공학석사학위논문

Mining Intangible Internal Resources from Employee Voice with Deep Learning

딥러닝과 직원 의견으로 파악한 조직의 무형내부자산

2021 년 2 월

서울대학교 대학원

산업공학과

박 서 영

Mining Intangible Internal Resources from Employee Voice with Deep Learning

딥러닝과 직원 의견으로 파악한 조직의 무형내부자산

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2021 년 1 월

서울대학교 대학원

산업공학과

박 서 영

박서영의 공학석사 학위논문을 인준함

2021 년 1 월

위 원 장 ______박 종 헌______(인)

부위원장 ______조 성 준______(인)

위　　원 ______이 경 식______(인)

Abstract

Mining Intangible Internal Resources from Employee Voice with Deep Learning

Elaine Seoyoung Pak

Department of Industrial Engineering

The Graduate School

Seoul National University

Intangible resources are non-physical firm resources that are critical to a firm's success. Among them, we refer to those that directly impact employee experience at work as intangible internal resources (IIR). We attempted to create a comprehensive list of IIR by applying a deep learning model to a large-scale company review dataset. We collected over 1.4 million company reviews written for S&P 500 firms from Glassdoor, one of the largest anonymous company rating and review website. Since Glassdoor reviews represent the collective employee voice, we hypothesized that prominent topics from the collective voice would represent different types of IIR. By applying a deep learning model to the review data, we discovered 24 resource types, among which 15 types such as "Atmosphere at Work," "Coworkers," and "Technological Resources" aligned with frameworks from the past literature. We then implemented a keyword extraction model to identify each firm's unique characteristics regarding different IIR types. We believe firms could utilize our findings to better understand and manage their strategic resources.

Keywords: Firm resources, Strategy, Deep learning, Text analytics

Student Number: 2019-24234

Contents

List of Tables

List of Figures

Chapter 1

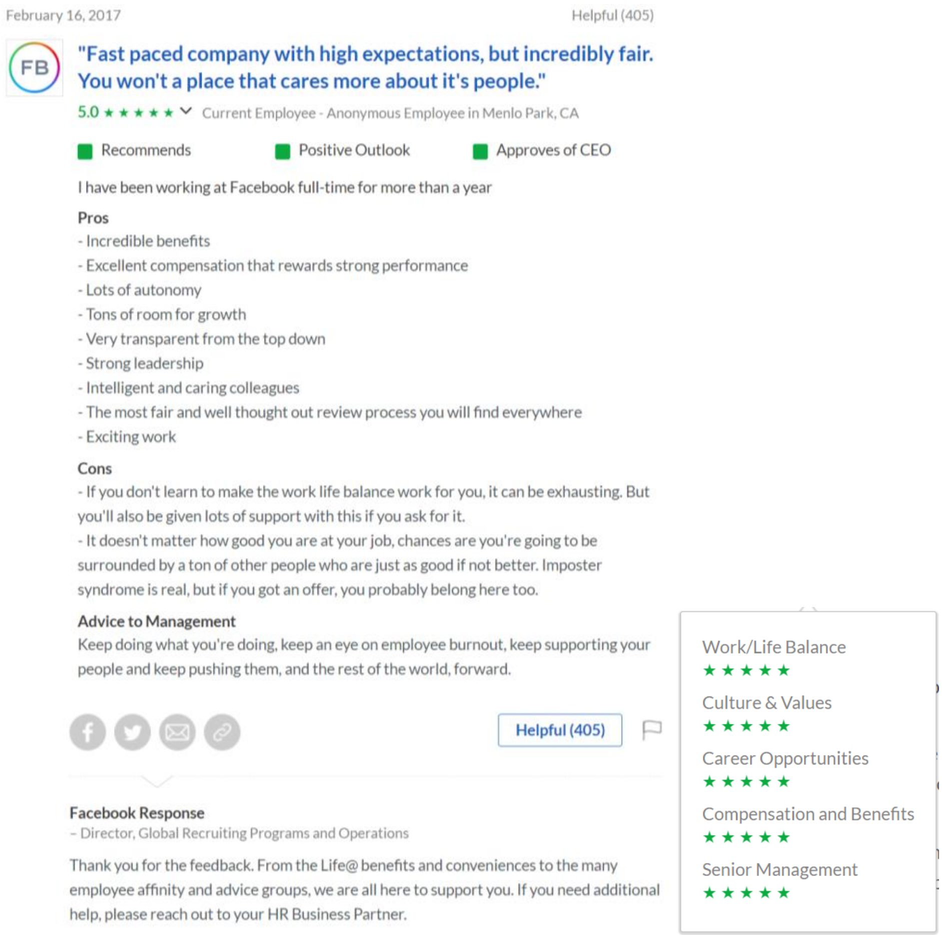
Introduction

Sustained competitive advantage is vital for firm survival, especially amid rapidly evolving modern economic landscape. Much research has been conducted in organization studies to identify the sources of firm's competitive advantage. In particular, the resource-based theory, which emerged in the 1990s, posited that firms within the same industry likely possess heterogeneous internal resources [1]; thus, organizations should understand and strategically utilize their unique internal resources to earn competitive advantage [2]. Intangible resources are firm resources that do not take a physical shape. They may include a wide variety of subjects ranging from firm reputation to organizational culture, customer relations, know-how, etc. [3]. Several works have already identified positive relationship between specific intangible resources, such as organizational reputation and celebrity, and firm performance [4, 5, 6].

But intangible resources are inherently difficult to measure and identify precisely due to its non-physical nature. In addition, the line between tangible and intangible resources is not always clear. For these reasons, to this date, there is no complete "master" list of intangible resources proposed from the organization studies perspective [7]. Moreover, previous works discussed intangible resources from a managerial perspective, often overlooking employee's voice. However, employee's opinions and perceptions are a valuable and reliable source of information that could guide firm strategies. Individual employees possess intimate knowledge about their firms, including but not limited to inner operations, rumors, business prospect, etc. Moreover, their contribution at work has been recognized as vital as early as in the 18th century by Adam Smith.

To address these points, we investigated a specific subset of firm resources that we call "intangible internal resources (IIR)," which refers to intangible resources related to employee experience. More specifically, we attempted to discover a comprehensive list of IIR types and propose ways for firms to understand and strategically manage their unique IIR. Methodologically, we explored IIR by relying on a plethora of employee-generated reviews from Glassdoor and data-driven text analytics approaches. Hence, the two specific questions that drove our research were the following: (1) Can employee-generated reviews reveal a reasonably comprehensive list of different types of IIR? (2) Is it possible to identify unique characteristics of each organization's IIR?

Glassdoor is one of the largest anonymous company rating and review websites. Each review is submitted by a former or current employee and consists of ratings and review text. Ratings are numerical values assigned to predefined categories such as *Overall* and *Work/Life Balance*, while review text is a written account of positive and negative work-related experiences. Employees could choose to additionally share their *Job Title* or *Location*. Figure 1.1 is an example review on Glassdoor left by a former employee of Facebook Inc (from hereon, Facebook).



February 16, 2017 Helpful (405)

FB "Fast paced company with high expectations, but incredibly fair. You won't a place that cares more about it's people."

5.0 ★★★★★ Current Employee - Anonymous Employee in Menlo Park, CA

🟩 Recommends 🟩 Positive Outlook 🟩 Approves of CEO

I have been working at Facebook full-time for more than a year

Pros

- Incredible benefits
- Excellent compensation that rewards strong performance
- Lots of autonomy
- Tons of room for growth
- Very transparent from the top down
- Strong leadership
- Intelligent and caring colleagues
- The most fair and well thought out review process you will find everywhere
- Exciting work

Cons

- If you don't learn to make the work life balance work for you, it can be exhausting. But you'll also be given lots of support with this if you ask for it.
- It doesn't matter how good you are at your job, chances are you're going to be surrounded by a ton of other people who are just as good if not better. Imposter syndrome is real, but if you got an offer, you probably belong here too.

Advice to Management

Keep doing what you're doing, keep an eye on employee burnout, keep supporting your people and keep pushing them, and the rest of the world, forward.

Helpful (405)

Facebook Response

– Director, Global Recruiting Programs and Operations

Thank you for the feedback. From the Life@ benefits and conveniences to the many employee affinity and advice groups, we are all here to support you. If you need additional help, please reach out to your HR Business Partner.

Work/Life Balance ★★★★★

Culture & Values ★★★★★

Career Opportunities ★★★★★

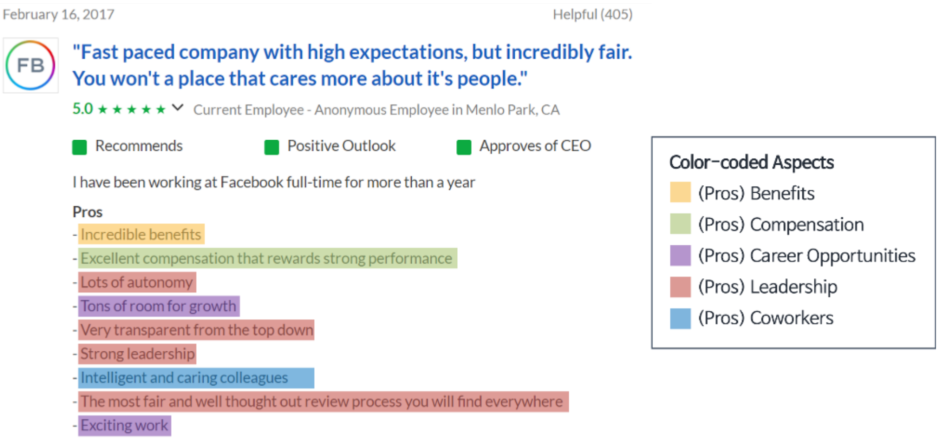
Compensation and Benefits ★★★★★

Senior Management ★★★★★

Figure 1.1: Example Glassdoor review by former Facebook employee. It includes the following textual information: *Title*, *Pros*, *Cons*, and *Advice to Management*. It also includes the following ratings: *Overall*, *Work/Life Balance*, *Culture & Values*, *Career Opportunities*, *Compensation and Benefits*, *Senior Management*, *Recommend to Friend*, *Positive Outlook*, and *Approval of CEO*. The reviewer also wrote about job location. 405 people said the review was helpful, and Facebook itself responded to the review.

Due to the multifaceted nature of Glassdoor reviews, we used them to extend our understanding of IIR. In the past, studying employee experience and opinions required costly and labor-intensive surveys. Therefore, the advent of anonymous company rating and review websites like Glassdoor presented an exciting opportunity for organizational theorists and strategists. Launched in 2008, Glassdoor boasts an unprecedented amount of large-scale data on employee opinions: tens of millions of reviews for approximately one million firms [8]. Unlike ratings, review text of Glassdoor is unstructured data. Due to the lack of tools suited for analyzing unstructured data, the previous literature on Glassdoor largely focused on its rating data. However, in the past few years, significant advances were made to deep learning models, allowing detailed and nuanced analysis of textual data.

We drew on the recent advances in deep learning to answer our two questions with Glassdoor's textual information: First, to identify different types of IIR, and second, to further identify organization's unique characteristics regarding each type. We hypothesized that different IIR types would be equivalent to some of the common topics, or *aspects*, that naturally occur within the voices of employees. For instance, Figure 1.2 shows that a single review may contain multiple aspects.



February 16, 2017 Helpful (405)

FB

"Fast paced company with high expectations, but incredibly fair. You won't a place that cares more about it's people."

5.0 ★★★★★ Current Employee - Anonymous Employee in Menlo Park, CA

🟩 Recommends 🟩 Positive Outlook 🟩 Approves of CEO

I have been working at Facebook full-time for more than a year

Pros
- Incredible benefits
- Excellent compensation that rewards strong performance
- Lots of autonomy
- Tons of room for growth
- Very transparent from the top down
- Strong leadership
- Intelligent and caring colleagues
- The most fair and well thought out review process you will find everywhere
- Exciting work

Color-coded Aspects
- 🟨 (Pros) Benefits
- 🟩 (Pros) Compensation
- 🟪 (Pros) Career Opportunities
- 🟥 (Pros) Leadership
- 🟦 (Pros) Coworkers

Figure 1.2: Example review with multiple aspects. Each aspect is color-coded.

Past works in management have identified these aspects based on exert opinions [9]. In contrast, we chose a data-driven approach. We implemented a deep-learning based model known as attention-based aspect extraction (ABAE) that automatically extracts common aspects among large-scale review data and annotates how each part of a review corresponds to such aspects [10]. With the help of the model, we identified 15 IIR types from the reviews that aligned well with intangible resources frameworks from the past works. To confirm that our model was truly reliable and trustworthy, we evaluated its outcomes with a mixed evaluation method.

Having discovered the comprehensive list of 15 IIR types, we went on to tackle the second question: Could we identify unique characteristics of each firm's IIR? To do so, we implemented a keyword extraction method called term frequency-inverse document frequency (TF-IDF) [11]. Keyword extraction is a task concerning extraction of words that best represent a given document's content compared to other documents' content. For example, in Figure 1.2, all the red parts represent sentences that discussed the *Leadership* aspect. If we selected only *Leadership* sentences from all reviews written for Facebook and treated them as a single "document," it would be possible to regard this document's keywords as equivalent to major characteristics of Facebook's *Leadership*—as seen from the collective employee perspective. When selecting keywords of Facebook's *Leadership* document, we compared Facebook's *Leadership* document to *Leadership* documents of other firms in the same industry of media and entertainment. Intuitively, comparing Facebook to firms in vastly different industries such as real estate or bank would yield less insight than comparing it to similar firms within the same industry. Figure 1.3 illustrates the overview of our approach in deriving IIR types and firm-specific characteristics.

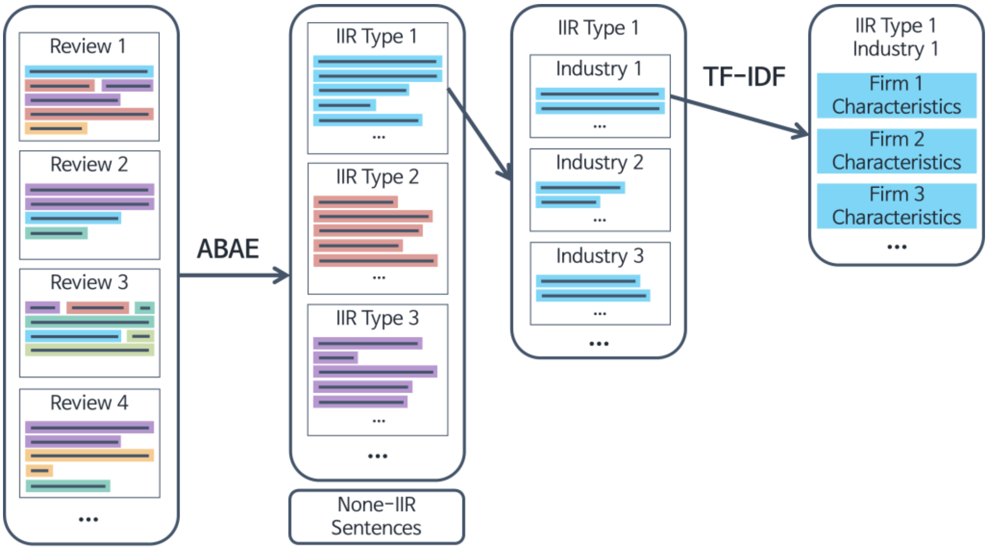


Figure 1.3: Overview of our approach to find IIR types and firm characteristics.

Our dissertation makes the following contributions. To our knowledge, this is the first attempt to identify a comprehensive set of intangible resources related to employee experience. Next, to do so, we combined the collective employee voice and deep learning techniques. Thirdly, unlike previous works, our work is completely open sourced. We released all relevant codes on our Github page[1] so that anyone interested in this topic could replicate our study and experiment with the results. Finally, our findings have managerial implications for organizations seeking to improve employee experience and gain competitive advantage. Using our models, firms can easily understand employee experience and perceptions in multiple dimensions. For instance, if a hypothetical firm A's *Culture & Values* keywords include "*fast-paced*" and "*stressful*" but lack "*supportive*," "*caring*," and "*listening*," it could improve its *Culture & Values* by launching internal campaigns to foster more supportive and family-like culture within the firm.

We organized our dissertation into six sections. In Section 2, we reviewed previous literature. In Section 3, we provided detailed explanation of our data, and in Section 4, we explained the theory and logic of our methods. In Section 5, we discussed our experiments and results. We concluded our dissertation in Section 6.

[1] https://github.com/elainespak/glassdoor_aspect_based_sentiment_analysis. The Python code for the ABAE model has been generously open sourced by the authors on their Github page: https://github.com/ruidan/Unsupervised-Aspect-Extraction [10]. We merely adjusted their codes for our research purposes. We do not claim any authorship or ownership to neither the ABAE model nor the relevant code.

Chapter 2

Literature Review

2.1 Intangible Resources

Prior to 1990s, the dominant perspective within the management literature assumed that firms in the same industry were identical in terms of strategic resources; also, since strategic resources were highly mobile, any heterogeneity among these firms would be temporary [12, 13, 14, 15, 1]. However, the resource-based theory in the 90s proposed a different perspective. It claimed not only that firms possessed heterogeneous resources, but also that the varying resources may not be perfectly mobile among firms. Therefore, to be successful, firms must strategically manage idiosyncratic and unique firm resources [1].

According to Barney [1], firm resources refer to "all assets, capabilities, organizational processes, firm attributes, information, knowledge, etc. controlled by a firm that enable the firm to conceive of and implement strategies that improve its efficiency and effectiveness." Intangible resources refer to non-physical elements among them. Because of the intensified business competition—catalyzed by globalization and deregulation—and proliferation of information technologies, intangible resources have been recognized as a critical driver of business, especially in developed economics [16].

Despite the ever-growing importance of intangible resources, in organization studies, what exactly constitutes intangible resources has been relatively unknown. In accounting, previous scholars such as Lev [16] have attempted to enumerate different types of intangible assets. In contrast, the past literature in management focused on either presenting theoretical frameworks detailing the characteristics of intangible resources or studying a specific type of intangible resource such as firm reputation or corporate culture. Notably, Hall [3] introduced the characteristics of intangible resources. Kryscynski et al. [17] created a typology of idiosyncratic resources commonly studied in the strategy literature. We did not include their typology in this paper because it was not explicitly labeled as "intangible," but their resource subcategories—reputation, mission/purpose, location, culture, etc.—indeed had intangible properties. Kaplan and Norton [18] categorized firm's intangible resources into human capital, information capital, and organization capital, listing their characteristics and illustrating how each capital could align to firm strategy (see Table 2.1; the content was directly sourced from the original paper). Diefenbach [7] presented a complete categorial system of intangible resources to systematically define and identify all intangible resources (see Table 2.2; the content was directly sourced from the original paper). These frameworks were created based on anecdotal, instead of empirical, evidence [7].

Table 2.1: Kaplan and Norton's [17] classification of intangible resources.

Category	*Description*
Human capital	- Strategic competencies – the availability of skills, talent, and know-how to perform activities required by the strategy.
Information capital	- strategic information – the availability of information systems and knowledge applications and infrastructure required to support the strategy.
Organization capital	- Culture – awareness and internalization of the shared mission, vision, and values needed to execute the strategy - Leadership – the availability of qualified leaders at all levels to mobilize the organizations toward their strategies - Alignment – alignment of goals and incentives with the strategy at all organization levels; teamwork – the sharing of knowledge and staff assets with strategic potential.

Table 2.2: Diefenbach's [7] classification of intangible resources.

Category	*Description*
Human capital	Tacit knowledge and individual competence for organizing oneself and for (inter-)acting within or with one's environment.
Social capital	Interpersonal relations and the aspects resulting from such relations for which there is no external reason (e.g. contractual or legal claim, social position).
Cultural capital	Official and informal norms, values and rules of a particular community (dyad, family, peer group, organization, society, nation, people, mankind).
Statutory capital	Person-independent positions in a social system and exclusive possibilities and responsibilities arising from or linked to such a position or role.
Informational and legal capital	Any explicit meaning of something that can be identified and demarcated individually without being necessarily internalized, shared or understood by one or more individuals
Embedded capital	Non-separable explicit knowledge embedded either in immaterial structures and processes or material goods ("artefacts").

Meanwhile, even more recent works have focused on analyzing specific types of intangible resources, such as internal reputation and organizational culture. They did not aim to comprehensively understand all types of intangible resources—rather, they deep-dived into specific types of intangible resources. Some of them specifically used large-scale data such as Glassdoor reviews. For instance, several past papers studied "employer branding," which is a strategic practice of forming a firm's identity as an employer to both external and internal audience [19]. Internal audience refers to employees. Dabirian et al.[20] applied IBM Watson's text mining capabilities on Glassdoor reviews to identify employer branding value propositions that employees considered important when evaluating employers. Similarly, Pitt et al. [21] used DICTION, a proprietary content analysis tool, to discover key drivers of employer branding and subsequently suggested appropriate brand engagement strategies. Kashive et al. [22] used SAS Visual Analytics Studio to identify employer value propositions and their managerial implications. These works illustrated the potential of Glassdoor reviews, but they all relied on proprietary analysis tools whose inner workings were black-box solutions. In other words, only the creators of these proprietary tools would know the architecture and the logic behind how the tools mine and process insights from textual data. This is problematic because how and why these tools chose certain employer branding value propositions from a given text would remain a mystery.

Another specific type of intangible resources that has been studied in-depth with Glassdoor review dataset is organizational culture. Several organizational theorists took advantage of the rich employee-generated text to discover latent factors that constitute culture at work and their managerial implications. Corritore et al. [23] applied Latent Dirichlet allocation (LDA) topic modelling method to Glassdoor reviews to measure intrapersonal and interpersonal cultural heterogeneity, and to examine its relationship with firm performance. Sull et al. [24] created a custom dictionary that mapped culture-related topics to vocabularies to measure organizational culture along these predefined topics. Both works yielded rich insights into components that make up organizational culture and showed the potential of measuring firm's performance along this new dimension. However, other potentially critical intangible resources were not examined in these studies. Also, neither work attempted to quantitatively evaluate the accuracy of findings.

Lastly, none of the works introduced in this section utilized the state-of-the-art deep learning concepts and architectures that have significantly advanced our ability to analyze textual data. Our work not only builds on the past works' theoretical frameworks and data-driven approaches but also utilizes a deep learning model to understand intangible resources in nuanced and transparent manner.

2.2 Glassdoor

Founded in 2007, Glassdoor Inc. launched its rating website Glassdoor in June 2008. Since then, Glassdoor has become one of the largest employer rating and review websites in the world. As of September 2020, it has accumulated tens of millions of reviews for around a million companies. As we noted before, employees may voluntarily and anonymously rate and write about their employers in various aspects on Glassdoor. They could even post information about salary, benefits, and interview questions. Firms, in turn, post job openings and accept applications via Glassdoor.

Glassdoor gained the reputation as a safe, reliable source of information for both employees and employers by guaranteeing anonymity of a reviewer and ensuring credibility of each review through rigorous internal process. To submit a review, one must create and activate a Glassdoor account by signing up with a social media account or email address [25]. Each reviewer may submit only one review per employer and per year [26]. Every submission is reviewed by the content management team at Glassdoor based in Ohio and Sausalito, whose job is to screen and remove posts that do not meet Glassdoor's Community Guidelines, such as those that include obscene or threatening language. Up to one out of ten submissions are rejected in this manner [27]. The screening and approval process takes up to 24 hours, after which the approved review is posted on Glassdoor. It is possible to edit the review within 30 days of submitting it, and it is also possible to delete it at any time [28].

Researchers took note of Glassdoor and its unprecedented amount of data on employee experience and perception of workplace. In finance and accounting, most academic works focused on the rating data of Glassdoor reviews and its predictive power of future events. Melián-González et al. [29] tested whether Glassdoor ratings are related to firm performance such as return over assets, operating margin, and revenue per employee. Huang et al. [30] examined corporate culture in family firms as seen from Glassdoor ratings and its implications for firm value. Farhadi and Nanda [31] studied whether Glassdoor ratings could predict post-IPO stock performance. Huang et al. [32] used Glassdoor ratings to study the effect of workplace environment's effect on auditor risk assessments and auditing outcomes.

Symitsi et al. [33] investigated the relationship between Glassdoor ratings and long-run equity returns. Hales et al. [34] examined whether rating data from Glassdoor is predictive of future corporate disclosures. Green et al. [35] showed that changes in ratings such as *Overall*, *Senior Management*, and *Career Opportunities* could be related to future financial indicators such as stock returns. Sheng [36] studied the potential of Glassdoor ratings for predicting stock returns and future trading activities by hedge funds and corporate insiders. Huang et al. [37] examined the effect of abnormal *Business Outlook* ratings from Glassdoor reviews in predicting future operating performance. Lastly, Ji et al. [38] analyzed Glassdoor rating data to investigate whether *Overall*, *Culture & Values*, and *Senior Management* ratings could indicate financial reporting quality.

Since the advent of online review platforms, various product/service reviews have received much attention across domains, including economics, marketing, psychology, and computer science. Examples of online product/service reviews include restaurant reviews, hotel reviews, movie reviews, book reviews, etc. Businesses have also regarded them as a valuable source of information for consumer opinions and brand perception. Recently, the rich literature concerning product/service reviews has actively applied the state-of-the-art deep learning models to derive even more meaningful and nuanced insights. Glassdoor's company reviews highly resemble product/service reviews in that they are both voluntarily and anonymously generated by users (employees) about products/services (firms) that they are familiar with and paid (worked) for in exchange for some value (income, personal fulfillment, etc.). However, most prior works related to Glassdoor heavily focused on the summary statistics of the rating data, and it is only in the recent years that the research involving Glassdoor's textual data began (see Section 2.1).

2.3 Unsupervised Aspect Extraction Methods

Aspect extraction is the task of detecting and extracting *aspect terms* that represent major aspects, or topics, on which opinions and sentiments are expressed [39, 40, 10]. Consider the following example: "The leadership is excellent at Microsoft." Here, "*leadership*" is the aspect term, and positive opinion ("*excellent*") is expressed on it. A group of similar aspect terms forms a single aspect. For example, we could group "*leadership*," "*management*," and "*managers*" into a single aspect called *Senior Management*.

Aspect extraction methods fall under two categories: Supervised and unsupervised learning methods. Supervised aspect extraction requires annotated dataset whose corpus is already annotated with corresponding "answer" aspect(s). See Table 2.3 for an example of a well-known annotated dataset called SentiHood [41].

Table 2.3: An example of an annotated review of the SentiHood dataset.

<table>
<tr><th colspan="3"><i>Example sentence</i>
LOCATION2 is <b>central London</b> so extremely <b>expensive</b>, LOCATION1 is often considered the <b>coolest area of London</b>.</th></tr>
<tr><td><i>Target</i></td><td><i>Aspect</i></td><td><i>Sentiment</i></td></tr>
<tr><td>LOC1</td><td>General</td><td><b>Positive</b></td></tr>
<tr><td>LOC2</td><td>General</td><td>None</td></tr>
<tr><td>LOC1</td><td>Price</td><td>None</td></tr>
<tr><td>LOC2</td><td>Price</td><td><b>Negative</b></td></tr>
<tr><td>LOC1</td><td>Safety</td><td>None</td></tr>
<tr><td>LOC2</td><td>Safety</td><td>None</td></tr>
<tr><td>LOC1</td><td>Transit-location</td><td>None</td></tr>
<tr><td>LOC2</td><td>Transit-location</td><td><b>Positive</b></td></tr>
</table>

Because supervised learning methods rely on domain-specific annotations, they tend to suffer from domain adaptation problems [10]. For example, supervised aspect extraction method built for and trained on restaurant reviews may not generalize well for Glassdoor reviews. Also, data annotation is labor-intensive and costly, and most real-world datasets are not annotated. Recent works that achieved state-of-the-art results on aspect extraction relied on a few annotated datasets like SentiHood and developed supervised learning methods specifically tailored for them [42, 43, 44].

Glassdoor reviews, like many real-world datasets, are not annotated. Each Glassdoor review consists of ratings for nine different aspects, but its text is categorized into only *Pros*, *Cons*, *Advice to Management*, and *Title* (see Figure 1.1). Thus, different parts of a review text are not annotated with more specific and meaningful aspects such as *Leadership* (see Figure 1.2). Therefore, we could not use supervised aspect extraction methods. Instead, we explored and experimented with various unsupervised aspect extraction methods. Unlike supervised learning, unsupervised learning methods have the advantage of being more versatile as they are not limited to a specific domain [10].

Early works on unsupervised aspect extraction relied on statistical models. For example, Blei et al. [45] proposed Latent Dirichlet Allocation (LDA), a topic modelling method that models topics (aspects) as distributions over words and each corpus as a mix of such topics (aspects). LDA is widely used and especially effective at extracting aspects of long documents like news articles. However, it usually yields poor performance on short, out-of-context text such as reviews. Wang et al. [46] developed Latent Aspect Rating Analysis (LARA), a probabilistic rating regression model that calculates reviewer's latent opinion on each aspect from a review text. LARA's aspect extraction process involves the following steps. First, a user must choose the total number of aspects, K, and designate several initial "seed" aspect terms for each aspect (e.g., "*compensation*" and "*benefits*" for the *Compensation and Benefits* aspect). Then, by calculating the dependencies between aspects and words via Chi-Square statistic, the model iteratively adds words with the highest dependencies to the corresponding aspect terms [47, 46]. This method is generally effective, but it also mistakenly categorizes words that are not aspect-specific, such as "*good*" and "*bad*," to specific aspects.

Recently, deep learning models proved to be highly effective at various natural language processing (NLP) tasks, including unsupervised aspect extraction. We used the attention-based aspect extraction (ABAE) model by He et al. [10] for our purposes. ABAE was specifically designed to work well on unannotated real-world review datasets. ABAE both automatically detects aspect terms from Glassdoor reviews (e.g., "*leadership*," "*management*," and "*managers*") and groups them into the most meaningful aspects (e.g., *Senior Management*). We hypothesized that these aspects would represent different types of firm resources and that most types would align with IIR types. Moreover, ABAE also predicts how different parts of a given review text corresponds to most relevant aspects, yielding results similar to Figure 1.2.

ABAE was built on some of the most important breakthrough concepts in NLP: pretrained embeddings and attention mechanism. Pretrained word embeddings were developed following distributional hypothesis, a linguistic theory that postulates that semantically similar words likely occur in similar linguistic contexts [48]. A word embedding refers to a word's numerical vector representation. Pretrained word embeddings, as the name suggests, are word embeddings calculated by training them on very large datasets so that they contain semantic and contextual meanings. Mikolov et al. [48] proposed pretrained word embeddings called word2vec, which were built on fast and effective architectures. Word2vec has been one of the most popular pretrained word embeddings to this date. Attention mechanism proposed by Bahdanau et al. [49] was another concept that revolutionized NLP research. As its name suggests, the goal of attention mechanism is to look for most important and relevant information within a given text.

We chose ABAE because it of the following two reasons. First, it overcame the shortcomings of other models like LDA and LARA and showed robust performance. ABAE's performance has been tested with two popular annotated datasets, and its superior results confirmed its effectiveness [10]. In addition, the authors generously open-sourced their codes on Github page.[2] We detailed the architecture and logic of ABAE in Section 4.1, and we listed the results from applying ABAE to Glassdoor reviews in Section 5.1.

[2] https://github.com/ruidan/Unsupervised-Aspect-Extraction

2.4 Unsupervised Keyword Extraction Methods

Keyword extraction refers to the task of automatically identifying and extracting terms that best represent a given document's topic. A document could be a news article, user-generated review, movie script, etc. In this dissertation, we defined our "document" as a collection of review sentences that discusses a single IIR type of a single organization. For example, among all reviews written by employees regarding Facebook, all sentences that spoke about Facebook's *Career Opportunities* type would constitute a single document, while sentences that discuss its *Work/Life Balance* type would constitute another document. Ultimately, keywords extracted from each document would be equivalent to words (characteristics) that best describe each firm's IIR types.

In NLP, text mining, and information retrieval, keyword extraction has been a crucial task [50]. Like aspect extraction, keyword extraction methods could be categorized into supervised and unsupervised methods. Supervised keyword extraction requires annotated datasets; namely, documents and their corresponding keywords. Therefore, supervised keyword extraction methods face similar problems as supervised aspect extraction methods; they require expensive, time-consuming data annotation and suffer from the domain adaptation problem [51, 52].

In contrast, unsupervised keyword extraction is advantageous in that it is domain independent and versatile. Most unsupervised keyword extraction methods have followed statistical approaches that consider the statistics of word occurrences within a document. For instance, the word frequency approach chooses the most frequent words as keywords [53]. But this approach tends to be flawed because oftentimes the most frequent words are not necessarily the most representative words. For instance, words like "is," "are," or "I" are often among the most frequent words in any document, but they clearly fail to represent any distinguishing characteristics. Another common approach is the word co-occurrence approach [54]. It generates co-occurrence distributions by counting co-occurrences between every word and frequent words. Based on the Chi-square measure, if a term is biased to a subset of frequent terms, then the term is regarded as a keyword.

We ultimately chose term frequency-inverse document frequency (TF-IDF) because it is highly effective at selecting unique keywords of a document when compared against other documents [11]. In Section 4.2, we explained its mechanism in detail.

Chapter 3

Glassdoor Data

3.1 Data Collection

We collected more than 1.4 million Glassdoor reviews from Glassdoor website with our custom Python web-scraper. Each review that we collected consisted of the following components. The first was a single mandatory 5-point Likert scale rating *Overall*, which represents the reviewer's overall feeling towards the employer. Next were optional 5-point Likert scale ratings of five aspects: *Compensation and Benefits*, *Work/Life Balance*, *Culture & Values*, *Senior Management*, and *Career Opportunities*. Next were optional 3-point ratings of three aspects: *Approval of CEO*, *Recommend to a Friend*, and *Business Outlook*. Again, these eight aspect-specific ratings were voluntary, so some reviews had missing values (for details, refer to Section 3.2). Next, there were three mandatory text sections, each limited to 5,000 characters: *Pros*, *Cons*, and *Title*. There was also an optional text section called *Advice to Management*. Each review also consisted of the time it was posted and the number of people who indicated that the review was *Helpful*. Additionally, each employee could choose to reveal more specific information such as job title, work location, duration of work, etc. Lastly, the employer could optionally choose to respond to the employee's review. Please refer to Figure 1.1 for an example review by a former Facebook employee.

We set the scope of our data to English reviews posted from June 2008 to June 2019 for organizations that belonged to the S&P 500 stock market index during that period for at least one quarter. Since Glassdoor website was launched in June 2008, it was the earliest period from which we could collect data, and June 2019 represented the month we began collecting data. Since the S&P 500 is used as the benchmark of the overall market, we assumed that the S&P 500 companies' reviews would serve well as the benchmark of overall Glassdoor reviews. While 771 firms satisfied this criterion, we were able to collect reviews for only 741 of them. 30 firms whose information were missing on Glassdoor were those that went bankrupt, were acquired by another firm, or merged with another firm (see Appendix 1 for details). To determine whether the collected review was written in English or not, we applied the fastText model, which is known for its robust performance in the language detection task [55]. Non-English reviews accounted for only 0.03% of the total reviews.

All in all, we obtained 1,401,126 English reviews for 741 firms.

3.2 Descriptive Statistics

In this section, we briefly described several summary statistics of our data. The average number of reviews per organization was about 1,900. We found that employees tended to write longer *Cons* than *Pros*. While the average number of characters of a text from *Cons* was 167, the average number of characters of a text from *Pros* was only 107.

Even though we collected over 1.4 million reviews, some reviews had missing values in the optional fields. The optional fields included *Advice to Management*, *Job Title*, *Location*, and all rating fields except for the *Overall* rating. Figure 3.1 illustrates the number of reviews missing for each optional rating field.

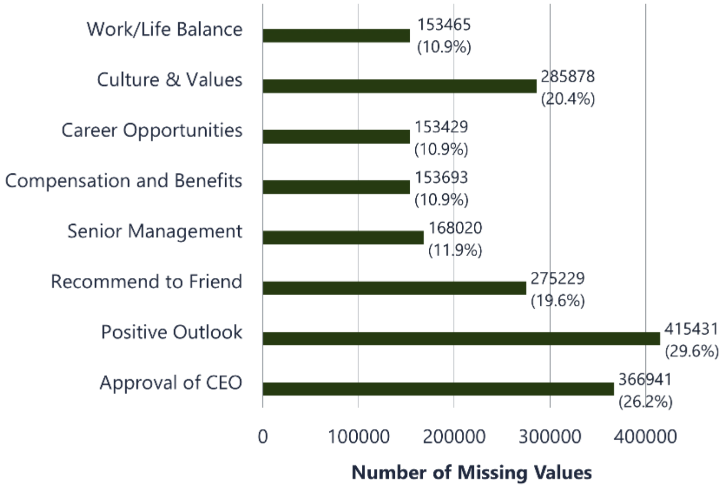


Figure 3.1: Visualization of the number of missing values for optional rating fields.

Only 58.5% of all reviews, or 820,037 reviews, did not miss any value for any rating field.

Finally, we explored whether any correlation existed among the rating fields. Figure 3.2 is a visualization of such correlations that we found from 820,037 reviews that did not have a missing rating value.

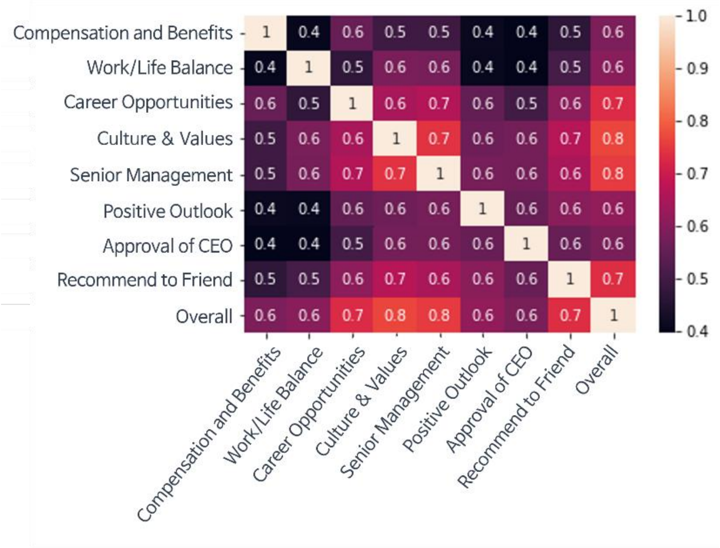


Figure 3.2: Correlation among all rating fields.

As seen from Figure 3.2, all fields have a relatively high correlation with each other, ranging from 0.4 to 0.8. Perhaps not surprisingly, the *Overall* rating maintained some of the highest correlations with other ratings. It was especially correlated with the *Culture & Values* rating and the *Senior Management* rating.

3.3 Text Preprocessing

Before implementing our models, it was necessary to transform all review text into a more structured format. We performed the following text preprocessing techniques: Sentence tokenization, stop words removal, word tokenization, lemmatization, and n-gram generation.

Sentence tokenization refers to the process of splitting a corpus into sentences. For instance, the following review, "you can expect good hike, cab facilities, shift allowance. Also opportunity to Switching technology," contains two sentences. We performed sentence tokenization with Python's sent_tokenize module from Python's nltk package [56]. Next, we removed stop words, which refer to words that appear frequently regardless of context without adding much value, such as "is," "are," "I," "we," and so on. We again used the nltk package for stop words removal. Since our models operate at a word level, we split each sentence into words through a process called word tokenization. We used the feature_extraction module from Python's sklearn package [57]. Afterwards, we conducted lemmatization, which refers to grouping words that appear in different inflected formats into their lemma. For instance, "facility" and "facilities" were grouped into "facility." We used the stem module available on nltk package [58]. Finally, we created n-grams, or continuous sequence of n words, with Python's genism package [59]. Specifically, we created bigrams or trigrams for sequences of two or three words that appeared often together. For instance, "cab" and "facility" were formed a bigram, "cab facility." Table 3.1 contains examples of Glassdoor review before and after applying text preprocessing.

Table 3.1: Examples of Glassdoor reviews before and after text preprocessing.

Before	*After*
you can expect good hike, cab facilities, shift allowance. Also opportunity to Switching technology	[['expect', 'good', 'hike', 'cab_facility', 'shift', 'allowance'], ['also', 'opportunity', 'switching', 'technology']]

Lack of Competitive pay, short staffed at the store levels, some stores have two CPP's and not enough RPP's, or sales persons and shorter Sunday hours or just be closed on Sunday. You have to do the work of 3 employees this makes customer service lag behind.	[['lack', 'competitive', 'pay', 'short_staffed', 'store', 'level', 'store', 'two', 'cpp', 'enough', 'rpp', 'sale', 'person', 'shorter', 'sunday', 'hour', 'closed', 'sunday'], ['work', 'employee', 'make', 'customer', 'service', 'lag_behind']]

Among the newly transformed text, we removed any sentence that contained less than two words since such sentences were too short to derive any insight from. For instance, a sentence such as [“okay”], which was originally “it is Okay,” was removed.

As a result of text preprocessing, 7,355,425 sentences were created. Among 7,355,425 sentences, 2,699,337 belonged to the *Pros* field, and 3,219,753 belonged to *Cons*. Given that *Cons* tended to have longer text than *Pros*, it seemed reasonable that a higher number of preprocessed sentences resulted in *Cons* than *Pros*.

Chapter 4

Unsupervised Methods for IIR and Firm Characteristic Analysis

4.1 ABAE Method for IIR Discovery

As we explained in Section 2.3, we chose the attention-based aspect extraction (ABAE) as opposed to LDA and LARA because it yielded much more meaningful and interpretable aspects (see Appendix 2.1 and 2.2 for how LDA and LARA respectively performed on Glassdoor reviews). ABAE is an unsupervised aspect extraction model that learns K number of aspect embeddings that exist in the same embedding space as word embeddings [10]. ABAE discovers aspect embeddings that can be interpreted with a list of nearest words in the embedding space. Based on this list, one could infer the name of the aspect embedding. Table 4.1 illustrates an example result by He et al. [10] from running the model on restaurant reviews to learn 14 aspect embeddings. The left column of Table 4.1 is a collection of most representative words, and the right column lists aspects manually inferred by He et al. [10]. For example, aspect terms that were closest to the first aspect embedding were "*beef*," "*duck*," "*pork*," "*mahi*," "*filet*," and "*veal*." Based on these aspect terms, the authors inferred the name of the first aspect embedding as *Main Dishes*.

Table 4.1: Example output of ABAE from the original paper.

Representative Aspect Terms	*Inferred Aspects*
beef, duck, pork, mahi, filet, veal	Main Dishes

gelato, banana, caramel, cheesecake, pudding, vanilla	Dessert
bottle, selection, cocktail, beverage, pinot, sangria	Drink
cucumber, scallion, smothered, stewed, chilli, cheddar	Ingredient
cooking, homestyle, traditional, cuisine, authentic, freshness	General
wall, lighting, ceiling, wood, lounge, floor	Physical Ambience
intimate, comfy, spacious, modern, relaxing, chic	Adjectives
waitstaff, server, staff, waitress, bartender, waiter	Staff
unprofessional, response, condescending, aggressive, behavior, rudeness	Service
charge, paid, bill, reservation, came, dollar	Price
celebrate, anniversary, wife, fiance, recently, wedding	Anecdotes
park, street, village, avenue, manhattan, brooklyn	Location
excellent, great, enjoyed, best, wonderful, fantastic	General
aged, reward, white, maison, mediocrity, principle	Other

The ABAE model consists of two steps: First, it assigns weights to important words within a sentence with an attention mechanism. Then, it reconstructs the said sentence embedding as a linear combination of aspect embedding matrix. Please refer to Table 4.2 for relevant notations.

Table 4.2: Parameters used in the ABAE model.

Variable	*Meaning*	*Parameter to be learned during training?*
s	Sentence in question	No
w_i	ith word of sentence s	No
e_{w_i}	Embedding of w_i	**Yes**
a_i	Attention weight of w_i	No
M	Transformation matrix of e_{w_i}	**Yes**
y_s	Global context embedding	No
z_s	Sentence embedding	No
T	Aspect embedding matrix	**Yes**
p_t	Probability that a sentence belongs to the tth aspect.	No
W	Weight matrix used to calculate p_t	**Yes**
b	Bias terms used to calculate p_t	**Yes**
r_s	Reconstructed sentence embedding	No
n_k	kth negative sample embedding	No
λ	The weight of orthogonal regularization	No

In the first stage, w_i is represented as e_{w_i}, which is initialized with pretrained word2vec embedding. Each word has an associated attention weight that represents the importance of the ith word in that sentence. The attention weight a_i is calculated as below:

$$a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)} \quad (4.1)$$

$$d_i = e_{w_i}^{T} \cdot \mathrm{M} \cdot y_s \quad (4.2)$$

$$y_s = \frac{1}{n}\sum_{i=1}^{n} e_{w_i} \quad (4.3)$$

Then, the input sentence z_s is calculated using the below equation:

$$z_s = \sum_{i=1}^{n} a_i \cdot e_{w_i} \quad (4.4)$$

In the second step, z_s is reconstructed to r_s as a linear combination of T and p_t:

$$r_s = \mathrm{T}^{\mathrm{T}} \cdot p_t \quad (4.5)$$

$$p_t = softmax(W \cdot z_s + b) \quad (4.6)$$

In other words, z_s goes through a dimension reduction process—from the original embedding dimension d to K, the number of aspect embeddings—which forces z_s to preserve as much information relevant to aspect embeddings as possible. To ensure this, a part of the overall training objective is set as the below contrastive max-margin objective function, which seeks to minimize the sentence reconstruction error. The objective function's goal is to make the reconstructed embedding r_s as like the original sentence z_s as possible by maximizing their inner product; simultaneously, it minimizes the inner product of r_s and m random negative samples to make sure r_s is as different from the negative samples as possible. Throughout this training process, important parameters are learned (see the last column of Table 4.2).

$$J(\theta) = \sum_{s \in D} \sum_{k=1}^{m} max(0, 1 - r_s z_s + r_s n_k) \quad (4.7)$$

$$\theta = \{E, T, M, W, b\}$$

Finally, to encourage the uniqueness of each aspect embedding, an orthogonal regularization term $U(\theta)$ is added. This encourages orthogonality among the rows of the aspect embedding matrix T and penalizes redundancy between different aspect embeddings, and is calculated as below:

$$U(\theta) = \|T_n \cdot T_n^T - I\| \qquad (4.8)$$

Hence, the final objective function is a combination of $J(\theta)$ and $\lambda U(\theta)$, where λ is a weight parameter controlling the weight of the regularization term:

$$L(\theta) = J(\theta) + \lambda U(\theta) \qquad (4.9)$$

4.2 TF-IDF Method for Firm Characteristic Discovery

TF-IDF consists of two elements: term frequency and inverse document frequency. Term frequency simply refers to how often a word appears in a document. Inverse document frequency measures how common a word appears in all documents [11]. For example, for word i in document j, (4.10) represents the TF-IDF value of i:

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (4.10)$$

The first element, $tf_{i,j}$ (term frequency of word i in document j), is a raw count of i within j. The second element, inverse document frequency, is a little more complex. It consists of N, the number of all documents, and df_i, the number of documents that contain i. Then, inverse document frequency is calculated by logarithmically scaling the inverse function of the ratio of df_i to N. In other words, the more documents contain the word i, the closer the ratio inside the logarithmic function gets to 1, making $tfidf_{i,j}$ closer to 0. Therefore, inverse document frequency penalizes words that commonly appear across documents such as "is," "are," and "I." It also favors words that uniquely appear in document j.

Looking at just term frequency or inverse document frequency creates the following problems. As we mentioned in Section 2.4, most frequent words tend to be those that frequently appear in any corpus regardless of topic. But considering only inverse document frequency engenders problems as well. Words with highest inverse document frequency tend to be extremely rare terms that fail to properly represent the overall theme of a document, such as rare names or typos.

To get the best of both worlds, it is essential to combine the two by multiplying them. Words with highest TF-IDF values of a given document capture meaningful and essential keywords that characterize the document's content.

There are variations of TF-IDF equation that differ from (4.10) according to how term frequency and inverse document frequency are calculated. For example, it is possible to log scale term frequency instead of using a raw count. It is also possible to log scale the inverse document frequency part. In this paper, however, we did not experiment with such variations. Our purpose was to show that applying the TF-IDF method after obtaining IIR types could lead to a discovery of unique firm characteristics, not to compare different variations of TF-IDF to select the best performing one. Therefore, we chose the simplest version of TF-IDF, which is represented by (4.10).

Chapter 5

Experimental Results

5.1 15 IIR Types from ABAE

We applied two separate ABAE models to the review corpus of *Pros* and the review corpus of *Cons*. As a result, we extracted a list of words related to positive employee experience and another list of words related to negative employee experience. Before running the ABAE model, one must choose the number of aspects K. We compared the results of ABAE with various numbers of K, ranging from $K = 10$ to $K = 30$. By manually comparing, we observed that $K = 12$ for both *Pros* corpus and *Cons* corpus yielded the most meaningful and coherent results (see Appendix 3 for hyperparameters). The first column of Table 5.1 represents the closest word embeddings, or aspect terms, of 12 aspect embeddings that ABAE detected from the *Pros* corpus. Similarly, the first column of Table 5.2 represents the closest aspect terms of 12 aspect embeddings that the second ABAE model found from *Cons*. When K was set to numbers from 10 to 30 other than 12, the ABAE model results were not as meaningful or coherent (see Appendix 4.1 and 4.2 for the results when $K = 10$; Appendix 4.3 and 4.4 for $K = 30$).

Having obtained the closest aspect terms for each aspect embedding, we then manually inferred aspects as He et al. [10] did in the original ABAE paper. We listed 12 inferred aspects of the *Pros* corpus in the second column of Table 5.1 and 12 inferred aspects of *Cons* in the second column of Table 5.2. Altogether, these resulted in 24 overall inferred aspects. The inferred aspects, in turn, would represent different types of firm resources. Hence, from here on, we use the word "inferred aspects" and "firm resource types" interchangeably.

Table 5.1: List of top representative words and 12 inferred aspects from *Pros* sentences. We manually assigned the names of the inferred aspects, which also represent firm resource types.

Top 40 Representative Words (=Aspect Terms)	*Inferred Aspects (=Firm Resource Types)*
benefit 401k_matching insurance eap benfits tuition_reimbursement medical_dental_vision benifits espp profit_sharing tuition_assistance medical_dental generous pension including fitness_reimbursement benifit matching adoption_assistance dental educational_assistance esop 401k includes maternity_paternity_leave tuition_reimbursment maternity_leave coverage fmla retirement match hsa plan maternity childcare_reimbursement pto stock_grant package paternity_leave reimbursement	(Pros) Benefits
employee leadership management input sincere constructive evident transparency subordinate strongly actively stakeholder equality candid inclusion organization diversity_inclusion ensuring thoughtful fairness responsive success feedback reinforced accountable direct_report eastman communicates shown genuine clearly departmental transparent demonstrated empowering communicative ensure compassion openly	(Pros) Leadership
hour schedule time scheduled weekend unpaid appointment day home weekday clock shift week remotely scheduling doctor_appointment log finish flexibly request needing workday workweek 4pm afternoon overtime calendar requested 2pm early_morning evening emergency schedual whenever late availability saturday_sunday adjust 40hrs leave	(Pros) Work Hours
pay wage salary payout compensation rate paying payouts commision hourly earning_potential considerably comission payscale salery commission comp income topped slightly cap commissioned grade minimum_wage decrease percentage capped equivalent bonus earnings payed ft considering ranking alright starting earning raise compared increment	(Pros) Compensation
breakroom catering cooky freebie salon merch themed wardrobing swag clothing giveaway merchandise gift vending_machine sample pastry movie cookout movie_ticket wardrobe gear clothes handbag voucher free complimentary candy toy furniture refreshment lobby theme_park catered_lunch jewelry breakfast_lunch kitchen goody grill ticket raffle	(Pros) Perks at Work
midtown location office centrally_located london hq conveniently_located located tampa boston geographical st_louis minnesota downtown raleigh suburban sf philadelphia atlanta nyc central pittsburgh downtown_seattle chennai headquarter dc central_london austin baltimore phoenix miami proximity irvine city spacious dallas la los_angeles env suburb	(Pros) Job Location

quit told didnt anyway wont sad eventually remember somewhere saying lie im unhappy likely exactly said truth complaining glad wait knew happened bc scared someone otherwise somebody stuck maybe might tell realize dont lazy would idiot upset disappointed incompetent stupid	(Pros) None
people colleague coworkers teammate atmosphere environment ppl collegues enviroment enviornment workmate peer enviorment mate culture coworker enthusiastic sociable engaging staff outgoing energetic environement folk cordial upbeat hardworking amazingly humble cooperative worker env alongside place hearted trench motivating surrounded personable overall	(Pros) Coworkers
research analysis analytics instrumentation analytic implementation database deployment computing instrument automation crm sophisticated application saas designing methodology testing coding construction utilizing packaging db broad_spectrum erp broad leveraging design troubleshooting software sap identifying mainframe proprietary printing deploy graphic intellectual_property manufacture architect	(Pros) Technological Resources
company compnay conglomerate financial_institution corporation telco comapny merck symantec firm pharma institution 3m amgen defense_contractor uhg ge rockwell grainger medtronic caterpillar qualcomm marketplace microsoft global_footprint cummins organisation allergan industry autodesk adobe stryker schlumberger halliburton publicly_traded biotech citigroup covidien intuit ecolab	(Pros) Company Itself
opportunity possibility opportunites opportunties opps avenue oportunities oppurtunities oppertunities oppurtunity oppotunities oportunity oppertunity chance opportunies room potential opprtunities opp scope opprotunities opportunitites upward ability oppty pathway vertically_horizontally skillsets lateral laterally horizontally path encouragement career role skillset horizontal allows diversify vertical	(Pros) Career Opportunities
ruin stunning stepped transformed ruined unprofessional reflection appeared mantra heaven de goog cry greed urgency nightmare greate succeeds tolerated unethical cozy toronto flaw dark trickle neutral forgotten satya_nadella war invented sits jeff commendable humane worst bullying harassment ensures questioned exemplary	(Pros) Atmosphere at Work

Table 5.2: List of top representative words and 12 inferred aspects from *Cons* sentences. We manually assigned the names of the inferred aspects, which also represent firm resource types.

Top 40 Representative Words (=Aspect Terms)	*Inferred Aspects (=Firm Resource Types)*
reorgs headcount_reduction orgs reorganization restructurings reorg restructures restructuring rifs reorganisation layoff restructure lay_offs knee_jerk_reaction redundancy downsizing spin_offs brain_drain realignments realignment instability upheaval rif org misstep wave	(Cons) Organization Restructuring

divestiture mass turmoil shakeup wfr disruption structuring occurring departure inertia transformation acquisition undergone analysis_paralysis	
company corporation qualcomm firm eastman hilton raytheon 3m institution compnay clorox corning disney expedia autodesk biotech boeing pepsico conglomerate merck nike financial_institution uhg general_mill microsoft hartford mckesson morgan_stanley kellogg kingsport caterpillar eaton pharma broadcom allstate staple bank_america ford ge slb	(Cons) Company Itself
business integrate creation executing operational architecture product implementation delivering execute automation deploy fundamental enterprise analysis research innovation functionality leveraging deliverable reliability feature platform leverage innovate capability establishing technology solution scalable ip design integration analytics simplify half_baked software initiative implement investing	(Cons) Business Innovation
hour weekday week work day noon workday shift peak_season monday_friday friday 3pm 7pm 30pm night scheduled midnight 8pm sat_sun twelve 4hrs overnights working sunday 5pm eight 6am 11pm 10pm 00_pm thursday summer 2am nine monday evening wk 6pm 12pm saturday weekend	(Cons) Work Hours
really think honestly laugh love gonna want know guess smile hey hate drink_kool_aid ya tell yes say mad thats something hear complain happy right scream gotta matter saying hang realise wanna believe one somebody surrounded try yell remember theyre thing	(Cons) Feeling
management managment mgmt mgt mangement managemnt leadership managament managerment managent manager mgrs egotistical mngt manger manangement aloof mgmnt asms unapproachable leader self_serving subordinate vindictive etls untrustworthy manipulative dismissive self_centered tyrannical uncaring rank_file callous lods pompous stl power_hungry condescending patronizing exec	(Cons) Leadership
pay salary wage 401k_matching compensation bonus payout profit_sharing substantially payouts measly meager paltry 5k premium offset benefit paying 10k stock_grant rsus incentive payment capped income percentage 20k raise 401k benifits commision rsu rate yearly comission comp annually deduct annual commission	(Cons) Compensation and Benefits
promotion opportunity skillset advancement mentoring upward candidate position role lateral qualified promoted career selected chosen lateral_movement fresher graduate mentor designation opportunites qualification deserving opportunties professional upward_mobility growth progression stagnate advancing within skill individual_contributor mentored recruited sponsor phd mba upward_movement path	(Cons) Career Opportunities
po register supposed submitted backup receipt confirmation verify ordered promo quiz phone instruction inform arrives password update install told ticket prompt explain approved photo signature authorization deleted submitting approve incomplete app submit appointment ask permission document wrote computer cash_register incorrect	(Cons) Operations
stressfull demanding strenuous exhausting physically_exhausting unfulfilling fast_paced unnecessarily hectic stressful beurocratic uptight	(Cons) Nature of Work

labor_intensive somtimes intense beaurocratic laborious tight_deadline repetitious overbearing repetitive grueling unpredictable tedious physically_demanding beurocracy tad burocratic multi_tasking monotonous paced unrewarding tiring pace bureacratic rigid extremly chaotic disorganised wacky	
employee passenger costumer customer associate employes baristas guest coworkers employess patient patron worker irate agent shopper pet_parent ungrateful painter teenager groomers receptionist spill staff swearing tax_preparers needy ppe impatient angry ordering animal clientele repairing resident obnoxious drive_thru bather folding_clothes caller	(Cons) Coworkers
resembles shabby describes transformed egypt disgraceful dot_com infested refers ultra_conservative linde infamous tolerates edward demoralising frightening represents cooper crumbling permeates russia gutted survives endemic lingering dismantled decimated horrifying turbulent deplorable frill plagued fort_worth reek abhorrent itw na_na_na_na reminiscent speculation racial_discrimination	(Cons) Atmosphere at Work

In addition, we obtained the most relevant aspect annotations for every sentence from *Pros* and *Cons*. In Section 4.1, we explained how the ABAE model not only discovers K aspect embedding but also predicts each input sentence's most relevant aspect by calculating p_t. Again, p_t represents the probability that the sentence belongs to the tth aspect. The ABAE automatically annotates the sentence with the aspect that has the highest probability. See Table 5.3 for examples. The first row contains an example sentence from Glassdoor reviews written for Amazon by former and current employees. More specifically, it was written in the *Pros* field instead of *Cons*. The ABAE model predicted that it is most relevant to the *(Pros) Leadership* resource type, and thus, annotated its aspect as *(Pros) Leadership*.

Table 5.3: Example sentences with annotations assigned by the ABAE model.

Organization	*Example Sentence*	*Aspect Annotation (=Firm Resource Type Annotation)*
Amazon.com Inc	Treating every employee equally.	(Pros) Leadership
Apple Inc	Strong product training for team members who choose to avail themselves of it.	(Pros) Technological Resources

Facebook Inc	- Gorgeous offices: HQ is like an amusement park and new SF office is in state of the art new building.	(Pros) Job Location
Netflix Inc	A free Netflix account at the highest plan	(Pros) Perks at Work
Expedia Group Inc	People are very kind.	(Pros) Coworkers
Microsoft Corp	Great support and benefit system for workers with family.	(Pros) Benefits
Walmart Inc	Great company to work for!	(Pros) Company Itself

The 24 firm resource types that we discovered using ABAE were much more granular and detailed than Glassdoor's predefined rating categories. To demonstrate the granularity of our 24 types, in Table 5.4, we compared them with Glassdoor's eight categories. We also compared our types with 20 categories defined by the Minnesota Satisfaction Questionnaire (MSQ) [9]. MSQ is a widely used job satisfaction scale that measures employee satisfaction level in 20 dimensions [60].

Table 5.4: Comparison of different firm resource types. 24 types found from the Glassdoor *Pros* and *Cons* text with ABAE, eight Glassdoor rating categories, and 20 MSQ categories.

Firm Resource Types from ABAE	*Glassdoor Rating Categories*	*MSQ Categories*
(Pros) Work Hours (Cons) Work Hours	Work/Life Balance	Working Conditions
(Pros) Atmosphere (Pros) Coworkers (Pros) Job Location (Cons) Atmosphere (Cons) Feeling (Cons) Nature of Work (Cons) People	Culture & Values	Co-workers Company Policies Independence Moral Values Supervision—Human Relations
(Pros) Career Opportunities (Pros) Technological Resources (Cons) Career Opportunities (Cons) Operations	Career Opportunities	Ability Utilization Achievement Advancement Recognition Supervision—Technical Variety
(Pros) Benefits (Pros) Compensation (Pros) Perks at Work	Compensation and Benefits	Compensation

(Cons) Compensation and Benefits		
(Pros) Leadership (Cons) Leadership	Senior Management	Authority
(Cons) Organization Restructuring		Security
(Cons) Business Innovation		Creativity
(Pros) Company Itself (Pros) None (Cons) Company Itself		
	Recommend to Friend	
	Positive Outlook	
	Approval of CEO	
		Activity Responsibility Social Service Social Status

Table 5.4 illustrates interesting insights. For instance, three separate aspects emerged from the *Pros* corpus regarding compensation and benefits: *(Pros) Compensation*, *(Pros) Benefits*, and *(Pros) Perks at Work*. But in contrast, Glassdoor instructs a user to rate opinion of compensation and benefits using a single rating category (*Compensation and Benefits*). This implies that most employees regarded compensation, benefits, and perks as distinct topics. Therefore, it would be reasonable to view the three as three distinct types of firm resources.

Among our 24 firm resource types, we searched for types that may belong to IIR. We did so by referring to the theoretical frameworks proposed by Kaplan and Norton [17] and Diefenbach [7]. As Table 5.5 indicates, except for *(Pros) Company Itself*, *(Cons) Company Itself*, and *(Pros) None*, all other types aligned with prior works' categories.

Table 5.5: Comparison with previous intangible resources frameworks.

Kaplan and Norton's Strategy Map	*Diefenbach's Categorial System*	*Firm Resource Types from ABAE*
Information Capital	Informational and legal capital	(Pros) Technological Resources (Cons) Business Innovation
	Embedded capital	(Cons) Operations

		(Cons) Organization Restructuring
Organization Capital	Social Capital	(Pros) Coworkers (Cons) People
	Statutory capital	(Pros) Career Opportunities (Pros) Leadership (Cons) Career Opportunities (Cons) Leadership
	Cultural capital	(Pros) Atmosphere at Work (Cons) Atmosphere at Work (Cons) Nature of Work
	Human Capital	(Pros) Benefits (Pros) Compensation (Pros) Job Location (Pros) Perks at Work (Pros) Work Hours (Cons) Compensation and Benefits (Cons) Feeling (Cons) Work Hours
Human Capital	Human Capital	
		(Pros) Company Itself (Pros) None (Cons) Company Itself

It is also worth noting from Table 5.5 that "human capital" from Kaplan and Norton's framework did *not* align with any of the types that we identified with ABAE. Similarly, from our list, eight types—*(Pros) Benefits, (Pros) Compensation, (Pros) Job Location, (Pros) Perks at Work, (Pros) Work Hours, (Cons) Compensation and Benefits, (Cons) Feeling, (Cons) Work Hours*—did not align with Kaplan and Norton's categories. We believe such discrepancies stemmed from the fact that Kaplan and Norton aimed to identify useful intangible resource types from a strictly managerial perspective. In contrast, we relied on employee voice to identify firm resource types with the goal of discovering IIR, which represent intangible resources that impact employee experience. Therefore, it seems reasonable that Kaplan and Norton's framework emphasized the value that employees bring to management—skills, talent, and knowledge—whereas our list emphasized the value that employees desire to receive from the firm, such as Benefits, Compensation, Job Location, etc.

Based on these comparisons, we identified 21 out of 24 all types as potential IIR types. The three types that we disregarded were *(Pros) Company Itself*, *(Pros) None*, and *(Cons) Company Itself*, which did not align with neither Kaplan and Norton's framework nor Diefenbach's framework.

Among the 21 potential IIR types, some overlapped with each other because similar types emerged from both *Pros* and *Cons*. Such examples include *(Pros) Leadership* and *(Cons) Leadership*, *(Pros) Career Opportunities* and *(Cons) Career Opportunities*, and *(Pros) Coworkers* and *(Cons) Coworkers*. We eliminated these duplicate types. Ultimately, we created a comprehensive list of 15 IIR types (see Table 5.6).

Table 5.6: The complete list of 15 IIR types as identified from the collective employee voice (in alphabetical order).

15 Intangible Internal Resource Types
Atmosphere at Work, Benefits, Business Innovation, Career Opportunities, Compensation, Coworkers, Feeling, Job Location, Leadership, Nature of Work, Operations, Organizations Restructuring, Perks at Work, Technological Resources, Work Hours

5.2 Unique Firm Characteristics from TF-IDF

By using ABAE on the textual data from Glassdoor's *Pros* and *Cons* fields, we obtained 15 types of IIR. Moreover, 6 million input sentences were automatically annotated with the most relevant IIR type. Based on these results, we next extracted unique characteristics of each organization with TF-IDF, a widely used unsupervised keyword extraction model. We took the following three steps. First, we defined the elements required for TF-IDF. Secondly, we calculated TF-IDF. Finally, we ranked the terms in the order of importance and chose terms with the highest rankings as keywords.

First, we defined the elements needed to run the TF-IDF equation: document, collection of documents, and term. Document and collection of documents are used for calculating the value for inverse document frequency, and term is required for calculating the value for term frequency. For our purposes, we defined document as a collection of a single organization's Glassdoor sentences annotated with a single IIR type; collection of documents as organizations that belong to the same Global Industry Classification Standard (GICS) Industry Group; and term as any word that went through the text preprocessing steps explained in Section 3.3. Let us use Facebook and Netflix Inc (Netflix) as example organizations. Both organizations belong to the same GICS Industry Group called "Media & Entertainment (code: 5020)." Among all English Glassdoor reviews written about Facebook between June 2008 and June 2019, 5,847 sentences came from the *Pros* field. Among them, 530 were annotated by ABAE as the *(Pros) Leadership* IIR type. These 530 sentences would constitute a single document, "Media & Entertainment-*(Pros) Leadership*-Facebook Inc." Meanwhile, among all reviews for Netflix, 2,535 sentences belonged to the *Pros* field. Among them, 220 sentences were denoted by ABAE as pertaining to *(Pros) Leadership* type. These 220 sentences constituted another document, "Media & Entertainment-*(Pros) Leadership*-Netflix Inc." Other than Facebook and Netflix, 24 organizations belong to the "Media & Entertainment" Industry Group. Therefore, there would be a total of 26 "Media & Entertainment-*(Pros) Leadership*" documents that constitute a single collection of documents.

We used the GICS Industry Group because we believed that comparing organizations within the same industry would be most meaningful and impactful. For example, obtaining unique firm characteristics of Facebook by comparing it to Netflix and 25 other “Media & Entertainment” organizations would have more relevant and actionable managerial implications than comparing it against organizations belonging to “Energy (code: 1010)” or “Real Estate (code: 6010).”

Next, having defined these elements, we implemented the TF-IDF method. To ensure that truly unique and distinct keywords are extracted from a document, we removed ten most frequent words in the collection of documents; we named these words “Industry-and-Resource-specific stop words.” For example, ten most frequent words that appeared in the entire corpus of 26 “Media & Entertainment-*(Pros) Leadership*” documents were “*employee*,” “*management*,” “*company*,” “*team*,” “*great*,” “*good*,” “*work*,” “*leadership*,” “*care*,” and “*manager*.” It was necessary to remove Industry-and-Resource-specific stop words because they appeared frequently regardless of document and thus lacked discriminatory power. Before we removed the additional stop words, there were 4,069 terms within the corpus of 26 “Media & Entertainment-*(Pros) Leadership*” documents; after removal, 3,969 terms remained. We then applied the TF-IDF method. For example, by applying TF-IDF to 26 documents, we obtained a TF-IDF value for every term in the corpus for every document. In the below equation, i represents the ith term, and j represents the jth document).

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Since there were 3,969 terms within the corpus, this meant that we obtained a 26×3969 matrix filled with TF-IDF value. Table 5.7 indicates that the TF-IDF value of “*mark_zuckerberg*” was positive for only the “Media & Entertainment-*(Pros) Leadership*-Facebook Inc” document. In all other documents, its value was zero. This should not come as a surprise because Mark Zuckerberg is Facebook’s CEO, and there is no reason for other organization’s employees to mention him when writing reviews about their organization’s leadership. In contrast, the TF-IDF value of “*transparency*” was bigger than 0 in ten organizations. One could infer that employees of these organizations would describe their leadership with the word "transparency."

Table 5.7: Example from the TF-IDF value matrix. We applied TF-IDF to 26 "Media & Entertainment-*(Pros) Leadership*" documents. Then, we calculated the TF-IDF value for "*mark_zuckerberg*" and "*transparency*" for each document.

Document (Firm)	*TF-IDF Value for "mark_zuckerberg"*	*TF-IDF Value for "transparency"*
Activision_Blizzard_Inc	0.000	0.000
Alphabet_Inc	0.000	**0.343**
Altice_usa_inc	0.000	0.000
CBS_Corp	0.000	0.000
Cars.com_Inc	0.000	0.000
Charter_Communications_Inc	0.000	**0.004**
Comcast_Corp	0.000	**0.002**
DISH_Network_Corp	0.000	**0.003**
Discovery_Inc	0.000	0.000
Electronic_Arts_Inc.	0.000	**0.012**
Facebook_Inc	**0.030**	**0.093**
Fox_Corp	0.000	**0.003**
Interpublic_Group_of_Cos_Inc_(The)	0.000	0.000
Meredith_Corp	0.000	**0.029**
Netflix_Inc	0.000	**0.095**
New_York_Times_Co_(The)	0.000	0.000
News_Corp	0.000	0.000
Omnicom_Group_Inc.	0.000	0.000
Scripps_(E.W.)_Co_(The)	0.000	0.000
TEGNA_Inc	0.000	0.000
Take-Two_Interactive_Software_Inc	0.000	0.000
TripAdvisor_Inc	0.000	**0.049**
Twitter_Inc	0.000	0.000
Viacom_Inc	0.000	0.000
Walt_Disney_Co_(The)	0.000	0.000
iHeartMedia_Inc	0.000	0.000

In Table 5.8, we show another TF-IDF value matrix example. This time, we applied TF-IDF on 26 "Media & Entertainment-*(Pros) Job Location*" documents. In this case, the Industry-and-Resource-specific stop words "*free*," "*discount*," "*great*," "*service*," "*perk*," "*food*," "*good*," "*cable*," "*get*," and "*phone*." 17 out of 26 documents observed the occurrence of the word "*snack.*" Those with especially high TF-IDF values for "*snack*" were mostly IT firms known for young and trendy corporate culture. Providing snacks at work as a perk is a relatively new phenomenon. Interestingly, a recent survey showed that employees receiving free snacks in the workplace enjoyed higher level of happiness [61]. More specifically, it seemed like millennials, who have become the largest generation in the U.S. workforce since 2016, valued snacks more than any other age group [62]. In the same survey, 66% of millennials not only valued snacks three times more than those who are 45 and older but also said that they would take another company's job offer if the company had better perks, including snacks [61].

Since some employees may have used a different word to describe snack-like perks, we made a list of synonyms of "*snack*" and obtained their TF-IDF values as well (see Table 5.8). The synonyms of "*snack*" were "*lunch*," "*morsel*," "*refreshment*," and "*tea*" according to Thesaurus [63]. Since "morsel" returned N/A for TF-IDF values, we did not include it in the results. Finally, please see Appendix 5.1 and 5.2 for TF-IDF values of these terms for firms in "Energy" and "Real Estate" industries, respectively.

Table 5.8: Another example from the TF-IDF value matrix. We applied TF-IDF to 26 "Media & Entertainment-*(Pros) Perks at Work*" documents. Then, we calculated the TF-IDF value for "*snack*" and its synonyms for each document.

Document (Firm)	*TF-IDF Value for "snack"*	*TF-IDF Value for "lunch"*	*TF-IDF Value for "refreshment"*	*TF-IDF Value for "tea"*
Activision_Blizzard_Inc	0.000	**0.047**	0.000	0.000
Alphabet_Inc	**0.122**	0.000	0.000	0.000
Altice_usa_inc	0.000	0.000	0.000	0.000
CBS_Corp	0.000	0.000	0.000	0.000
Cars.com_Inc	**0.062**	**0.274**	0.000	0.000
Charter_Communications_Inc	**0.083**	**0.033**	**0.004**	**0.06**
Comcast_Corp	**0.013**	**0.025**	0.000	0.000
DISH_Network_Corp	**0.012**	**0.053**	0.000	0.000
Discovery_Inc	**0.044**	**0.059**	0.000	0.000

Electronic_Arts_Inc.	**0.036**	**0.028**	0.000	**0.012**
Facebook_Inc	**0.111**	**0.016**	**0.007**	0.000
Fox_Corp	**0.024**	**0.217**	0.000	**0.041**
Interpublic_Group_of_Cos_Inc_(The)	0.000	**0.109**	0.000	0.000
Meredith_Corp	**0.044**	**0.039**	0.000	0.000
Netflix_Inc	**0.172**	**0.118**	**0.010**	**0.016**
New_York_Times_Co_(The)	0.000	**0.034**	0.000	0.065
News_Corp	0.000	**0.052**	0.000	0.000
Omnicom_Group_Inc.	0.000	**0.107**	0.000	0.000
Scripps_(E.W.)_Co_(The)	0.000	0.000	0.000	0.000
TEGNA_Inc	**0.181**	0.000	0.000	0.000
Take-Two_Interactive_Software_Inc	**0.187**	0.000	0.000	0.000
TripAdvisor_Inc	**0.368**	**0.401**	0.000	0.000
Twitter_Inc	0.000	**0.007**	0.000	0.000
Viacom_Inc	**0.019**	**0.034**	0.000	0.000
Walt_Disney_Co_(The)	**0.020**	**0.015**	0.000	**0.002**
iHeartMedia_Inc	**0.014**	**0.032**	0.000	0.000

As the final step, we selected keywords for each document based on a ranking by TF-IDF values. For each document, we chose 25 terms with the highest TF-IDF values as its keywords. As an example, we listed unique characteristics of Facebook and Netflix regarding *(Pros) Leadership* and *(Pros) Perks at Work* in Table 5.9. We boldfaced the keywords, or characteristics, only if they did not appear for both companies. For instance, regarding *(Pros) Leadership*, both firms' characteristics overlapped in terms of "*transparent*," "*transparency*," "*feedback*," "*open*," "*culture*," and "*level*." But all other characteristics appeared to be unique to either Facebook or Netflix. Some characteristics of Facebook's *(Pros) Leadership* that stood out in contrast to those of Netflix were mission-related keywords such as "*mission*," "*impact*," and "*focus*," and supportive culture such as "*support*," "*encouraged*," and "*trust*." *(Pros) Leadership* characteristics of Netflix, when compared to those of Facebook, included management-related terms like "*performance*" and "*strategy*," emphasis on communication such as "*communication*" and "*context*," and terms such as "*freedom*."

Table 5.9: Comparison of Facebook and Netflix regarding certain IIR types, namely, *(Pros) Leadership* and *(Pros) Perks at Work*.

	(Pros) Leadership		*(Pros) Perks at Work*	
Ranking	*Facebook*	*Netflix*	*Facebook*	*Netflix*
1	**facebook**	**netflix**	**laundry**	**netflix**
2	transparent	feedback	**etc**	snack
3	open	**adult**	**benefit**	**movie**
4	culture	**communication**	**amazing**	**subscription**
5	**openness**	culture	**gym**	**lunch**
6	**fast**	**candid**	**shuttle**	**account**
7	feedback	**performance**	snack	**nice**
8	transparency	transparent	**breakfast_lunch_dinner**	**break**
9	**really**	transparency	**campus**	**swag**
10	**impact**	open	**awesome**	**time**
11	decision	**clear**	**meal**	**coffee**
12	**focus**	**context**	**day**	**room**
13	**strong**	**like**	**micro**	**plan**
14	**amazing**	**best**	**office**	**pay**
15	**encouraged**	**high**	work	drink
16	**truly**	**strategy**	**facebook**	**oatmeal**
17	level	**well**	lot	lot
18	**support**	**leader**	**dry_cleaning**	**stocked**
19	**mission**	**everyone**	drink	**employee**
20	**trust**	**freedom**	**transportation**	work
21	**value**	level	**kitchen**	**unlimited**
22	**take**	decision	**health**	**call**
23	**openly**	**process**	**gourmet**	**banana**
24	**move**	**lot**	**really**	**center**
25	**feel**	**make**	**spa**	**fruit**

5.3 Managerial Implications

Our findings illustrate the current state of each organization's unique characteristics regarding different IIR. Note that our intention was not to judge whether certain characteristics or firms were superior to others. Instead, we aimed to present a data-driven methodology for organizations to accurately assess their status quo in multiple dimensions. Due to the descriptive and interpretable nature of our methods and findings, firms could replicate our study and plan detailed roadmaps on each dimension.

For instance, a company could compare its keywords most associated with the *(Pros) Leadership* IIR type with a company's mission statement regarding leadership, observing whether the two are aligned with each other. However, if there are alarming gaps between the two, it could mean that it is time for the organization to reevaluate its leadership strategies. Let us assume, for instance, that Facebook's goal is to have an open, transparent, and trustworthy leadership. Then, the results in Table 5.9 indicate that it is doing a great job at achieving that vision.

In addition, organizations could devise specific plans to emulate their role models or competitors using our methods. If a firm in the media and entertainment industry would like to benchmark the perks provided by Netflix for its employees, it could start by reviewing Netflix's unique characteristics in terms of *(Pros) Perks at Work*.

5.4 Evaluation of ABAE

The downside of using unsupervised models on unannotated real-world datasets is that it is difficult to measure how well the models performed. To tackle this issue, we examined whether the ABAE model accurately annotated Glassdoor sentences with the most relevant IIR type. Since there were about 6 million annotated sentences, it would have been impossible—both timewise and resource-wise—to check every sentence and its annotation. Therefore, we came up with a mixed evaluation method.

The mixed method involved three human coders who manually graded the accuracy of the randomly chosen sentences and their annotations. All in all, the method consisted of four steps. First, using Table 5.4, we re-annotated 6 million sentences with Glassdoor's predefined categories. In other words, if a sentence was originally annotated by the ABAE model as *(Pros) Work Hours*, this time, we re-annotated it as *Work/Life Balance*, since the two types were similar. As a result, all 6 million sentences were re-annotated with Glassdoor's five predefined categories.

Next, we randomly selected 100 sentences from each category. Since there were five categories, this provided us with 500 randomly selected sentences.

Then, we asked three human coders to indicate whether they believed each sentence's annotation accurately reflected the content of the sentence. We asked that they grade the sentence as "accurate" if they thought the annotation was accurate, regardless of whether it included other annotations' contents as well; and "inaccurate" if the annotation was wrong. For example, all three coders gave 1 to the following sentence, which was annotated as *Compensation and Benefits*: "The scheduling is variable, and the pay just isn't enough for a long term job." Even though this sentence also included content more relevant to *Work/Life Balance*, it still talked about the compensation type, so all three coders regarded the annotation as "accurate."

As the fourth and final step, we calculated the accuracy score and the inter-rater reliability measure. To calculate the accuracy score, we averaged each coder's accuracy, which is the number of "accurate" grades divided by 500. To calculate the inter-rater reliability, we used Fleiss' kappa, a statistical measure that determines the reliability of agreement among numerous raters who assign ordinal ratings [64]. See Table 5.10 for the summary statistics of our evaluation results.

Table 5.10: Summary of human coders' evaluations of the ABAE model annotations.

Glassdoor's Predefined Categories	*Accuracy Score*	*Fleiss' Kappa*
Work/Life Balance	74.3%	0.66
Culture & Values	75.3%	0.39
Career Opportunities	65%	0.65
Compensation and Benefits	90.7%	0.49
Senior Management	81.3%	0.63
Total	**77.3%**	**0.59**

A higher accuracy score represents that the ABAE model correctly annotated the input sentences. Due to the lack of comparable datasets or metrics, it was difficult to determine whether 77.3% was a satisfactory result or not. Therefore, we compared the accuracy score with a different model as similar as possible. The current state-of-the-art *supervised* aspect extraction model achieved an accuracy score of 79.8% on the SentiHood dataset [43]. While 79.8% was slightly higher than 77.3%, the SentiHood dataset contained only four, instead of five, aspects, so it would have been much easier to achieve a higher accuracy score with the SentiHood dataset. Moreover, because supervised models are dataset-specific and trained with annotations, they are known to yield better results than unsupervised models. Thus, we concluded that given the 77.3% accuracy score, our ABAE model was indeed robust and reliable.

Similarly, a higher Fleiss' kappa indicates that rating by human coders is consistent across coders. Therefore, a higher kappa would indicate that human coders' judgement is reliable. It has been acknowledged that Fleiss' kappa between 0.41 and 0.60 indicates moderate agreement among raters, while a value from 0.61 to 0.80 indicates substantial agreement [65]. As seen from Table 13, in our case, all but *Culture & Values* and *Compensation and Benefits* indicated substantial agreement across three coders. The two categories' relatively lower values may be explained by a lack of agreement on what counts as culture, values, or benefits. For instance, one may regard perks at work, such as free snacks, as relevant to organizational culture, whereas another may regard snacks as benefits. Therefore, we concluded that the average score of 0.59 for Fleiss' kappa was satisfactory.

Chapter 6

Conclusion

In organization studies, the previous literature either analyzed characteristics of intangible resources using anecdotal evidence or focused on a specific type of intangible resource such as culture or reputation. Moreover, what we named as "intangible internal resources (IIR)," which refer to intangible resources that impact and shape employee experience, have not been explored by prior works. Therefore, we built on previous frameworks to create a holistic and comprehensive list of IIR—as exhaustively as possible. We took an empirical approach, utilizing large-scale Glassdoor review data and deep learning model. Furthermore, we implemented the keyword extraction method to discover unique firm characteristics regarding each IIR type.

Our work is limited in the following ways. First, because we used the unsupervised model to detect prominent aspects, or firm resource types, on the unannotated dataset, it was difficult to accurately assess how reliable our model is. To overcome this issue, we proposed a mixed evaluation method that involved three human coders. But in doing so, we randomly sampled only 500 sentences out of approximately 6 million sentences. Our second limitation is like the first in that we utilized TF-IDF, the unsupervised keyword extraction method, to detect unique firm characteristics; thus, it was difficult to provide a way to assess whether the method was trustworthy or not. We tried to evaluate the method using a qualitative evaluation method, namely, by comparing Facebook and Netflix on two aspects—*(Pros) Leadership* and *(Pros) Perks at Work*. But due to the sheer volume of data, it was impossible to verify the method's reliability in every scenario.

However, our work has shown that by using large-scale data and deep learning model, it is indeed possible to answer some of the questions in management and organizational theory that have remained unanswered for a long time. Despite the limitations, we have successfully demonstrated that the IIR types from the collective employee voice closely aligned with the previous work's framework. We believe that the question on strategic firm resources would continue to be of great interest to future scholars. We hope that our work, along with our open-sourced materials, would inform and guide future works.

Bibliography

[1] J. Barney, *Firm Resources and Sustained Competitive Advantage*, Journal of Management, 17 (1991), pp. 99 ~ 120.

[2] R. Makadok, *Toward a Synthesis of the Resource-Based and Dynamic-Capability Views of Rent Creation*, Strategic Management Journal, 22 (2001), pp. 387~401.

[3] R. Hall, *The Strategic Analysis of Intangible Resources*, Strategic Management Journal, 13 (1992), pp. 135 ~ 144.

[4] E. L. Black, T. A. Carnes, and V. J. Richardson, *The Market Valuation of Corporate Reputation, Corporate Reputation Review*, 3 (2000), pp. 31~42.

[5] J. Anderson and G. Smith, *A Great Company Can Be a Great Investment*, Financial Analysts Journal, 62 (2006), pp. 86~93.

[6] M. D. Pfarrer, T. G. Pollock, and V. P. Rindova, *A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions*, Academy of Management Journal, 53 (2010), pp. 1131~1152.

[7] T. Diefenbach, *Intangible resources: a categorical system of knowledge and other intangible assets*, Journal of Intellectual Capital, 7 (2006), pp. 406~420.

[8] Glassdoor, About Us. https://www.glassdoor.com/about-us/, 2020.

[9] D. J. Weiss, R. V. Dawis, G. W. England, and L. H. Lofquist, *Manual for the Minnesota Satisfaction Questionnaire. Vol. 22*, Minnesota Studies in Vocational Rehabilitation, Minneapolis: University of Minnesota, Industrial Relations Center, 1967.

[10] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, *An Unsupervised Neural Attention*

Model for Aspect Extraction, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 1 (2017), pp. 388~397.

[11] K. S. Jones, *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, Journal of Documentation, 28 (1972), pp. 11~21.

[12] M. Porter, *The contributions of industrial organization to strategic management*, Academy of Management Review, 6 (1981), pp. 609~620.

[13] R. Rumelt, *Towards a strategic theory of the firm*, in R. Lamb (Ed.), *Competitive Strategic Management*, Englewood Cliffs, NJ: Prentice-Hall, 1984.

[14] E. M. Scherer*, Industrial market structure and economic performance* (2nd ed.), Boston: Houghton-Mifflin, 1980.

[15] J. Hirschliefer, *Price theory and applications* (2nd ed.), Englewood Cliffs, NJ: Prentice-Hall, 1980.

[16] B. Lev, *Intangibles: Management, Measurement, and Reporting*, Brookings Institution Press, 2001.

[17] D. Kryscynski, R. Coff, and B. Campbell, *Charting a path between firm-specific incentives and human capital-based competitive advantage*, Strategic Management Journal, Forthcoming.

[18] R. S. Kaplan and D. P. Norton, *The strategy map: Guide to aligning intangible assets*, Strategy & Leadership, 32 (2004), pp. 10~17.

[19] K. Backhaus, and S. Tikoo, *Conceptualizing and Researching Employer Branding*, Career Development International, 9 (2004), pp. 501~517.

[20] A. Dabirian, J. Kietzmann, and H. Diba. *A great place to work!? Understanding crowdsourced employer branding*, Business Horizons, 60 (2017), 197-205.

[21] C. S. Pitt, E. Botha, J. J. Ferreira, and J. Kietzmann, *Employee brand engagement on social media: Managing optimism and commonality*, Business Horizon, 61 (2018), pp.

635~642.

[22] N. Kashive, V. T. Khanna, and M. N. Bharthi, *Employer branding through crowdsourcing: understanding the sentiments of employees*, Journal of Indian Business Research, 12 (2020), pp. 93~111.

[23] M. Corritore, A. Goldberg, and S. B. Srivastava, *Duality in Diversity: How Intrapersonal and Interpersonal Cultural Heterogeneity Relate to Firm Performance*, Administrative Science Quarterly, 65 (2019), pp.359~394.

[24] D. Sull, C. Sull, and A. Chamberlain, Measuring Culture in Leading Companies. https://sloanreview.mit.edu/projects/measuring-culture-in-leading-companies/?og=culture500, 2019.

[25] Glassdoor, Create a Glassdoor user account. https://help.glassdoor.com/article/Create-a-Glassdoor-user-account/en_US, 2020.

[26] Glassdoor, Writing a company review. https://help.glassdoor.com/article/Writing-a-company-review/en_US/, 2020.

[27] Z. Henry, Secrets of a very opaque Glassdoor. www.inc.com/magazine/ 201412/zoe-henry/a-very-opaque-glassdoor.html, 2014.

[28] Glassdoor, Edit or Delete My Contribution/Review. https://help.glassdoor.com/article/Edit-or-delete-my-contribution/en_US/, 2020.

[29] S. Melián-González, J. Bulchand-Gidumal, and B. G. López-Valcárcel, *New evidence of the relationship between employee satisfaction and firm economic performance*, Personnel Review, 44 (2015), pp. 906~929.

[30] M. Huang, P. Li, F. Meschke, and J. Guthrie, *Family Firms, Employee Satisfaction, and Company Performance*, Journal of Corporate Finance, 34 (2015), pp. 108~127.

[31] R. Farhadi and V. Nanda, *What Do Employees Know? Employee Opinions in Firms Going Public*, 2017.

[32] M. Huang, A. Masli, F. Meschke, and J. P. Guthrie, *Clients' Workplace Environment and Corporate Audits*, Auditing: A Journal of Practice & Theory, 36 (2017), pp. 89~113.

[33] E. Symitsi, P. Stamolampros, and G. Daskalakis, *Employees' online reviews and equity prices*, Economic Letters, 162 (2018), pp. 53~55.

[34] J. Hales, J. R. Moon Jr., and L. A. Swenson, *A new era of voluntary disclosure? Empirical evidence on how employee postings on social media relate to future corporate disclosures*, Accounting, Organizations and Society, 68-69 (2018), pp. 88~108.

[35] T. C. Green, R. Huang, Q. Wen, and D. Zhou, *Crowdsourced employer reviews and stock returns*, Journal of Financial Economics, 134 (2019), pp. 236~251.

[36] J. Sheng, *Asset Pricing in the Information Age: Employee Expectations and Stock Returns*, 2019.

[37] K. Huang, M. Li, and S. Markov, *What Do Employees Know? Evidence from a Social Media*, The Accounting Review, 95 (2020), pp. 199~226.

[38] Y. Ji, O. Rozenbaum, and K. Welch, *Corporate Culture and Financial Reporting Risk: Looking Through the Glassdoor*, 2017.

[39] M. Hu and B. Liu, *Mining and summarizing customer reviews*, In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[40] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool publishers, 2012.

[41] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, *SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods*, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016.

[42] X. Li, L. Bing, W. Zhang, and W. Lam, *Exploiting BERT for End-to-End Aspect-based Sentiment Analysis*, Proceedings of the 5th Workshop on Noisy User-generated Text

(W-NUT 2019), 2019, pp. 34~41.

[43] H. Xu, B. Liu, L. Shu, and P. S. Yu, *BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019.

[44] C. Sun, L. Huang, and X. Qiu, *Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019.

[45] D. Blei, A. Ng, and M. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research, 3 (2003), pp. 993~1022.

[46] H. Wang, Y. Lu, and C. Zhai, *Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach*, 2010.

[47] Y. Yang and J. O. Pederson, *A Comparative Study on Feature Selection in Text Categorization*, Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 412~420.

[48] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *Distributed Representations of Words and Phrases and their Compositionality*, Proceedings of the 26th International Conference on Neural Information Processing Systems, 2 (2013), pp. 3111~3119.

[49] D. Bahdanau, K. H. Cho, and Y. Bengio, *Neural Machine Translation by Jointly Learning to Align and Translate*, 3rd International Conference on Learning Representations, 2015.

[50] S. Beliga, A. Meštrović, S. Martinčić-Ipšić, *An Overview of Graph-Based Keyword Extraction Methods and Approaches*, Journal of Information and Organizational Sciences, 39 (2015), pp. 1~20.

[51] K. S. Jones, *Information retrieval and artificial intelligence*, Artificial Intelligence, 114

(1999), pp. 257~281.

[52] F. Sebastiani, *Machine learning in automated text categorisation*, ACM Computing Survays, 34 (2002), pp. 1~47.

[53] H. P. Luhn, *A statistical approach to mechanized encoding and searching of literary information*, IBM Journal of Research & Development, 1957.

[54] Y. Matsuo and M. Ishizuka, *Keyword extraction from a single document using word co-occurrence statistical information*, International Journal of Artificial Intelligence Tools, 4 (2004).

[55] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, *Bag of Tricks for Efficient Text Classification*, 2016.

[56] nltk, nltk.tokenize package, https://www.nltk.org/api/nltk.tokenize.html.

[57] scikit-learn, 6.2. Feature extraction. https://scikit-learn.org/stable/modules/feature_extraction.html.

[58] nltk, nltk.stem package, http://www.nltk.org/api/nltk.stem.html?highlight=lemmatizer.

[59] gensim, models.phrases. https://radimrehurek.com/gensim/models/phrases.html.

[60] D. L. Fields, *Taking the Measure of Work: A Guide to Validated Scales for Organizational Research and Diagnosis*, Thousand Oaks, CA: Sage, 2002.

[61] Peapod, Happiest Office Workers Are Those Who Get Free Snacks. https://www.prnewswire.com/news-releases/happiest-office-workers-are-those-who-get-free-snacks-300144291.html#:~:text=In%20a%20recent%20survey%20with,a%20big%20impact%20on%20morale., 2015.

[62] R. Fry, Millennials are the largest generation in the U.S. labor force. https://www.pewresearch.org/fact-tank/2018/04/11/millennials-largest-generation-us-labor-force/, 2018.

[63] Thesaurus, snack. https://www.thesaurus.com/browse/snack, 2020.

[64] J. L. Fleiss, *Measuring nominal scale agreement among many raters*, Psychological Bulletin, 76 (1971), pp. 378~382.

[65] J. R. Landis and G. G. Koch, *The measurement of observer agreement for categorical data*, Biometrics, 33 (1977), pp. 159~174.

Appendix

Appendix 1.

List of 30 employers that were missing reviews on Glassdoor.

Employer Name	*Likely Reason for Lacking Glassdoor Reviews*
Baxalta Inc.	Acquired by Shire in 2016
Wm. Wrigley Jr. Co	Acquired by Mars in 2008; Mars formed a new subsidiary, Mars Wrigley Confectionery, in 2016
Massey Energy Co	Acquired by Alpha Natural Resources in 2011
Plum Creek Timber Co Inc.	Acquired by Weyerhaeuser in 2016
GGP Inc	Acquired by Brookfield Property Partners in 2018
UST Inc.	Acquired by Altria in 2009
CSRA Inc	Acquired by General Dynamics in 2018
CareFusion Corp	Acquired by BD in 2015
Bemis Co Inc	Acquired by Amcor in 2019
Dr Pepper Snapple Group Inc	Acquired by Keurig Green Mountain in 2018
Time Warner Cable Inc	Acquired by Charter Communications in 2016
Genzyme Corp	Acquired by Sanofi in 2011
Novell Inc.	Acquired by The Attachmate Group in 2011
Goodrich Corp	Acquired by United Technologies Corporation in 2012
Schering-Plough	Merck & Co. merged with Schering-Plough in 2009
Quality Care Properties Inc	Unsure (possibly due to a mismatch in name)
Lehman Brothers Holdings Inc	Filed for Chapter 11 bankruptcy in 2008
Bard (C.R.) Inc	Acquired by Becton Dickinson in 2017
MeadWestvaco Corp	Merged with RockTenn in 2015
Columbia Pipeline Group Inc	Unsure
DIRECTV	Acquired by AT&T in 2015
Questar Corp	Unsure (possibly due to a mismatch in name)
Sears Holdings Corp	Filed for Chapter 11 bankruptcy in 2018
Electronic Data Systems Corp	Acquired by Hewlett-Packard
NYSE Euronext	Unsure (possibly due to a mismatch in name)
Countrywide Financial Corp	Acquired by Bank of America in 2008
RS Legacy Corp	Unsure
Washington Mutual Inc	Filed for Chapter 11 voluntary bankruptcy in 2009
Novellus Systems Inc.	Acquired by Lam Research in 2011

Appendix 2.

Appendix 2.1: Results from applying LDA on *Pros* and *Cons*. For fair comparison with ABAE results, we set the number of topics as 12.

Field	*Topic*	*Words*
Pros	0	0.138*"employee" + 0.074*"well" + 0.056*"really" + 0.047*"company" + 0.034*"care" + 0.026*"people" + 0.018*"helpful" + 0.017*"associate" + 0.017*"worked" + 0.016*"customer"
	1	0.101*"time" + 0.091*"hour" + 0.068*"flexible" + 0.051*"work" + 0.046*"schedule" + 0.042*"day" + 0.039*"manager" + 0.029*"paid" + 0.022*"part" + 0.018*"week"
	2	0.097*"lot" + 0.060*"many" + 0.052*"learn" + 0.042*"people" + 0.033*"different" + 0.026*"department" + 0.024*"opportunity" + 0.022*"meet" + 0.021*"skill" + 0.019*"business"
	3	0.053*"new" + 0.045*"product" + 0.042*"company" + 0.028*"pro" + 0.026*"love" + 0.024*"strong" + 0.024*"leadership" + 0.024*"value" + 0.021*"technology" + 0.017*"leader"
	4	0.180*"good" + 0.091*"environment" + 0.084*"work" + 0.079*"people" + 0.058*"nice" + 0.048*"working" + 0.046*"friendly" + 0.039*"fun" + 0.025*"coworkers" + 0.023*"location"
	5	0.119*"opportunity" + 0.069*"training" + 0.067*"company" + 0.048*"great" + 0.042*"career" + 0.041*"growth" + 0.036*"advancement" + 0.034*"lot" + 0.025*"within" + 0.025*"room"
	6	0.046*"year" + 0.035*"service" + 0.031*"make" + 0.025*"much" + 0.025*"money" + 0.022*"company" + 0.022*"customer" + 0.018*"long" + 0.014*"made" + 0.013*"market"
	7	0.239*"work" + 0.231*"great" + 0.061*"people" + 0.060*"place" + 0.057*"life" + 0.049*"good" + 0.047*"balance" + 0.042*"company" + 0.027*"culture" + 0.018*"amazing"
	8	0.038*"best" + 0.037*"experience" + 0.037*"always" + 0.031*"help" + 0.028*"job" + 0.026*"like" + 0.026*"one" + 0.022*"want" + 0.018*"working" + 0.015*"company"
	9	0.188*"benefit" + 0.127*"pay" + 0.120*"good" + 0.074*"great" + 0.042*"decent" + 0.028*"salary" + 0.023*"bonus" + 0.021*"food" + 0.020*"health" + 0.016*"competitive"
	10	0.107*"job" + 0.105*"get" + 0.061*"easy" + 0.049*"team" + 0.049*"management" + 0.027*"position" + 0.025*"member" + 0.025*"level" + 0.019*"support" + 0.017*"move"
	11	0.086*"discount" + 0.071*"worker" + 0.061*"co" + 0.054*"store" + 0.036*"customer" + 0.036*"pretty" + 0.036*"free" + 0.032*"sale" + 0.022*"family" + 0.020*"ok"

Cons	0	0.137*"hour" + 0.066*"working" + 0.063*"long" + 0.061*"low" + 0.039*"pay" + 0.033*"shift" + 0.026*"schedule" + 0.025*"time" + 0.024*"enough" + 0.023*"week"
	1	0.307*"work" + 0.054*"life" + 0.042*"environment" + 0.039*"place" + 0.037*"balance" + 0.020*"retail" + 0.020*"hard" + 0.020*"stressful" + 0.019*"expectation" + 0.018*"family"
	2	0.081*"pay" + 0.042*"benefit" + 0.030*"raise" + 0.026*"le" + 0.026*"better" + 0.025*"salary" + 0.021*"bonus" + 0.020*"could" + 0.019*"money" + 0.018*"cut"
	3	0.046*"department" + 0.025*"office" + 0.024*"company" + 0.024*"issue" + 0.023*"different" + 0.023*"horrible" + 0.022*"area" + 0.020*"lot" + 0.020*"problem" + 0.018*"location"
	4	0.083*"get" + 0.054*"time" + 0.026*"people" + 0.026*"make" + 0.024*"thing" + 0.022*"hard" + 0.022*"know" + 0.021*"job" + 0.017*"good" + 0.017*"never"
	5	0.088*"company" + 0.041*"opportunity" + 0.033*"growth" + 0.032*"con" + 0.030*"career" + 0.028*"promotion" + 0.028*"advancement" + 0.026*"think" + 0.022*"slow" + 0.020*"big"
	6	0.089*"customer" + 0.064*"sale" + 0.053*"store" + 0.048*"manager" + 0.038*"bad" + 0.035*"team" + 0.030*"goal" + 0.027*"service" + 0.018*"product" + 0.018*"member"
	7	0.055*"day" + 0.032*"one" + 0.023*"call" + 0.022*"would" + 0.022*"even" + 0.020*"worked" + 0.019*"back" + 0.018*"every" + 0.018*"time" + 0.018*"nothing"
	8	0.045*"management" + 0.034*"level" + 0.025*"business" + 0.020*"manager" + 0.018*"many" + 0.017*"team" + 0.016*"politics" + 0.016*"people" + 0.015*"decision" + 0.014*"leader"
	9	0.167*"management" + 0.042*"poor" + 0.041*"lack" + 0.029*"change" + 0.026*"upper" + 0.021*"leadership" + 0.020*"communication" + 0.020*"policy" + 0.015*"system" + 0.014*"constant"
	10	0.132*"employee" + 0.045*"like" + 0.034*"company" + 0.032*"care" + 0.021*"great" + 0.018*"number" + 0.018*"high" + 0.017*"well" + 0.014*"people" + 0.013*"feel"
	11	0.061*"year" + 0.044*"job" + 0.035*"new" + 0.024*"training" + 0.020*"position" + 0.020*"everything" + 0.020*"company" + 0.015*"month" + 0.015*"people" + 0.013*"hire"

Some topics were represented by words that formed meaningful and coherent topics. For example, Topic 5 of *Pros* and *Cons* each coherently talked about career advancement at work. However, most topics were incoherent. For example, Topic 7 of *Pros* contained mixture of words pointing to culture and words speaking to work-life balance. Also, in both *Pros* and *Cons*, words indicative of leadership appeared across multiple topics.

Appendix 2.2: Results from applying LARA on *Pros* and *Cons*. LARA requires the user to input seed words in addition to specifying the number of aspects.

Field	*Seed Words*	*Aspect Terms (Words Added by LARA)*	*Inferred Aspect*
Pros	compensation, benefit	pay, good, salari, work, great, packag, place, environ, decent, competit, differ, within, averag, 401k, meet, encourag, chang, take, pension, way, person, hard, knowledg, pretti, across, bonus, custom, match, invest, given	Compensation and Benefits
	life_balance, life, balance	worklif, 9_80, home, flexibl, lot, schedul, hour, time, freedom, every_friday, sheet, telecommut, vacat, week, day, long, flex, full, holiday, sick, paid, 9, 15, 3_week, weekend, extra, get, 16, unlimited_sick, 8	Work/Life Balance
	culture, value	safeti, ethic, freedom_respons, compani, strong, deck, big, larg, focu, stabl, divers, employe, care, well, treat, integr, global, respect, truli, establish, adult, well_known, perform, import, high, like_adult, feel, focus, footprint, taken_seri	Culture & Values
	career, opportunities, career_path	advanc, growth, learn, develop, mani, grow, plenti, skill, move, new, busi, room, technolog, move_around, willing_reloc, experi, intern, provid, gain, train, potenti, thing, divis, different_area, program, rotat, function, chanc, transfer, quickli	Career Opportunities
	senior, leadership, management	team, support, micro, level, help, top, director, need, approach, alway, execut, mentor, listen, middl, question, understand, issu, everyon, direct, will, feedback, ask, know, transpar, talk, advic, advisor, everyth, design, assist	Senior Management
	coworker, co_worker, colleague	friendli, smart, nice, fun, talent, brilliant, peopl, offic, staff, campu, surround, creativ, intellig, passion, hardwork, around, willing_help, motiv, realli, atmospher, met, brightest, incred, amaz, awesom, enjoy, dedic, cool, super, bright	Coworkers
	facility, pipeline, resources	access, beach, research, human, avail, gym, site, tool, volleybal, basketball_court, outdoor, soccer_field, indoor, basketbal, cafeteria, court, workout, juno, starbucks_sit, food,	Resources

		hockey, cafe, trainer, soccer_pitch, therapi, coupon, onsit, physio, membership, dispos	
Cons	compensation, benefit	pay, salari, compar, averag, full_tim, packag, low, lower, competit, rais, increas, less, wage, bonus, bonu, industri, offer, higher, commiss, low_compar, year, rate, starting_salari, industry_standard, cut, market_averag, base_salari, 401k, merit_increas, last	Compensation and Benefits
	life_balance, life, balance	work_lif, famili, long_hour, personal_lif, work, challeng, hour, hard, weekend, environ, day, home, long, week, overtim, shift, repetit, expect, time, bore, schedul, stress, 12_hour, night, sometim, vacat, flexibl, monoton, get, spend	Work/Life Balance
	culture, value	corpor, compani, bank, toxic, conserv, sharehold, larg, big, size, within, red_tap, chang, bureaucraci, huge, organ, polit, bureaucrat, slow_mov, process, navig, typical_larg, bank_america, get_lost, easy_get, large_corpor, innov, procedur, lost, old_school, things_don	Culture & Values
	career, opportunities, career_path	limit, advanc, growth, move, grow, career_advanc, slow, unless, career_develop, mobil, little_room, career_growth, promot, reloc, limited_upward, progress, learn, room, chanc, movement, posit, path, growth_opportun, ladder, willing_reloc, vertic, someon, forward, move_around, upward	Career Opportunities
	senior, leadership, management	level, team, care, project, poor, micro, commun, execut, decis, director, incompet, leader, hr, lack, top, support, direct, lead, vision, employe, skill, peopl, staff, train, mani, develop, treat, con, help, new	Senior Management
	coworker, co_worker, colleague	stun, 00, tear, camaraderi, fellow, architectur, sweat, arriv, 06, poignant, miseri, tardi, 00am, homophob, 46, suprvisor, douchebag, standoffish, veto, righteou, burst, rave, preparatori, destitut, wallow, clerk, lawer, sci, scroog, companhi	Coworkers
	'facil', 'pipelin', 'resourc'	manufactur, jackson, cafeteria, alloc, rom, working_condit, food, plant, engin, v, menu, rent, petroleum, closur, cafe, canteen, powertrain, rural, microwav, kitchen, gca, directionless, acronym, mexico, itoc, locat, gym, cutback, area, chevi	Resources

LARA showed much more coherent and consistent topics than LDA. However, LARA's results were extremely susceptible to the changes in seed words. Even changing one word would yield vastly different results. Furthermore, it would mistakenly assign words that were not aspect-specific to a specific aspect. For instance, "good," "great," and "decent" were all assigned to the *Pros-Compensation and Benefits* aspect. This is problematic because LARA is designed to annotate a sentence as a certain aspect if it contains an aspect term from that aspect. For instance, both "The pay is good" and "The leaders are good" would automatically get categorized as *Pros-Compensation and Benefits*.

Appendix 3.

Hyperparameters of the ABAE model.

Hyperparameter	*Value*
Embedding size of e_{w_i}	200
Batch size	50
Vocabulary size	12000
Number of epochs	15
Number of negative sample embedding n_k	20
The weight of orthogonal regularization λ	0.1

Appendix 4.

Appendix 4.1: ABAE results when $K = 10$ for *Pros*.

Top 20 Representative Words	*Inferred Aspects (=Firm Resource Types)*
gym_membership voucher movie breakfast_lunch massage breakfast shuttle_bus gym fitness_center movie_ticket cooky leftover fitness bike free ticket sporting_event spa commuter fitness_reimbursement	
somewhere probably realize truth saying msft sad perhaps gonna wal_mart hope certainly tomorrow remember couldnt quit die someone somewhere_else unemployed	
alignment neutral differentiated underlying hw focusing v sensible aimed defining effectiveness aligned integration fintech perceived consultancy transformed creation strictly tactical	
opportunity possibility opportunites oppurtunities opportunties opps scope oportunities oppotunities oppertunities ability avenue oppurtunity oportunity chance assignment allow desired varied explore opportunitites	
manager associate supervisor asking coach asked asm direct_report subordinate manger guide showed approve told management timely_manner dm report director respond	
company organization eastman organisation intuit corporation equality diversity_inclusion cummins quest inclusion eaton philanthropy sustainability promoting evident volunteerism pnc pvh whirlpool	
hour weekday weekend rush scheduled schedual shift overnight schedule overnights scheduling saturday ish dress_code workday night season clock saturday_sunday afternoon	
colleague people teammate coworkers collegues ppl atmosphere environment workmate enviornment enviorment enviroment poeple coworker env culture sociable peer environement mate	
engine technique printing platform graphic introduction designing industrial computing crm functionality packaging ecommerce advertising logistics machine film content b2b commerce	
pay compensation salary benefit 401k_matching profit_sharing benfits wage benifits package pto_accrual espp tuition_assistance tution_reimbursement benifit medical_dental stock_grant medical_dental_vision tuition_reimbursement rsu	

Some topics were represented by words that formed meaningful and coherent topics. For example, Topic 5 of *Pros* and *Cons* each coherently talked about career advancement at work. However, most topics were incoherent. For example, Topic 7 of *Pros* contained mixture of words pointing to culture and words speaking to work-life balance. Also, in both

Pros and *Cons*, words indicative of leadership appeared across multiple topics.

Appendix 4.2: ABAE results when $K = 10$ for *Cons*.

Top 20 Representative Words	*Inferred Aspects (=Firm Resource Types)*
management managment mgmt mangement mgt managemnt leadership managament managerment managent egotistical mgrs manager mngt aloof mgmnt self_serving unapproachable manger vindictive	
capability architecture automation functionality implementation operational research fundamental analysis deploy analytics technology business creation executing platform product leverage enhancement execute	
costumer customer needy angry shopper greasy smile guest cranky mad upselling passenger yell picky irritable annoying spill sell sweat hey	
shabby resembles describes blah infested fort_worth frill primitive horrifying deplorable linde hyd thee egypt represents lively bland transformed synopsys disgrace	
promoted move brown_nose promotion liked promote shine kiss_butt climb_ladder advance grow superstar kiss_as advancing deserve join friend career fit_mold socialize	
pay salary wage compensation 401k_matching profit_sharing bonus payout substantially payouts meager paltry measly benefit stock_grant capped benifits offset incentive 5k	
company merck pfizer medtronic abbott corporation xilinx symantec amgen ge intuit hpe hp_inc qualcomm caterpillar firm ihs autodesk broadcom covidien	
uncoordinated unnecessarily beurocracy tight_deadline disorganization beauracratic beaurocratic repetitious beurocratic burocratic rigidity strenuous unstructured somtimes chaotic tedious hierarchial overbearing stressfull ambiguity	
told supposed register wrote receipt phone confirmation ordered incorrectly file accidentally verify ticket immediately month reprimanded asked alarm submitted called	
weekday hour workday work peak_season noon week day overnights 3pm shift monday_friday night 7pm 4hrs friday sat_sun 30pm scheduled twelve	

Appendix 4.3: ABAE results when $K = 30$ for *Pros*.

Top 20 Representative Words	*Inferred Aspects (=Firm Resource Types)*

swag raffle goody ice_cream sporting_event cooky giveaway movie_ticket ticket soda breakroom catered_lunch pizza movie random beer_bash popcorn breakfast candy vending_machine	
payout payouts earnings salary earning_potential spiff pay uncapped_commission bonus incentive comp compensation commision quarterly cap sti monthly commission myshare capped	
appreciates strives cast valued recognizes happiness inspires appreciated strive recognize empower empowered inspired empowers recognizing team dedication motivates empowering understands	
manager management manger mgmt mgt dm staff director supervisor vice_president tl vp vps exec managment boss folk md president mgr	
wanna bad suck want dont somewhere kill hell gonna die ur retire somewhere_else butt fine place lie settle coast wont	
sustainability transparency philosophy equality integrity diversity_inclusion admirable inclusion honesty_integrity belief emphasizes operational_excellence stakeholder accountability instills credo honesty fairness evident reflected	
hated joined began interviewed happened vastly ive worked former started talked totally ruined heard drastically shocked screwed luckily awful appeared	
company firm institution footprint financial_institution corporation comapny conglomerate manufacturer industry organisation broker pharma supplier franchise provider sector powerhouse organization brand	
varied various numerous multiple across_globe multitude overseas vast many myriad travel mobility wide_range nationwide different large cross_functional across geographic networking	
environment atmosphere enviornment enviroment enviorment environement energetic envrionment cordial inviting envirnoment evironment env envirnment atmospher upbeat cheerful sociable workmate lively	
company intuit hartford boeing corporation quest anthem schwab uhg eastman ti amgen schlumberger eaton comcast cummins chevron caterpillar adp symantec	
climb_ladder move advance grow pursue excel rotate stretch explore jump abound shine seek upward_movement advancement grab climb progress presented define	
people ppl surrounded amazingly collegues colleague folk surround clever attracts poeple filled hardworking funny humble diligent incredibly enthusiastic teammate passionate	
retiree graduating uni enrolled aged h partial asu studying highschool graduated gig secondary finishing asu_online retired undergrad bachelor_degree semester employment	
publishing magazine australia stunning white represents highly_regarded association underlying nasdaq iconic owns pedigree prominent toronto legendary historical image admired award_winning	

workload independence deadline load downtime physical_labor supervision autonomy pressure oversight micromanagement hectic ambiguity red_tape demanding monotonous micro_managing overwhelming mindless workday	
canteen cab shuttle_bus car_wash cab_facility transport pantry shuttle subsidized cafeteria dry_cleaning subsidised amenity premise transportation subsidy pickup_drop café laundry indoor	
training traning classroom onboarding documentation induction boarding thorough detailed material academy ojt manual mentoring module mentorship coaching recruiting trainning licensing	
benefit insurance benifits benfits coverage 401k_matching medical_dental_vision plan medical_dental profit_sharing pension dental maternity_paternity_leave hsa espp adoption_assistance retirement tuition_reimbursment fmla healthcare	
pay salary wage payscale comparatively remuneration compensation comparison salery paymaster comparing compare renumeration compared payer imo slightly comparable relative considering	
arrangement schedule remotely remote timing wfh telecommuting scheduling home availability balance life schedual telecommute accomodate accommodates accomodating request accommodating teleworking	
innovating disruptive adapting innovative technologically aggressively forefront bold transformation technological transforming reinvent evolving adopting transformational cloud_computing digital_transformation rapidly exciting iot	
technology instrument automation software erp methodology design coding application architecture analytics testing java hardware implementation database instrumentation domain sap feature	
termination firing severe harsh threat incompetent paperwork caused unnecessary delay prevent wont consequence nepotism petty error monitored fear favoritism failed	
merchandise jewelry cosmetic salon grocery discount makeup clothes ulta item electronics apparel macy macys ten_percent outlet first_dibs fragrance gratis beauty	
customer pet_parent patient resolve knowing guest costumer pet educate medication smile physician stranger upset situation explain daily_basis animal assisting asking	
opportunity opportunites oppurtunity oppurtunities possibility opportunties oppertunity oportunities avenue chance oppertunities opps opp opprotunities oportunity oppotunities opportunies opprtunities room oppty	
suburb located city headquarters charlotte atlanta philadelphia north nj dc downtown suburban hq bellevue boston los_angeles raleigh colorado centrally_located nyc	
job culture feeling workplace motto atmosphere brand ibmer impression environment attitude vibe thing moment environement name cache enviornment sense_belonging heritage	

hour week day december month saturday cent wk min year 30pm roughly minute quarter saturday_sunday april 2pm january november monday	

Appendix 4.4: ABAE results when $K = 30$ for *Cons*.

Top 20 Representative Words	*Inferred Aspects (=Firm Resource Types)*
salary pay compensation increment wage paltry raise meager 401k_matching nominal rsus stock_grant remuneration bonus comparatively payscale profit_sharing measly rsu cola	
warranty extended_warranty dealer homeowner add_ons debit_card membership fee geek_squad_protection accessory selling commision attachment upselling ppp sell customer pricing catalog bundle	
il location ohio memphis headquarter minneapolis nj denver charlotte located dallas iowa headquarters st_louis north miami virginia milwaukee hq atlanta	
method methodology sop policy system procedure technique philosophy rule guideline process protocol rule_regulation framework model directive formula workflow strategy structure	
weekday hour monday_friday weekend evening night overtime holiday saturday_sunday 10pm schedule ot 24hrs work midnight 4hrs overnights peak_season sat_sun mon_fri	
hubris institutional overemphasis intellectual shaped survives mixture rooted invented_syndrome ethos institutionalized ultra_conservative velocity holistic dot_com analog incestuous transformed defines characteristic	
roughly wk yr approximately month whopping year week approx cent 15k mo averaged lt workday percent 11_00 75k 30k dec	
writes explain think pull admit tell complain complains listen argue fix explaining know hear communicate agree fault apologize asks telling	
mentorship empowerment encouragement vision leadership guidance accountability insight empathy mentoring cohesion positive_reinforcement communication foresight motivation acknowledgement courage autonomy engagement acknowledgment	
saas commerce cloud_computing aws cloud innovator enterprise business mainstream segment portfolio ecommerce emerging firm g commercial digital hybrid consultancy marketplace	
maximizing stockholder squeezing loyalty destroying loyal sacrificing reducing bottomline negatively_impacting easily_replaceable sacrificed killing eps expendable stock_holder poor maximize disposable hurting	
stressfull strenuous stressful exhausting physically_exhausting demanding taxing tense stress draining labor_intensive grueling emotionally_draining physical_labor physically_demanding tiring fast_paced unrewarding physically_mentally unfulfilling	

incident confirmation anonymous eeoc investigation exit_interview caller grievance retaliated inquiry verbally disciplinary_action retaliation observation confirm escalated indicated verbal letter infraction	
year february month april october september june august decade january feb march november jan couple cancelled three two announced switched	
hit_miss differs vary_greatly varies_greatly severely_lacking oversee varies overseeing poor varies_widely vary_wildly lacked lacking functional inconsistency coordinator uneven respective differ	
nothing guy bos honestly people everybody really hey ppl yeah wanna hell anything awesome ya fine laugh thing excited theyre	
managent managment management mangement managerment manipulative unapproachable dismissive uncaring mgmt borderline deceitful managament egotistical vindictive mgt cocky patronizing condescending untrustworthy	
employee people folk personnel staff worker workforce engineer employes contractor talent ftes employess systematically skilled college_grad fte immigrant ppl underperformer	
register drive_thru aisle cashier cashiering stocking backroom cash_register photo_lab guest store self_checkout zoning unloading_truck salesfloor floor bagger station ring ringing	
brown_nosing inner_circle butt_kissing favourite cronyism friendship brown_nosers teacher_pet favouritism nepotism politics favored kiss_as promotion favoritism promoted clique popularity favortism click	
realignments reorganization lay_offs reorgs restructuring turmoil restructures upheaval reorg structuring reorganisation instability restructurings downsizing layoff rifs restructure surplus sizing reshuffling	
sexual_harassment rein insisted banned ada assault d combat absentee slander disgrace harrassment literal smh harassment visiting spy bulling facial_hair affair	
bureaucratic cumbersome convoluted beaurocratic burdensome complicated burocratic bogged red_tape rigid overkill burocracy bureaucracy bloated complicate beurocratic bureacracy inefficient bog beuracracy	
grow advance move progress pursue climb_ladder elevate evolve climb develop establish succeed carve advancing progressing assimilate explore advancement shine navigate	
overqualified graduate qualification recent_college_grad advancement mba interviewed hired college_grad graduation credential grad mit advanced training shadowing bachelor_degree applying candidate fresh_graduate	
kpi quota goal attainment metric scorecard kpis target number monthly projection forecast quarterly unrealistically sph revenue percentage earnings stats sla	
company esi corporation anthem eastman bank_america merck stryker lockheed compnay american_express mckesson wal_mart uhg lockheed_martin pepsico zimmer dollar_tree fis hilton	

testing software database buggy programming automation tooling crm installation troubleshooting cad desktop documentation configuration sql coding manual_testing library portal custom	
want able need ready <unk> hard wont decide enough wil willing asap actually somewhere_else quicker faster dont wanted good wanting	
chicken smelly flower coat stained ice wet upstairs greasy towel tray stinky dusty refrigerator napkin grease fry cup breakfast fridge	

Appendix 5.

Appendix 5.1: We applied TF-IDF to 40 "Energy-*(Pros) Perks at Work*" documents (firms). Then, we calculated the TF-IDF value for "snack" and its synonyms.

Document (Firm)	*TF-IDF Value for "snack"*	*TF-IDF Value for "lunch"*	*TF-IDF Value for "refreshment"*	*TF-IDF Value for "tea"*
Apache_Corp	0.000	0.000	N/A	0.000
Apergy_Corp	0.000	**0.144**	N/A	0.000
Baker_Hughes_a_GE_Co	**0.129**	0.000	N/A	**0.166**
Cabot_Oil_&_Gas_Corp	0.000	0.000	N/A	**0.237**
Chesapeake_Energy_Corp	**0.015**	**0.028**	N/A	0.000
Chevron_Corp	**0.147**	**0.047**	N/A	0.000
Concho_Resources_Inc	**0.102**	0.000	N/A	0.000
Conocophillips	0.000	**0.042**	N/A	0.000
Denbury_Resources_Inc.	0.000	0.000	N/A	0.000
Devon_Energy_Corp	0.000	**0.144**	N/A	0.000
EOP_Resources_Inc.	0.000	0.000	N/A	0.000
EQT_Corp	0.000	0.000	N/A	0.000
Exxon_Mobil_Corp	**0.025**	**0.063**	N/A	**0.032**
Halliburton_Co	0.000	**0.021**	N/A	0.000
Helmerich_&_Payne_Inc.	0.000	0.000	N/A	0.000
Hess_Corp	0.000	**0.081**	N/A	0.000
HollyFrontier_Corp	0.000	**0.119**	N/A	0.000
Kinder_Morgan_Inc.	0.000	**0.164**	N/A	0.000
Marathon_Oil_Corp	0.000	**0.092**	N/A	0.000
Marathon_Petroleum_Corp	0.000	**0.092**	N/A	0.000
Murphy_Oil_Corp	0.000	0.000	N/A	0.000
Nabors_Industries_Ltd	0.000	**0.075**	N/A	0.000
National_Oilwell_Varco_Inc	**0.021**	**0.027**	N/A	0.000
Noble_Corp_plc	0.000	0.000	N/A	0.000
Noble_Energy_Inc	**0.105**	0.000	N/A	0.000
ONEOK_Inc.	0.000	0.000	N/A	0.000
Occidental_Petroleum_Corp	0.000	**0.075**	N/A	0.000

Peabody_Energy_Corp	0.000	0.000	N/A	0.000
Phillips_66	0.000	**0.043**	N/A	0.000
Pioneer_Natural_Resources_Co	0.000	**0.374**	N/A	0.000
QEP_Resources_Inc	0.000	0.000	N/A	0.000
Range_Resources_Corp.	0.000	0.000	N/A	0.000
Schlumberger_Ltd	0.000	**0.053**	N/A	0.000
Southwestern_Energy_Co	0.000	**0.078**	N/A	**0.079**
TechnipFMC_plc	**0.079**	**0.050**	N/A	0.000
Transocean_Ltd	0.000	0.000	N/A	0.000
Valero_Energy_Corp	0.000	**0.144**	N/A	0.000
WPX_Energy_Inc	0.000	**0.221**	N/A	0.000
Weatherford_International_plc	**0.051**	**0.129**	N/A	0.000
Williams_Cos_Inc._(The)	0.000	0.000	N/A	0.000

Appendix 5.2: We applied TF-IDF to 31 "Real Estate-*(Pros) Perks at Work*" documents (firms). Then, we calculated the TF-IDF value for "snack" and its synonyms.

Document (Firm)	*TF-IDF Value for "snack"*	*TF-IDF Value for "lunch"*	*TF-IDF Value for "refreshment"*	*TF-IDF Value for "tea"*
Alexandria_Real_Estate_Equities_Inc.	0.136	**0.273**	0.000	N/A
American_Tower_Corp	0.292	0.000	0.000	N/A
Apartment_Investment_and_Management_Co	0.000	0.056	0.000	N/A
AvalonBay_Communities_Inc.	0.000	0.000	0.000	N/A
Boston_Properties_Inc	0.187	**0.280**	0.000	N/A
CBRE_Group_Inc	0.000	**0.173**	0.000	N/A
Crown_Castle_International_Corp	0.134	**0.179**	0.000	N/A
Digital_Realty_Trust_Inc	0.396	**0.099**	0.000	N/A
Equinix_Inc	0.130	**0.174**	0.000	N/A
Equity_Residential	0.033	0.000	0.000	N/A
Essex_Property_Trust_Inc.	0.000	0.000	0.000	N/A

Extra_Space_Storage_Inc	0.013	**0.025**	0.000	N/A
Federal_Realty_Investment_Trust	0.000	0.000	**0.134**	N/A
Four_Corners_Property_Trust_Inc	0.000	0.000	0.000	N/A
HCP_Inc	0.000	**0.121**	0.000	N/A
Host_Hotels_&_Resorts_Inc	0.000	0.000	0.000	N/A
Iron_Mountain_Inc	0.029	**0.088**	0.000	N/A
JBG_SMITH_Properties	0.000	**0.128**	0.000	N/A
Kimco_Realty_Corp	0.000	0.000	0.000	N/A
Macerich_Co_(The)	0.000	0.000	0.000	N/A
Mid-America_Apartment_Communities_Inc	0.042	0.000	0.000	N/A
Public_Storage	0.000	**0.015**	0.000	N/A
Realty_Income_Corp.	0.000	**0.098**	0.000	N/A
Regency_Centers_Corp.	0.213	0.000	0.000	N/A
SBA_Communications_Corp	0.000	0.000	0.000	N/A
Simon_Property_Group_Inc.	0.000	0.000	0.000	N/A
Site_Centers_Corp	0.213	0.000	0.000	N/A
UDR_Inc	0.038	**0.038**	0.000	N/A
Vornado_Realty_Trust	0.380	0.000	0.000	N/A
Welltower_Inc	0.000	0.000	0.000	N/A
Weyerhaeuser_Co	0.000	0.000	0.000	N/A

As seen from the above charts, compared to the “Media & Entertainment” industry, a much smaller number of firms in “Energy” and “Real Estate” have positive TF-IDF values for “*snack*” or its synonyms.

국문초록

무형자산이란 조직이 보유한 자산 중 형태가 없는 자산을 뜻하며, 최근 들어 유형자산처럼 기업의 성과에 기여하는 동력 중 하나로 주목받고 있다. 그런데 정작 무엇이 무형자산인지, 무형자산의 종류에는 무엇이 있는지에 대한 연구는 활발하게 진행되어오지 않은 실정이다. 특히 직원의 관점에서 바라본 무형자산, 즉 "무형내부자산"에 대한 연구 역시 이론에 기반한 프레임워크 이상으로 이루어지지 않았다. 본 연구는 대량의 회사 리뷰 데이터에 딥러닝을 접목시켜 무형내부자산의 종류를 포괄적으로 파악하고자 했다. 이를 위해 세계 최대 회사 평점 및 리뷰 사이트인 글래스도어에서 S&P 500 회사에 대해 게재된 140만 개 이상의 리뷰 데이터를 수집했다. 방대한 양의 직원의 목소리에서 자주 등장하는 주제가 무형내부자산의 종류와 일치할 것이라고 가정한 것이다. 해당 데이터에 어텐션 기반의 뉴럴 네트워크 모델을 적용하여 24개의 주제를 추출하였고, 이 중 "직장 분위기," "동료," "기술적인 자원" 등 15개의 주제가 기존 문헌에서 언급되어온 무형자산 종류와 일치했음을 확인했다. 이후 키워드 추출 방법을 적용해 회사별로 보유한 각 무형내부자산의 특징을 파악했다. 본 연구가 제시한 방법론을 통해 회사들이 전략적인 자산을 보다 잘 이해하고 활용할 수 있을 것으로 사료된다.

주요어: 조직자산, 조직문화, 딥러닝, 텍스트 마이닝, 자연어 처리

학번: 2019-24234

감사의 글

조성준 교수님의 연구 지도, 아낌없는 조언, 그리고 격려에 깊이 감사드립니다. 석사과정을 시작하기 전부터 흥미를 가지고 있었던 주제를 연구하고 논문으로 결실을 맺을 수 있어 매우 운이 좋았던 것 같습니다. 연구의 기회를 제공해주시고 연구를 지도해주신 조성준 교수님께 큰 감사를 드립니다.

바쁘신 와중에도 시간을 내주시고 흔쾌히 논문 심사에 참여해주신 박종헌, 이경식 교수님께도 감사를 드립니다. 두 분의 수업에서 얻은 귀중한 경험과 지식은 본 연구를 진행하는데 크나큰 도움이 되었습니다.

마지막으로 데이터마이닝 연구실의 식구들께 감사의 인사를 드립니다.