senior leadership, and corporate culture and values, we argue that they are still critical factors in that their minimum level can increase dissatisfaction, but their maximum level alone will not generate high levels of satisfaction. In line with the two-factor motivation theory (Herzberg et al., 2011), the former factors are considered as “hygienes” and the latter as “motivators”. These asymmetries in the way different factors impact satisfaction and dissatisfaction of employees are also reflected in the topic analysis when assessing the advantages and disadvantages of working with an employer. Not surprisingly, tangible job satisfaction aspects, such as compensation and benefits, prevail in topic proportion; however, the findings of the topic analysis converge with the regression analysis when we examine how topics evolve among more satisfied and more dissatisfied employees.

Our second contribution is to the literature related to employee turnover determinants. We argue that employee big data that arrive from both current and former employees provide an extremely useful and reliable source of information for an otherwise difficult and expensive dataset to form by surveys. The Linear Discriminant Analysis demonstrates the information content and the significance of coupling numerical and textual job elements with considerable implications for research.

From the operational planners’ perspective, and in relation to the important operational construct of workforce planning, our study has significant managerial implications. Companies that have an established appraisal cycle take into account both a measurement perspective as well as a qualitative perspective reflected in employee appraisal documents. Our results demonstrate that the qualitative information encapsulated on an employee’s narrative regarding her experience with the company is of significant value and therefore, a holistic approach of extracting features from this relatively unused source of information can bring significant predictability benefits, in the same way that we demonstrate with online review data. Nonetheless, contractual relations in an organization vary in terms of tenure and contract type (e.g., fixed-term contract, sub-contractor contract). Therefore, different levels of employee experience with the company may exhibit heterogeneous beliefs about the company environment and these reflections, when utilized in a topic model, outperform numerical ratings. As such, organizations that have not yet engaged in systematic employee satisfaction measurement frameworks can utilize existing sources of unstructured data to plan, recruit and retain talent. Especially in industries of high employee turnover, such as those in high contact services (Stamolampros, Korfiatis, Chalvatzis & Buhalis, 2019), this information can be further utilized for capacity planning, especially in periods of high operational load.

Third, we also contribute to the literature that associates employee satisfaction, human capital and intangibles with superior firm profitability. In particular, we document that UK firms rated highly by their current employees in terms of satisfaction achieve superior financial performance (ROA) compared to those rated poorly. This finding, which is robust to controlling for firm characteristics, industry and time fixed-effects and endogeneity concerns, favors the human-centered view of the firm. The significant positive relationship between employee satisfaction rating and profitability indicates that online employee reviews can be used to forecast financial profitability. Importantly, we demonstrate how the models improve when we decompose the overall satisfaction to additional employee satisfaction elements, or employ information derived from textual analysis. As the tested period is characterized by turmoil in the UK labor market and increased uncertainty for UK firms due to the Brexit referendum and its outcome, our findings are robust to volatile market conditions suggesting that employee big data can be a source of significant competitive advantage for firms.

Our results also show the value relevance of employee online information for investors, supporting the view that non-financial information is of key importance for security valuation as it addresses the inability of standard accounting measures to capture investments in intangibles (e.g., Amir and Lev 1996). Abnormal returns of comparable magnitude are also obtained when we account for portfolio risk with popular asset pricing models, corroborating the link between employee satisfaction and corporate performance and suggesting that intangibles are not fully priced in the market. Moreover, since online employee reviews are publicly and freely available, the latter can be attributed to the failure of market participants to recognize the value of satisfied staff for firms, rather than to lack of information.

The aforementioned applications show some of the firm benefits of harnessing the potential of big data. Informed management decision support systems and machine learning workflows can be effectively utilized making use of existing structured and unstructured data that is available at the company systems and servers. Figure 3 provides an example of a machine learning (ML) pipeline utilizing company data to combine unstructured and structured data in predicting employee turnover and financial performance. Information, such as the employee annual performance reviews (which are standardized across many organizations), as well as past employee satisfaction surveys and measures (if available), along with a topic modeling application can be used in a similar way that was demonstrated in this study, considering the document input and its associated steps in conjunction with the structured data that contain important meta-data about employees such as demographics and job-related variables (e.g., bonus and remuneration data). These can be fed-in to train and test a classification algorithm that can be evaluated with both in-sample and out-of-sample data with hyperparameter optimization and tuning added as an iterative feedback stage for both topic modeling and model building.

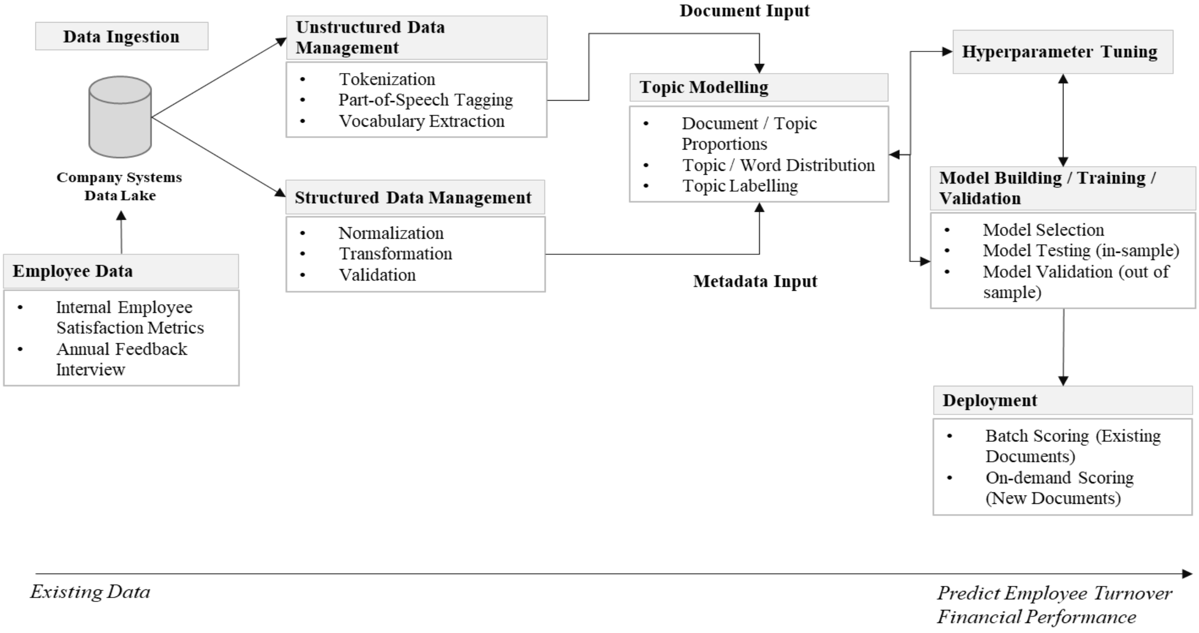

Data Ingestion
Company Systems Data Lake
Employee Data
Internal Employee Satisfaction Metrics
Annual Feedback Interview
Unstructured Data Management
Tokenization
Part-of-Speech Tagging
Vocabulary Extraction
Structured Data Management
Normalization
Transformation
Validation
Document Input
Metadata Input
Topic Modelling
Document / Topic Proportions
Topic / Word Distribution
Topic Labelling
Hyperparameter Tuning
Model Building / Training / Validation
Model Selection
Model Testing (in-sample)
Model Validation (out of sample)
Deployment
Batch Scoring (Existing Documents)
On-demand Scoring (New Documents)
Existing Data
Predict Employee Turnover
Financial Performance

Figure 3*: Machine Learning (ML) Scoring pipeline for predicting the likelihood of employee turnover and financial performance*

The resulting scoring pipeline can be used to either evaluate existing employee data (batch scoring), for example, in cases where operational planning is highly dependent on employees, such as company growth, or to examine whether labour market competition for talent hinders the success of replacement vacancies. Furthermore, in the case of employees that are considered key employees for a particular business unit, risk assessment can be incorporated in profiling their likelihood to leave or stay with the company, thus enabling better operational planning. Depending on their relevance, the prevalence of specific topics can be used to evaluate whether the company is still an attractive place to work or not. In addition, restructuring decisions about the performance of particular business units can be also supported using ML scoring pipelines as the one described above, thus enabling higher efficiencies through better workforce planning.

This analysis is mainly subject to limitations that are inherently associated with online reviews. For example, the literature unveils several biases that govern consumer responses when providing online reviews, such as self-selection and response biases (e.g., a U-shape review distribution) (Li & Hitt, 2008; Hu, Zhang, & Pavlou, 2009). Interestingly, our sample does not follow a standard U-shaped pattern indicating that it may be less exposed to the latter bias. Moreover, the pre-determined job satisfaction measures, namely career opportunities, compensation and benefits, senior management, work/life balance, and culture and values, have been found to be significant determinants of employees' overall satisfaction with an employer. However, these job elements do not exhaust the factors that may induce satisfaction or dissatisfaction among employees. For example, the work environment, role ambiguity and role conflicts are variables that are not captured in the existing measurement scales of Glassdoor. To address this limitation, we employ unsupervised textual analytics methods.

9. Conclusions

This paper empirically investigates the informational gain of exploiting structured and unstructured data in several applications, alongside offering managerial guidance on machine learning techniques that could inform decision support systems. In particular, we explore a novel big dataset consisting of employee e-WOM. This dataset offers access to both numerical ratings and review text processed using probabilistic topic analytics. We find that standard data analysis techniques and models could be greatly benefited by accommodating information generated from unsupervised textual techniques that allow data to "speak for itself", unveiling latent factors that determine key operational and financial indicators, such as job satisfaction, employee turnover and financial performance. Our research makes contributions to several streams in the literature and demonstrates how big data analytics novelties can generate competitive advantage and informational gains to managerial practice.

References

Amir, E., & Lev, B. (1996). Value-relevance of nonfinancial information: The wireless communications industry. *Journal of Accounting and Economics*, *22*(1–3), 3–30.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.

Bernhardt, K. L., Donthu, N., & Kennett, P. A. (2000). A longitudinal analysis of satisfaction and profitability. *Journal of Business Research*, *47*(2), 161–171.

Bertels, K., Jacques, J.-M., Neuberg, L., & Gatot, L. (1999). Qualitative company performance evaluation: Linear discriminant analysis and neural network models. *European Journal of Operational Research*, *115*(3), 608–615.

Bordoloi, S. K., & Matsuo, H. (2001). Human resource planning in knowledge-intensive operations: A model for learning with stochastic turnover. *European Journal of Operational Research*, *130*(1), 169–189.

Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-based management: Concept cleanup time? *Academy of Management Perspectives*, *23*(4), 19–32.

Cascio, W. F. (1991). *Costing human resources: The financial impact of behavior in organizations* (3rd ed.). PWS-Kent.

Chen, H., De, P., Hu, Y. J., & Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, *27*(5), 1367–1403.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944–957.

Cohen, G., Blake, R. S., & Goodman, D. (2016). Does turnover intention matter? Evaluating the usefulness of turnover intention rate as a predictor of actual turnover rate. *Review of Public Personnel Administration*, *36*(3), 240–263.

Corominas, A., Lusa, A., & Olivella, J. (2012). A detailed workforce planning model including non-linear dependence of capacity on the size of the staff and cash management. *European Journal of Operational Research*, *216*(2), 445–458.

Darmon, R. Y. (2004). Controlling sales force turnover costs through optimal recruiting and training policies. *European Journal of Operational Research*, *154*(1), 291–303.

De Bruecker, P., Van den Bergh, J., Beliën, J., & Demeulemeester, E. (2015). Workforce planning incorporating skills: State of the art. *European Journal of Operational Research*, *243*(1), 1–16.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS One*, *8*(11), e79449.

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial Economics*, *101*(3), 621–640.

Fabozzi, F. J., Neave, E. H., & Zhou, G. (2011). *Financial economics*. John Wiley & Sons, Inc.

Filbeck, G., & Preece, D. (2003). Fortune's best 100 companies to work for in America: Do they work for shareholders? *Journal of Business Finance & Accounting, 30*(5–6), 771–797.

Flannery, M. J., & Hankins, K. W. (2013). Estimating dynamic panel models in corporate finance. *Journal of Corporate Finance, 19*, 1–19.

Griffeth, R. W., Hom, P. W., & Gaertner, S. (2000). A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. *Journal of Management, 26*(3), 463–488.

Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, *59*, 467–483.

Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. *Journal of Management*, *39*(3), 573–603.

Hausknecht, J. P., Rodda, J., & Howard, M. J. (2009). Targeted employee retention: Performance-based and job-related differences in reported reasons for staying. *Human Resource Management*, *48*(2), 269–288.

Herzberg, F., Mausner, B., & Snyderman, B. B. (2011). *The motivation to work* (Vol. 1). Transaction publishers.

Heskett, J. L., Jones, T. O., Loveman, G. W., Sasser, W. E., & Schlesinger, L. A. (1994). Putting the service-profit chain to work. *Harvard Business Review, 72*(2), 164–174.

Holland, P., Cooper, B. K., & Hecker, R. (2016). Use of social media at work: A new form of employee voice? *The International Journal of Human Resource Management*, *27*(21), 2621–2634.

Hom, P. W., & Kinicki, A. J. (2001). Toward a greater understanding of how dissatisfaction drives employee turnover. *Academy of Management Journal*, *44*(5), 975–987.

Hom, P. W., Lee, T. W., Shaw, J. D., & Hausknecht, J. P. (2017). One hundred years of employee turnover theory and research. *Journal of Applied Psychology, 102*(3), 530.

Huang, M., Li, P., Meschke, F., & Guthrie, J. P. (2015). Family firms, employee satisfaction, and corporate performance. *Journal of Corporate Finance, 34*, 108–127.

Jung, T., Scott, T., Davies, H. T., Bower, P., Whalley, D., McNally, R., & Mannion, R. (2009). Instruments for exploring organizational culture: A review of the literature. *Public Administration Review, 69*(6), 1087–1096.

Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, *116,* 472–486.

Kulkarni, G., Kannan, P., & Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support Systems*, *52*(3), 604–611.

Kunc, M., & O'brien, F. A. (2019). The role of business analytics in supporting strategy processes: Opportunities and limitations. *Journal of the Operational Research Society, 70*(6), 974–985.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review, 52*(2), 21.

Lee, M., & Mimno, D. (2014). Low-dimensional Embeddings for Interpretable Anchor-based Topic Inference. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–1328.

Lee, T. W., Hom, P. W., Eberly, M. B., Li, J. (Jason), & Mitchell, T. R. (2017). On the next decade of research in voluntary employee turnover. *Academy of Management Perspectives*, *31*(3), 201–221.

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456–474.

Maertz Jr, C. P., & Griffeth, R. W. (2004). Eight motivational forces and voluntary turnover: A theoretical synthesis with implications for research. *Journal of Management*, *30*(5), 667–683.

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review, 90*(10), 60–68.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Mitchell, T. R., Holtom, B. C., Lee, T. W., Sablynski, C. J., & Erez, M. (2001). Why people stay: Using job embeddedness to predict voluntary turnover. *Academy of Management Journal*, *44*(6), 1102–1121.

Mohr, D. C., Young, G. J., & Burgess, J. F., Jr. (2012). Employee turnover and operational performance: The moderating effect of group-oriented organisational culture. *Human Resource Management Journal*, *22*(2), 216–233.

Mortenson, M. J., Doherty, N. F., & Robinson, S. (2015). Operational research from Taylorism to Terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, *241*(3), 583–595.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2017). *stm: R Package for Structural Topic Models*. http://www.structuraltopicmodel.com

Rubenstein, A. L., Eberly, M. B., Lee, T. W., & Mitchell, T. R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, *71*(1), 23–65.

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, *32*(2), 243–256.

Song, H., & Huang, H.-C. (2008). A successive convex approximation method for multistage workforce capacity planning problem with turnover. *European Journal of Operational Research*, *188*(1), 29–48.

Stamolampros, P., Korfiatis, N., Chalvatzis, K., & Buhalis, D. (2019). Job satisfaction and employee turnover determinants in high contact services: Insights from employees' online reviews. *Tourism Management*, *75*, 130–147.

Stamolampros, P., Korfiatis, N., Kourouthanassis, P., & Symitsi, E. (2019). Flying to Quality: Cultural Influences on Online Reviews. *Journal of Travel Research*, *58*(3).

Steel, R. P. (2002). Turnover theory at the empirical interface: Problems of fit and function. *Academy of Management Review*, *27*(3), 346–360.

Symitsi, E., Stamolampros, P., & Daskalakis, G. (2018). Employees' online reviews and equity prices. *Economics Letters*, *162*, 53–55.

Tirunillai, S., & Tellis, G. J. (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–479.

Ton, Z., & Huckman, R. S. (2008). Managing the impact of employee turnover on performance: The role of process conformance. *Organization Science*, *19*(1), 56–68.

Wintoki, M. B., Linck, J. S., & Netter, J. M. (2012). Endogeneity and the dynamics of internal corporate governance. *Journal of Financial Economics*, *105*(3), 581–606.

Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, *69*(5), 1562–1566.

Zhan, Y., & Tan, K. H. (2020). An analytic infrastructure for harvesting big data to enhance supply chain performance. *European Journal of Operational Research*, *281*(3), 559–574.

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.

Zingales, L. (2000). In search of new foundations. *The Journal of Finance, 55*(4), 1623–1653.

Appendix: Structural Topic Models for analyzing employee reviews

The STM process for the employee review text is graphically depicted in Figure using plate notation. The steps are described as follows: Let us assume a corpus of R reviews with each review r indexed as $r_i \in (1, \ldots. R)$ containing $\boldsymbol{w}$ observed words which are indexed as $\mathrm{n} \in (1, \ldots, N_{\mathrm{r}})$. Each word, denoted by $w_{r,n}$, is part of the general vocabulary of the review corpus with each term denoted by $v \ \in (1, \ldots, V)$ a word in the vocabulary. The primary input variable in a topic model is the number of topics K: $\mathrm{k} \ \in (1, \ldots \mathrm{K})$ drawn from a distribution and this should be defined at the beginning of the estimation process. There are various ways to identify the number of topics in a corpus, such as using the concepts of held-out likelihood and a combination of qualitative criteria which may require input by domain experts. These are outlined in the relevant section in the manuscript. For our analysis we did not use the topical content approach (which would consider a simultaneous estimation of the review-topic and topic-word distribution for positive and negative text) but repeated the process for positive and negative content in order to allow for the possibility that the optimal number of topics would be different for positive and negative text.

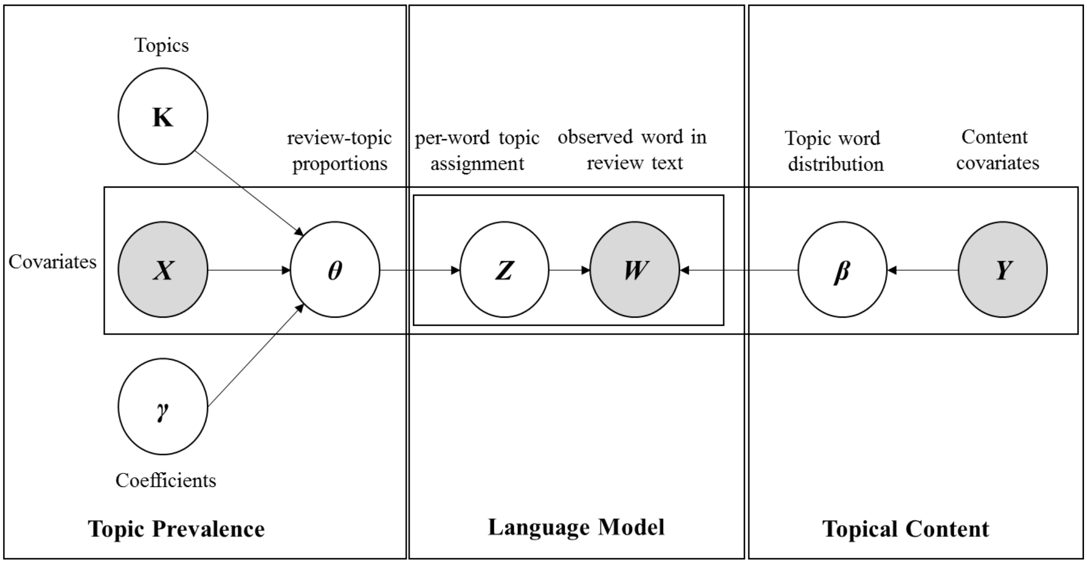


Figure A1: Structural topic model process using plate notation (adopted by Roberts et al., 2016).

The distribution is determined by topic prevalence covariates which are specified in a $p \times 1$ vector X_r. When no topic prevalence covariates are defined, then the STM process works in the same way as Latent Dirichlet allocation by using Gibbs sampling to draw the topics from a Dirichlet distribution. In our case X_r contains three review-based covariates that affect the dominance of a topic k_i for each review r_i. The covariates that were used were: (a) the overall score of the review, (b) the employee status (current vs. former) and (c) the Glassdoor provided sector where this company is part of. The first two covariates are central to the research questions that we aim to address in this paper and which function as a proxy for employee satisfaction and turnover intention. The third covariate (Sector) is used as a control.

The process runs in three steps as follows:

First, the review-level relation to each topic k is drawn from a logistic normal generalized linear model based on covariates and a set of priors as shown in Equation (1).

$$\vec{\theta}_{\gamma} \,|X_{r\gamma}, \textstyle\sum \sim LogisticNormal(\mu = X_{r\gamma}, \textstyle\sum), \tag{1}$$

$$X_{r\gamma} = [OverallRating, isFormer, Sector]$$

where γ represents a $p \times (K-1)$ matrix of coefficients drawn from a Normal distribution for each k ($k = 1, \dots, K-1$) with the other *K-1* topics to provide bivariate dependence between topics. $\sum$ is a $(K-1) \times \ (K-1)$ covariance matrix.

Second, using the review-specific distribution over words initially attributed to each topic (k) by the log frequency distribution (m) of the vocabulary vector, a topic-specific deviation from the initial stage κ_{k} as well as a covariate for group deviation κ_{g} and an interaction term κ_{i} between them can be modeled as:

$$\beta_{\mathrm{r,k},v} \propto exp\big(m + \kappa_{\mathrm{k},v} + \kappa_{\mathrm{g},v} + \kappa_{\mathrm{i=(k,g_r},w\,)}\big), \tag{2}$$

where $m, \kappa_{\mathrm{k,w}}, \kappa_{\mathrm{g,w,}}\ \kappa_{\mathrm{i}}$ are vectors (V-length) that contain one input per word *(w)* in the vocabulary.

Finally, for each n$^{\mathrm{th}}$ observed word n $\in (1, \dots, N_{\mathrm{r}})$ in a review text r_i the word-specific topic assignment $z_{r,n}$ can be modelled based on the review-specific distribution over the given finite set of topics as:

$$z_{r,n} |\, \vec{\theta}_{\gamma} \sim Multinomial(\vec{\theta}_{\gamma}) \tag{3}$$

The probability of an observed word attributed to this topic is given by:

$$w_{r,n}|z_{r,n}, \beta_{\mathrm{r,k}} = Z_{r,n} \sim Multinomial(\beta_{\mathrm{r,k}} = Z_{r,n}) \tag{4}$$

The model is then fitted using a semiparametric estimation from a semi Expectation – Maximization algorithm (Blei & Lafferty, 2007; Wang & Blei, 2013) which upon convergence identifies the topic-specific proportions $\theta_{r,k|X}$ of a review using information from the vector of covariates provided in the initial stages of the estimation. The process is repeated two times, one for the positive and one for the negative corpus.

We also estimated time effects for the topic prevalence by allowing the topic membership function *(θ)* to fluctuate by time as: $Y_k = \{Y_{kt}: t \in T\}$ where T is the bandwidth of the period (in days) that is available in our dataset. We test the impact of time on the prevalence of the k$^{\text{th}}$ topic prevalence with the following model:

$$Y_{kt} = \begin{bmatrix} Prevalence_1 \\ \vdots \\ Prevalence_k \end{bmatrix} = \alpha_k + \mathrm{\beta}_{ik} \begin{bmatrix} \mathrm{rating}_{1t} \\ \vdots \\ \mathrm{rating}_{iT} \end{bmatrix} + \gamma_i F + \delta_c S + d_{1t}, \tag{5}$$

where a_k is a constant, β_{ik} indexes the influence of the rating covariate on the topic prevalence, γ and δ_c indexes the employee tenure (F=1 if it is former) for the reviewer *i* and sector *S* for the company *c*, respectively.Variable *d* is a smoothing covariate to account for time effects. For each topic we can estimate the influence of these covariates on the topic distribution and, in particular, how the topic fluctuates over the time-based covariate. Figures A2 and A3 provide an overview of the fluctuation of topics over time. The plots consider the effect of the time covariate which has been transformed using a non-linear (spline) smoothing. For presentation reasons we have segmented the bandwidth labels to 113 points with the first point corresponding to May 2008 and last point to December 2017. We also added representative tick-marks for a six month interval (Month – Year).

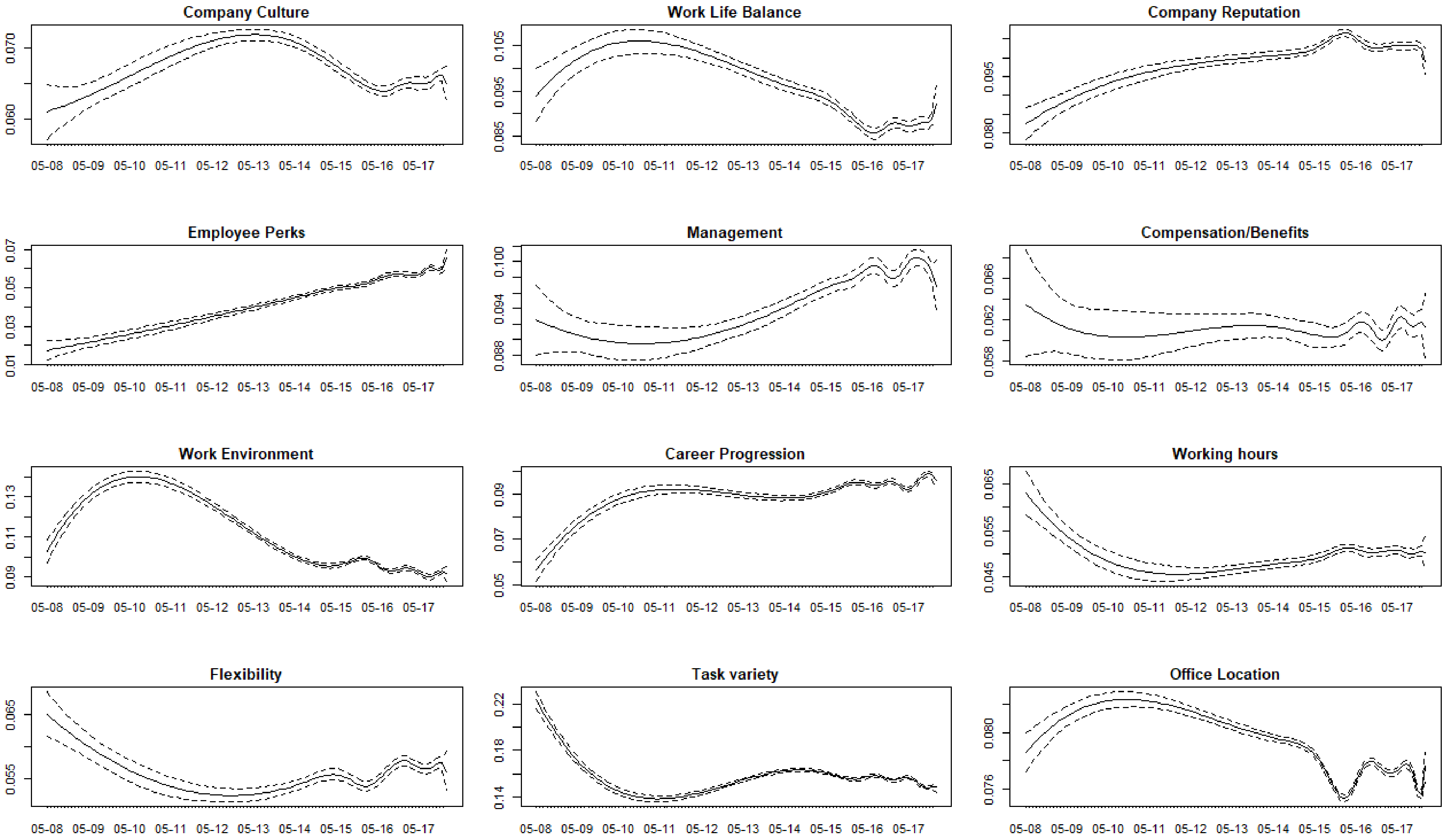


Figure A2: *Fluctuation of positive topics over time.*

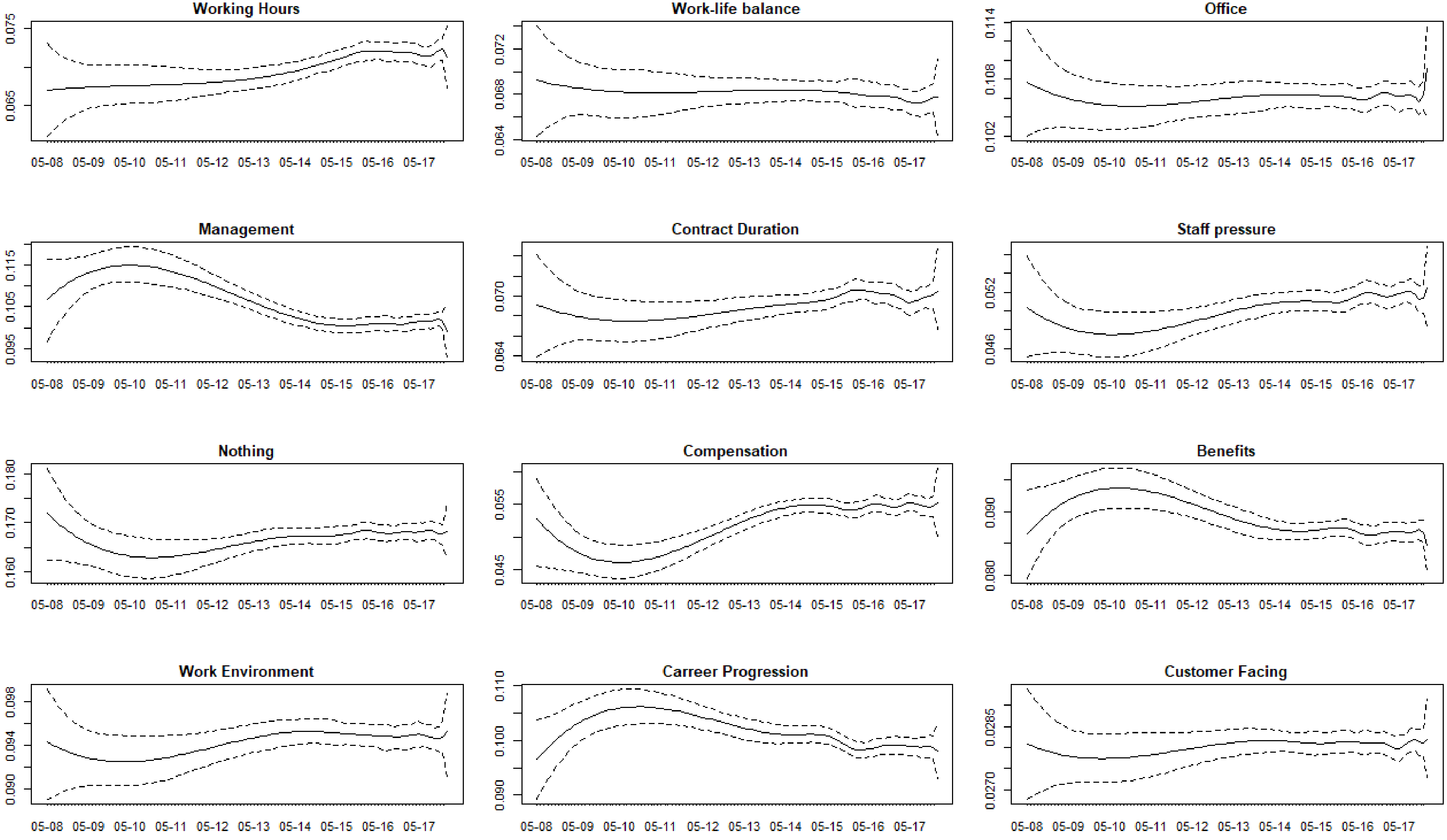
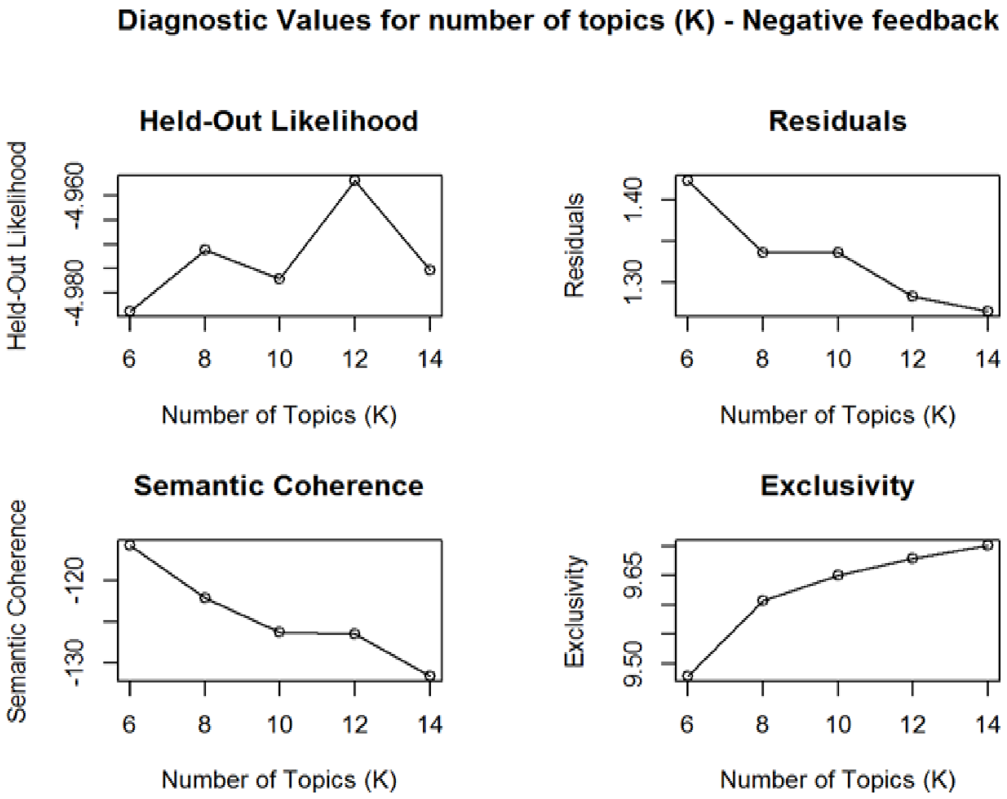


Figure A3: *Fluctuation of negative topics over time.*

Appendix B: Diagnostic values for the number of topics in the topic solution



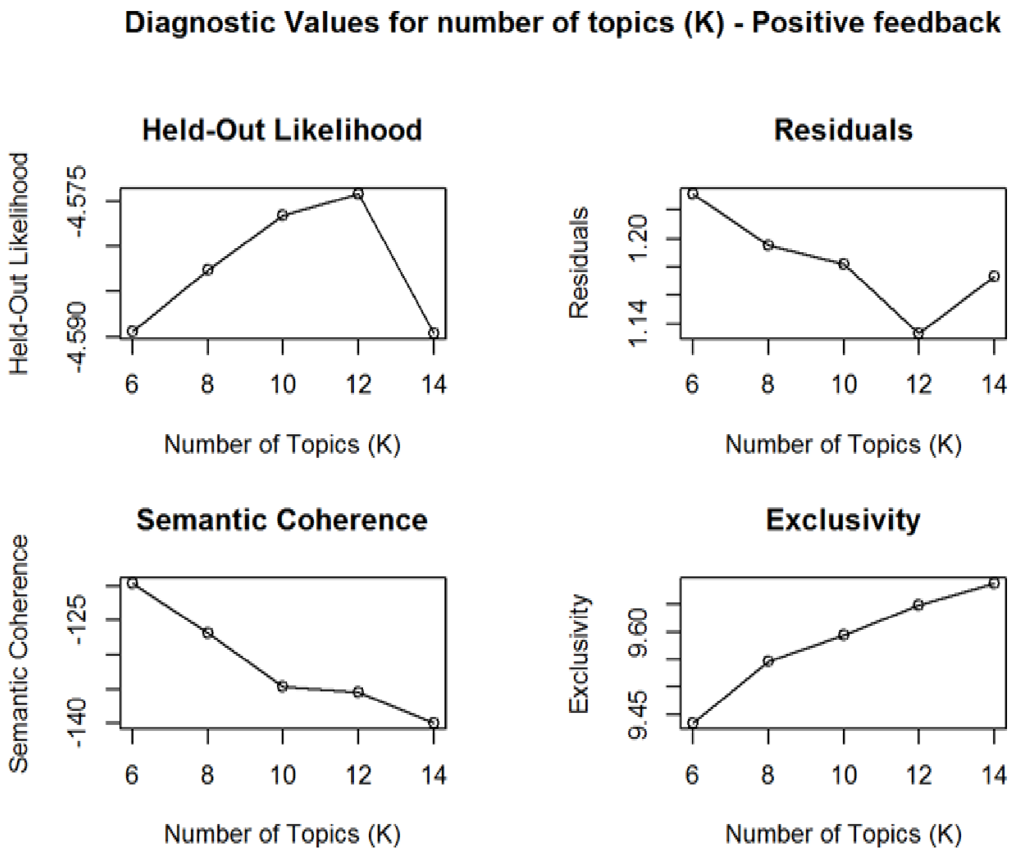


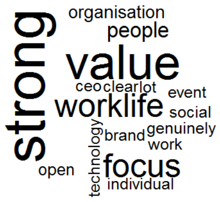
Figure B1: *The plot illustrates the diagnostic values in terms of held-out likelihood, semantic coherence, lower bound for word importance and the residuals obtained for the full model. The best combination is achieved when the number of topics (K) is 12, as this provides the best relationship between the held-out likelihood and semantic coherence.*

Table B1: *Model selection criteria for competing values of the number of topics.*

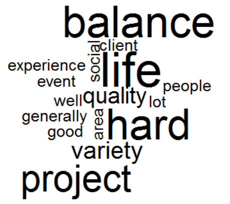
# Topics	Held out Likelihood		Ratio of Semantic Coherence to Exclusivity	
	Positive	Negative	Positive	Negative
K=6	-4.589	-4.984	-12.682	-12.215
K=8	-4.583	-4.971	-13.286	-12.711
K=10	-4.577	-4.977	-14.028	-13.083
K=12	**-4.574**	**-4.957**	**-14.034**	**-13.076**
K=14	-4.590	-4.975	-14.451	-13.571

Appendix C:

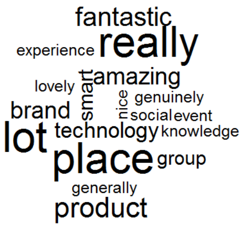
Word clouds for Positive Topics



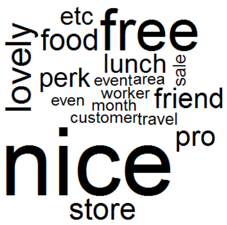
***(1)** Company Culture*



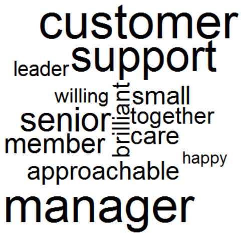
***(2)** Work-life Balance*



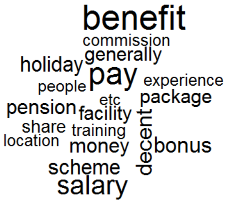
***(3)** Company Reputation*



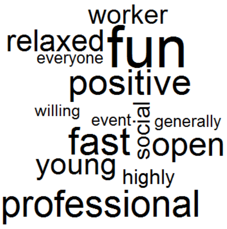
***(4)** Employee Perks*



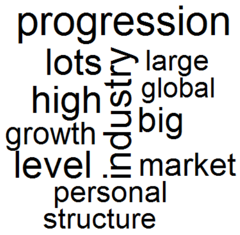
***(5)** Management*



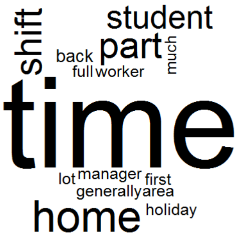
***(6)** Compensation/Benefits*



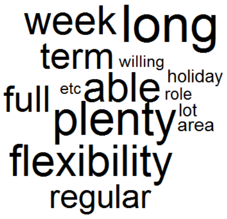
***(7)** Work Environment*



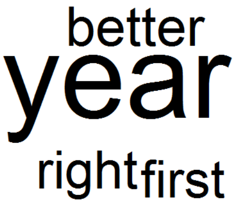
***(8)** Career Progression*



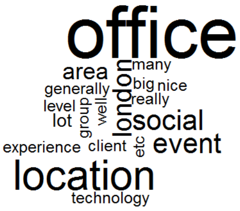
***(9)** Working hours*



(10)Flexibility

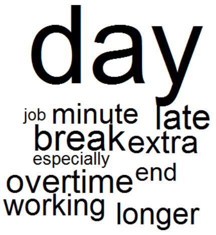


(11) Task Variety

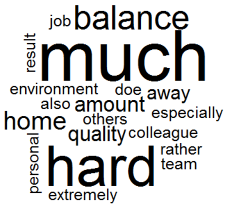


(12) Office Location

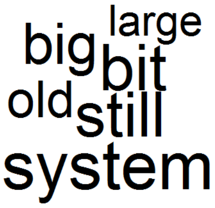
Word clouds for Negative Topics



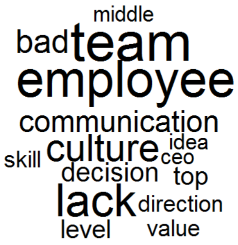
(1) Working hours



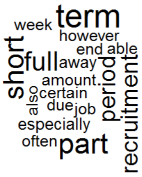
(2) Work/life balance



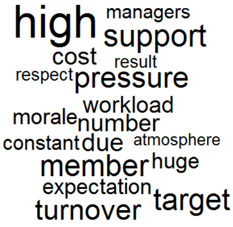
(3) Office/Premises



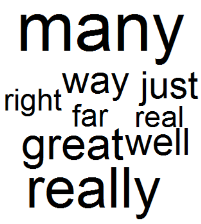
(4) Management/Leadership



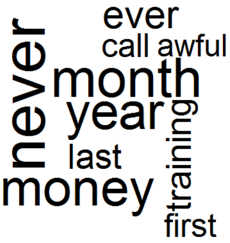
(5) Recruitment



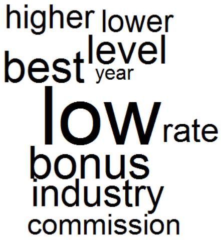
(6) Staff Pressure



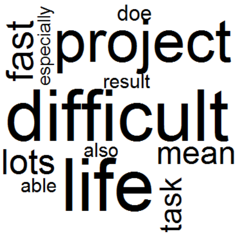
(7) No Negatives



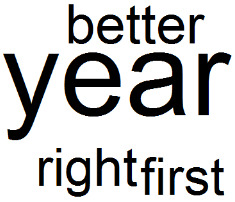
(8) Compensation



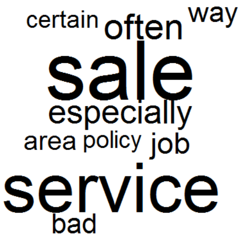
(9) Benefits



(10) Job Role



(11) Career Progression



(12) Customer Facing