12-2020

Job Satisfaction and Employee Turnover Determinants in Fortune 50 Companies: Insights from Employee Reviews from Indeed.com

Bishal Sainju
Utah State University

JOB SATISFACTION AND EMPLOYEE TURNOVER DETERMINANTS IN FORTUNE 50 COMPANIES: INSIGHTS FROM EMPLOYEE REVIEWS FROM INDEED.COM

by

Bishal Sainju

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Computer Science

Approved:

John Edwards, Ph.D. Major Professor	Christopher Hartwell, Ph.D. Committee Member
Curtis Dyreson, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2020

ABSTRACT

Job satisfaction and employee turnover determinants in Fortune 50 companies: Insights from employee reviews from Indeed.com

by

Bishal Sainju, Master of Science

Utah State University, 2020

Major Professor: John Edwards, Ph.D.
Department: Computer Science

We explored 682176 employee reviews of Fortune 50 companies from Indeed.com using topic discovery techniques like Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM) to identify salient aspects in employee reviews and automatically infer latent topics that tend to drive employee satisfaction. We also studied how various satisfaction factors could be related to employee turnover. We discovered important topics in the reviews, including *Management and Leadership*, *Advancement Opportunity*, *Pay and Benefits*, *Work-Life Balance*, and *Culture*, which we compare to the five Job Descriptive Index (JDI) facets. Both LDA and STM discovered well-separated and distinguishable topics. We also incorporated a "Job Status" covariate in STM, which helped distinguish between what topics were talked about most by "Former" vs "Current" employees, and consequently helped us analyze the factors that could have caused employee turnover. We found that *Leadership and Management* and *Overwork and Stressful Environment* were the dominant factors contrasting between former and current employees, suggesting that they might be a leading cause of employee turnover. Furthermore, we post-processed the topic probability result from the STM model and analyzed it to determine sector-wise topic contribution for each topic, and also analyzed the company-wise topic contribution in each sector. We

found that Retail sectors talked the most about *Pay and Benefits* and *Length of Breaks*, whereas the Technology sector's employees were more concerned about the *Work-Life Balance* issue. Our results are directly usable to support company behavioral management decision makers to conceive and evaluate initiatives intended to enhance employee satisfaction. Furthermore, our techniques, including a novel visualization of topic composition and quality, are generalizable to any setting that uses topic discovery from unstructured text, and especially those comparing topics across entities.

(68 pages)

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. John Edwards, for the continuous support of my Master's study, research and scholarship. Dr. Edwards was always approachable whenever I ran into a trouble spot or had a question about my research. His guidance helped in all the time of research and writing of this thesis.

I am also equally grateful to Dr. Christopher Hartwell, for providing me with this wonderful opportunity to work on this excellent project from the Management Department. This project turned out to be a great collaboration between Management and Computer Science department, and helped me expand my knowledge on both domains. Dr. Hartwell was really helpful, and his management insights really helped me make progress in this research.

Similarly, I would also like to thank Dr. Curtis Dyreson for his encouragement and insightful comments.

My sincere thanks also goes to Genie Hanson, Vicki Anderson, Kaitlyn Fjeldsted and Cora Price for providing constant help and support whenever I ran into departmental issues, these are the people who bind the Computer Science department at Utah State University together.

Last but not the least, I am extremely grateful to my family and my friends here at Utah State University for their love, prayers, and sacrifices for educating me and preparing me for my future.

Bishal Sainju

CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Employee job satisfaction and retention are some of the important human factors affecting a company's operating and financial performance. Thus, it is important to understand and analyze these satisfaction aspects. While such examinations have a long history in the literature, this project moves beyond traditional means of employee surveys and instead looks at proactive employee comments. In this project, we used *Indeed.com* (hereafter referred to as *Indeed*) reviews of current and former employees of Fortune 50 companies to extract latent satisfaction categories and analyzed the data to better understand the drivers affecting the employee job satisfaction and turnover.

1.1 Motivation

The primary motivation for this research is the plethora of opportunities that online review sites like *Indeed* provide for the companies to discover new and latent satisfaction aspects that companies can utilize in order to direct effective use of human capital and positively impact organizational outcomes.

Traditional methods for measuring job satisfaction, like questionnaires and surveys, provide limited capabilities as the user provides their opinions on a limited set of topics typically developed and delivered by the employer. However, an online platform like *Indeed* contains millions of free-form employee reviews that can be analyzed to extract unrestricted critique on a variety of topics, which may allow for a truer gauge of employee job satisfaction.

Also, with increasing computational resources, topic modeling algorithms like Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM) can be leveraged to mine large corpora of textual data. Thus, thousands of employee reviews can be analyzed with relative ease.

In addition to studying job satisfaction determinants, the focus of this research extends

to employee turnover analysis, motivated by a desire to better understand the reasons behind employee turnover and perhaps lessen the costly consequences that turnover incurs to the firms. Prior models fail to incorporate important constructs that explain employee turnover due to methodological constraints. However, with the online proliferation of first-hand information from both current and former employees, turnover implications can be more readily made.

Also, we further investigated the sector-wise and company-wise topic differentiation on both positive and negative feedback, thus demonstrating to employers one way to evaluate and analyze their company's comparative satisfaction aspects in a sleek and discrete manner. This could help companies compare different satisfaction facets with their competition, and provide perspective on what they need to focus on to enhance their employees' satisfaction and reduce turnover, in turn driving company performance.

1.2 Contribution

In this project, we make the following contributions: In this project, we make the following contributions:

- We use machine learning approaches to mine latent job satisfaction topics in the large corpus of employee reviews and analyze the strengths and weaknesses of each model. We find that LDA performs relatively better than STM in topic discovery, whereas STM is needed for incorporating covariate information.

- We draw upon the most ubiquitous job satisfaction framework, the Job Descriptive Index [JDI] [1], but also discover novel job satisfaction aspects that provide additional breadth and depth to the concept of job satisfaction.

- We present a visualization that allows a person to clearly distinguish between positive and negative satisfaction drivers (what drives dissatisfaction and satisfaction) and also distinguish between former and current employees on satisfaction factors, allowing a conceptual understanding of what drives employee turnover.

- Finally, we implement a method to compare and analyze company-wise and sector-wise topic contributions, thus providing comparisons across sectors and within each sector's companies, regarding common positive and negative job satisfaction facets.

CHAPTER 2
BACKGROUND & RELATED WORK

2.1 Employee Job Satisfaction and Job Descriptive Index (JDI)

Employee job satisfaction is an important correlate of both individual employee performance [2,3] and organizational success [4,5]. Traditionally, employee surveys have been the main method used to evaluate job satisfaction. Of these, the Job Descriptive Index (JDI) is the most common measure, and has demonstrated strong reliability and validity [6]. The JDI comprises five facets, including satisfaction with: coworkers, the work itself, pay, opportunities for promotion, and supervision. These facets have shown good independence [7], but some research has shown that breaking job satisfaction into more than five dimensions may be appropriate [8].

When trying to develop and understand drivers of job satisfaction, there are some drawbacks to the JDI and similar measures. These types of survey methods have limited validation, are restricted to a relatively small set of topics/questions, and demonstrate large amounts of method and error variance [6, 9]. Thus, it is difficult to identify new factors or gauge employees' independent thoughts. And, although employees are often told their survey responses are anonymous, their answers may still be biased by social desirability based on fears of repercussions when the survey is developed, delivered, and/or sponsored by their employer [10].

Job sites like *Glassdoor.com* and *Indeed.com* provide outlets for employees to proactively express their opinions about their current and former employers anonymously and in an open-ended fashion, thus allowing employees' opinions to cover an infinite range of subjects, rather than a restricted set of topics. This allows the employees to express a much broader and unfiltered opinion of their employers. One more advantage that these platforms have is that employees who have left the company can also leave their comments,

thus providing information about dissatisfaction with their previous employer. These online satisfaction ratings have demonstrated good construct validity in prior research [11].

Indeed allows employees to provide overall satisfaction ratings, which can provide a general view of their job experience with the company. Users can also rate in 5 different dimensions: i) Work-Life Balance, ii) Compensation / Benefits, iii) Job Security / Advancement, iv) Management, and v) Culture, thus allowing further depth. Some of these five factors are similar to the facets of the JDI [1], such as satisfaction with pay, opportunities for promotion, and supervision. Finally, *Indeed* users can provide overall comments about their job experience, and can provide comments about the *pros* and *cons* of their current and former jobs. These free-form comments provide insights about what employees see as the most salient positive and negative aspects of their jobs, and our research focuses on these comments in an attempt to use clustering and job modeling techniques to identify the positive and negative drivers of job satisfaction, and to see whether the emerging clusters match the dimensions of the JDI. Because users are self-identified as current or former employees, we also examine what drives retention (the *pros* identified by current employees) and turnover (the *cons* identified by former employees).

2.2 Clustering and Topic Modeling

2.2.1 Clustering

Clustering is an unsupervised machine learning algorithm, in which data points segregates into a different number of clusters such that items within the same cluster are similar to each other compared to items in different clusters. As it is an unsupervised learning algorithm, the dataset does not contain a label, whereby, the task of this kind of algorithm in the machine learning field would be to find out hidden patterns in the data, which are not explicitly identified. Thus, clustering is an important machine learning algorithm, and this can also be used in the field of document clustering, so as to cluster similar documents together.

Basically, there are 2 methods of clustering:

1. Hard Clustering
2. Soft Clustering

Hard Clustering

In hard clustering, each data point belongs to only one cluster, and that cluster only. In our case of document (review) clustering, the task is to cluster the similar documents together, so we can apply the hard clustering on these documents with the assumption that a particular document/review talks about one particular topic/cluster only. Since the length of the review is less than 20 terms, it can be assumed that the reviews are so small to be talking about multiple topics. Hence, our assumption should hold true and we will be able to apply hard clustering algorithms on our dataset (reviews).

Following is the hard clustering algorithm that we will use:

1. K-means clustering: Kmeans algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

   However, k-means do suffer from the curse of dimensionality. As the number of dimensions tends to infinity the distance between any two points in the dataset converges. This means the maximum distance and minimum distance between any two points of the dataset will be the same. This is a big problem when using the euclidean distance in K-Means.

   And, as our dataset is text reviews and the features being each term, there are lots of unique tokens/dimensions, which can cause our model to suffer from high dimensionality curse, which is a big problem. So to get around this problem spherical k-means

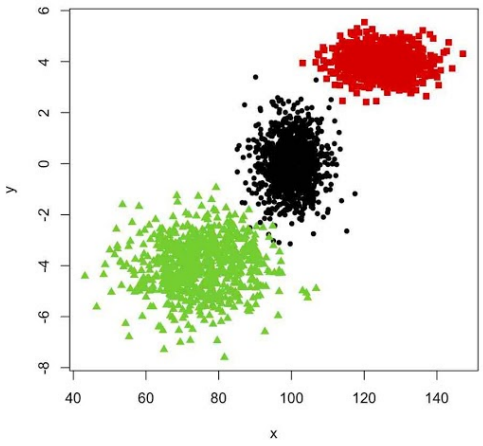


Fig. 2.1: K-Means Clustering

can be used, based on the cosine distance instead of Euclidean distance with thousands of features without any problems. Dimensionality reduction technique could also have been applied before applying k-means clustering, however, we haven't applied dimensionality reduction in our case, because spherical k-means itself gave a pretty reasonable result for our dataset.

We will primarily be applying k-means clustering to our dataset. The "k" value (number of cluster / topics) will be chosen according to "elbow method". According to this method, for various values of k, Sum of Squared (SSE) is calculated, and at whatever value of k, the SSE does not decrease significantly, we pick that value of k, as this is also known as the elbow point as the "k" value that should be the optimal number of cluster. In our case we want the "k" value to be as minimum as possible.

Soft Clustering

In soft clustering, each data point belongs to all of the cluster with some membership probability associated with being in that cluster. In our case, each document / reviews might be discussing about one or more issues (topics) i.e. our assumption that one document talks about only one topic might be false. Although document length are pretty small, it can be possible that one or more topics or issues are being talked about by a user, in their reviews. For this reason, soft clustering are used to exploit this factor that a document might belong

to more than one cluster.

2.2.2 Topic Modeling

Topic modelling refers to the task of identifying topics that best describes a set of documents. Some of the most popular topic modeling algorithm are soft clustering algorithms. They are:

1. Latent Dirichlet Allocation (LDA)
2. Structural Topic Modeling (STM)

Topic modeling is a type of statistical modeling for discovering the abstract "topics" that occur in a collection of documents. A document can be a part of multiple topics, similar to fuzzy clustering (soft clustering), in which each data point belongs to more than one cluster with some membership probability.

LDA

Latent Dirichlet allocation (LDA) [12] is a topic modeling algorithm in which set of topics are extracted from documents such that each word in the corpus is assigned with a probability of being in a particular topic, and each document has probabilities of being in each topic/cluster. Documents can be viewed like a mixture of topics. In LDA, the topic distribution is assumed to have a sparse Dirichlet prior. This assumption encodes the intuition that documents cover only a small set of topics and that topics use only a small set of words frequently. In practice, this results in a better disambiguation of words and a more precise assignment of documents to topics.

Latent Dirichlet allocation (LDA), originally introduced by Blei et al.(2003) [12], is a generative model for text. In this model, a "topic" t is a discrete distribution over words with probability vector ϕ_t. Dirichlet priors, with concentration parameter β and base measure n, are placed over the topics $\phi = \{\phi_1, ...\phi_T\}$:

$$P(\phi) = \prod_t Dir(\phi_t; \beta n). \tag{2.1}$$

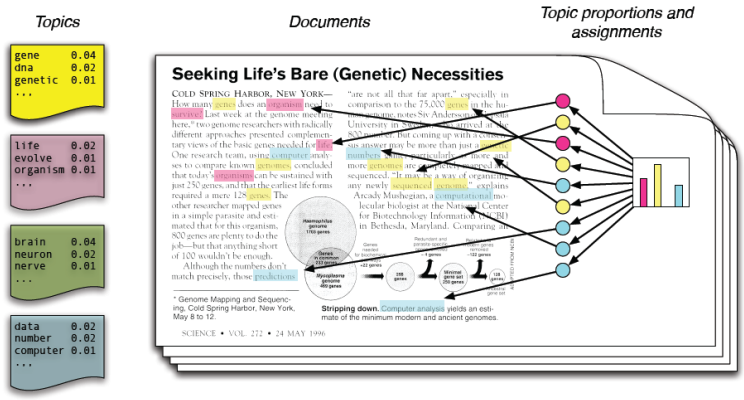


Fig. 2.2: Latent Dirichlet Allocation (LDA)

LDA looks for patterns of co-occurences and makes guesses about sets of themes. Passing through each document it first randomly assigns probability of a document containing particular theme ("topic") among various topics and then iterates to improve the classification of the probability of document belonging to one of these hypothetical topics. At the end, we get topic-term and document-topic distribution matrices.

1. *Document-Topic Distribution* : A probability matrix that gives us the probability value of any particular topic belonging to a particular document. For a given document, the probability over all topics sums to 1.

2. *Topic-Term Distribution* : A probability matrix that gives us the probability of a particular term belonging to a particular topic, defining a topic. The probability of all terms for a topic sums to 1.

STM

The Structural Topic Model [13] is a general framework for topic modeling with document-level covariate information. The covariates can improve inference and qualitative interpretability and are allowed to affect topical prevalence, topical content or both. STM is basically an extension of LDA, incorporating the additional information about the structure of the corpus into the model by altering the prior distributions to partially pool information amongst similar documents. Numerous special cases of this framework have been developed for particular types of corpus structure affecting both topic prevalence and topical content.

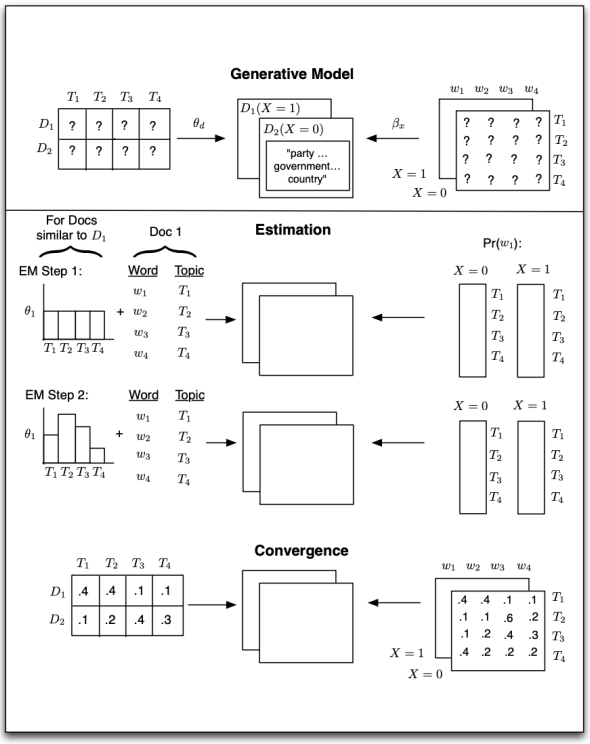


Figure 1: Heuristic description of generative process and estimation of the STM.

Fig. 2.3: Structural Topic Modeling (STM)

STM is basically the combination and the extension of existing models: the correlated topic model (CTM) [14], the Dirichlet-Multinomial regression (DMR) topic model [15] and the Spase Additive Generative (SAGE) topic model [16].

STM provides a general way to incorporate corpus structure or document metadata (information about the document such as in our case Employee Status, whether the review was written by “former” or “current” employee) into standard topic model. Using STM we can observe how topical prevalence varies on the basis of a certain covariate information, by inclusion of interest into the prior distributions for document-topic proportions and topic-word distributions. For example, using STM, we can observe what topics “Former Employees” are talking most about vs what topics “Current Employees” are talking about.

2.2.3 Related Work

Several studies have applied text mining approaches to online employee reviews. Luo et al. (2016) [5] used Glassdoor's employee reviews to build a model which found the correlation between employee satisfaction and company performance. Lee and Kang [17] performed topic modeling using LDA by adopting n-gram technique on the employee reviews obtained from *Glassdoor.com*. They then conducted dominance analysis to examine the relative importance of job factors. They found that *culture and value*, and *senior management* had the highest influence on both retention and turnover groups. Similarly, Jung and Suh [18] used LDA to extract job satisfaction factors from *jobplanet.co.kr*. They then measured sentiment and importance of each job satisfaction factor at industry, company, group, and chronological levels, using the dominance and correspondence analysis. They found that *Senior Management* and *Benefit and Compensation* had the highest importance on overall job satisfaction. Stamolampros et al. [19] used 297,933 online employee reviews from US tourism and hospitality firms to study the determinants of job satisfaction and employee turnover. They found that *leadership* and *cultural values* are better predictors of high employee satisfaction, while *career progression* is a critical predictor of employee turnover.

In our work we first perform comprehensive text analysis to extract latent satisfaction factors, and using STM distinguish these factors between former and current employees, providing a basis to infer employee turnover factors. Prior works had only used LDA, which does not support covariates, as a means to extract satisfaction factors. In our research we distinguish between factors dominant in former versus current employees by employing STM. Similarly, we also identify what factors are dominant across which sector and which company in that sector contributed to such dominance, which helps us compare topics across companies and sectors to a degree not achieved in prior work.

CHAPTER 3

DATA COLLECTION & PREPROCESSING

3.1 Data Collection

Indeed is the number one job site (https://www.reviews.com/job-sites/), and has the most listings compared to its other competitors (e.g. Glassdoor.com). Furthermore, with its easy-to-use user interface and extensive features it has become one of the best platforms for the employees to express their opinions regarding the companies that they work for (current employees) or that they previously worked for (former employees). Moreover, Indeed has one of the highest count of the number of reviews available, which are readily served for analysis.

Indeed's employee reviews were used for the purpose of our analysis. We got the approval from Indeed for using their employee reviews. They provided us with all the reviews of Fortune 50 companies. The list of Fortune 50 companies were obtained from "https://fortune.com/fortune500/".

The Fortune 500 is an annual list compiled and published by Fortune magazine that ranks 500 of the largest United States corporations by total revenue for their respective fiscal years. The list includes publicly held companies, along with privately held companies for which revenues are publicly available. The Fortune 500 is more commonly used than its subset Fortune 100 or superset Fortune 1000.

In our experiment, only Fortune 50 companies were used for analyses, as a preliminary step into analyzing employee review. Since, we wanted to understand what aspects the employees in topmost companies are satisfied with, we just used Fortune 50 companies as a first step towards understanding employee satisfaction aspects.

For our analysis purpose, only the attributes like "Review Title", "Reviewer Job Status", "Review Text", "Pros Texts", "Cons Texts", "Ratings" were used. Furthermore,

Indeed also provided a platform for users to rate in five of the most general satisfaction aspects (similar to "JDI" facets) which are Job Work/Life Balance, Compensation/Benefits, Job Security/Advancement, Management, and Job Culture. Most of the employees have given their ratings on these dimensions as well, however, only a few of them did not bother to rate in these dimensions. For our analysis, we had enough data in these dimensions as well to carry out quantitative analyses. Similarly, not all users had provided "pros" and "cons" reviews separately, however, we had enough dataset to carry out our analyses.

3.2 Data Preprocessing

We gathered the following information from *Indeed*: "Review Title", "Reviewer Job Status", "Review Text", "Pros Text", "Cons Text", and "Ratings" - both overall ratings and the five sub-dimensions sub-ratings (*5 sub-dimensions are: Work-Life Balance, Benefits, Job Advancement, Management, Culture*). Because this study is concerned with understanding the salient positive and negative aspects related to job satisfaction, we focused on the "Pros Text" and "Cons Text" in our analyses. There are a total of 675,117 total *pro* and *con* reviews combined. Among them are 344,573 *pro* reviews and 330,544 *con* reviews that were gathered. For each of the Fortune 50 companies, the following steps were taken for both the positive (*pro*) feedback and negative (*con*) feedback (see Figure 3.1 for an example of processing):

1. *Data Cleaning*: Data cleaning was done to remove the URL, @ mentions, hashtags, punctuation marks, and letter repititions.

2. *Upper to Lowercase*: Each of the terms was lowercased.

3. *Tokenization*: Each of the documents (reviews) was tokenized.

4. *Stop Word Removal*: Stop words were removed from each of the documents.

5. *Stemming*: Stemming was done on each of the tokens using the Porter Stemmer algorithm.

	Review_Text	Px_Texts	Tknz_Texts
	hour lunch, friendly co-workers.	cowork friendli hour lunch friendli_cowork hou...	['cowork', 'friendli', 'hour', 'lunch', 'frien...
	benefits, Medical, dental, myshare, 401k, stocks	benefit dental medic myshar stock	['benefit', 'dental', 'medic', 'myshar', 'stock']
	discount card . schedule 3 on 3 off	card discount schedul discount_card	['card', 'discount', 'schedul', 'discount_card']
	good break lengths and plenty of hours	break hour length plenti plenti_hour	['break', 'hour', 'length', 'plenti', 'plenti_...

Fig. 3.1: Text Preprocessing.

6. *N-Gram Creation and Addition*: Bigrams and Trigrams were generated using words that appeared together and added to the document.

7. *Stop Word Removal*: Stop words were again removed after the text had been stemmed and bigrams and trigrams generated.

8. *Pruning*: Terms that did not appear in the top 1000 unigrams, top 500 bigrams, or top 300 trigrams for each company were pruned.

We randomly sampled 1000 reviews from each company for both the positive (*pro*) and the negative (*con*) feedback, so that companies with the largest volume of reviews (e.g., Walmart) would not dominate the results. Some companies had fewer than 1000 reviews, in which case we used all reviews. The fewest reviews for a single company was 125. All reviews from each company were merged to form two large groups, one for the positive text and one for negative text. After this, each of the documents that had less than 3 terms were removed, and modeling was done on the remaining data. Thus, of the 1000 original reviews for each company, some documents didn't have enough terms and were discarded.

CHAPTER 4

PROPOSED MODEL

Figure 4.1 shows various steps of the proposed model, which consists of Data Collection, Data Pre-processing, Topic Modeling, and Topic Evaluation, which will be discussed in more detail in the following sub-sections.

4.1 Data Collection

Data Collection was discussed in detail in Section 3.1

4.2 Data Preprocessing

Data Preprocessing was discussed in detail in Section 3.2

4.3 Topic Modeling

Figure 4.2 shows various topic modeling algorithms that will be applied. We first started with a hard clustering algorithm like the k-means algorithm to get the general idea of what kind of topics will come out.

Then we moved onto soft clustering algorithms like LDA and STM. LDA has various variations and there are various papers which by incurring minor changes in the algorithm has significantly improved this algorithm. LDA's implementation in Python and R programming language varies. Python's gensim library provides gensim implementation of LDA, and mallet package provides mallet implementation of LDA in python, which had performed much better than the gensim's implementation. However, to incorporate the effects of covariates in the topic modeling, Structural Topic Modeling (STM) was most widely used in various disciplines, ranging from Business and Management to Psychological domains. So, we applied STM in order to incorporate the effect of Job Status on employee satisfaction aspects.

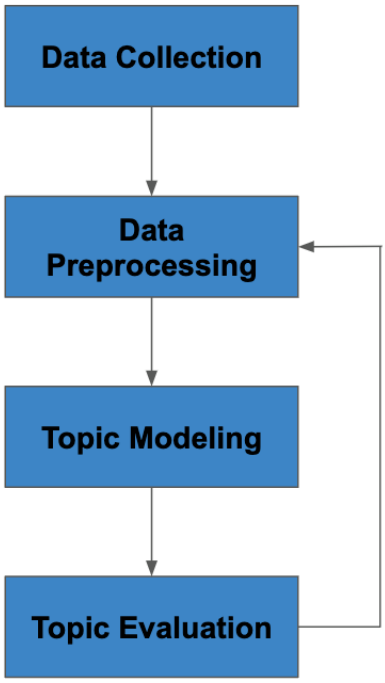


Fig. 4.1: Proposed Methodology

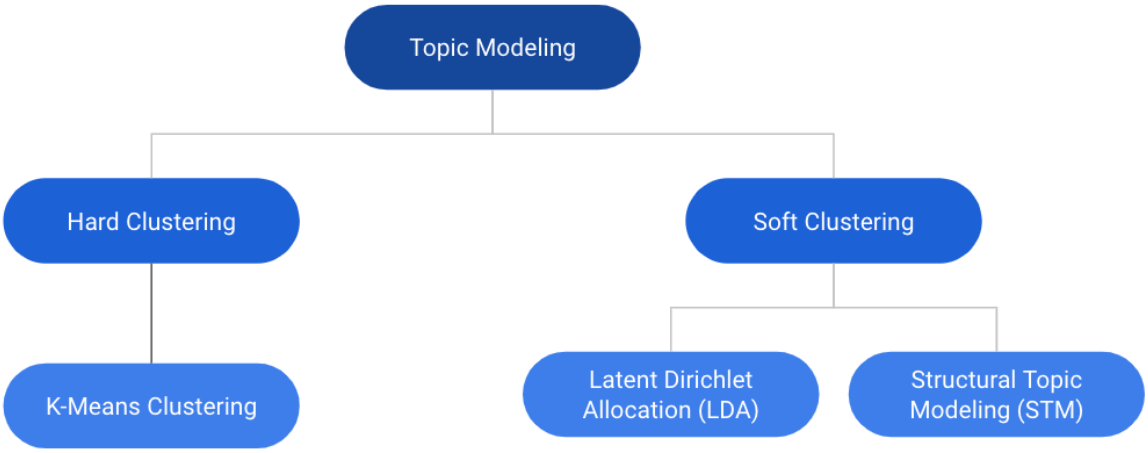


Fig. 4.2: Proposed Models

The main purpose of Topic Modeling in our research analysis is to discover topics that are most widely talked across or expressed in the employee reviews across Fortune 50 companies. This method of discovering latent topics in the employee reviews, certainly proves to be beneficial, as they help us discover latent topics that might have not been considered before now. The traditional method of figuring out which metrics best influence employee satisfaction is limited in that surveys and questionnaires rely on fixed sets of dimensions to quantify employee satisfaction. Although Indeed does provide employees 5 other domains to rate in, besides overall rating, employees have no obligation to rate, and they can express their reviews in the form of texts, as they deem necessary, since their satisfaction is not just bounded upon those 5 categories. They might have other reasons to be satisfied or dissatisfied with their current or former employer. Indeed thus provides a platform for users to freely express their employee satisfaction in whatever dimension they wish. In order to discover such latent metrics of satisfaction, we used k-means clustering algorithm, LDA, and STM.

We also go one extra step and look at what factors differentiates between the former employee and the current employee, thus reflecting upon employee turnover determinants in Fortune 50 companies. It would be really interesting to look at what factors influenced the employees to leave that company, and what's making the current employees stick with their current employer. So, a covariate of Employee Status (Former/Current) is introduced, which can help to discover an important relationship between the former and current employees, which we aim to discover with our analysis. Thus, naive LDA's model is inadequate in providing this flexibility, which was why Structural Topic Modeling (STM) was introduced, which can incorporate covariates to perform analysis on the basis of these covariates. There is no Structural Topic Modeling package provided in Python, so we used R programming to perform this task, using R's STM package.

For LDA, semantic coherence was evaluated for models with topics ranging from 2 to 40, and a graph was plotted to figure out the best topic model for our document. Similarly for STM, various metrics like Semantic Coherence, Exclusivity, and Held-Out Likelihood

were used to evaluate the efficient number of topics.

Also we processed the document topic probability, topic term probability matrix which was outputted from our model, to analyze sector wise topic proportion and in each sector company wise topic contribution, which would help us compare topic distribution across various sectors and companies.

4.4 Topic Evaluation

Evaluation is an important issue: the unsupervised nature of topic models makes model selection difficult. Topic evaluation is a universal way to generalize the efficacy of the topic model in a way that is accurate, computationally efficient, and independent of any specific application. With this metric, we will be able to compare one model with another. Some of the topic evaluation metrics that will be used in this paper to evaluate and compare various models are presented in the following sub-sections:

4.4.1 Topic Coherence

The evaluation of statistical topic models has traditionally been dominated by either extrinsic methods (i.e., using the inferred topics to perform some external task such as information retrieval (Wei and Croft, 2006 [20])) or quantitative intrinsic methods, such as computing the probability of held-out documents (Wallach et al., 2009 [21]). Recent work has focused on the evaluation of topics as semantically coherent concepts. For example, Chang et al. (2009) [22] found that the probability of held-out documents is not always a good predictor of human judgments.

Semantic coherence is a criterion developed by Mimno et al. (2011) [23] and is closely related to pointwise mutual information (Newman et al. 2010 [24]): it is maximized when the most probable words in a given topic frequently co-occur together. Mimno et al. (2011) [23] show that the metric correlates well with the human judgment of topic quality. Formally, let $D(v_i, v_j)$ be the number of times that words v_i and v_j appear together in a document. Then for a list of the M most probable words in topic k, the semantic coherence for topic k is given as

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log \left(\frac{D(v_i, v_j) + 1}{D(v_j)} \right) \tag{4.1}$$

High semantic coherence can be easily obtained by having a few topics dominated by very common words, as was pointed out in Roberts et al. (2014) [25]. That is why other metrics need to be taken into consideration for evaluating various topic models.

Since these scores are log probabilities they are negative. Large negative values indicate words that don't co-occur often; values closer to zero indicate that words tend to co-occur more often.

For LDA, the "gensim" package provides the implementation of coherence score evaluation, where the coherence score is normalized such that they have positive values and higher the coherence score indicates better the model.

For STM, however, "stm" package in R provides the implementation of semantic coherence score evaluation, where score closer to zero indicates higher coherence, and higher negative values mean that the top terms in the topic don't occur coherently.

For each model, an overall coherence score is calculated by calculating the topic coherence for each topic individually and then averaging these values.

4.4.2 Exclusivity

There are various ways to define the theme of the topics. However, the most general way of defining the core concept of the topic is by the highest probable words in a topic. However, it is not always sufficient that the most probable terms in a topic are always the best definer of a topic, as the terms can be rather frequently occurring in the whole corpus, and can be occurring equally frequently in other topics as well. So, it is also important to understand if the most probable terms in the topic in question are relatively exclusive to that particular topic only, and not common in other topics or not. So, exclusivity is basically the measure of the extent to which the top words for this topic do not appear as top words in other topics – i.e., the extent to which its top words are 'exclusive'. The value is basically the average, over each top word, of the probability of that word in the

topic divided by the sum of the probabilities of that word in all topics. The FREX metric (Bischof and Airoldi 2012 [26]; Airoldi and Bischof 2016) is used to measure the exclusivity in a way that balances word frequency. FREX is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency.

$$FREX_{k,v} = \left(\frac{\omega}{ECDF(\beta_{k,v} / \sum_{j=1}^{K} \beta_{j,v})} + \frac{1-\omega}{ECDF(\beta_{k,v})} \right)^{-1} \tag{4.2}$$

where $ECDF$ is the empirical CDF and ω is the weight which is set to .7 to favor exclusivity.

Both term frequency and term exclusivity are informative: non-exclusive words are less likely to carry topic-specific content, while infrequent words occur too rarely to form the semantic core of a topic.

Both topic coherence and exclusivity are calculated for each topic of a model and then averaged over all the topics to get the score for the model. Therefore, a model with higher exclusivity and semantic coherence is generally preferred (i.e., models with average scores towards the upper right side of the diagnostic plot).

4.4.3 Held-out Likelihood

One of the oldest evaluation methods for statistical topic modeling is held-out likelihood, developed by Wallach (2009) [21], which is basically the probability of generating unseen held-out documents given a trained model. Better models on average tend to have a higher probability of held-out documents. A better model will give rise to a higher probability of held-out documents, on average.

Held-out likelihood is generally measured by splitting the dataset into two parts: one for training, the other for testing. For LDA, a test set is a collection of unseen documents w_d, and the model is described by the topic matrix Φ and the hyperparameter α for topic-distribution of documents. The LDA parameters Θ is not taken into consideration as it represents the topic-distributions for the documents of the training set, and can therefore be ignored to compute the likelihood of unseen documents.

$$\mathcal{L}(w) = \sum_{d} \log p(w_d|\Theta, \alpha) \tag{4.3}$$

Held-out likelihood is the measure of the predictive power of the model and does not infer the latent structure of the model. Furthermore, Chang et al. (2009) [22] found that the probability of held-out documents is not always a good predictor of human judgments.

Therefore, recent work has focused more on the evaluation of topics as semantically coherence concepts, rather than as a held-out likelihood of the documents, which is why we focus more on the "semantic coherence" and "exclusivity" more than the "held-out likelihood" while evaluating the topic models.

CHAPTER 5

PRELIMINARY ANALYSIS

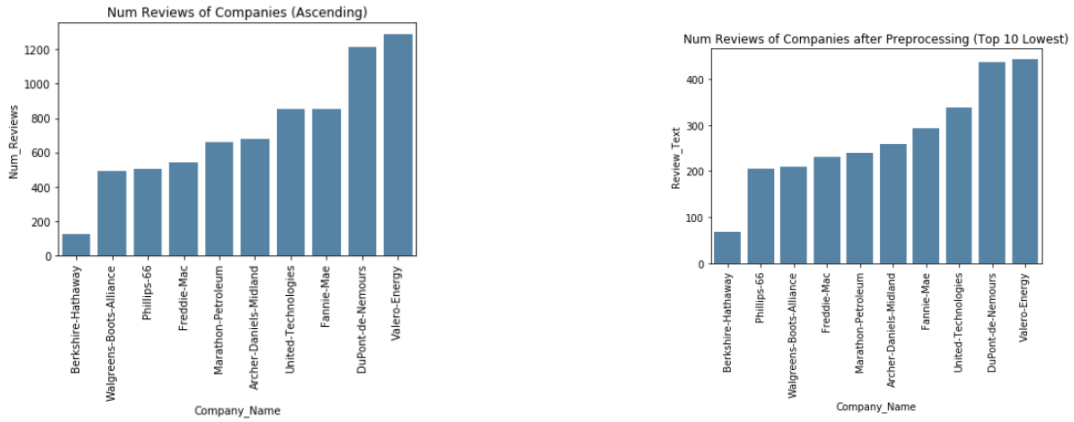
There are a total of 675117 total pros and cons reviews combined. Among them are 344573 pros reviews and 330544 cons reviews that we were provided from Indeed of Fortune 50 companies. Although we had an average of 6891 pros reviews, and 6610 cons reviews of each company. Walmart had the highest number of reviews of 159328, while, Berkshire Hathaway had the lowest number of reviews of 125. The top 10 lowest number of reviews is shown in Figure 5.1a.

Most of the companies have more than 1000 reviews (pros and cons combined).

After preprocessing the dataset, we tokenized each review, and only those documents that had more than 3 tokens in it were taken, which reduced our total review document size from "675117" to "215452". Top 10 lowest number of reviews after preprocessing and elimination is in Figure 5.1b.

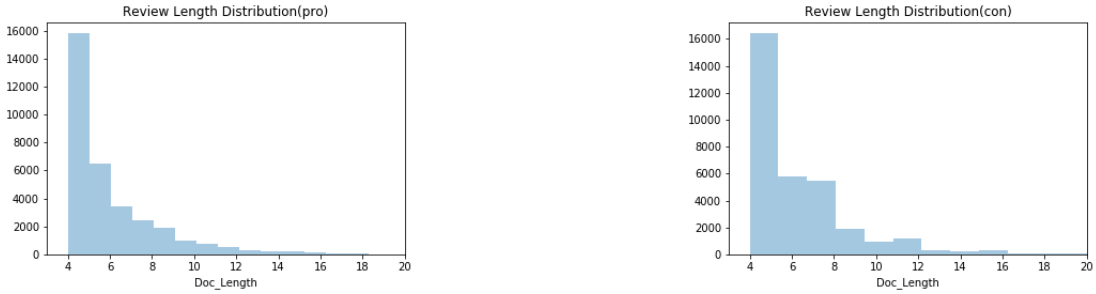
Most of the companies still had a reasonable number of reviews. We had 107954 pros document, and 107498 cons documents. We then sampled at most 1000 pros and 1000 cons reviews from each of the Fortune 50 companies, so as to distribute the effect of one company overpowering our topic model. So, we ended up with 2 datasets, one for pros and one for cons. Pros had 33624 reviews while cons had 32988 reviews. Review Length Distribution in Figure 5.2a and Figure 5.2b for pros and cons shows that the number of tokens in the preprocessed reviews are mostly less than 10. So, these reviews are quite short. The actual review might have been quite long, but after the preprocessing, the number of tokens in each review might have been reduced drastically. The lengthiest topic is actually just 72 tokens long after preprocessing.

We used these sampled datasets for further analysis.



(a) Before Preprocessing (b) After Preprocessing

Fig. 5.1: Total Number of Reviews in each company



(a) Pro (b) Con

Fig. 5.2: Document Length Distribution

CHAPTER 6

ANALYSIS & RESULTS *for technical interpretation*

6.1 Hard Clustering

6.1.1 K-Means Clustering

K-means algorithm was first applied to get a general idea about the kind of topics that would come up in the topic modeling process. Reviews generally contain multiple topics, i.e. employees generally talk about multiple topics in one single review, however, from our preliminary analysis in Figure 5.2a and Figure 5.2b, it can be observed that the length of the reviews was very small. Most of the reviews were less than 20 tokens long after pre-processing, that, it can be concluded that not many topics are talked about by a single employee. In fact, for k-means, we assume that only one topic is being talked about in a single review. Hence, with this assumption, the k-means clustering algorithm was applied to the reviews.

K-means clustering was applied on the sampled corpus which contained around 33624 reviews for "pros" and 32988 reviews for "cons". The total number of features (unique words/terms/tokens) in the corpus were 14889.

The time complexity of k-means algorithm is $O(t * k * n * d)$, where

t = num of iterations

k = num of clusters

n = num of data points (num of documents in our case)

d = num of dimensions (num of features / num of unique words in the corpus)

For our clustering analysis, the data points are basically each processed reviews, each of dimension d indicating d number of features, where features are basically the terms/tokens

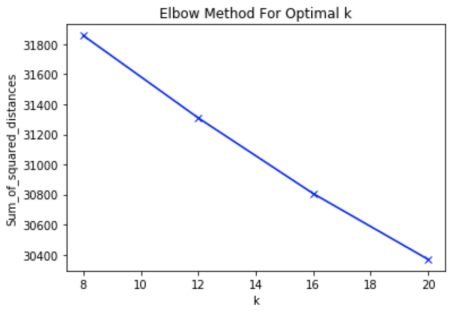


Fig. 6.1: Elbow Method for optimal "k" (Pro)

in the whole corpus (bag of reviews). Tf-idf values for each of those tokens are calculated for each of the reviews. And these d-dimensional tf-idf values becomes a single point describing a particular review.

The time complexity is pretty reasonable. However, increasing the number of clusters would lose the significance and interpretability of the topics. For this reason, the number of clusters/topics was kept relatively low, i.e. less than 20 topics.

From preliminary modeling, it was observed that setting a number of topics/clusters greater than 8 and less than 20, would produce better clustering results. Hence, setting the k (Number of Topics/Clusters) = [8, 12, 16, 20] as the number of clusters to work on, Sum of Squared Error (SSE) was calculated for each "k" value. Using "elbow method", the value of "k" to work with would be determined.

The elbow plot for "pro" is shown in Figure 6.1. A similar plot is obtained for "con".

The elbowing effect was not observed. This indicated that we needed to increase the value of k furthermore, however, increasing the number of k would significantly reduce the interpretability of the topics, so it was decided that "20" was the maximum number of clusters that we would work with. Maybe if we increased the number of topics further, the graph might have elbowed, but choosing a large number of topics would lose interpretability, so we stopped at 20 topics.

Each of the models with a value of k = [8, 12, 16, 20] was studied qualitatively since the elbow plot did not give a significant conclusion. An 8 topic model gave "less coherent" and "meaningful" topics, while a 20 topic model gave "well separated', "exclusive", and "less interpretable" topics. An 8 topic model is presented in Figure 6.2a and Figure 6.2b

for both “pros” and “cons”.

The top 10 values across the cluster centroid’s dimension were taken to describe a particular cluster. Each dimension is a token, since, we are passing n documents/reviews as a data point, and the dimension of each of those reviews is the total number of unique tokens in the whole corpus. Hence, taking the top 10 values across these dimensions for each centroid defines a particular cluster or topic.

From the figure, it can be observed that some interesting topics comes out of the clustering. For the “pro” model, topics like: *opportunity to advance/growth*, *free lunch*, *flexibility in schedule*, *management and co-workers*, *decent pay and benefits*, *meeting nice people*, and *paid vacations* comes out. Similarly for the “con” model, topics like: *poor work-life balance*, *short breaks*, *low pay and benefits*, *lack of advancement opportunities*, and *poor management* come out. There are some jargon topics and topics which are a mixture of multiple concepts as well, but overall the topics that k-means clustering has generated were cohesive and interpretable.

Hence, employees do emphasize on JDI facets for their satisfactions. Also, some other factors like *meeting nice people*, *nice co-workers*, *paid vacations*, *benefit packages*, *free lunches and longer breaks*, *schedule flexibility* are some of the other factors that influence employee satisfaction.

6.2 Soft Clustering

6.2.1 Latent Dirichlet Allocation (LDA)

LDA was applied using both python’s “gensim” and “mallet” package on both “pros” and “cons” reviews. However, better models were observed using “mallet” package, hence, the model created from “gensim” package was not used for analysis and evaluation. Both “pros” and “cons” were modeled using LDA, which will be discussed in each of the following sections.

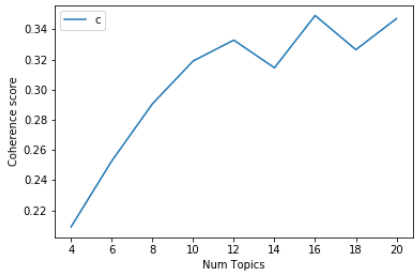
```
K-Means Clustering
Top terms per cluster (pro/8 topics):

Cluster 1:
opportun, advanc, opportun_advanc, advanc_opportun, growth, learn, benefit, career, opportun_learn, lot

Cluster 2:
free_lunch, lunch, free, sometim, free_lunch_sometim, lunch_sometim, occasion, time, occasion_free_lunch, holiday

Cluster 3:
flexibl, hour, hour_lunch, schedul, flexibl_schedul, flexibl_hour, lunch, break, 1_hour_lunch, benefit

Cluster 4:
manag, cowork, team, benefit, manag_team, support, pay, environ, cowork_manag, benefit_manag

Cluster 5:
benefit, free, employe, environ, discount, health, break, custom, food, learn

Cluster 6:
pay, benefit, pay_benefit, decent, decent_pay, benefit_pay, hour, excel, excel_pay, cowork

Cluster 7:
peopl, meet, meet_peopl, nice_peopl, nice, benefit, help, lot, help_peopl, fun

Cluster 8:
paid, vacat, time, paid_time, vacat_time, benefit, paid_vacat, holiday, sick, benefit_paid
```

(a) Pro

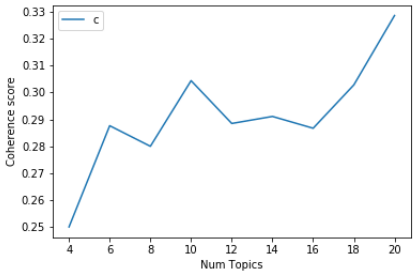
```
K-Means Clustering
Top terms per cluster (Con/8 topics):

Cluster 1:
hour, pay, employe, day, benefit, custom, chang, lack, stress, schedul

Cluster 2:
balanc, life_balanc, life, worklif_balanc, worklif, poor, manag, poor_worklif_balanc, poor_worklif, hour

Cluster 3:
short_break, short, break, hour, short_break_hour, break_hour, manag, short_break_lunch, break_lunch, lunch

Cluster 4:
low_pay, low, pay, pay_low, hour, manag, benefit, advanc, break_low_pay, manag_low_pay

Cluster 5:
advanc, opportun, career, advanc_opportun, career_advanc, opportun_advanc, lack, limit, manag, lack_advanc

Cluster 6:
time, hour, break, famili, manag, stress, day, hard, stress_time, schedul

Cluster 7:
lunch, lunch_break, short_lunch, break, short, short_lunch_break, minut, minut_lunch, 30_minut_lunch, break_lunch

Cluster 8:
manag, poor, poor_manag, bad, upper, upper_manag, lack, bad_manag, commun, hour
```

(b) Con

Fig. 6.2: Top 10 terms across each topics



(a) Pro



(b) Con

Fig. 6.3: Coherence score for various number of topics using LDA

Pro

LDA was run on sampled "pros" reviews, and the coherence score for topics from 4 to 20 was plotted, as shown in Figure 6.3a.

For LDA, the coherence score is normalized such that they are positive value. In this case, the higher the value of the coherence score, better the model. The coherence score increases with an increasing number of topics. However, we limited the number of topics to 20 as more jargon topics come up if we increase the topic number to more than 20. From the plot, a 16 topic model was chosen since it had the highest coherence score. A 16 topic model is shown in Figure 6.4

The stacked barplot presented in Figure 6.4 indicates for each topic, how dominant a particular term is. In other words, each bar in the plot for each topic is composite of multiple smaller bars of terms, size of which represents the probability of the term being in a particular topic (let's say topic 1). Hence with this graph, we can easily observe what a particular topic is about looking at the dominant terms used for that topic. Furthermore, we have included just the top 10 terms for each topic, as other terms have a lesser probability of being in that topic, and are insignificant for identifying the concept that a particular topic is conveying. A horizontal dotted line at probability .5, denotes a threshold, below which if the top ten terms of a particular topic fall is considered a bad quality topic because every term has an equal probability of being on that topic. And since a topic has a few dominant terms, it will be hard for us to identify what a topic is about, however, this is a

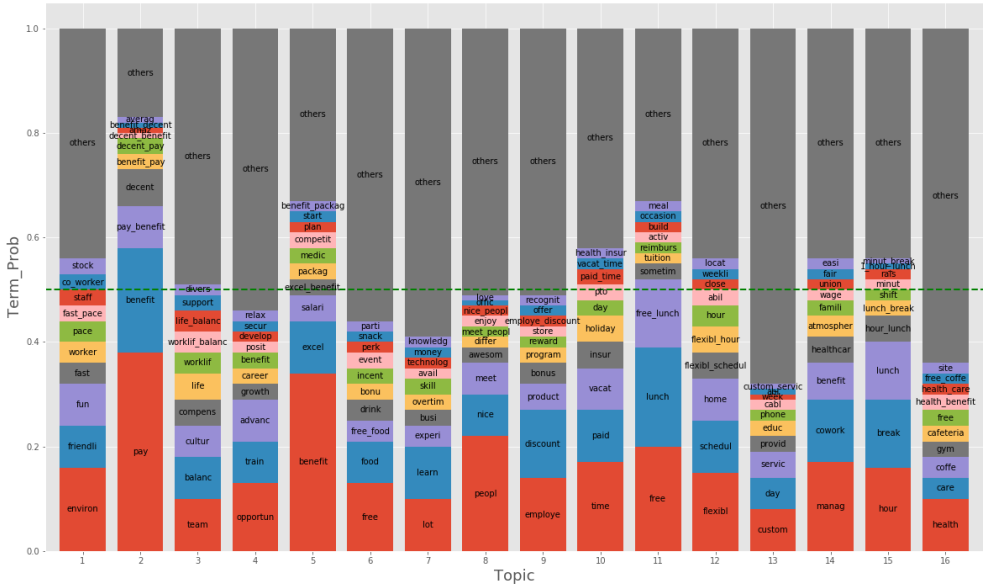


Fig. 6.4: Top 10 terms across each topics (Pro) using LDA.
(*The dotted line indicates 50% probability.*)

just qualitative interpretation of the quality of the topic, and it should be considered with caution, as there can be times when top ten terms falls below .5 threshold probability and still can be considered a good quality topics. However, in most cases, if the top ten terms falls below .5 probability threshold, it was seen that the topics were of bad quaity.

Looking at the bar plots we can observe some interesting topics are coming out from LDA as well. For the "pros" model, topics like: *fast and friendly environment*, *pay and benefits*, *work-life balance*, *advancement opportunities*, *free foods and long breaks*, *learning opportunities*, *meeting nice people*, *paid vacations*, *flexibility in schedule*, *management and leadership*, *nice co-workers*, and various *benefit packages* comes up. However, some topics have greater significance than others as the dotted line shows the threshold below which if the top 10 terms probability falls are considered poor quality topics. So, it can be inferred from the graph that topics 1, 2, 5, 10, 11, 12, 14, and 15 are relatively better topics than other topics. And it can also be qualitatively observed that these topics have their ideas easily conveyed compared to other bad quality topics. However if we look at topic

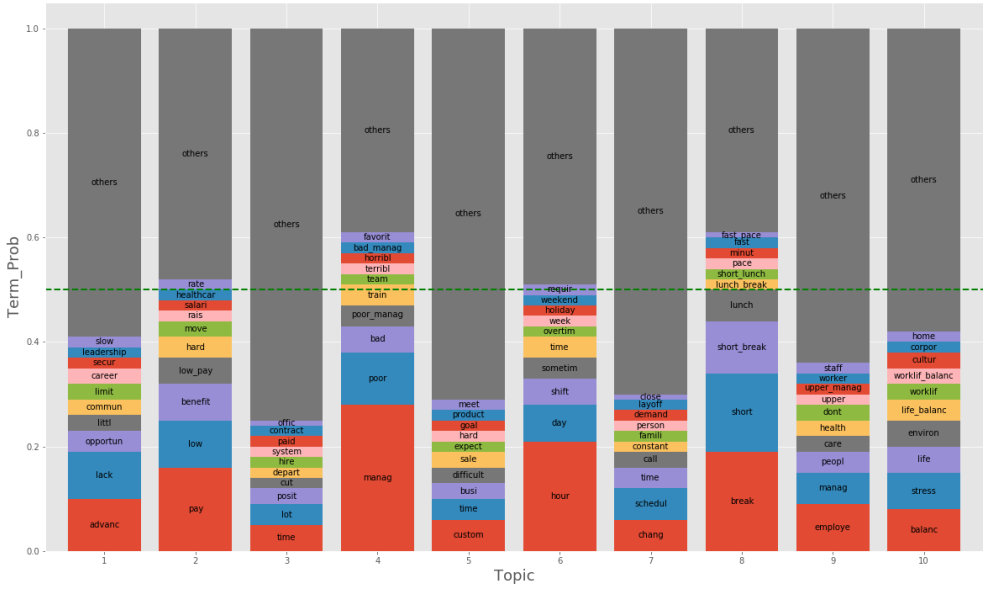


Fig. 6.5: Top 10 terms across each topics (Con) using LDA.
(The dotted line indicates 50% probability.)

8, although the top ten terms have probabilities below .5 threshold, it still is a very good topic, as it clearly conceives the idea of *meeting people*.

So, we obtained some good and well-separated topics that come from LDA model, which are consistent with the topics that we obtained from k-means. Most of the JDI facets are considered satisfaction aspects. Besides these facets, some other topics also come out such as: *nice co-workers*, *meeting nice people*, various *benefit packages*, *free foods and long breaks* and so on. Hence, focusing on these aspects will help uplift company performance in terms of employee satisfaction.

Con

LDA was then run on sampled "con" reviews, and the coherence score for topics from 4 to 20 was plotted, as shown in Figure 6.3b.

From the plot, the 10 topic model was chosen since it had the highest coherence score. A 10 topic "con" model is shown in Figure 6.5

From the figure, it can be observed that employees made negative comments about various aspects like: *lack of advancement opportunities*, *low pay and benefits*, *poor management and leadership*, *short breaks*, *poor and stressful work-life balance*, *problem in managing schedules* and so on. Compared to the "pro" model, there are many bad quality topics. Only topics 2, 4, 6 and 8 have highly probable terms, however other topics have less probable terms, thus they are less qualitative topics.

6.2.2 Structural Topic Modeling (STM)

To observe the effect of the covariate (in our case Job Status being the covariate), STM was applied on both "pros" and "cons" reviews separately, and on "pros" and "cons" reviews combined as well.

Pro

1. **Model Selection**: The evaluation score for various evaluation metrics were plotted for topics from 4 to 20, and the plot in Figure 6.6 was obtained. Various evaluation metrics were discussed in Section 4.4. Only "Semantic Coherence", "Exclusivity" and "Held-Out Likelihood" were used as an evaluation metrics to further our model selection process, as these are the only robust evaluation metrics, and other evaluation metrics did not provide much distinction.

   However, it's not quite clear from the diagnosis plot which model to select. A model with high semantic coherence, high exclusivity, and high held-out likelihood is generally preferred. In our diagnosis, we can observe that exclusivity and held-out likelihood increased and the semantic coherence decreased with the increasing number of topics. Since the diagnosis produced unclear results, 14 to 20 topic models were built individually, and qualitatively analyzed. A 16 topic model was determined to be the best model, after being qualitatively analyzed. Hence, further analyses were done using the 16 topic model.

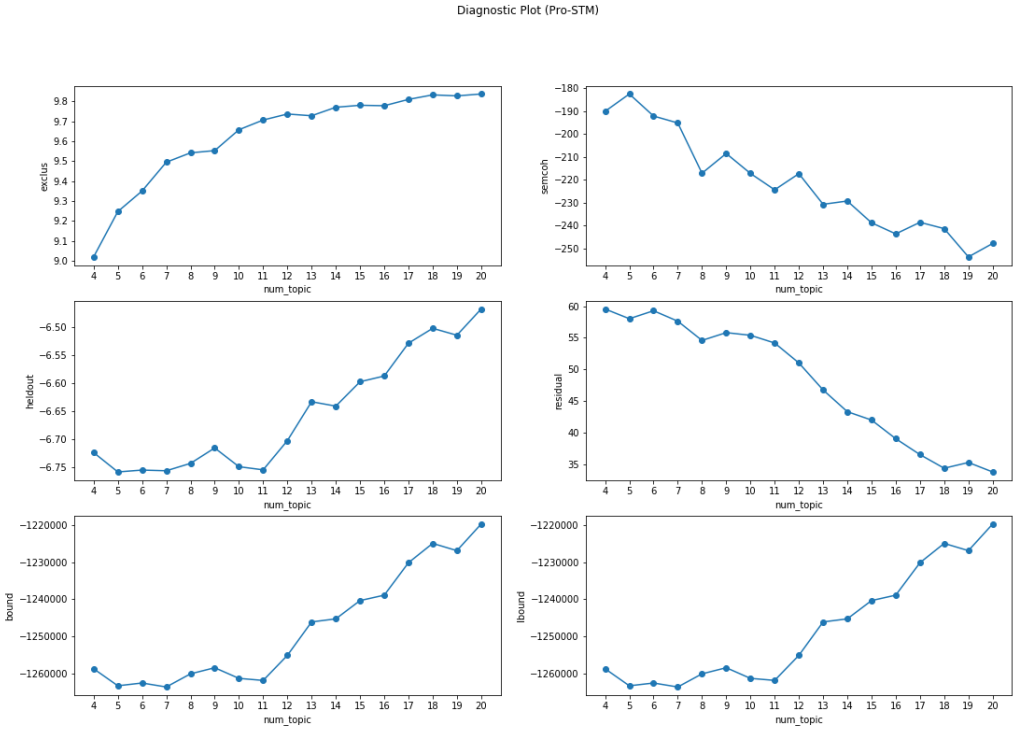


Fig. 6.6: Diagnostic Plot (Pro) for STM

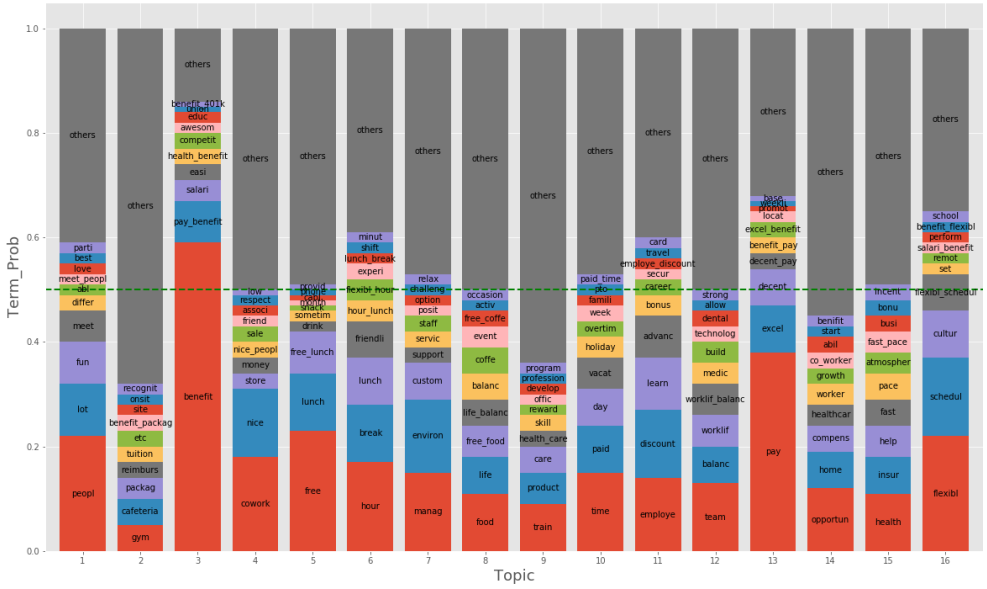


Fig. 6.7: Top 10 terms across each topics (Pro) using STM.
textit(The dotted line indicates 50% probability.)

2. **Model Definition/ Topic Labeling**: The stacked barplot in Figure 6.7 indicates the dominance of specific terms within each topic. In other words, each bar in the plot for each topic is composite of multiple smaller bars of terms, the size of which represents the probability of the term being in a particular topic. For example, Topic 2 is dominated by the "pay" and "benefit" terms, which together account for nearly 60% of the probability. Hence with this graph, we can easily observe what a particular topic is about by analyzing the term probabilities. We have included the top 10 terms for each topic, as these dominant terms are most likely to convey the topic meaning most clearly. We have qualitatively established a .5 threshold to help differentiate topic quality. If the top 10 terms compose a total of more than .5 probability, these terms dominate a majority of the topic, and the topic should be relatively strong. If the ten topics do not sum to at least .5 probability, the topic should be viewed with some caution, and may be considered weaker.

   Looking at the bar plots for *pro* in Figure 6.7 identifies that 13 of the 16 topics

(81.25%) are above the .5 threshold. Some clearly defined topics come out from this analysis. For the *pro* model, topics like: *fun people* (Topic 1), *pay and benefits* (Topic 3), *nice co-workers* (Topic 4), *free lunch* (Topic 5), *long breaks* (Topic 6), *paid time off* (Topic 10), and *flexible schedules* (Topic 16) emerge. Some topics comprise two interrelated sub-topics, such as *management and work environment* (Topic 7), *free food and work-life balance* (Topic 8), *work teams and work-life balance* (Topic 12), and *health insurance and fast-paced work* (Topic 15). Some topics have greater significance than others as the dotted line shows the threshold below which if the top 10 terms probability falls are considered poor quality topics. Thus, topics 2, 9, and 14 are relatively weaker topics than the others.

Many topics do make sense, however, there are also many jargon topics compared to "LDA" that does not make sense at all, and there are topics that are a mixture of multiple ideas as well. So we need to further investigate the quality of each topic individually, to infer which topics are good and which are not, which we have conducted in the subsequent sections.

3. **Model Evaluation**:

   (a) *Topic Proportion*: This plot shows the proportion of each topic in the whole corpus. This graph helps us determine which topic is the most dominant across all the documents, and which topics are the least dominant ones.

   Figure 6.8 shows that topic 5 relating to *free foods'* is the most dominant topic among all, and topic that about *nice co-worker* is the least dominant one, and the least talked about topic.

   (b) *Topic Quality*: Topic quality graph was plotted to observe which topics were of good quality and which ones were bad. A plot of "Semantic Coherence" vs "Exclusivity", should give us the visualization for good quality topics vs bad quality topics. Topic having higher semantic coherence would mean that the top terms in the topics are much more likely to co-occur across all the documents where that particular topic has a higher probability. Similarly, a topic having

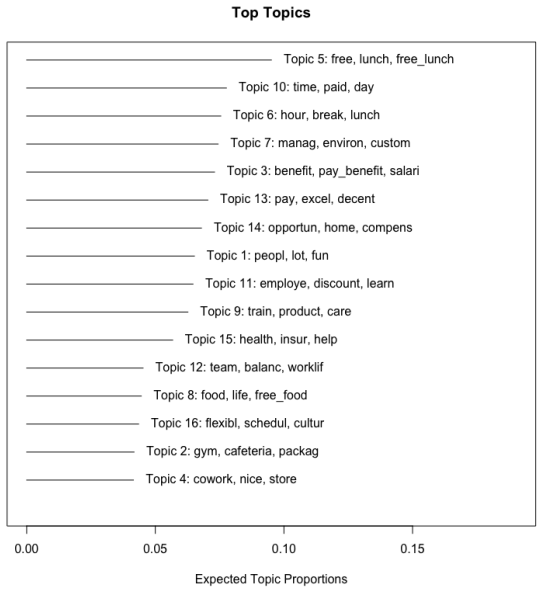


Fig. 6.8: Topic Proportion (Pro) using STM

higher exclusivity would signify that the top terms that belong to that particular topic are exclusive to that particular topic only and are not general across all the topics.

Figure 6.9 shows the topic quality plot. It can be observed that "Topic 5", which discusses *free food* has a very low exclusivity, although it has a pretty comparable semantic coherence score. So, what we can conclude from this is that the *free food* topic is not exclusive to certain documents only, rather it is general, meaning, almost every employee is talking about this issue and seems to like this a lot. So, companies should focus on this issue, to satisfy their employees and improving the employee rating. Similarly, "Topic 4" seems to have the lowest semantic coherence and can be considered as a bad quality topic. "Topic 4" talks about *nice co-workers and nice people*, however, since this topic has low semantic coherence, it can be concluded that the top terms used to describe this topic do not co-occur coherently in many documents.

(c) *Topic Correlation*: Next, we looked at the correlation graph. This graph shows the connection between those topics which are most similar to each other. Using this graph, we can understand which topics are similar to each other, and which

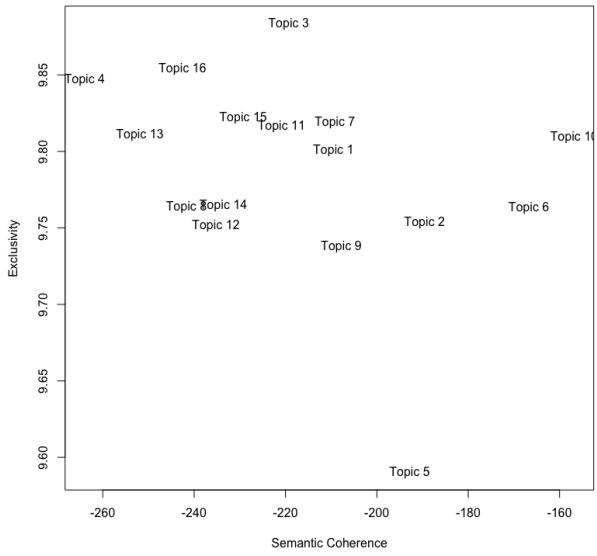


Fig. 6.9: Topic Quality (Pro) using STM

other topics are distinct from each other.

Figure 6.10 shows the topic correlation plot for our model. As can be seen from the figure, "Topic 3", "Topic 6" and "Topic 14", which all discusses *benefits* are correlated, however, it's quite surprising to have "Topic 6" correlated with "Topic 3" and "Topic 14" as "Topic 6" is observed to convey a different concept. Similarly, "Topics 11, 9, 7, 5, 10", are observed to be correlated, although those topics are talking about entirely different concepts. So it is not always wise to stick to these observations, and these observations should be consolidated with proper qualitative evaluation.

4. **Covariate Effect (Effect of Job Status on Topic Distribution)**: Figure 6.11 show the distribution of topics across the former and current employees. A vertical bar in the middle separates the topics that are most talked about by the former employees versus the topics that are most talked about by the current employees. Further is the topic from the vertical bar, more significantly a particular topic is dominant in particular covariate ("Former/Current"). From Figure 6.11, it can be observed that the current employees are satisfied by the aspects like: *pay and benefits*,

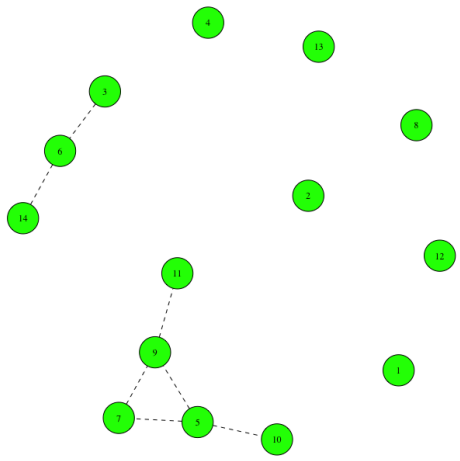


Fig. 6.10: Topic Correlation (Pro) using STM

paid vacations, *work-life balance*, and *flexible schedules*, whereas the former employees were satisfied by aspects like: *meeting people*, *nice co-workers*, *free foods*, and *good management*.

It can be observed in the figure that the scale on the x-axis is very small (one hundredth), hence, indicating very less discriminative power, so the result should be interpreted with precaution.

Con

1. **Model Selection**: Figure 6.12 shows the diagnostic plot for "cons" reviews for topics from 4 to 20.

   For "cons" as well the plot does not provide a clear distinction of which optimal model to choose, hence, qualitative analysis was performed for topics 14 to 20. A 17 topic model was chosen after being qualitatively evaluated.

2. **Model Definition/ Topic Labeling**: A 17 topic model for "con" is shown in Figure 6.13.

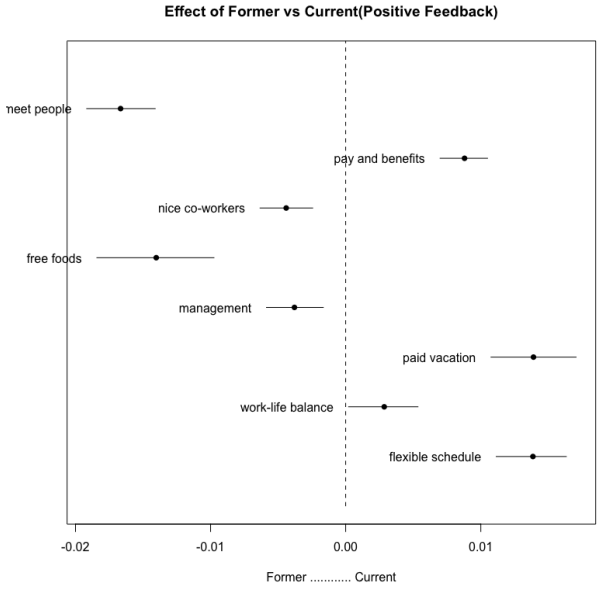


Fig. 6.11: Effect of Job Status on Topics (Pro) using STM

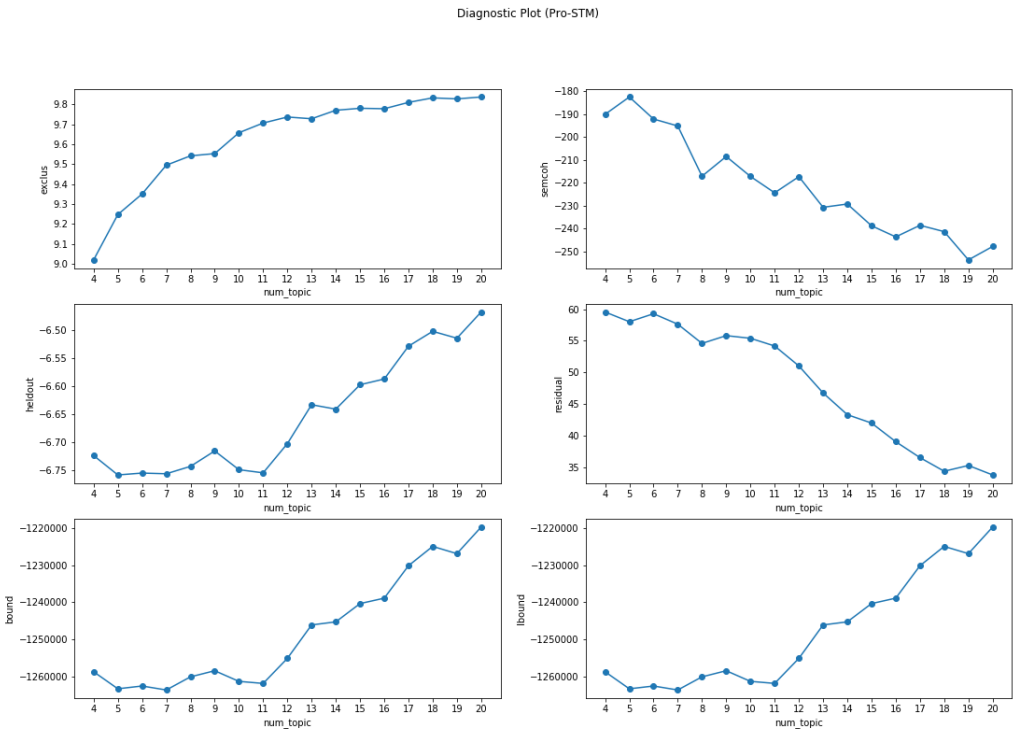


Fig. 6.12: Diagnostic Plot (Con) using STM

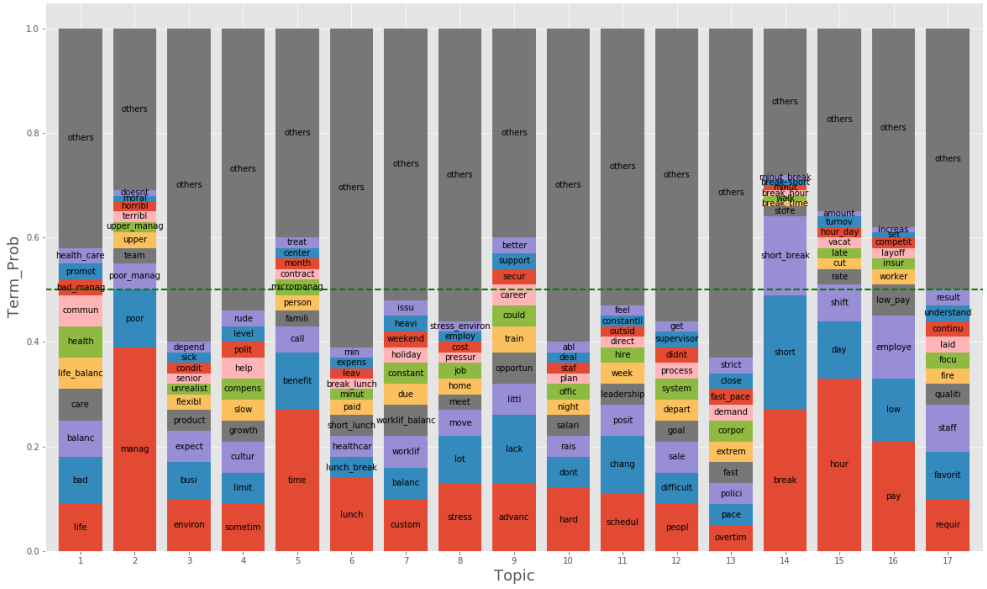


Fig. 6.13: Top 10 terms across each topics (Con) using STM.
(The dotted line indicates 50% probability.)

Figure 6.9 highlights the negative comments made by employees about various aspects, such as *poor work-life balance* (Topic 1), *poor management* (Topic 2), *stressful environment* (Topic 8), *lack of advancement opportunities* (Topic 9), *short breaks* (Topic 14), *work schedule* (Topic 15), and *low pay* (Topic 16). As with the *pro* topics, some of the *con* topics also cover two interrelated sub-topics. These include *customers and work-life balance* (Topic 7) and *hard work and low pay* (Topic 10). Compared to the "pro" model, there are many lower quality topics, as only eight of 17 topics (47%) have at least 50% probability with the top ten terms. Each topic is further investigated in more detail in subsequent sections.

3. **Model Evaluation**:

   (a) *Topic Proportion*: Figure 6.14 shows that "Topic 2" relating to *poor management* is the most dominant topic among all.

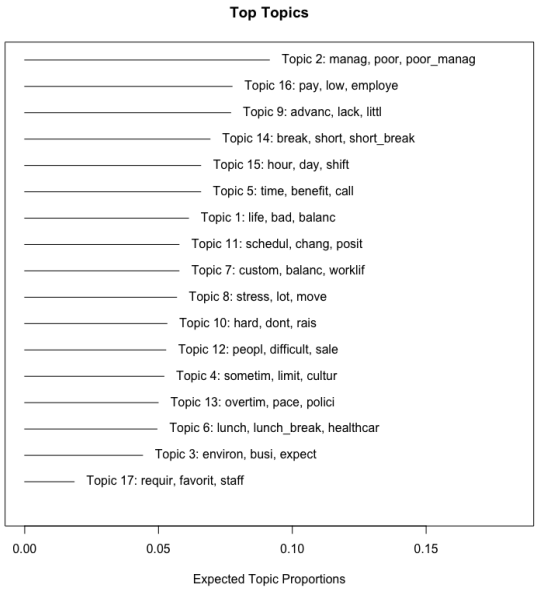


Fig. 6.14: Topic Proportion (Con) using STM

(b) *Topic Quality*: Figure 6.15 shows the topic quality plot. It can be observed that "Topic 8", which discusses lots of different ideas (esp. *stressful environment*), is a bad quality topic as it has a very low exclusivity score. "Topic 10" has a low coherence score, and does not make sense, so, it is also a bad quality topic.

(c) *Topic Correlation*: Figure 6.16 shows the topic correlation plot for our model. As can be seen from the figure, "Topic 6" and "Topic 14", which discusses *short breaks*, are correlated, however, it's quite surprising to have "Topic 3" correlated with "Topic 6 and 14", as "Topic 3" is a bad quality topic.

4. **Covariate Effect (Effect of Job Status on Topic Distribution)**: When it comes to the negative aspect of the job, it can be observed from Figure 6.17 that current employees are dissatisfied by aspects like *lack of advancement opportunity* and *low pay*, while for the former employees, the negative topics discussed tends to cluster around aspects like *management issues* and *benefits*.

   It can be observed from the figure that the scale on the x-axis is very small (one thousandth), smaller than that in the case of "pro", hence the result should be interpreted with caution. This indicates that the differentiation of topics between 2 co-variate has

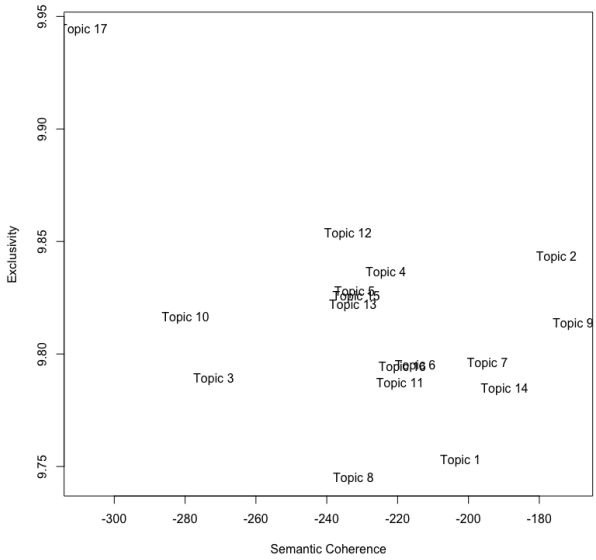


Fig. 6.15: Topic Quality (Con) using STM

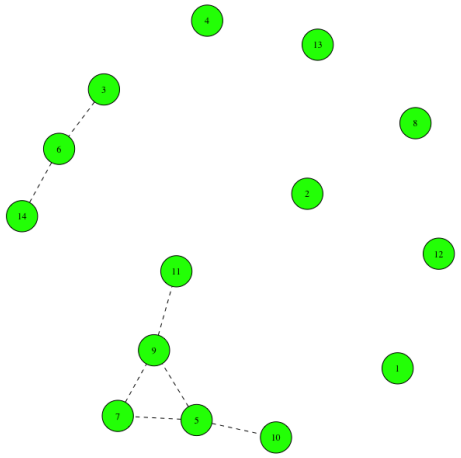


Fig. 6.16: Topic Correlation (Con) using STM

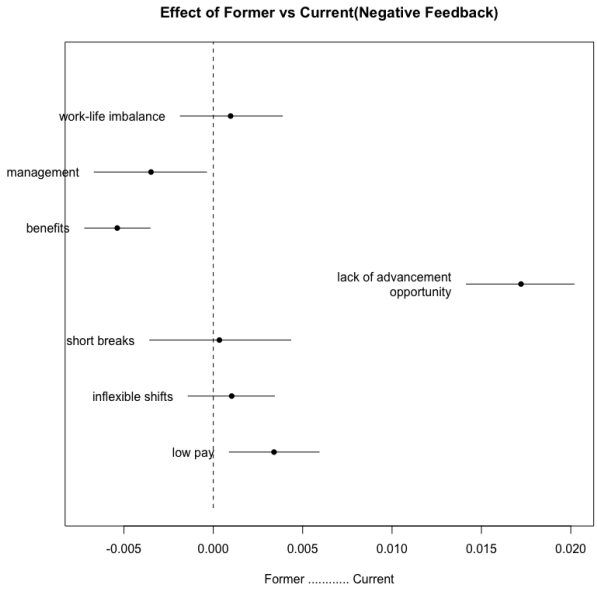


Fig. 6.17: Effect of Job Status on Topics (Con) using STM

a very low discriminative power, and any interpretation should be consolidated with qualitative analysis.

Pro Con Combined

Next, we combined the "pros" and "cons" reviews and modeled them together. Rather than analyzing them separately, we wanted to observe the model when combining them.

1. **Model Selection**: Figure 6.18 shows the diagnostic plot for the "combined" reviews for topics from 8 to 20.

   For "combined reviews" as well the plot does not provide a clear distinction of which optimal model to choose, hence, qualitative analysis was performed for topics from 4 to 20. After qualitatively analyzing each of the models, and scoring and reviewing by three experts, it was decided that the 14 topic model be chosen for further investigation and analysis since this model was well-separated, much distinct, and much more understandable. A 16 topic model was another good model, however, we decided to stick with a 14 topic model as it had a better score from the graders.

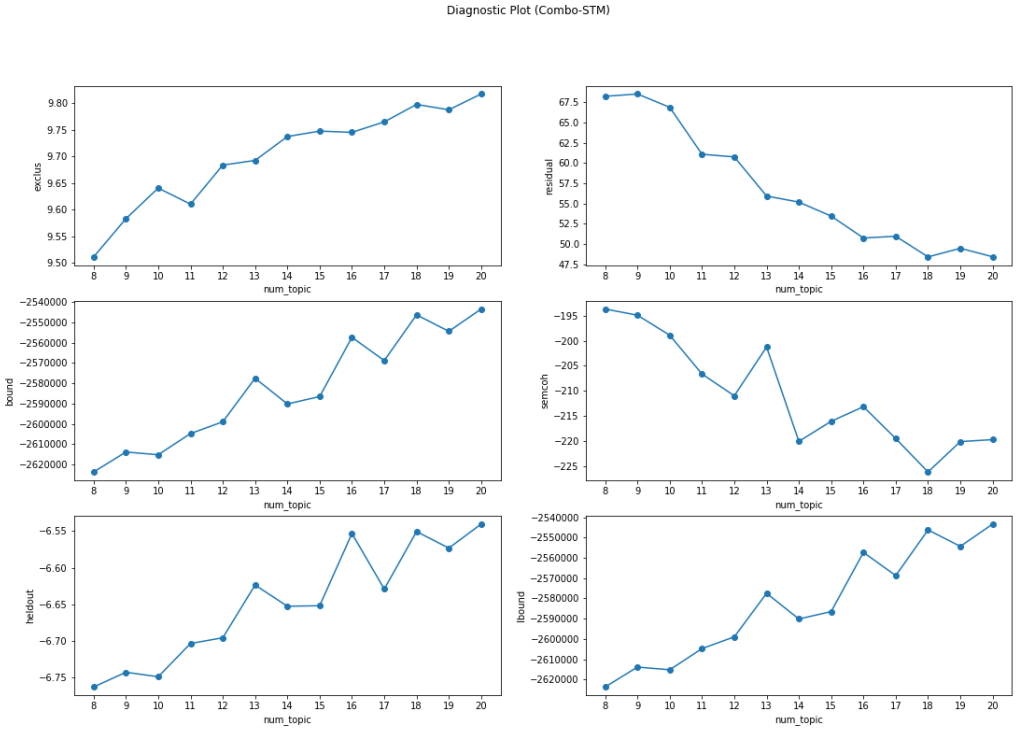


Fig. 6.18: Diagnostic Plot (Pro & Con combined) using STM

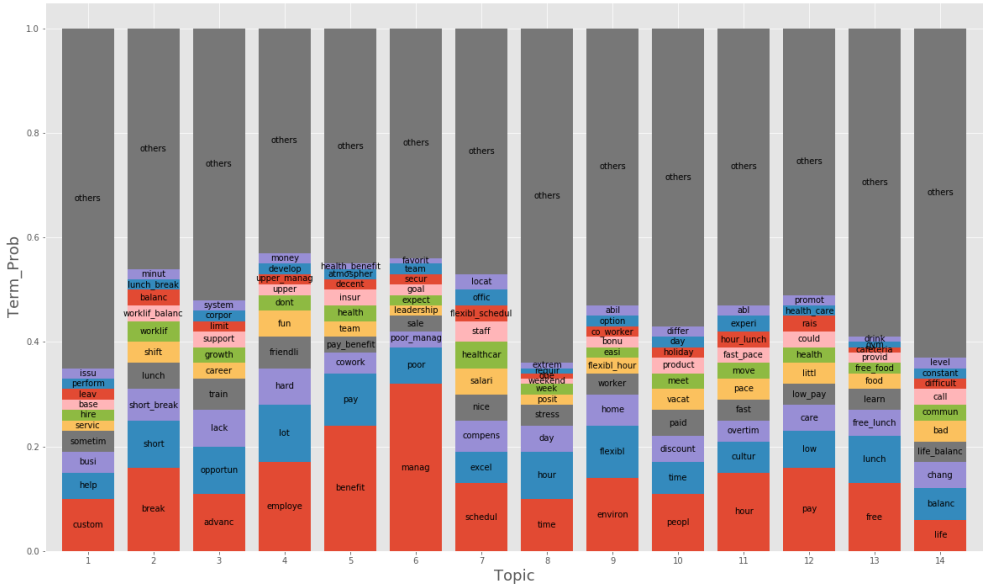


Fig. 6.19: Top 10 terms across each topics (Pro & Con combined) using STM.
(The dotted line indicates 50% probability.)

2. **Model Definition/ Topic Labeling**: A 14 topic model for "combined reviews" is shown in Figure 6.19.

   Employees are talking about topics like: free food and short/long breaks, advancement opportunity, pay and benefits, management and leadership, schedule flexibility, people and culture, and work-life balance, which is consistent with the prior models using "pros" and "cons" reviews separately. Since we mixed 2 different reviews, topics are much more unclear and are a combination of multiple ideas. These topics will be elaborated in the following subsections furthermore.

3. **Model Evaluation**:

   (a) *Topic Proportion*: Figure 6.20 shows that "Topics 2" relating to *pay and benefit* is the most dominant topic among all. And *schedule flexibility* is talked about the least.

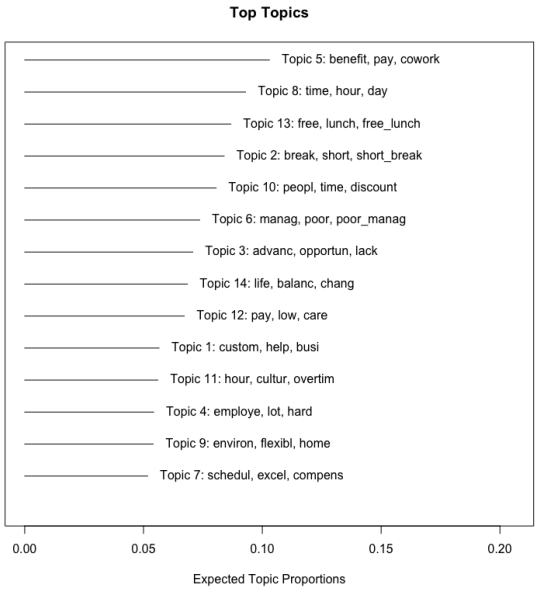


Fig. 6.20: Topic Proportion (Pro & Con combined) using STM

(b) *Topic Quality*: Figure 6.21 shows the topic quality plot. It can be observed that "Topic 13", which discusses *free food* has the lowest exclusivity, meaning employees talk about *free foods* everywhere. "Topic 4" is a bad quality topic and can be ignored since it has the lowest semantic coherence score.

(c) *Topic Correlation*: Figure 6.22 shows the topic correlation plot.

4. **Covariate Effect (Effect of Job Status on Topic Distribution)**: From Figure 6.23, it can be observed that the current employees are most satisfied with various satisfaction aspects like: *benefits and pay*, *flexible schedule*, and *meeting people*, however, they are dissatisfied by factors like: *short breaks*, *lack of opportunity* for *growth and advancement*, *low pay*, and *difficulty for life balance*.

   Similarly, the former employees were mostly satisfied by factors like: *fast pace environment and culture*, and *free foods*, however, they were dissatisfied because of aspects like: *poor management and leadership*, and *extreme stress, and overworks*.

   Although topics are easily separated with large values in the "Pro_Con" dimension, they are not separated by larger values in "Former_Current" dimension. Therefore, although topic distribution across positive and negative feedbacks are prominent, topic

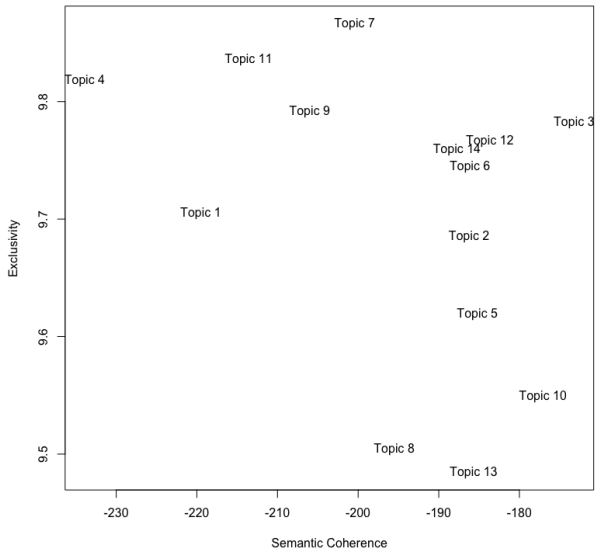


Fig. 6.21: Topic Quality (Pro Con combined) using STM

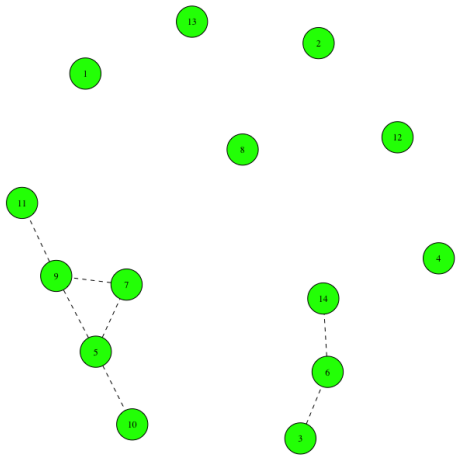


Fig. 6.22: Topic Correlation (Pro & Con combined) using STM

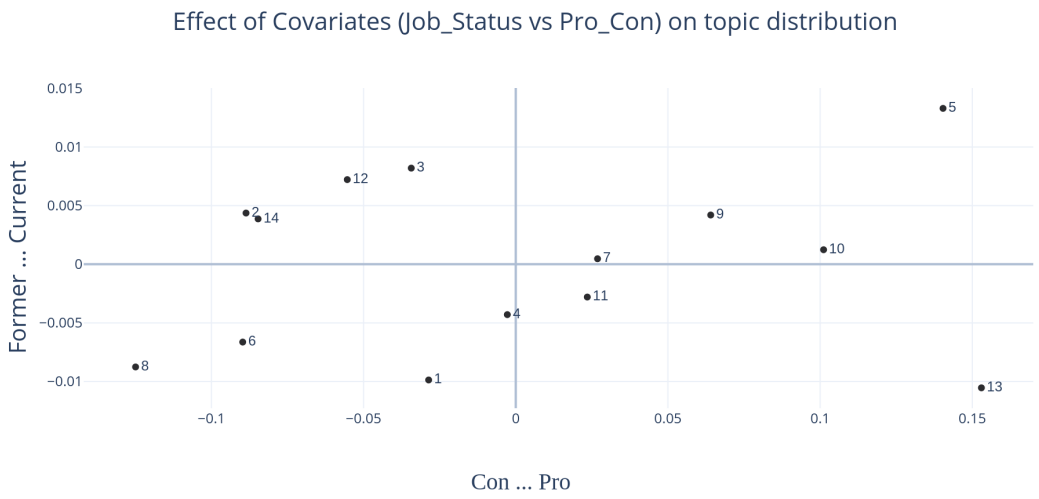


Fig. 6.23: Effect of Job Status on Topics (Pro & Con Combined) using STM

distribution across former and current employee domains is not much significant, to have any significance to drive any managerial or executive decision.

CHAPTER 7

ANALYSIS & RESULTS *for management interpretation*

7.1 K-Means Clustering

Figure 6.2a and 6.2b shows the topics that were discovered using k-means clustering on "pros" and "cons" reviews respectively. From the figure it can be seen that employees are mostly satisfied by aspects like: *opportunity to advance*, *free foods*, *schedule flexibility*, *good management*, *nice co-workers*, *decent pay and benefits*, *meeting nice people*, *vacations (paid)*, whereas, they are dissatisfied by aspects like: *inflexible schedules*, *poor work-life balance*, *short breaks*, *low pay*, *lack of opportunity to advance*, *stress*, *poor management and leadership*.

Companies, therefore, needs to focus on aspects like: *management and leadership*, *pay and benefits*, *work-life balance*, *advancement opportunity*, which are also the four of the five "JDI" facets. *Work Culture*, which is also one of the "JDI" facets, however, does not seem to be talked a lot by employees. Besides these factors, *schedule flexibility*, *meeting nice people and nice co-workers*, *free foods and long breaks* are also important areas for the employees to work on, to keep their employees happy. Similarly, companies should make the environment as *less stressful* for the employees to work.

7.2 LDA

With "LDA" as well similar topics like that from "k-means clustering" comes out, as can seen in 6.4 and 6.5 for "pros" and "cons" reviews respectively. Some other topics that comes out of "LDA" besides that from "k-means" are : *fast and friendly environment*, *career growth*, *opportunities to learn new skills/technology and gain experiences*, and various types of benefits, specifically *employee discounts and bonuses*, *paid vacations*, *free foods*, *health benefits, gym, cafeteria* and so on using "pros", and *difficult goals and expectations*,

requiring to work overtime and on weekends as well, *layoffs*, *unstable schedules*, *no care from upper management* using "cons".

So, companies need to focus on aspects like: *maintaing good environment*, *teacihng new skills/technology*, *various benefits like: discounts, bonuses, health care, gym, cafeteria, free foods* to keep their employees happy. Also, companies should *ease their expectation/goals* to make their employees less stressful, and also *maintain a stable schedule*. Upper management should provide constant care and support for the employees to feel motivated and happy.

7.3 STM

Similarly, we applied STM to incorporate "Job_Status" co-variate information. There were many bad quality topics with "STM" compared to "k-means" and "LDA". As shown in Figure 6.7, we could infer topics like: *meeting people*, *nice co-workers*, *work-life balance*, *management*, *decent pay*, *schedule flexibility*, *growth opportunities* and *various benefits like: gym, cafeteria, tuition reimbursement, health benefits, free foods, paid vacation and so on* using "pro" reviews, which is quite similar to topics discovered from "k-means" and "LDA". Similarly, using "con" reviews, topics like: *poor life balance*, *poor management and leadership*, *unrealistic expectations and extreme pressures*, *limited growth opportunities*, *short breaks*, *lack of advancement opportunities*, *schedule problems* and *low pay* were discovered as shown in Figure 6.13. Similarly, we also combined "pros" and "cons" reviews together, and formed a model by combining both the reviews together using STM. The model that we obtained had topics like: *short breaks*, *lack of advancement opportunities*, *pay and benefits*, *schedule flexibiliity*, *work-life balance*, *free foods* and *culture*, which are quite similar to prior models, however, since we mixed 2 different kind of reviews "pros" and "cons" together, the topics were much more congested and noisy.

Similarly, we applied covariate information of "Job_Status" on all 3 models using "pros", "cons" and "combined", and obtained the "Effect of Job Status on Topics" as shown in Figure 6.11, Figure 6.17 and Figure 6.23. It can be seen from those figures that "Former Employees" are motivated by factors like: *meeting new people*, *free foods*, *better management*, *nice co-workers* and so on. However, they are demotivated by factors like:

bad management, *lack of benefits*, *stressful environments* and so on. Similarly, "Current Employees" are motivated by factors like: *pay and benefits*, *paid vacations*, *schedule flexibility* and *better work-life balance*, and are demotivated by factors like: *low pay*, *lack of advancement opportunities* and *work-life imbalance*.

Therefore, companies need to focus on improving their *management and leadership*, *providing free foods and other benefits*, *creating a culture for meeting new people and nice co-workers* and *creating stress-free environment for their employee to work on* if they want to retain their employees. Similarly, to keep their current employees happy, the companies need to focus more so on *decent and competitive pays and benefits*, *creating good work-life balance for employees*, *creating opportunities for employees to advance/grow* and *maintaining a flexible schedule for employees*.

7.4 Company-wise and Sector-wise Analysis

The output from the optimal "pro" and "con" model using STM was used for this analysis. The document-topic distribution matrix and the topic-term distribution matrix was used for creating this visualization. Each of the companies belonged to a certain sector, the information which was provided by Fortune 500 website. Each review (document) belonged to a certain company, and the document-topic distribution matrix provides us with the probability of any particular document (review) having a certain probability of belonging to a particular topic. Aggregating and normalizing this over each of the Fortune 50 companies gives us the probability of the company having a particular topic, and further aggregating over sector gives us the probability of that sector having a specific topic. Similarly, what a topic is can be recalled from the topic-term distribution and taking the top 10 probable terms to define any particular topic, will help us identify the general idea of that topic.

Figure 7.1 shows the analysis result for the topic *Pay and Benefits*.

The figure shows the grouped stacked bar chart, indicating which sector talks about which topic the most, and in each sector which companies contribute to the most topic proportion. In each sector, we take the top 3 companies contributing to a particular topic.

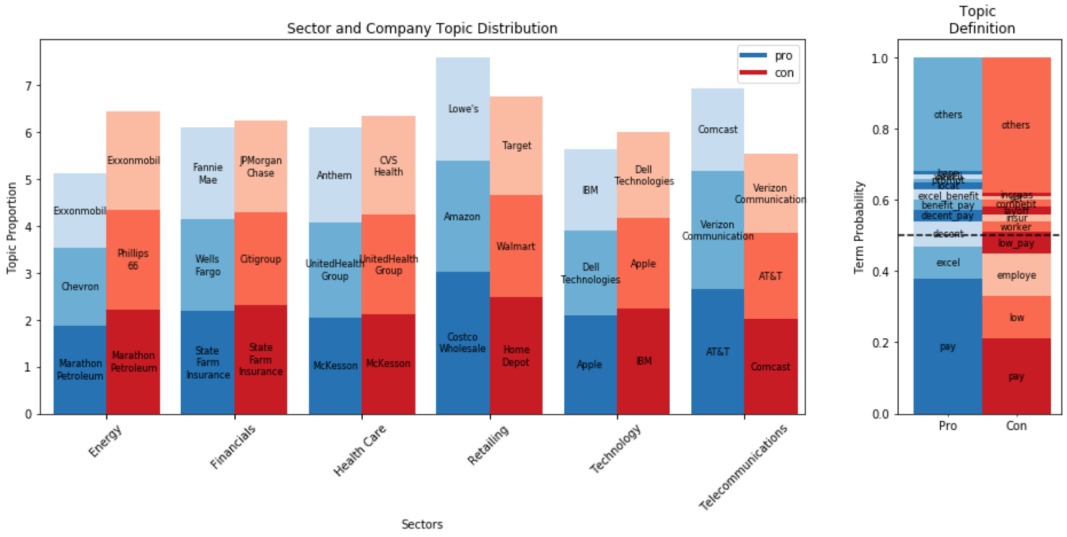


Fig. 7.1: Sector-wise Topic Distribution and Company-wise Topic Distribution

We take the top 3 companies only, for comparability between various sectors, as some sectors have less than 3 companies in the Fortune 50 company list. We ignore all those sectors that have less than 3 companies. There are 2 bars for each sector, blue one indicating topic proportion on the pro side, and the red one indicating topic proportion on the con side. On the right subplot is the topic description for positive and negative feedback. A blue bar indicates "pro" topic description, whereas a red bar indicates "con" topic description. For comparison purposes, we qualitatively choose similar "pro" and "con" topics, and place them together. Each bar is a stacked bar chart, indicating which companies contribute to what proportion of the topic for that sector. These stacked bars are arranged in the descending order of the topic proportion from bottom to top.

From Figure 7.1 it can be seen that the Retailing sector talks the most about the pay on the positive side as well as the negative side. Retailing sector employees are very satisfied with the *pay* compared to the Financial and Technology sector, which is kind of surprising, since we assume that Financial and Technology people have better earning, however, we see Retailing people commenting more on decent pay and benefits. It might be because the employees of Retailing sectors have lesser expectations compared to the other sectors. Particularly, Costco, Amazon, and Lowe's employees are more satisfied with the *pay* within the Retailing sector. Similarly, on the "con" side as well Retailing sectors employees are

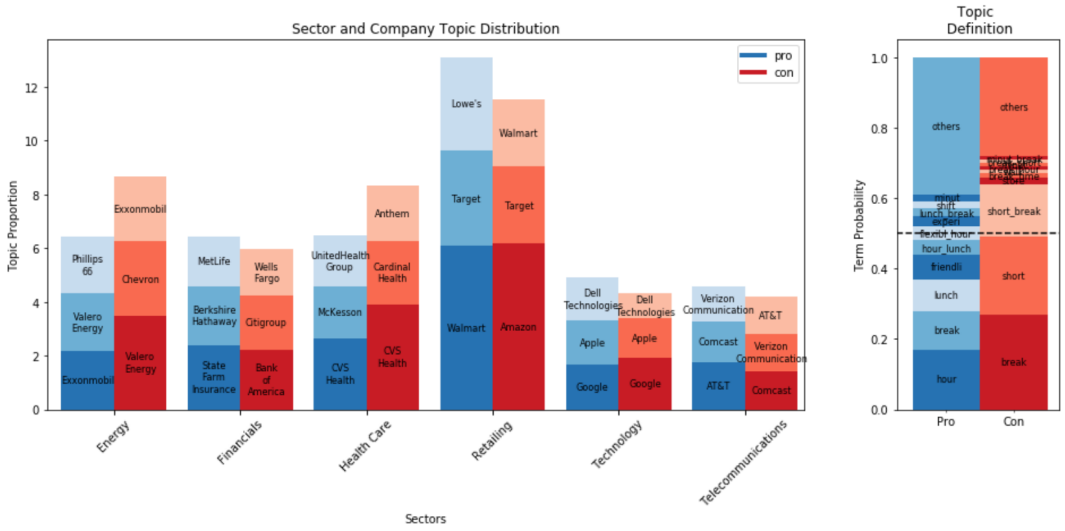


Fig. 7.2: Sector-wise Topic Distribution and Company-wise Topic Distribution

complaining, this time, contributed by companies like Home Depot, Walmart, and Target. Similarly, in the Technology sector, Apple employees are most satisfied with pay while IBM's employees are the least satisfied.

From Figure 7.3 it can be seen that again Retailing people seem to be more concerned about the duration of the lunch breaks, compared to people in other areas. Technology people don't care about the duration of the breaks. In Retailing, employees from Walmart are happy that they are given longer breaks, whereas, Amazon employees are complaining about the short breaks. Similarly, CVS health in the Health care sector, and Valero Energy in the Energy sector are dissatisfied with the shorter duration of the break. Other companies don't care about the duration of the break that much.

Similarly, from the Figure 7.3 it can be seen that Technology sector employees are the most satisfied by the *Work-life Balance*, due to companies like IBM, Dell, whereas they are also least satisfied due to companies like Microsoft and Intel. All sectors seem to have pretty even *Poor Work-life Balance*, so every sector needs to work on providing better work-life balance to their employees. In the financial sector, Freddie Mac, Fannie Mae, and Prudential Financial also seem to have a good work-life balance.

In this project, we made the following contributions:

- We proposed a novel text pre-processing approach for short reviews that of length less

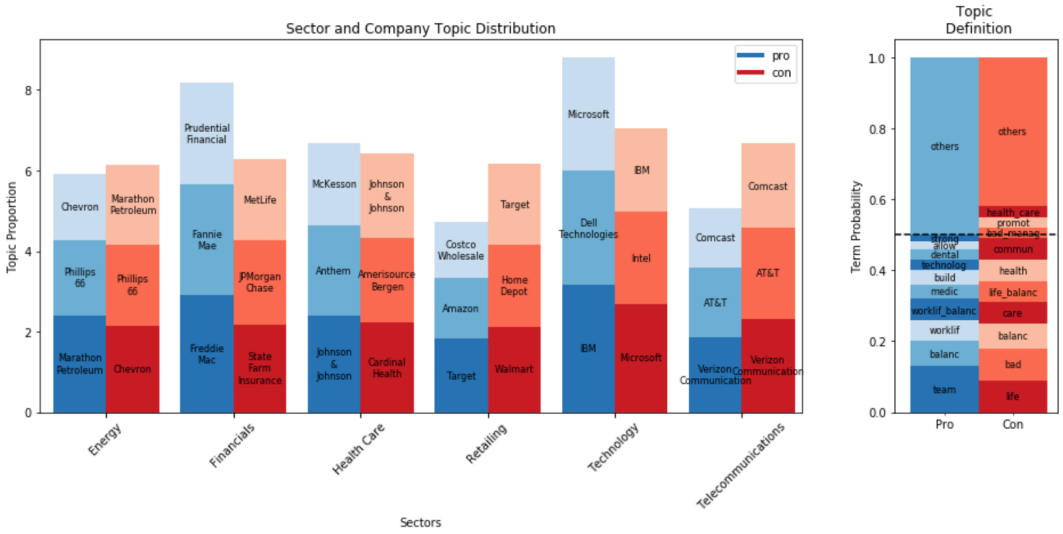


Fig. 7.3: Sector-wise Topic Distribution and Company-wise Topic Distribution

than 20 words, which is very effective for various topic modeling algorithm approaches like LDA and STM.

- We extended the work of [27] by employing the analyses on Indeed.com's reviews of Fortune 50 companies.
- We experimented and validated various machine learning approaches to mining latent topics in the large corpus of short reviews and analyze the efficacy of each model.
- We incorporated an important topic visualization to clearly distinguish between positive and negative satisfaction aspects and also between former and current employees discrepancy on satisfaction factors.
- We found that a simple k-means algorithm can perform very well when the reviews are short.
- We found that LDA and k-means perform better in topic discovery, whereas STM is needed for incorporating covariate information.
- We found that various aspects like: *free foods and long breaks*, *meeting people*, *benefits and packages* are also important besides 5 JDI facets for satisfying employees.

- We found that former employees were dissatisfied by *leadership and management issues*, *work-life imbalance*, and were mostly satisfied by aspects like: *meeting people*, *free foods* and *nice co-workers*, whereas current employees were satisfied by aspects like: *pay and benefits*, *scedule flexiility*, and dissatisfied by *lack of opportunity to advance* and indicating that *management and leadership issues*, *work-life balance issues*, need to be resolved in order to retain their current employee.

- We provided an elegant visualization to compare sector-wise and company-wise topic proportions. It was found from this visualization that Retailing sectors were the most concerned about the payments on both the positive and the negative sides. And they were also very concerned about the length of the break. However, the Technology sector was the most concerned about work-life balance.

CHAPTER 8

CONCLUSION

In this paper, we have presented novel approaches for leveraging open-source reviews to extract latent (hidden) satisfaction aspects of the employees in Fortune 50 companies, using LDA and STM. We have also analyzed these factors and distinguished between former versus current employees' contributions to the topics, and have come up with suggestions for employee turnover, which could potentially help companies improve employee retention and long-term organizational success.

It was found that *free food* and *long breaks* were some of the factors that were talked about by most of the employees, which is somewhat surprising since little prior research focuses on these specific factors. We specifically point out these factors because they are so common in the reviews that we analyzed. This could imply that if the companies focused more on these aspects they may see gains in employee satisfaction. These topics also deserve more attention in academic research.

Many topics that we discovered were similar to the JDI facets, albeit often more specific. Thus, this project highlights the factors in a particular facet that may contribute most strongly to employee satisfaction. For example: *Fun People*, *Nice Co-workers* and *Teamwork* were some of the aspects that contributed for the satisfaction of employee with co-workers, while *Difficult People* dissatisfied them. This helped us break down specific areas to focus on in satisfying employees. In the above example, hiring people that are "fun" and "nice" and that know how to work in a team may be a productive focus.

By analyzing sector-wise topic distributions, we see that the *work-life balance* topic may apply more heavily in Technology and Financial sectors and *pay* may apply more in the Retail industry, indicating the relative interests of employees in those sectors. In addition, we found that *breaks* are also topics discussed heavily by employees in the Retail sector. In a general sense, this indicates that pay and breaks may be highly salient for Retail employees,

while employees in other sectors are most focused on other aspects like personal satisfaction of a job and maintaining a work-life balance.

Our visualization of topic composition (Figures 6.7 and 6.13), along with our probability-based measure of topic quality (see the 50% threshold in the figures) are generalizable to topic discovery from unstructured text in settings including customer reviews, critic reviews, and social media comments. Our company-wise and sector-wise topic distribution analyses are also applicable to these settings, especially when topics need to be compared across attributes. And finally, the visualization of topic composition (again, Figures 6.7 and 6.13) can be a way to gain additional intuition into traditional measures of topic quality (e.g. coherence and exclusivity).

8.1 Future Work

This work opens the door to new areas of study. For example, analyzing direct correspondence of employee satisfaction to company revenues, and how employee satisfaction impacts company profits. Further investigation could be done to find the difference in the satisfaction needs between employers in low performing companies and high performing companies. Similarly, other algorithms like hierarchical clustering and other variants of LDA can be used to compare resulting models and topics. Finally, job satisfaction research may benefit from drilling down to understand what specific components drive satisfaction of the five general facets, and the JDI may benefit from updating to include other relevant job satisfaction facets like *work-life balance*).

REFERENCES

[1] P. C. Smith, L. M. Kendall, and C. L. Hulin, *The measurement of satisfaction in work and retirement: A strategy for the study of attitudes.* Rand McNally, 1969.

[2] M. T. Iaffaldano and P. M. Muchinsky, "Job satisfaction and job performance: A meta-analysis." *Psychological bulletin*, vol. 97, no. 2, p. 251, 1985.

[3] T. A. Judge, C. J. Thoresen, J. E. Bono, and G. K. Patton, "The job satisfaction–job performance relationship: A qualitative and quantitative review." *Psychological bulletin*, vol. 127, no. 3, p. 376, 2001.

[4] J. K. Harter, F. L. Schmidt, and T. L. Hayes, "Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis." *Journal of applied psychology*, vol. 87, no. 2, p. 268, 2002.

[5] N. Luo, Y. Zhou, and J. Shon, "Employee satisfaction and corporate performance: Mining employee reviews on glassdoor.com," in *ICIS*, 2016.

[6] A. J. Kinicki, F. M. McKee-Ryan, C. A. Schriesheim, and K. P. Carson, "Assessing the construct validity of the job descriptive index: A review and meta-analysis." *Journal of applied psychology*, vol. 87, no. 1, p. 14, 2002.

[7] B. Schneider and H. P. Dachler, "A note on the stability of the job descriptive index." *Journal of Applied Psychology*, vol. 63, no. 5, p. 650, 1978.

[8] S. J. Yeager, "Dimensionality of the job descriptive index," *Academy of Management Journal*, vol. 24, no. 1, pp. 205–212, 1981.

[9] M. R. Buckley, S. M. Carraher, and J. A. Cote, "Measurement issues concerning the use of inventories of job satisfaction," *Educational and Psychological Measurement*, vol. 52, no. 3, pp. 529–543, 1992.

[10] R. H. Moorman and P. M. Podsakoff, "A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research," *Journal of Occupational and Organizational Psychology*, vol. 65, no. 2, pp. 131–149, 1992.

[11] R. N. Landers, R. C. Brusso, and E. M. Auer, "Crowdsourcing job satisfaction data: Examining the construct validity of glassdoor. com ratings," *Personnel Assessment and Decisions*, vol. 5, no. 3, p. 6, 2019.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 993–1022, Mar. 2003.

[13] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: An R package for structural topic models," *Journal of Statistical Software*, vol. 91, no. 2, pp. 1–40, 2019.

[14] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *Ann. Appl. Stat.*, vol. 1, no. 1, pp. 17–35, 06 2007. [Online]. Available: https://doi.org/10.1214/07-AOAS114

[15] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'08. Arlington, Virginia, USA: AUAI Press, 2008, p. 411–418.

[16] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11. Madison, WI, USA: Omnipress, 2011, p. 1041–1048.

[17] J. Lee and J. Kang, "A study on job satisfaction factors in retention and turnover groups using dominance analysis and lda topic modeling with employee reviews on glassdoor.com," in *ICIS*, 2017.

[18] Y. Jung and Y. Suh, "Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews," *Decision Support Systems*, vol. 123, p. 113074, 06 2019.

[19] P. Stamolampros, N. Korfiatis, K. Chalvatzis, and D. Buhalis, "Job satisfaction and employee turnover determinants in high contact services: Insights from employees'online reviews," *Tourism Management*, vol. 75, 04 2019.

[20] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 178–185. [Online]. Available: https://doi.org/10.1145/1148170.1148204

[21] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 1105–1112. [Online]. Available: https://doi.org/10.1145/1553374.1553515

[22] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," vol. 32, 01 2009, pp. 288–296.

[23] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. USA: Association for Computational Linguistics, 2011, p. 262–272.

[24] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 530–539. [Online]. Available: https://www.aclweb.org/anthology/E14-1056

[25] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103

[26] J. M. Bischof and E. M. Airoldi, "Summarizing topical content with word frequency and exclusivity," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 9–16.

[27] P. Stamolampros, N. Korfiatis, K. Chalvatzis, and D. Buhalis, "Job satisfaction and employee turnover determinants in high contact services: Insights from employees'online reviews," *Tourism Management*, vol. 75, 04 2019.