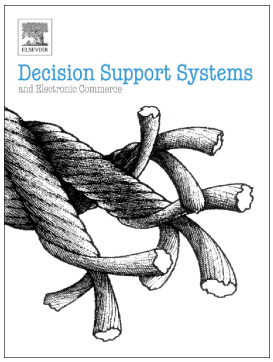

Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews

Yeonjae Jung, Yongmoo Suh

PII: S0167-9236(19)30103-4
DOI: https://doi.org/10.1016/j.dss.2019.113074
Article Number: 113074
Reference: DECSUP 113074

To appear in: *Decision Support Systems*

Received date: 14 January 2019
Revised date: 7 June 2019
Accepted date: 7 June 2019

Please cite this article as: Y. Jung and Y. Suh, Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews, Decision Support Systems, https://doi.org/10.1016/j.dss.2019.113074

Mining the Voice of Employees: A Text Mining Approach to Identifying and Analyzing Job Satisfaction Factors from Online Employee Reviews

Yeonjae Jung, Yongmoo Suh*

Business School, Korea University, 145, Anam-Ro, Seongbuk-Gu, Seoul, 02841, Republic of Korea

{yonjaejung@korea.ac.kr, ymsuh@korea.ac.kr}

First Author:
Yeonjae Jung, Researcher, Business School, Korea University
Tel: +8210-2929-3034
Fax: +82-2-922-7220
E-mail: yonjaejung@korea.ac.kr
Address: Business School, Korea University, 145, Anam-Ro, Seongbuk-Gu, Seoul, 02841, Republic of Korea

*Corresponding Author:
Yongmoo Suh, Professor of MIS, Business School of Korea University
Tel: +82-2-3290-1945
Fax: +82-2-922-7220
E-mail: ymsuh@korea.ac.kr
Address: Business School, Korea University, 145, Anam-Ro, Seongbuk-Gu, Seoul, 02841, Republic of Korea

Abstract

Online reviews have become a significant information source for business practitioners to know about customers' opinions of their products or services. Previous studies examined product or service satisfaction factors of customers by analyzing online consumer reviews. However, examining job satisfaction factors of employees through online employee reviews has rarely been studied. In this study, we first identified job satisfaction factors from 35,063 online employee reviews posted on jobplanet.co.kr using Latent Dirichlet Allocation (LDA). Then, we conducted a series of analyses based on the factors. We measured the sentiment and importance of each job satisfaction factor at industry, company, group, and chronological levels. Dominance analysis examined the relative

importance of each star-rated job satisfaction factor on overall job satisfaction. Further, the association strength between each job satisfaction factor and overall job satisfaction is computed from correspondence analysis. The results from this study will provide business managers with profound insights into making decisions on managing job satisfaction of their employees in various aspects.

Keywords**:** Online employee reviews, Job satisfaction, Latent Dirichlet Allocation, Sentiment analysis, Dominance analysis, Correspondence analysis

1. Introduction

In recent years, advancement in information and communication technologies (ICTs) and explosive proliferation of web 2.0 applications have changed the way how people interact and communicate with each other. Especially, a myriad of users of social media have been generating innumerable user-generated contents (UGCs), which are voluntarily described data, information, or media created by people, generally available on the Web [25], and containing originator's interests, opinions, experiences, etc. UGC is one of the most rapidly growing sources of information and its most prevalent type is online review.

Since online reviews contain the voice of the customers (VOC), they play a significant role in allowing us to understand the factors that their reviewers deem most important [17] and capture the degree of satisfaction in each factor. This kind of reviews convey useful and critical meanings to their readers, such as business managers and customers. For business managers, they can be a key to figuring out a market response, namely satisfaction or dissatisfaction with their products and services [28]. Also, they can help customers avoid uncertainty before making a purchasing decision [37]. Yet, it is virtually impossible for a human to read a voluminous amount of reviews.

Owing to the opportunity and difficulty of analyzing the online reviews, various attempts have been made by researchers and business practitioners to mine reviews for useful knowledge using a text mining approach. However, most of the prior studies have focused on analyzing the product or service

satisfaction of *external customers* from online consumer reviews [1, 14, 56, 58]. Only a few studies have an interest in analyzing the job satisfaction of *internal customers* (i.e., employees). Since job satisfaction is related to employee motivation, performance, absenteeism, and turnover [24], maintaining employees to be in a high level of job satisfaction is important for achieving a competitive advantage. Prior studies analyzing online employee reviews tend to focus on relationships between job satisfaction factors and firm performance [35] or between job satisfaction factors and employee retention/turnover [29], on the value proposition for job seekers [12], and on discovering negative reviews [16]. However, there is still a lack of studies on identifying employees' job satisfaction factors from online employee reviews and deriving managerial insights from those factors. If online reviews contain what employees like or dislike about their companies, they can be a great source for identifying job satisfaction factors and analyzing employees' thoughts about these factors. Moreover, mining a large number of online reviews seems to have the effect of alleviating some limitations (e.g., finite questionnaire items, generalizability, etc.) of traditional survey methods (e.g., MSQ [54], JDI [44], JSS [46], JDS [18], etc.) which have been widely used to measure job satisfaction [29, 35].

Having realized that it is valuable to analyze the employees' reviews in diverse aspects for the better management of human resources, we propose a research framework (see Fig. 1) for identifying job satisfaction factors and evaluating employees' opinions about the factors. More specifically, we aim to provide answers to the following questions:

RQ1. What are the key factors of the job satisfaction expressed in reviews?

RQ2. What are the sentiment and the relative importance associated with each of these factors?

RQ3. What are the most important factors influencing overall job satisfaction?

RQ4. How do these factors vary across the overall star-rates on job satisfaction?

To provide answers to the research questions above, we first collect and extract employees' job satisfaction factors from online employee reviews posted on Jobplanet.co.kr[1], the largest online

[1] https://www.jobplanet.co.kr/

company review site by employees in South Korea. Then, we perform diverse analyses, including sentiment, importance, dominance, and correspondence analyses.

Contributions of this paper can be summarized as follows. First, we suggested a way of determining the proper number of topics and validated their quality by a reliability test and external validity test. Second, the job satisfaction factors we provided and the results of diverse analyses we ran based on the factors will enable the managers of HR departments to plan, design and implement HR related activities. Each of those activities will be supposed to lead employees to have better job satisfaction than before, and they will eventually change their companies more competitive. Finally, we found five new job satisfaction factors which were not considered in the literature.

The rest of this paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the details of the research framework and methods utilized in our study. Section 4 includes experimental results. In Section 5, we discuss research results in terms of implications and future work. Finally, we conclude our paper in Section 6.

2. Related works

2.1 Data mining in human resource management

Like other functional departments, business managers of HR departments should make decisions on various issues in their departments. As more data, including textual data, are generated and accumulated inside or outside of their companies, they can make use of the actionable knowledge obtained from such data using data mining techniques for their decision making. Recently, a rapidly increasing number of HRM-related research adopted data mining, including text mining, and suggested a new paradigm for producing advanced information for decision support [47].

Strohmeier and Piazza [47] conducted the comprehensive literature review on HRM-related research which uses data mining techniques and found *staffing*, *development*, *performance management*, and *compensation* are the most frequently researched subcategories. Moreover, they found that only a few research papers utilized textual data, which may include hidden values for HRM

and argued that there still remain substantially potential contributions of data mining for the knowledge in HR domain. Strohmeier and Piazza [48] suggested general scenarios of adopting data mining techniques for several issues in HRM. For example, authors showed that text mining approach to examining web-based documents on employer ratings can be utilized to know the sentiment of employees about several HRM-relevant aspects (e.g., *compensation ratio*, *career possibilities*, *quality of training*, *leadership style*, *work climate*, etc.).

Several studies applied text mining approach to online employee reviews [12, 16, 29, 35]. Luo et al. [35] built a corporate performance prediction model using OLS regression and over 250,000 online employee reviews posted on Glassdoor.com. They found higher employee satisfaction tends to increase higher corporate performance. Dabirian et al. [12] collected 38,000 online reviews of Glassdoor.com's best 10 and worst 10 firms to work for and used IBM Watson to analyze the data. Having found seven employer branding value propositions (i.e., *social value*, *interest value*, *application value*, *development value*, *economic value*, *management value*, and *work-life balance value*), they compared relative valences (i.e., positive or negative) and weights (i.e., importance) of them across the best and the worst firms. Lee and Kang [29] performed topic modeling using LDA on the online employee reviews obtained from Glassdoor.com. They found job satisfaction factors positively related to the retention group and those negatively related to the turnover group. In addition, they evaluated the relative importance of job satisfaction factors for each group and found *culture and value*, and *senior management* are the most influential job satisfaction factors on retention and turnover groups, respectively. Goldberg and Zaman [16] designed and implemented an HR domain-specific text analytics tool for predicting dissatisfied employee reviews and prioritizing those reviews to discover the most urgent issues of employees' dissatisfaction. They used 200,000 randomly selected reviews obtained from Indeed.com for generating smoke words for predicting dissatisfied reviews and evaluating the efficacy of those words to rank the most urgent reviews.

However, the above studies using online employee reviews focused on a narrow analysis to find relationships between job satisfaction factors and a few detailed research topics, (e.g., *firm*

performance, *employee retention* or *turnover*, *value proposition*, etc.). We believe that it is worthwhile to derive more comprehensive knowledge from what employees have in their mind regarding their company, thereby providing insights into many issues in HRM, which is the objective of our study.

2.2 Topic modeling

As the number of social media platforms increases which produce lots of textual data, a traditional manual approach to analyzing textual data has faded out and is being replaced with text mining techniques, including text summarization, keyword extraction, sentiment analysis, topic modeling, etc. Among these, topic modeling is used to capture the hidden structures (e.g., per-document topic distribution, per-topic word distribution, etc.) in documents. Latent Dirichlet Allocation (LDA), one of the topic modeling algorithms, assumes that each document consists of probabilistically distributed topics and each topic can be represented by probabilistically distributed words [4], to infer the hidden structures using Bayesian inference, given a set of documents [5].

One of the open questions in using LDA is how to determine an appropriate number of topics from a corpus [5, 7]. If an inappropriate number, either too big or too small, is chosen, the interpretation of the topics can be hampered. Chang et al. [9] showed that the human interpretability of topics is not always necessary when LDA is used to vectorize the documents for a predictive model, while the interpretability of topics needs to be secured when LDA is used to understand the contents of documents. For the latter case, Boyd-Graber et al. [7] suggested to check whether the individual topics and their assignments to documents are meaningful and coherent. Debortoli et al. [13] recommended to vary the number of topics from 10 to 50 and assess the interpretability of generated topics to decide the appropriate number of topics. This method is quite an effective way, but relatively more time-consuming. Blei et al. [5] proposed a perplexity measure to estimate the suitable number of topics, with lower perplexity measure indicating better topic modeling. However, a low perplexity score may not always improve the human interpretability of extracted topics [3, 9]. Guo et al. [17] determined the quality of topics extracted from online hotel reviews by evaluating face validity and external validity.

The former is evaluated by comparing topics extracted from LDA and topics generated by human reviewers, while the latter by comparing topics extracted from LDA and items appeared in questionnaires from prior literature. Another approach to determining the quality of topics is exploiting secondary data, such as Wikipedia or news articles [27, 36, 40]. If the most probable keywords of each topic obtained from LDA co-occur in the article found from Wikipedia or news using the topic as a keyword, then the quality of the topic is considered high. In sum, there is no best practice yet to determine the appropriate number of topics and measure the quality of topics. In our study, we extract the number of topics via using both the perplexity score and the hierarchical clustering analysis. Afterward, the quality of the extracted topics is assessed by the reliability test and external validity test.

LDA has been used widely in prior studies, to discover key dimensions of hotel service from online review [17], to identify influential subjects from tweets about Uber experience [42], to derive a set of variables to predict the success of crowdfunding project [57], to capture key aspects of smartphones from the reviews [28], to cluster landmarks to recommend and optimize tourists' traveling plans [49] etc.

3. Research framework

In this section, we explain our research framework depicted as Fig. 1. The framework contains the preprocessing of the online reviews, including POS tagging, especially the treatment of neologisms and compound nouns which are frequently used on online, topic modeling and a series of analyses based on the derived topics.

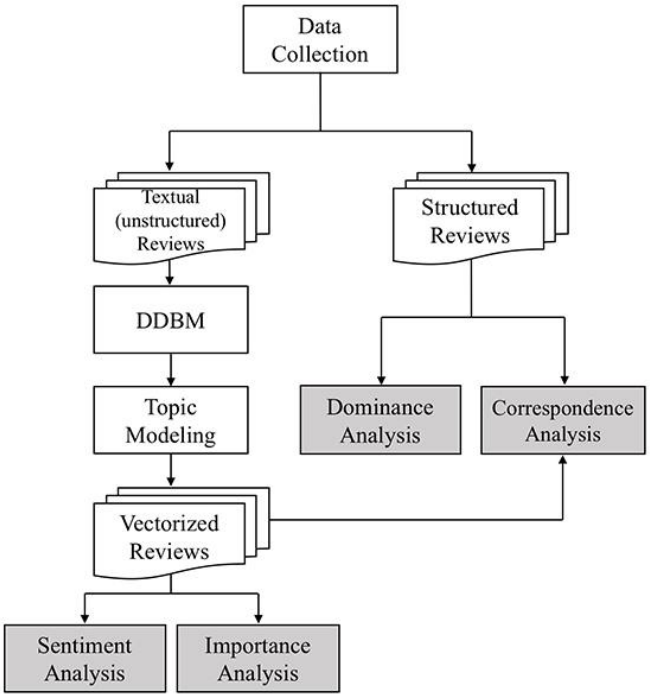

Fig. 1. Proposed research framework.

3.1 Domain Dictionary Building Module for POS tagging

Though documents contain thousands of words, it is not always valuable to use all of them in text mining because of the '*curse of dimensionality*'. Thus, it is essential to remove irrelevant words for better analytical results. In this study, we only use nouns for analyses because they are the most representative tokens in a document [33, 39]. However, extracting nouns from the corpus is still challenging because of the '*out-of-dictionary problem*[2]', which occurs frequently in POS tagging for online reviews due to the Internet neologisms (e.g., offitics[3], etc.) and compound nouns (e.g., working condition, retirement annuity, etc.). Since the morphological analyzers for Korean (e.g., KOMORAN, KKMA, Hannanum, MeCab-ko, etc.) draw upon only the system dictionary[4], we devise and implement the domain dictionary building module (DDBM), which automatically captures neologism and compound nouns from online reviews and build a domain dictionary for POS tagging with a little

[2] Some words do not appear in a dictionary that we use for tokenizing and POS tagging.

[3] Office politics

[4] An embedded dictionary which the morphological analyzers use for POS tagging

human intervention. Fig. 2 shows the blueprint of the DDBM (left) and the example of how DDBM works (right). The process of DDBM's generating the domain dictionary is described step by step as follows:

Step 1. Rectify spacing errors in the corpus (using Soyspacing[5]).

Step 2. Repeat the following steps until neither new compound nouns nor neologisms are added to the domain dictionary.

Step 2.a. Conduct POS-tagging on each review in the corpus (using KOMORAN[6]).

Step 2.b. Extract nouns from the corpus.

Step 2.c. Perform a bigram collocation analysis (using NLTK[7]) on extracted nouns. Afterward, let human experts examine the list of bigrams to determine compound nouns and neologisms.

Step 2.d. Build the domain dictionary by adding compound nouns and neologisms.

Step 3. Conduct POS tagging on the corpus, whose spacing is rectified after Step 1, using KOMORAN with both the system dictionary and the domain dictionary.

Step 4. Extract only nouns from the POS tagged corpus for topic modeling.

[5] Heuristic algorithm for correcting errors in spacing words (rf. https://github.com/lovit/soyspacing)

[6] https://github.com/shineware/komoran-2.0

[7] https://www.nltk.org

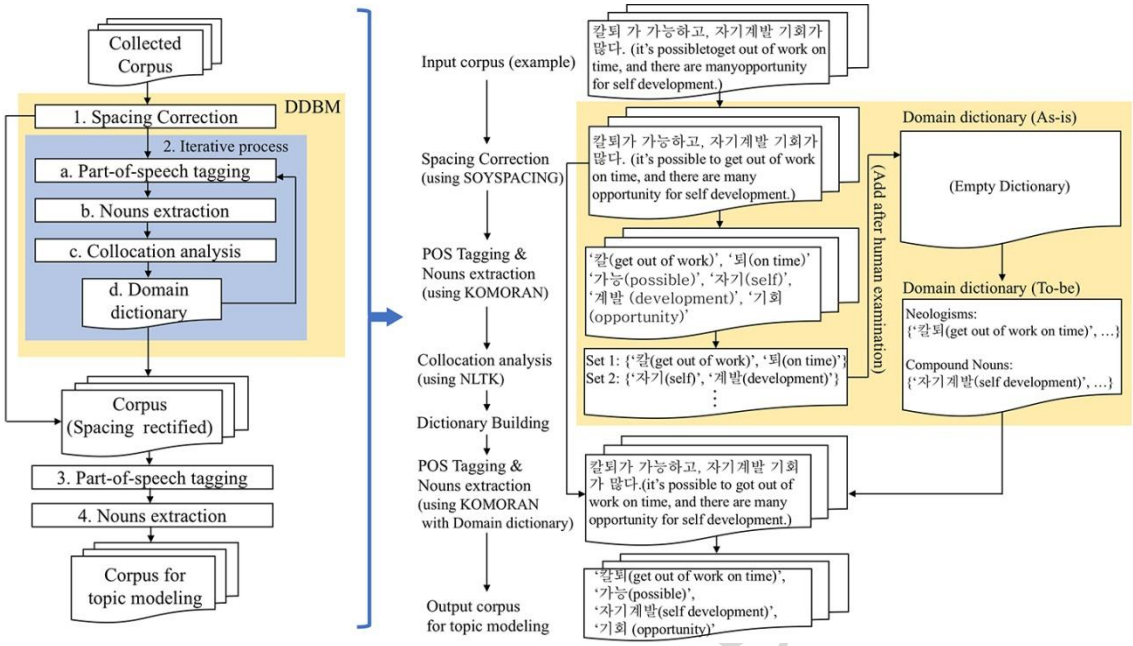

Fig. 2. DDBM for POS tagging (left) and the example (right).

3.2 Topic modeling for review vectorization

We use LDA to grasp the hidden structures of online employee reviews. In other words, we discover per-factor word distributions (i.e., per-topic word distributions) for identifying job satisfaction factors and per-review factor distributions (i.e., per-document topic distributions) for vectorizing each of reviews for further analyses. Note that topics extracted from LDA correspond to employees' job satisfaction factors.

As shown in Fig. 3, users are allowed to write 'Pros' and 'Cons' of a firm in separate sections in Jobplanet.co.kr. Thus, we perform topic modeling using both of them. Each 'Pros' and 'Cons' section in a single review is vectorized as follows:

$$Pr_i = \{T_{1,i}^P,\ T_{2,i}^P, \dots,\ T_{|F|,i}^P\} \times n(Pr_i) = \{P_{1,i},\ P_{2,i}, \dots,\ P_{|F|,i}\} \quad (i = 1, 2, \dots,\ n), \tag{1}$$

$$Cr_i = \{T_{1,i}^C,\ T_{2,i}^C, \dots,\ T_{|F|,i}^C\} \times n(Cr_i) = \{C_{1,i},\ C_{2,i}, \dots,\ C_{|F|,i}\} \quad (i = 1, 2, \dots,\ n), \tag{2}$$

where Pr_i (Cr_i) is a 'Pros' ('Cons') section vector in the i-th review R_i, n is the number of collected reviews, $T_{k,i}^P$ ($T_{k,i}^C$) is an occurrence probability of the k-th job satisfaction factor in Pr_i (Cr_i), $n(Pr_i)$ ($n(Cr_i)$) is the number of nouns in 'Pros' ('Cons') section, and $|F|$ is the number of job satisfaction

factors. Since each $T_{k,i}^{P}$ ($T_{k,i}^{C}$) only represents the relative occurrence probability of each job satisfaction factor in Pr_i (Cr_i), we multiply $n(Pr_i)$ ($n(Cr_i)$), to each review vector to obtain the number of nouns allocated for each $T_{k,i}^{P}$ ($T_{k,i}^{C}$). Afterward, we vectorize each review by concatenating Pr_i and Cr_i. Thus, each review R_i is vectorized as follows:

$$R_i = \{P_{1,i},\ P_{2,i}, \dots,\ P_{|F|,i},\ C_{1,i},\ C_{2,i}, \dots,\ C_{|F|,i}\} \qquad (i = 1, 2, \dots, n), \qquad (3)$$

where $P_{k,i}$ ($C_{k,i}$) is the number of words assigned to the k-th job satisfaction factor in 'Pros' ('Cons') section of an i-th review R_i.

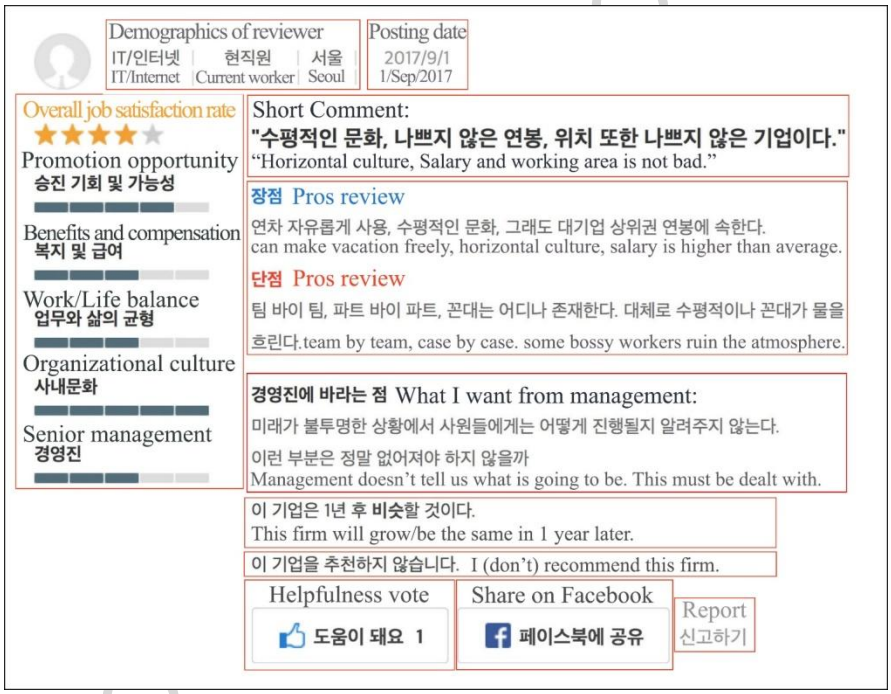

Fig. 3. An example of the review from Jobplanet.co.kr.

Meanwhile, it is crucial to determine the number of factors. We try to reduce human efforts in determining the number of factors by using the perplexity score [5]. The lower perplexity score implies the lower uncertainty of resulting factors. However, the perplexity score tends to be lowered as the number of factors increase. In other words, depending only upon the perplexity score could lead to generating numerous meaningfully overlapping factors which may decrease the human interpretability. Thus, we use a hierarchical clustering analysis with Ward's method [53] to merge the overlapping

factors and enhance the interpretability. Then, to evaluate the reliability of the final set of factors, we compared them with the topics assigned by humans. And to evaluate the external validity of the factors in the set, we compared them with the factors used in the literature.

3.3 Sentiment and importance of job satisfaction factors

We analyze the sentiment and relative importance of job satisfaction factors obtained from online employee reviews at several levels: industry level, company level, group level, and chronological level. We consider each level of analyses may help business managers in HR departments monitor how employees think about their company and make a decision on which factors should be managed for better satisfaction.

In this study, we devised Eq. (4) in order to estimate the sentiment of each job satisfaction factor. We use the frequency of nouns describing certain job satisfaction factor for calculating the sentiment. Assuming that a reviewer is satisfied (dissatisfied) with a certain job satisfaction factor if the words associated with the factor appears in 'Pros' ('Cons') section, we calculate the sentiment of a job satisfaction factor f_k in reviews as follows:

$$Sentiment(f_k) = \frac{\sum_{i=1}^{n}(P_{k,i} - C_{k,i})}{\sum_{i=1}^{n}(P_{k,i} + C_{k,i})}, \quad (4)$$

where n, $P_{k,i}$ and $C_{k,i}$ have the same meanings as those in Eq. (3). $Sentiment(f_k)$ is more positive as it is closer to 1 and more negative as it is closer to -1.

It is also worthwhile to examine important job satisfaction factors because they are the factors to which the management of the corresponding company has to pay attention to get a better evaluation of the company from employees. Assuming that a certain job satisfaction factor is important if words associated with the factor appears frequently in the reviews, we define the importance of a *k*-th job satisfaction factor f_k as follows:

$$Importance(f_k) = \frac{\sum_{i=1}^{n}(P_{k,i} + C_{k,i})}{\sum_{j=1}^{|F|}\sum_{i=1}^{n}(P_{j,i} + C_{j,i})}, \quad (5)$$

where $|F|$, n, $P_{k,i}$ and $C_{k,i}$ have the same meanings as those in Eq. (3). $Importance(f_k)$ is higher as it is closer to 1.

3.4 Dominance analysis

Understanding which factors are most influential on the overall job satisfaction rate may provide managerial insights for business managers in the HR departments. For example, business managers may prioritize the factors based on the relative importance of each factor. Guo et al. [17] used multiple regression analysis to discover the relative importance of variables in online hotel reviews. However, utilizing the regression coefficients is not suitable for determining relative importance when independent variables have correlations with each other [51]. Since dominance analysis can be utilized in such a case, we conduct not only regression analysis but also dominance analysis to examine which factors are most influential on the overall job satisfaction rate. We use each of the five specific rates obtained from the employees' reviews as the independent variable and the overall job satisfaction rate as the dependent variable in the online employee reviews (see Fig. 3).

In the dominance analysis, the relative importance of a particular variable is measured by the average increment in R^2 when the variable is added to every possible regression model built without it [8]. Lee and Kang [29] used dominance analysis in their study on job satisfaction. However, they did not make judgments on the statistical significance of the relative importance. In this study, we adopt a bootstrap method in the study of Tonidandel and LeBreton [51] to determine the statistical significance of relative importance. Moreover, since we have a relatively large dataset (i.e., over 35,000 reviews), this method is also effective in regression analysis to avoid a pitfall of the '*p*-value problem': as the size of data is increased, *p*-values go quickly to zero [31]. In other words, depending only on *p*-values for hypothesis testing with a large dataset could yield the problem of supporting the hypotheses of very small or no practical significance [31]. The bootstrap method for dominance analysis and regression analysis are conducted as follows. First, we add a randomly generated variable to our original dataset. This variable has no effect on a dependent variable (i.e., the overall job

satisfaction rate) in the regression model; thus, its relative importance is zero. Second, we produce plentiful resampled datasets through sampling with replacement of the original dataset with the randomly generated variable. In this study, we generate 2,000 resampled datasets, each with 200 instances. Third, for each resampled dataset, we conduct regression and dominance analysis to uncover the coefficients and relative importance of each of the five specific rates. Finally, for determining the statistical significance of relative importance, we compare the relative importance of each of the five specific rates with that of the randomly generated variable; and for regression analysis, we average out statistics such as regression coefficients, *p*-values, adj. R^2, etc.

3.5 Correspondence analysis

The correspondence analysis (CA) is a technique for exploring the relationships among categorical variables [20]. We perform CA to investigate the relationship between the overall job satisfaction rate and each job satisfaction factor obtained from textual reviews. The CA may help business managers in HR departments learn how the most important factors vary across employees' star-rates (from 1 to 5) about overall satisfaction. We used the overall job satisfaction rate and the most important job satisfaction factor in each review for CA. The result of CA is presented graphically to simplify its interpretation. Distances on the graph (see Fig. 10) between the rates and the factors determine the association strength between them.

4. Experiments and results

4.1 Data collection from Jobplanet.co.kr

Our online employee review dataset was retrieved from one of the most popular company review sites in South Korea, namely Jobplanet.co.kr, using a Python crawler we developed. We gathered a total of 232,400 reviews from 2014 to 2017. After we eliminated firms that had reviews less than 10,

our final dataset includes 204,659 reviews from 4,347 firms in 10 industries. Table 1 shows the summary statistics of the final dataset.

Table 1

Summary statistics of the final dataset.

Industry	Number of firms	Number of reviews	Mean length of 'Pros' section in words	Mean length of 'Cons' section in words	Mean of overall job satisfaction rate
IT	844	35,110	61.4	76.5	2.8
Finance	208	12,118	58.4	71.0	3.1
Construction	162	6,003	59.4	74.5	2.9
Education	156	8,154	64.2	84.7	2.6
Logistics	566	28,112	59.7	76.9	2.7
Manufacturing	1,282	64,245	58.4	74.2	2.7
Service	290	14,659	61.0	77.4	2.7
Medical	202	8,346	56.7	72.4	2.8
Media	356	13,859	65.4	84.3	2.7
Organization	281	14,053	66.9	77.7	3.1

4.2 Textual review vectorization

In this study, vectorized forms of textual reviews serve as the foundation of all experiments we conducted. Employees engaging in the same industry tend to show resemblance in terms of task characteristics, job characteristics, organizational cultures, interest, etc. Therefore, we make a hypothesis that topics and words presented on the online employee reviews will be different across industries. Accordingly, in our study, reviews about the companies in the IT industry were chosen for the vectorization and further analyses.

4.2.1 Domain Dictionary Building Module

As a result of implementing DDBM for POS tagging on textual reviews, a list of thousands of candidate neologisms and compound nouns was returned in descending Pointwise Mutual Information (PMI) order. Three graduate students who major in information systems thoroughly examined top 6,000 words in the list to pick words to build a domain dictionary. We only used words which had

been selected unanimously by the three graduate students. During the first iteration, 713 neologisms and compound nouns were found to be included in the domain dictionary. 20 compound nouns were additionally added to the dictionary in the second iteration. In this iteration, some trigram compound nouns (e.g., case by case, work-life balance, Pangyo Techno Valley[8], etc.) were also found, part of which were detected as bigram compound nouns in the first iteration.

After the second iteration, no words were additionally found in the next iteration. As a result, a total of 733 words was registered in the domain dictionary. Table 2 presents a part of the list of the neologisms and the compound nouns included in the domain dictionary. Utilizing both the domain dictionary and the system dictionary, the DDBM puts the corresponding POS as a tag on each word in the corpus.

Table 2

A part of a list of words registered in the domain dictionary.

Number of iteration	Neologism	Compound nouns
1	샤뱌샤뱌(flattering), 복불복(taking pot luck), 칼퇴(leaving on time), 워라밸(work-life balance), 케바케(case by case), 여초(women outnumbered men), 열정페이(paying small but forcing for working passionately), 월급도둑(getting paid without working), 복붙(copy and paste), 대기업코스프레(pretending to be a large company), etc.	기혼여성(married women), 단체생활(group life), 퇴직연금(retirement annuity), 성과측정(result evaluation), 직업특성(job related characteristics), 근무여건(working condition), 교육컨텐츠(education contents), 모바일광고(mobile advertisement), 여성친화(female-friendly), 통신요금(telecommunication fee), 고위직급(upper management), 데이터분석(data analysis), 편의시설(amenity), 직무변경(department transfer), 승진기회(promotion opportunity), 탁상행정(bureaucracy), etc.
2		포털서비스(portal service), 음악서비스(music streaming service), 이동통신사(mobile carrier), 케이스바이케이스(case by case), 모래시계구조(sandglass structure),

[8] An IT cluster in South Korea

	개인정보보호(private information protection), 판교테크노밸리(Pangyo Techno Valley), 워크라이프밸런스(work-life balance), etc.

4.2.2 Topic modeling

We only extracted noun words in the corpus for topic modeling. Afterward, stop words were also eliminated and reviews containing no nouns were excluded from the corpus. As a result, a total of 13,631 nouns from 35,063 reviews remained for topic modeling. Prior to building a topic model from the reviews, we performed a grid search to determine the optimal number of topics extracted from the corpus through LDA and found that the perplexity score becomes the lowest (451.6) when the number of topics is 65. Since only a very small difference in perplexity score is expected after 65, we set and obtained 65 job satisfaction factors from the corpus through LDA.

Unsurprisingly, as the number of the extracted topics was relatively big, a number of meaningfully overlapping factors were found and as such the interpretability of those factors was relatively low. In order to combine the overlapping factors, thereby enhancing interpretability, we exploited a hierarchical clustering analysis. The three graduate students who worked together to build the domain dictionary scrutinized and labeled each clustered job satisfaction factor (i.e., topic). As a result, we obtained 30 job satisfaction factors from online employee reviews, as is shown in Table 3.

Table 3

Job satisfaction factors obtained from textual reviews.

No.	Factors	Keywords
1	Vacation	휴가(vacation), 연차(annual leave), 여름휴가(summer vacation), 리프레쉬(refresh), 월차(monthly leave)
2	Organizational culture	조직문화(organizational culture), 기업문화(firm culture), 분위기(atmosphere), 자유(free), 강요(oppression)
3	Work intensity & efficiency	업무(task), 강도(intensity), 효율(efficiency), 과도(excess), 야근(overtime work)
4	Working hour	근무시간(working hour), 퇴근(leaving work), 출근(going to work), 야근(overtime work), 주말출근(working on the weekend)
5	Project	프로젝트(project), 투입인력(labor input), 개발일정(schedule),

		관리(management), 수주(win a contract)
6	Self-development	발전(advancement), 기회(opportunity), 경험(experience), 성장(growth), 승진(promotion)
7	Operating procedure	업무보고(operational report), 업무프로세스(business process), 체계(system), 절차(procedure), 처리(process)
8	Work-life balance	삶(life), 업무(work), 균형(balance), 개인생활(private life), 가정(family)
9	Software development	개발(development), 디자인(design), 서비스(service), 솔루션(solution), 유지보수(maintenance)
10	Inter-firm relationship	고객사(client company), 갑을관계(contract relation), 상대그룹(opponent company), 그룹사(affiliate), 본사(headquarter)
11	Working area	지역(area), 사이트(site), 위치(location), 접근성(accessibility), 근무지(working area)
12	Growth & profitability	성장(growth), 매출(sales), 신규투자(new investment), 정체(stagnation), 안정(stability)
13	Reputation	인지도(awareness), 시장점유(market share), 업계(industry), 최고(top), 자부심(pride)
14	Marketing	마케팅(marketing), 광고(advertisement), 홍보(promotion), 미디어(media), 브랜드(brand)
15	Overseas business	글로벌(global), 외국(foreign), 해외(overseas), 해외출장(overseas business trip), 영업(sales)
16	Firm image	이미지(image), 대외(external), 외부(outside), 이름(name), 인지도(awareness)
17	Attitude to change	비전(vision), 미래(future), 변화(change), 보수(conservatism), 정체(stagnation)
18	Sales & performance pressure	영업(sales), 고객(client), 실적(performance), 실적압박(performance pressure), 스트레스(stress)
19	Supervisor competency	팀장(team leader), 능력(ability), 임원(executive), 리더(leader), 실력(competency)
20	Decision making	의사결정(decision making), 의견(suggestion), 말(comment), 결정(decision), 권한(authority)
21	Organizational politics	사내정치(politics in the firm), 정치(politics), 라인(line), 파벌(party), 싸움(conflict)
22	Form of employment	계약직(contract worker), 정규직(permanent worker), 비정규직(temporary worker), 인턴(intern), 채용(recruitment)
23	Organizational structure	수평(horizontality), 수직(verticality), 조직(organization), 구조(structure), 문화(culture)
24	Supervisor (human relations)	경영진(management), 마인드(mind), 생각(think), 배려(consideration), 소모품취급(treating as consumable goods)
25	General welfare	복지(welfare), 이벤트(event), 동호회(club), 카페테리아(cafeteria), 헬스장(fitness center)
26	Welfare for women	여자(woman), 여성(female), 결혼(marriage), 육아휴직(parental leave), 출산휴가(maternity leave)
27	Financial	제공(offer), 지원(support), 지급(payment), 교통비(transportation

	support	fee), 식대(food expenses), 의료비(medical expenses), 도서구입비(book cost), 통신비(telecommunication expenses), 학자금(educational expenses)
28	Salary	월급(monthly salary), 연봉(annual salary), 급여(pay), 임금(wage), 보수(remuneration)
29	Human resource	채용(recruitment), 승진(promotion), 직급(position), 입사(joining a company), 퇴사(resignation)
30	The others	직원(employee), 시간(time), 상태(state), 매출(sales), 문화(culture), etc.

We evaluated the reliability of our job satisfaction factors to judge whether our factors are trustworthy or not by calculating an inter-rater agreement using Cohen's Kappa coefficient [11, 26]. We asked two graduate students (hereafter, 'student A' and 'student B') who are independent from our study to extract job satisfaction factors in each of the 300 randomly sampled reviews. They were not given any prior knowledge about the list of factors and how LDA classified each review. Also, student A did not know how student B labeled each review and *vice versa*.

From 300 reviews, students A and B identified 27 and 24 factors, respectively. 25 factors are common to student A and LDA, and 21 factors to student B and LDA. Only student A found *Department* and *Training*; and only student B found *Business environment*, *Coworker*, and *Building* (e.g. old building) as job satisfaction factors from the sampled reviews. We selected the most common nine factors that the two students identified in their labeling task to test the reliability of LDA, the result of which is given in Table 4. As shown in Table 4, the degrees of the agreement between students A and B are relatively high for all of the nine factors. That is, the degrees of the agreement are almost perfect for six factors and substantial for three factors. Although the degrees of the agreement between LDA and student A and between LDA and student B are relatively lower than those between two students, they are fair or moderate in most cases. Fig. 4. represents the number of reviews, in which each of the most common nine factors is identified. Since LDA is an unsupervised topic modeling algorithm, the Kappa coefficients in Table 4 indicate that LDA is reliable to some extent, and LDA is useful especially when dealing with a huge number of reviews.

Table 4

The result of the reliability test of job satisfaction factors.

Factor	LDA-Student A		LDA- Student B		Student A- Student B	
	Overlap	Kappa	Overlap	Kappa	Overlap	Kappa
Vacation	268 (20)	0.496 (Moderate)	263 (11)	0.558 (Moderate)	289 (31)	0.896 (Almost)
Organizational culture	223 (104)	0.487 (Moderate)	213 (93)	0.415 (Moderate)	270 (130)	0.800 (Almost)
Work intensity & efficiency	241 (33)	0.404 (Moderate)	221 (36)	0.311 (Fair)	258 (57)	0.640 (Substantial)
Working hour	230 (26)	0.341 (Fair)	229 (27)	0.337 (Fair)	281 (84)	0.852 (Almost)
Self-development	219 (15)	0.155 (Slight)	218 (15)	0.151 (Slight)	277 (65)	0.798 (Substantial)
Organizational politics	273 (5)	0.233 (Fair)	270 (4)	0.211 (Fair)	295 (25)	0.900 (Almost)
General welfare	195 (33)	0.301 (Fair)	187 (30)	0.329 (Fair)	278 (110)	0.848 (Almost)
Salary	197 (36)	0.204 (Fair)	196 (37)	0.207 (Fair)	279 (107)	0.853 (Almost)
Human resource	206 (39)	0.355 (Fair)	205 (39)	0.319 (Fair)	273 (43)	0.706 (Substantial)

Overlap: the number of overlapping reviews (the number of overlapping reviews of each of the most common nine factors); Kappa: Cohen's Kappa coefficient (the interpretation of the Kappa coefficient).

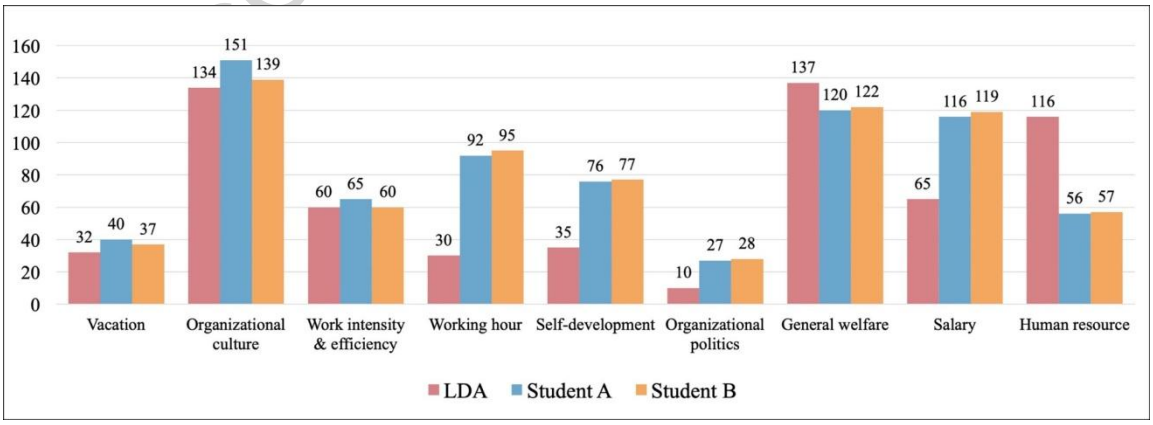

Fig. 4. The number of reviews in which each of the most common nine factors is identified.

To perform the external validity test of the job satisfaction factors from LDA, we compared them with the job satisfaction factors examined in the literature related with HRM [2, 6, 10, 15, 19, 21, 22, 23, 30, 32, 34, 38, 41, 43, 45, 46, 50, 52, 54, 55]. Among the 30 job satisfaction factors, five factors (i.e., *Project*, *Software development*, *Inter-firm relationship*, *Marketing*, and *Overseas business*) were not found in the literature. However, factors such as *Coworker*, *Autonomy*, and *Training* from the literature were not identified through LDA since the occurrence probabilities of those factors are lower than that of each of the 30 job satisfaction factors in our collected reviews.

4.3 Sentiment and importance analysis

Having extracted 30 job satisfaction factors from the online employee reviews, we analyzed the sentiment and the importance of each factor at four levels: industry level, company level, group level, and chronological level. For the better readability and interpretability, we only present some of the most common nine job satisfaction factors, when necessary, in some figures below.

4.3.1 Industry level

Industry level analysis uses all the vectorized reviews on companies belonging to the IT industry, containing 844 firms and 35,063 reviews. The result of this level of analysis shows the overall sentiment and importance of each job satisfaction factor in the IT industry in South Korea, which is shown in Fig. 5. The horizontal and the vertical axes in the figure represent the degree of sentiment and importance of each factor, respectively. The gray horizontal line in the figure represents the median of relative importance. The size of each bubble indicates the relative number of the reviews mentioned each job satisfaction factor. According to Fig. 5, *Organizational culture* is relatively more positive and important than the other factors. In other words, employees in the IT industry consider those factors to be important and are satisfied with those factors. On the other hand, *Human resource* is perceived as important but negative.

Chronological dynamics of job satisfaction factors were also analyzed. This analysis shows how the degree of sentiment and importance of each factor varies by year in Fig. 6. According to the figure, *Organizational culture* (importance: $F = 92.389$, $p < 0.001$) is more important in 2016 than in 2015 and in 2017. *Vacation* (importance: $F = 448.9$, $p < 0.001$) remains important every year. *Human resource* (sentiment: $F = 5.473$, $p = 0.019$; importance: $F = 20.287$, $p < 0.001$) is more negative and important in 2017 than in 2015 and in 2016. However, as the yearly change in each factor is marginal, changing at the industry level seems to be monotonous.

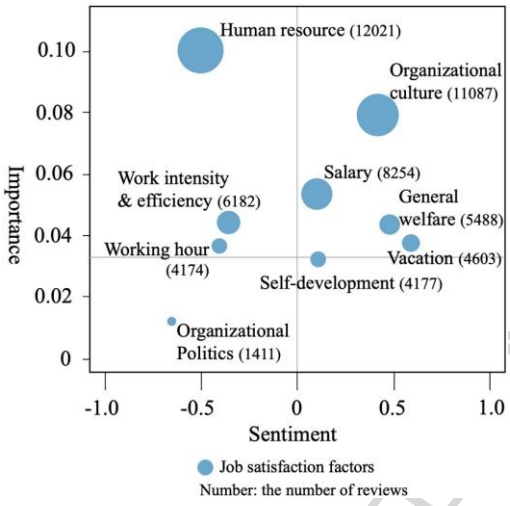

Fig. 5. Sentiment and importance of the factors in the industry level.

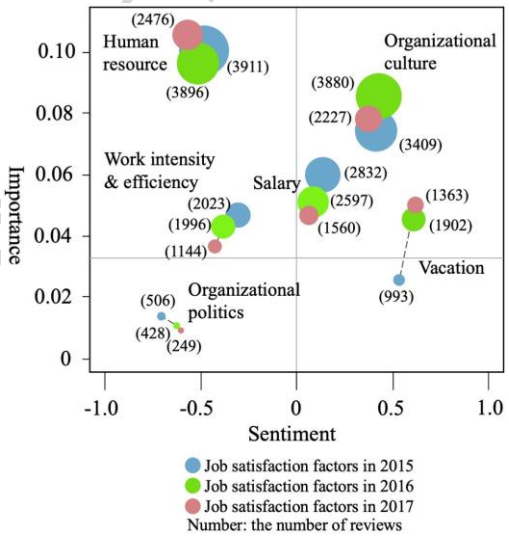

Fig. 6. Chronological dynamics of the job satisfaction factors in the industry level.

4.3.2 Company level

Company level analysis calculates the sentiment and the importance of each job satisfaction factor by a firm, thereby making it possible to compare the differences in the job satisfaction factors among firms. For this level of analysis, we selected the reviews of two large firms (hereafter, 'company A' and 'company B') in South Korea. They are in a competitive relationship. We obtained 746 and 705 reviews for company A and company B, respectively. Fig. 7 illustrates the result of the company level analysis. As shown in the figure, *Salary* (sentiment: $t = 10.026$, $p < 0.001$) is shown as relatively more

positive in company A than in company B. The sentiment and the importance of *Human resource* (sentiment: t = 10.582, p < 0.001; importance: t = -107.29, p < 0.001) are shown as relatively more negative and important in company B than in company A.

The analysis of the chronological dynamics at the company level was also done. For this analysis, we divided the reviews on company A by year. Fig. 8 represents the result of a chronological analysis of company A. *Vacation* (sentiment: F = 10.646, p = 0.001; importance: F= 15.251, p < 0.001) is noticeably positive and important every year. Although *Organizational culture* (sentiment: F = 0.014, p = 0.903; importance: F= 0.643, p < 0.423) is more positive and important in 2016 than in 2015, it is significantly negative and less important in 2017 than in 2016. However, the differences were statistically insignificant.

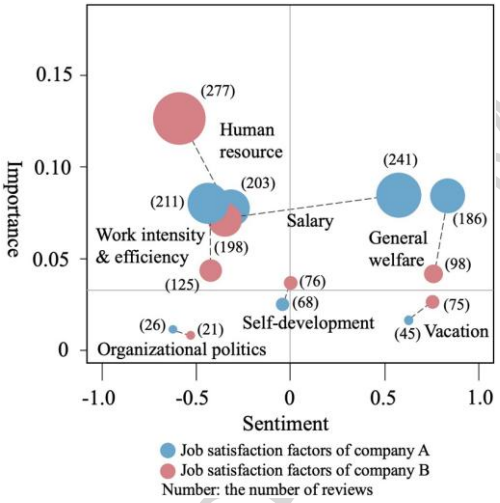

Fig. 7. Sentiment and importance of the factors in the company level.

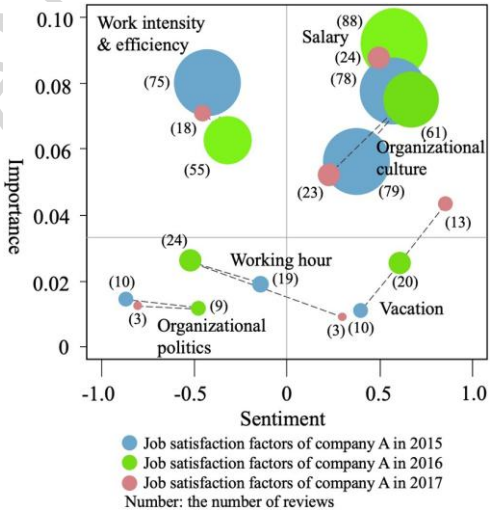

Fig. 8. The chronological dynamics of the factors in company A.

4.3.3 Group level

Group level of analysis is beneficial in comparing the differences in the job satisfaction factors among groups. In our study, we defined two distinctive groups: a current workers group and a former workers group. Fig. 9 shows the sentiment and the importance of the job satisfaction factors of the two groups. As shown in

Fig. 9, eight out of nine factors are more positive in the current workers group than in the former workers group. Only one factor (i.e., *Human resource*) is more negative in current workers group. The result of a *t*-test showed that *Human resource* (t = 2.739, p < 0.006) is more satisfactory in former workers group. On the other hand, the differences in the sentiments of the following factors were statistically insignificant at the 95% level of significance: *Organizational culture* (t = 1.151, p = 0.249) and *Salary* (t = -1.416, p = 0.156).

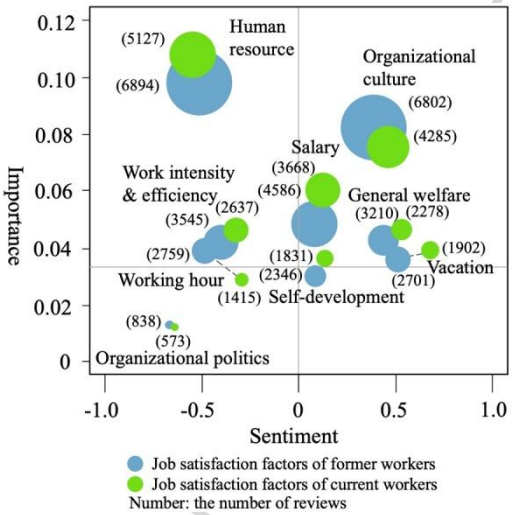

Fig. 9. Sentiment and importance of the factors in the group level.

4.4 Dominance analysis

We ran dominance analysis and regression analysis to identify relative importance and causal effects of each of the five independent variables (i.e., *Promotion opportunity*, *Benefits and compensation*, *Work-life balance*, *Organizational culture*, and *Senior management*) on the overall job satisfaction rate. The result of the analysis is given in Table 5. The regression model is statistically significant (F = 86.2682, p < 0.001) at the 95% level of significance; the five independent variables are also statistically significant except the 'Intercept' at the same level of significance. The five independent variables explain 67.57% of variance in the overall job satisfaction (Adj.R^2 = .6757). The relative importance of each independent variable is statistically significant at the 95% level of significance. Among the five independent variables, *Senior management* has the highest importance

(0.1695) on the overall job satisfaction, followed by *Benefit and compensation* (0.1474). Considering the result from topic modeling, we conclude that the job satisfaction factors such as *Supervisor (human relations)* and *Supervisor competency* can improve the satisfaction level of *Senior management*. Similarly, job satisfaction factors such as *General welfare*, *Welfare for women*, *Financial support*, and *Salary* can enhance the satisfaction level of *Benefit and compensation*.

Table 5

The result of multiple regression and dominance analysis.

Variable	Coefficient		t-statistic	Relative importance
	Beta	(S.E.)		
(Intercept)	0.0118	(0.1531)	0.062	
Promotion opportunity	0.2248**	(0.0538)	4.210	0.1323***
Benefit and compensation	0.2426**	(0.0494)	4.936	0.1474***
Work-life balance	0.1413*	(0.0447)	3.184	0.0941***
Organizational culture	0.1833**	(0.0514)	3.586	0.1393***
Senior management	0.2548**	(0.0554)	4.618	0.1695***

*p < .05; **p < .01; ***p < .001; F = 86.2682; p-value < .001; Adj. R^2 = .6757.

4.5 Correspondence analysis

We performed CA to investigate how the most important factors vary across the overall satisfaction rates. That is, we examined what factors are most important in reviews of each satisfaction rate. CA squashed 30 job satisfaction factors and overall satisfaction rates into a two-dimensional space. According to Fig. 10 in which we represent only the most important nine factors, *Self-development* and *General welfare* are most important in the reviews rated as 5 (represented as 'R5' in Fig. 10), *Organizational culture* in the reviews rated as 4, *Vacation*, *Work intensity & efficiency* and *Salary* in the reviews rated as 3, *Organizational politics* in the reviews rated as either 2 or 3, and *Human resource* and *Working hour* in the reviews rated as either 1 or 2.

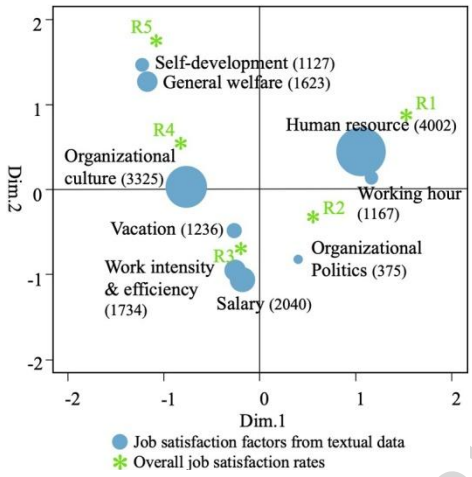

Fig. 10. Correspondence analysis of online employee reviews.

5. Discussion

It might be an important issue for business managers in HR departments to manage their employees' job satisfaction. This study proposes a novel approach to supporting the business managers to make decisions by providing a comprehensive view on job satisfaction factors. This study utilizes a relatively large number of samples collected from a company review site in South Korea, allowing us to secure more generalizability and reliability than prior studies which use traditional survey methods. A series of analyses on the voluminous online reviews give diverse insights into employees' job satisfaction.

The findings of our study may provide several implications for academic researchers and business managers of HR departments. First, they need to pay attention to the 30 fine-grained job satisfaction factors derived from LDA and the corresponding keywords, in order to lead their employees to have better job satisfaction. We consider that the factors and the corresponding keywords may be used for designing survey questionnaires to measure the degree of job satisfaction of employees. Note that five new job satisfaction factors (i.e., *Project*, *Software development*, *Inter-firm relationship*, *Marketing*, and *Overseas business*) were found that had not been considered in the literature. Some of them have a good reason for their inclusion. Because *Project* or *Software development* are common in the IT industry, employees frequently expressed their satisfaction with their *Project* and *Software*

development in online employee reviews. Employees also mentioned the *Inter-firm relationship* between their organization and the other (e.g., a parent company, a subsidiary company, a client firm, etc.). Second, the results from the sentiment and importance analyses conducted at the industry level would let business managers recognize the general degree of job satisfaction in the industry to which their company belongs. This analysis may help business managers diagnose the industry level competitiveness of their companies regarding HRM. Third, from the company level analysis, business managers could notice what they need to improve against their competitors to attain a low turnover rate of their staffs. Fourth, the results from the chronological analysis will serve as a basis for assessing the performances of specific HRM activities designed to improve employees' job satisfaction. Fifth, the dominance analysis reveals that *Senior management* has the highest importance on the overall job satisfaction, implying that they play an important role in developing job satisfaction of employees. Finally, the result of correspondence analysis indicates which factors need to be taken into more consideration by business managers according to their company's average rate of overall job satisfaction.

For the better results of extracting nouns from online reviews through POS tagging, we designed the DDBM to capture compound nouns and neologisms with less human intervention. Moreover, to determine the number of topics to be extracted using LDA, we first use the perplexity score and then cluster them to derive the final set of topics. Through the inter-rater reliability test and external validity evaluation, we validate the reliability of job satisfaction factors identified from online employee reviews.

Our study has some limitations. First, although the topic modeling approach (i.e., LDA) shows acceptable levels of agreement in the reliability test, it still has relatively low levels of agreement compared to humans. As an unsupervised learning algorithm, LDA inherently has shortcomings in fully understanding natural languages but it requires no human intervention. Future research may use supervised learning algorithms for identifying job satisfaction factors from online employee reviews. Because supervised learning algorithms require human intervention (i.e., human labeling of reviews),

they may effectively identify job satisfaction factors from the reviews. However, they can only detect the existence of each factor in 'Pros' and 'Cons' sections of a single review, and cannot identify job satisfaction factors that have not been labeled by humans. Second, it might be necessary to collect and use more data in future research to obtain more generalized findings. The reviews used in our study were collected from Jobplanet.co.kr and the analyses were conducted based on the reviews belonging only to the IT industry in South Korea. While human reviewers (i.e., student A and student B) identified *Coworker* and *Training* as job satisfaction factors from the reviews, the topic modeling approach did not identify a few factors such as *Coworker*, *Autonomy*, and *Training* which were considered in the literature. Finally, it may be worthwhile to extend our study so that it includes other industries, other text mining techniques, and other reviews (e.g., reviews from specific regions or from specific job positions, etc.) as future research for more specific analytical results on employees' job satisfaction.

6. Conclusions

These days, as more and more user-generated contents are accumulated as textual data, researchers and business practitioners have been granted a new opportunity to mine valuable meanings from them. Various text mining techniques are usually adopted to obtain actionable knowledge from such textual data. We also used topic modeling technique, one of the typical text mining techniques, to investigate online employees' reviews on their companies and identify job satisfaction factors hidden in the reviews. From the diverse analyses based on the factors, we were able to obtain useful insights into making decisions on leading employees to a higher level of job satisfaction than before.

Acknowledgments

This work was supported by the Korea University Business School Research Grant.

References

[1] A.S. Abrahams, W. Fan, G.A. Wang, Z. Zhang, J. Jiao, An integrated text analytic framework for product defect discovery, Production and Operations Management, 24(6) (2015) 975-990.

[2] A.J. Antoncic, B. Antoncic, Employee satisfaction, intrapreneurship and firm growth: a model, Industrial Management & Data Systems, 111(4) (2011) 589-607.

[3] P. Anupriya, S. Karpagavalli, LDA based topic modeling of journal abstracts, Proceedings of 2015 International Conference on Advanced Computing and Communication Systems, (2015), 1-5.

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

[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 3(4-5) (2003) 993-1022.

[6] P. Boxall, K. Macky, High-involvement work processes, work intensification and employee well-being, Work, Employment and Society, 28(6) (2014) 963-984.

[7] J. Boyd-Graber, D. Mimno, D. Newman, Care and feeding of topic models: Problems, diagnostics, and improvements, in: Handbook of mixed membership models and their applications, (CRC Press, 2014), pp. 225-254.

[8] D.V. Budescu, Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression, Psychological Bulletin, 114(3) (1993) 542-551.

[9] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, D.M. Blei, Reading tea leaves: How humans interpret topic models, Proceedings of Advances in Neural Information Processing Systems, (2009), 288-296.

[10] W.C.K. Chiu, C.W. Ng, Women-friendly HRM and organizational commitment: A study among women and men of organizations in Hong Kong, Journal of Occupational and Organizational Psychology, 72(1999) 485-502.

[11] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement, 20(1) (1960) 37-46.

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

[13] S. Debortoli, O. Müller, I. Junglas, J. vom Brocke, Text mining for information systems researchers: An annotated topic modeling tutorial, Communications of the Association for Information Systems, 39(7) (2016).

[14] W. Duan, B. Gu, A.B. Whinston, The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry, Journal of Retailing, 84(2) (2008) 233-242.

[15] S.D. Galup, G. Klein, J.J. Jiang, The impacts of job characteristics on IS employee satisfaction: A comparison between permanent and temporary employees, Journal of Computer Information Systems, 48(4) (2008) 58-68.

[16] D.M. Goldberg, N. Zaman, Text analytics for employee dissatisfaction in human resources management, Proceedings of the 24th Americas Conference on Information Systems, (2018).

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

[18] J.R. Hackman, G.R. Oldham, Development of the job diagnostic survey, Journal of Applied Psychology, 60(2) (1975) 159-170.

[19] O. Herrbach, K. Mignonac, How organisational image affects employee attitudes, Human Resource Management Journal, 14(4) (2004) 76-88.

[20] D.L. Hoffman, G.R. Franke, Correspondence analysis: Graphical representation of categorical data in marketing research, Journal of Marketing Research, 23(3) (1986) 213-227.

[21] M.T. Iaffaldano, P.M. Muchinsky, Job satisfaction and job performance: A meta-analysis, Psychological Bulletin, 97(2) (1985) 251-273.

[22] J.M. Ivancevich, R. Konopaske, R.S. Defrank, Business travel stress: A model, propositions and managerial implications, Work & Stress, 17(2) (2003) 138-157.

[23] N. Kinnie, S. Hutchinson, J. Purcell, B. Rayton, J. Swart, Satisfaction with HR practices and commitment to the organisation: Why one size does not fit all, Human Resource Management Journal, 15(4) (2005) 9-29.

[24] H.C. Koh, E.f.H.Y. Boo, The link between organizational ethics and job satisfaction: A study of managers in Singapore, Journal of Business Ethics, 29(4) (2001) 309-324.

[25] J. Krumm, N. Davies, C. Narayanaswami, User-generated content, IEEE Pervasive Computing, 7(4) (2008) 10-11.

[26] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics, 33(1) (1977) 159-174.

[27] J.H. Lau, D. Newman, T. Baldwin, Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, (2014), 530-539.

[28] A.J.T. Lee, F.-C. Yang, C.-H. Chen, C.-S. Wang, C.-Y. Sun, Mining perceptual maps from consumer reviews, Decision Support Systems, 82(2016) 12-25.

[29] J. Lee, J. Kang, A study on job satisfaction factors in retention and turnover groups using dominance analysis and LDA topic modeling with employee reviews on Glassdoor.com, Proceedings of the 38th International Conference on Information Systems, (2017).

[30] R. Lee, E.R. Wilbur, Age, education, job tenure, salary, job characteristics, and job satisfaction: A multivariate analysis, Human Relations, 38(8) (1985) 781-791.

[31] M. Lin, H.C. Lucas, Jr., G. Shmueli, Research commentary—too big to fail: large samples and the p-value problem, Information Systems Research, 24(4) (2013) 906-917.

[32] J.W. Lounsbury, L.L. Hoopes, A vacation from work: Changes in work and nonwork outcomes, Journal of Applied Psychology, 71(3) (1986) 392-401.

[33] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, Proceedings of the 17th International Conference on World Wide Web, (2008), 121-130.

[34] D.B. Lund, Organizational culture and job satisfaction, Journal of Business & Industrial Marketing, 18(3) (2003) 219-236.

[35] N. Luo, Y. Zhou, J.J. Shon, Employee satisfaction and corporate performance: Mining employee reviews on glassdoor. com, Proceedings of the 37th International Conference on Information Systems, (2016).

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

[37] S.M. Mudambi, D. Schuff, Research note: What makes a helpful online review? a study of customer reviews on Amazon.com, MIS Quarterly, 34(1) (2010) 185-200.

[38] J.P. Near, R.W. Rice, R.G. Hunt, Work and extra-work correlates of life and job satisfaction, Academy of Management Journal, 21(2) (1978) 248-264.

[39] D. Newman, C. Chemudugunta, P. Smyth, Statistical entity-topic models, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2006), 680-686.

[40] D. Newman, J.H. Lau, K. Grieser, T. Baldwin, Automatic evaluation of topic coherence, Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, (2010), 100-108.

[41] J.L. Pierce, J.W. Newstrom, Toward a conceptual clarification of employee responses to flexible working hours: A work adjustment approach, Journal of Management, 6(2) (1980) 117-134.

[42] D.E. Pournarakis, D.N. Sotiropoulos, G.M. Giaglis, A computational model for mining consumer perceptions in social media, Decision Support Systems, 93(2017) 98-110.

[43] B. Scott-Ladd, A. Travaglione, V. Marshall, Causal inferences between participation in decision making, task attributes, work effort, rewards, job satisfaction and commitment, Leadership & Organization Development Journal, 27(5) (2006) 399-414.

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

[45] K.L. Sommer, M. Kulkarni, Does constructive performance feedback improve citizenship intentions and job satisfaction? The roles of perceived opportunities for advancement, respect, and mood, Human Resource Development Quarterly, 23(2) (2012) 177-201.

[46] P.E. Spector, Measurement of human service staff satisfaction: Development of the Job satisfaction survey, American Journal of Community Psychology, 13(6) (1985) 693-713.

[47] S. Strohmeier, F. Piazza, Domain driven data mining in human resource management: A review of current research, Expert Systems with Applications, 40(7) (2013) 2410-2420.

[48] S. Strohmeier, F. Piazza, Artificial intelligence techniques in human resource management—A conceptual exploration, in: Intelligent techniques in engineering management, (Springer, 2015), pp. 149-172.

[49] C.-Y. Sun, A.J.T. Lee, Tour recommendations by mining photo sharing social media, Decision Support Systems, 101(2017) 28-39.

[50] M.R. Testa, Satisfaction with organizational vision, job satisfaction and service efforts: An empirical investigation, Leadership & Organization Development Journal, 20(3) (1999) 154-161.

[51] S. Tonidandel, J.M. LeBreton, Relative Importance Analysis: A Useful Supplement to Regression Analysis, Journal of Business and Psychology, 26(1) (2011) 1-9.

[52] E. Vigoda, Organizational politics, job attitudes, and work outcomes: Exploration and implications for the public sector, Journal of Vocational Behavior, 57(2000) 326-347.

[53] J.H. Ward, Jr., Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, 58(301) (1963) 236-244.

[54] D.J. Weiss, R.V. Dawis, G.W. England, Manual for the Minnesota satisfaction questionnaire, University of Minnesota, Minneapolis, 1967.

[55] M.L. Williams, S.B. Malos, D.K. Palmer, Benefit system and benefit level satisfaction: An expanded model of antecedents and consequences, Journal of Management, 28(2) (2002) 195-215.

[56] Z. Yang, X. Fang, Online service quality dimensions and their relationships with satisfaction: A content analysis of customer reviews of securities brokerage services, International Journal of Service Industry Management, 15(3) (2004) 302-326.

[57] H. Yuan, R.Y.K. Lau, W. Xu, The determinants of crowdfunding success: A semantic text analytics approach, Decision Support Systems, 91(2016) 67-76.

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

Yeonjae Jung is a Master's student in MIS at Korea University Business School, Republic of Korea. He received his Bachelor degree in MIS at Keimyung University. His Current research interests include business intelligence, big data analytics, machine learning, and decision support systems.

Yongmoo Suh is a professor of MIS in Korea University Business School, Republic of Korea. He received his PhD from the University of Texas at Austin, an MS in Computer Science from the Korea Advanced Institute of Science and a BS from the Seoul National University. He has presented his research at various areas, such as object-oriented systems, collaboration, business process management, data mining, etc. His current research interests include use of ontology in business domain, recommendation, social data analysis, data mining and text mining.

Highlights

- Identification of fine-grained job satisfaction factors from online employee reviews.
- Suggesting a heuristic way of determining the number of topics from LDA.
- Measuring sentiment and importance of each job satisfaction factor.
- *Senior management* is the most important factor for overall job satisfaction.
- Uncovering key factors with regard to overall job satisfaction rates.