frontiers | Frontiers in Physics

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fphy.2022.822351

A Joint Model for Extracting Latent Aspects and Their Ratings From Online Employee Reviews

Zhuo-Ming Ren, Yue Zheng, Wen-Li Du and Xiao Pan**

Alibaba Business School, Hangzhou Normal University, Hangzhou, China

The personal description of a company associated with job satisfaction, company culture, and opinions of senior leadership is available on workplace community websites. However, it is almost impossible to read all of the different and possibly even contradictory reviews and make an accurate overall rating. Therefore, extracting aspects or sentiments from online reviews and the corresponding ratings is an important challenge. We collect online anonymous employees' reviews from Glassdoor.com which allows people to evaluate and review the companies they have worked for or are working for. Here, we propose a joint rules-based model which combines the numerical evaluation reflected in the form of 1–5 stars, and the reviewed context to extract aspects. The model first inputs the five aspects with the initial word sets that are manually screened, and expands the aspect keyword sets through bootstrapping semi-supervised learning, and then uses latent rating regression to obtain the aspect score and aspect weight to update the corresponding score. Our experimental evaluation has shown better results as compared with an unsupervised learning of the latent Dirichlet allocation. The results could not only help companies understand their strengths and weaknesses, but also help job seekers apply for companies.

Keywords: aspect-based sentiment analysis, aspect extraction, bootstrapping algorithm, latent rating regression, online employ review

OPEN ACCESS

Edited by:
Ali Mehri,
Babol Noshirvani University of Technology, Iran

Reviewed by:
Jie Cao,
Nanjing University of Finance and Economics, China
Xiaoqiu Shi,
Southwest University of Science and Technology, China

Correspondence:
Zhuo-Ming Ren
zhuoming.ren@hznu.edu.cn
Xiao Pan
pan.xiao@hznu.edu.cn

Specialty section:
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

Received: *25 November 2021*
Accepted: *24 March 2022*
Published: *26 April 2022*

Citation:
Ren Z-M, Zheng Y, Du W-L and Pan X (2022) A Joint Model for Extracting Latent Aspects and Their Ratings From Online Employee Reviews. Front. Phys. 10:822351. doi: 10.3389/fphy.2022.822351

1 INTRODUCTION

Online reviews representing users' subjective opinions and insights have become incredibly easy to find, which have gradually become valuable references to help users choose and improve the objects being reviewed. Therefore, sentiment analysis of online reviews known as opinion mining has attracted increasing interest and even achieved good results in company evaluation associated with job satisfaction, company culture, and opinions of senior leadership [1–3]. For example, Glassdoor.com is one of the largest jobs and recruiting websites in the world, which covers more than 700,000 global companies and has provided nearly 33 million anonymous salary reports and employees' reviews since 2008 [4]. Glassdoor.com allows people to evaluate the companies they have worked for or are working for. Employees are free to rate their companies with 1–5 stars and text comments. This information not only allows those who are looking for a job to know more information about the company they may join in the future, but also provides the company with real feedback from its employees. However, it is almost impossible to read all of the different and possibly even contradictory reviews and make an accurate overall rating. Therefore, extracting aspects of sentiments from online reviews and the corresponding ratings is important and valuable. Sentiment analysis is an analysis method to obtain the opinion polarity in a given text and the types of emotions in various aspects of a subject for the emotion [5]. Opinions, attitudes, thoughts, and judgments are the private states of individual subjective expression [6, 7]. Therefore, objective observation or verification is not carried out. With the need for fine-grained emotional division of research objects in review texts, aspect-level sentiment analysis of review texts has become a hot topic. Thet et al. [8] proposed a new concept of aspect-based sentiment analysis (ABSA), where aspect refers to the attribute or component of the review object. Liu et al. [9] further gave the perspective of aspect-level sentiment analysis, which pointed out the direction of the research in this field, that is, the goal of the research is to find the aspect information of the review object and the emotional polarity of each aspect.

Aspect extraction is a major task in the work of aspect sentiment analysis [10]. The extracted aspects not only need to contain as many reviewers' concerns as possible, but also require less semantic overlap among different aspects. "The existing work of aspect extraction can be divided into three categories", only two categories are listed, please confirm which part of the following paragraph refers to the third category. The first is extraction based on a deep neural network model. Since Poria and Wang et al. [11, 12] set the extraction of fine-grained aspects as the target, He et al. [13] subsequently proposed a neural network-based model for identifying aspect terms, and yet the aspect outputted was too fine-grained to be viewed as an aspect. The second category is automatic extraction based on the topic model, that is, modeling the topic according to the review text and extracting aspects from it. The latent Dirichlet allocation (LDA) model [14] is a topic model widely adapted to text classification, Blei et al. then introduced the Bayesian framework into the probabilistic latent semantic analysis (PLSA) model [15, 16]. Shams [17] innovatively added the co-occurrence relationship on the basis of the LDA model and regarded it as an a priori domain. Furthermore, Lin [18] and Moghaddam et al. [19] studied the dependence between potential aspects and scores and performed modeling analysis. Nowadays, the mining of potential aspects has become an important part of the field of aspect mining. Wang et al. [20] defined this problem as a new type of text mining problem of latent aspect rating analysis (LARA). Subsequently, Wang et al. [21] proposed a generative model, which added topic modeling technology to the latent rating regression (LRR) model. The LRR is based on association rules, that is, most of them use manual rules to extract potential aspects and use clustering algorithms for them. Poria et al. [22] were successful in extracting products using a hand-designed mining rule. Qiu et al. [23] utilized this rule while using a bidirectional propagation method to better connect emotions and aspects. Gindl et al. [24] implemented the bidirectional propagation method with anaphora resolution to identify co-reference, so as to improve accuracy. Su et al. [25] mapped the implicit aspect to the explicit aspects through the clustering algorithm. Rana et al. [26] proposed a two-fold rule-based model for aspect extraction defined by a sequential pattern. Wang et al. [20] applied the bootstrapping method to the extraction of aspects for the first time, whose results were remarkable.

The success of aspect extraction lays a good foundation for aspect review mining and analysis, that is, judging the emotional polarity of words. The current common methods include: an aspect-emotion hybrid model and aspect-level sentiment analysis method based on emotional words. The aspect-emotion hybrid model synthesizes the topic model and various factors to construct a new model to mine multiple viewpoints of a given review text. Lu et al. [27] thought that this problem aims to decompose the overall rating of a large number of short reviews into main aspect ratings, and used eBay seller feedback review data to verify the feasibility of automatically generating aspect ratings. Titov et al. [28] proposed a joint model which combined aspects and sentiment rating to extract aspect and corresponding score for given comment texts. For the sake of probing potential aspects from online product review texts, Moghaddam et al. [19] investigated the ILDA model of interdependence, the model was based on aspect extraction and generated its corresponding opinions. Aspect sentiment methods construct positive and negative polarity dictionaries artificially to judge the sentiment tendency of aspect opinion words chiefly. However, each subject field keeps exclusive emotion words, that is, the method possesses domain-specific properties, which makes it challenging to construct a general sentiment dictionary. Apart from manual construction, initializing the keyword sets with identification of emotional polarity and analyzing the emotional polarity of undefined words through a bootstrapping algorithm have been also universally recognized. The bulk of previous sentiment analysis research essentially focused on dividing the overall sentiment of a review text into positive and negative sides, and those failed to probe the degree of emotional orientation of different aspects of a given comment text. With the in-depth study of mining and analysis based on aspect views, the research has developed into a multi-level rating analysis, which uses a certain range of values to indicate the degree of emotional polarity, such as a score from 1 to 5 stars. Effectually, research based on aspect sentiment analysis remains at a theoretical level, lacking an assessment of the company as a whole, and it does not focus on the emotional polarity of the employees to the companies. In other words, it does not take into account the employees' concern and emphasis on different aspects of the company's job opportunities, salary, and so on. Meanwhile it is short of comprehensive analysis combining text comments and total scores. Therefore, we extend the LRR model [20, 21] and apply it to the field of comments from company employees, that is, adopting bootstrapping for anonymous employee reviews on the Glassdoor.com platform to obtain aspects and digging out the emotional tendency of the employees in the text data. We use the LRR model to convert these emotional text data into a digital expression and revise and enrich the initial aspect score based on the employee's overall score given to the company. This contributes to finding a company that meets their own preferences for job seekers, alternatively discovering employee opinions for subsequent company structure optimization. The five aspects including work/life balance, culture and values, senior management, career opportunities, and salary and benefits have been confirmed in the data collected from Glassdoor.com, there is no need to automatically explore the aspects, so the following work only relies on the bootstrapping algorithm and LRR model.

2 METHODS AND METRICS

The bootstrapping (semi-supervised learning) algorithm has been favored in the research on the extraction of review texts in the hotel field (known as "cleanliness", "location", "service", and other specific information). We apply the model to the company's sentiment analysis research for the first time. The first step is to identify related sentences in different aspects for a given set $D = \{d_1, d_2, \ldots, d_{|D|}\}$ of employees' review texts.

2.1 Aspect Extraction Based on Bootstrapping Algorithm

The aspects of the company may be stressed in each sentence of the employees' reviews, thus the purpose of extracting aspects based on the bootstrapping algorithm is to map the sentences in the review text to a subset corresponding to each aspect. The cardinal procedure can be summarized as manually extracting aspects for the text collection and designing keywords to describe this aspect, and using the bootstrapping algorithm to match the sentence in the review with the relevant subset of each aspect to obtain more relevant words to expand each aspect, until all the comment texts have been run. The last is to assign the review texts to the aspects. According to the matching degree, the aspect of the sentence is determined to realize the aspect recognition and extraction.

The algorithm process is depicted as follows. For given comment datasets X and Y (X is a dataset containing a set of k aspect keywords, Y is a dataset with unlabeled information), we gradually expand the collection of keywords of each aspect of dataset X by means of dataset Y. The given input data includes the collection $D = \{d_1, d_2, \ldots, d_{|D|}\}$ of review texts, the collection $\{T_1, T_2, \ldots, T_k\}$ of aspect keywords, a vocabulary V for a given review text, iteration step limit I, and thresholds p. The algorithm has the following steps.

1) Group employee review text D into sentence set $S = \{sentence_1, sentence_2, \ldots, sentence_m\}$ in one unit. Set an initial collection $\{T_1, T_2, \ldots, T_k\}$ of keywords for k aspects of the review text D based on manual analysis.
2) Match the word of each sentence in S with the keyword set $\{T_1, T_2, \ldots, T_k\}$ of k aspects and assign it to the aspect with the highest matching degree. If there is more than one aspect with the highest matching degree, label the sentence with these aspects at the same time.
3) Based on this initial aspect labeling, we calculate the value χ^2 of each word and k aspect in the vocabulary V of the company's employee review text according to the formula of the Chi square statistics (χ^2) [29] of word w and aspect k. Then for each aspect, sort the value χ^2 from large to small (the larger the χ^2, the greater the correlation), and then add the top p words to the corresponding aspect keyword set T_i.

$$\chi^2(w, A_i) = \frac{C \times (C_1C_4 - C_2C_3)^2}{(C_1 + C_3) \times (C_2 + C_4) \times (C_1 + C_2) \times (C_3 + C_4)}, \tag{1}$$

C_1 represents the number of occurrences of word w in sentences belonging to aspect A_i. C_2 represents the number of occurrences of w in sentences that do not belong to aspect A_i. C_3 represents the number of sentences belonging to aspect A_i but without w. C_4 represents the number of sentences that neither belong to aspect A_i nor word w. C represents the total number of occurrences of w.

4) Repeat (2) and (3) until the aspect keyword list is no longer growing or the iteration step limit I is reached. After running the bootstrapping algorithm, the review text D will be divided into a set of sentences marking the corresponding aspects. Therefore, we can count the frequencies of the word in the k aspect sentences to get the correlation between the word and each aspect. By counting the number of occurrences of each word in the vocabulary list V in various aspects of the sentence, the word frequency feature matrix of the word representing the review text in the vocabulary is obtained.

2.2 Aspect Score and Aspect Weight Analysis of LRR Model

LRR model definition. The phenomenon of accidentally or deliberately giving low evaluation and high score always exists in real data. It is particularly vital to determine the aspects the review content actually refers to. The LRR model infers the aspect score and aspect weight of each aspect based on the content of the text review and the related overall score. Therefore, the word frequency matrix W_d can be used as the input of the model, and the known overall score r is regarded as the dependent variable. The process variables of the aspect score S_d and aspect weight α_d in the k aspects can be deduced backward.

1) Aspect score: Aspect score S_{di} can be expressed as a linear combination of sentiment polarity β_i and word frequency feature matrix W_{di}:

$$S_{di} = \sum_{j=1}^{n} \beta_{ij} W_{di}, \tag{2}$$

$\beta_{ij} \in \mathfrak{R}$ is the sentiment polarity of word j in aspect A_i.

2) Aspect weight: The overall score r is generated by the weighted sum of aspect weight α_d and aspect score S_{di}, in accordance with $\alpha_d^T S_d = \sum_{i=1}^{k} \alpha_{di} S_{di}$, we assume that r is a sample drawn from a Gaussian distribution with mean $\alpha_d^T S_d$ and variance δ^2, then the relationship of r_d and α_d:

$$r_d \sim N\left(\sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{ij} W_{dij}, \delta^2\right). \tag{3}$$

Different employees always have different focal points on the evaluation of the same company. Taking into account the diversity of employees' preferences for different aspects, we

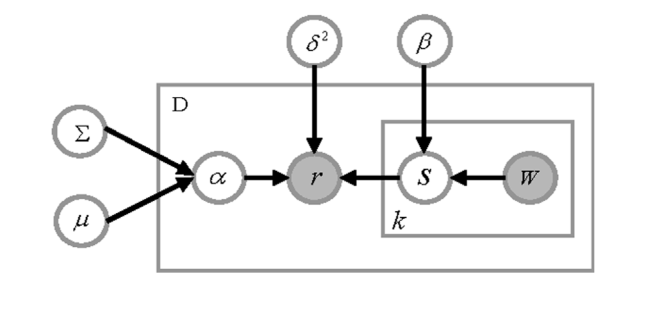


FIGURE 1 | The LRR model flow chart.

take the multivariate Gaussian distribution as the prior distribution of the aspect weight α_d of each review text d and treat aspect weight α_d as a group of random variables extracted from the prior distribution of corpus d, namely:

$$\alpha_d \sim N(\mu, \Sigma), \tag{4}$$

μ and Σ are respectively the mean and variance of the multivariate Gaussian distribution. Combining **Equation 3** and **Equation 4**, a Bayesian linear regression problem can be obtained. Based on a given company employees' review text d, the LRR model defines the probability r of the overall score as follows:

$$P(r|d) = P(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \tag{5}$$

$$= \int p(\alpha_d|\mu, \Sigma) p\left(r_d| \sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{dij} W_{dij}, \delta^2\right) d\alpha_d \tag{6}$$

The paper assumes that δ^2 and β do not rely on a single company employee, which are corpus-level model parameters, and $\theta = (\mu, \Sigma, \delta^2, \beta)$ is a collection of corpus-level model parameters. The graphical model diagram of the LRR model is shown in **Figure 1**. The outer box represents the comment, and the inner box represents the composition of the potential aspect score and word description in the comment. δ^2 and β are assumed to not depend on the comments of a single company employee and $\theta = (\mu, \Sigma, \delta^2, \beta)$ are parameters of the corpus-level model.

Aspect score and aspect weight analysis based on LRR model. Based on the LRR model, this section provides the model parameters $\theta = (\mu, \Sigma, \delta^2, \beta)$ to obtain the aspect score and aspect weight in each company's employee review text. The detailed process is as follows: 1) Generate aspect score S_{di} for each aspect A_i in the employees' review text d. 2) Retrieve the optimal value of aspect weight α_d by maximum a posteriori (MAP) estimation. The objective function of the MAP method is defined as follows.

$$\varsigma(d) = \log p(\alpha_d|\mu, \Sigma)\left(r_d| \sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{dij} W_{dij}, \delta^2\right). \tag{7}$$

This objective function is then extended and all the items in each company's employee comment text are converted to the expression of α_d: $\varsigma(d)$. We obtain the estimated values and take its derivative. The estimate value α_d is defined as follows.

$$\hat{\alpha}_d = \arg\max \varsigma(\alpha_d) \tag{8}$$

$$= \arg\max\left[-\frac{(r - \alpha_d^T S_d)^2}{2\delta^2} - \frac{1}{2}(\alpha_d - \mu)^T \Sigma^{-1} (\alpha_d - \mu)\right], \tag{9}$$

α_d is not only satisfied $\sum_{i=1}^{k} \alpha_{di} = 1$, but also $0 < \alpha_{di} < 1$ and $i = 1, 2, \ldots, k$.

Parameter estimation of LRR model. In order to make the aspect score S_d and aspect weight α_d inferred by the LRR model more accurate, we introduce maximum likelihood estimation (ML) to hunt for the best model parameter $\hat{\theta} = (\mu, \Sigma, \delta^2, \beta)$, and the log-likelihood function of the whole set of employee's comments is:

$$\varsigma(D) = \sum_{d \in D} \log p(r_d|\mu, \Sigma, \delta^2, \beta, W_d). \tag{10}$$

The maximum likelihood estimation of several parameters in the model is as follows:

$$\hat{\theta} = \arg\max_{\theta} \sum_{d \in D} \log p(r_d|\mu, \Sigma, \delta^2, \beta, W_d). \tag{11}$$

We then randomly initialize all model parameters of the above formula to obtain $\theta_{(0)}$, and use the expectation maximization (EM) algorithm to update and optimize parameters during each iteration of alternately executing the E-step and M-step. The current model parameter of the i-th iteration is $\theta_{(t)}$. For each comment in a given comment text set, the aspect score S_d and aspect weight α_d can be deduced according to **Equation 2** and **Equation 9**. When the overall score of all company employee review texts is r_d, the probability of the obtained aspect score S_d and aspect weight α_d is maximized. At this time, the new parameter value $\theta_{(t+1)}$ is composed of μ, Σ, δ^2, $and\beta$, and the model parameter update process of the LRR model is as follows. Model parameters μ and Σ: It is necessary to maximize the occurrence probability of all α_d inferred in the observation E-step, so as to update the model parameters μ and Σ. For a given set of employees' reviews of the entire company, there is the following update equation based on the maximum likelihood estimation of the multivariate Gaussian distribution:

$$\mu_{(t+1)} = \arg\max_{\mu} - \sum_{d \in D} (\alpha_d - \mu)^T \Sigma^{-1} (\alpha_d - \mu) \tag{12}$$

$$= \frac{1}{|D|} \sum_{d \in D} \alpha_d. \tag{13}$$

$\Sigma_{(t+1)}$ is:

$$\Sigma_{(t+1)} = \arg\max_{\Sigma} \left[-|D| \log \Sigma - \sum_{d \in D} \left(\alpha_d - \mu_{(t+1)}\right)^T \Sigma^{-1} \left(\alpha_d - \mu_{(t+1)}\right)\right] \tag{14}$$

$$= \frac{1}{|D|} \sum_{d \in D} \left(\alpha_d - \mu_{(t+1)}\right)\left(\alpha_d - \mu_{(t+1)}\right)^T. \tag{15}$$

Model parameters δ^2 and β: Based on the aspect score S_d and aspect weight α_d obtained from the E-step, we can maximize $P(r_d|\alpha_d, \delta^2, \beta, W_d)$ to update the model parameters δ^2 and β. See the updated **Eqs 17**, **18** for details:

$$\delta^2_{(t+1)} = \underset{\delta^2}{\arg\max}\left[-|D|\log\delta^2 - \frac{\sum_{d\in D}\left(r_d - \alpha_d^T S_d\right)^2}{\delta^2}\right] \tag{16}$$

$$= \frac{1}{|D|}\sum\nolimits_{d\in D}\left(r_d - \alpha_d^T S_d\right)^2. \tag{17}$$

$$\beta_{(t+1)} = \underset{\beta}{\arg\max}\sum_{d\in D} -\frac{\left(r_d - \sum_{i=1}^{k}\alpha_{di}\beta_i^T W_{di}\right)^2}{2\delta^2_{(t+1)}}. \tag{18}$$

In this way, an "E-M-step" cycle has been completed. Then repeating the above two steps until the likelihood of **Eq. 10** converges, and the optimal model parameter $\hat{\theta} = (\hat{\mu}, \hat{\Sigma}, \hat{\delta}^2, \hat{\beta})$ is obtained.

2.3 "Bootstrapping + LRR" Model

The aspect sentiment analysis of company employee review texts mainly includes two steps: 1) Aspect extraction of employee's review text and 2) analysis of aspect scoring and aspect weight. Firstly, this paper utilizes the bootstrapping algorithm to mine the aspects of the review text and output a word frequency feature matrix W_d. Next, the LRR model uses W_d obtained in the first step, based on the overall score and text evaluation content in d, to infer potential aspect score S_d and aspect weight α_d related to different aspects of the company. We call the combination of the bootstrapping algorithm and the LRR model the joint model, and the basic flow of the model is shown in Algorithm 1.

Algorithm 1. The framework of the joint model.

```
Input: Collection $D = \{d_1, d_2, \cdots, d_{|D|}\}$ of comment texts, initial aspect keywords $\{T_1, T_2, \cdots T_k\}$ of $k$ aspects, company employee comment text vocabulary $V$, threshold $p$, iteration step limit $p$.
Output: Aspect score $S_d$ and aspect weight $\alpha_d$ of comment text $d$.
1: Divide the review texts in set $D$ into sentence set $S = \{sentence_1, sentence_2, \cdots, sentence_m\}$ of sentences, and set an initial keyword set $\{T_1, T_2, \cdots T_k\}$ for $k$ aspects of $D$;
2: Match the word of each sentence in $S$ with $\{T_1, T_2, \cdots T_k\}$ and assign the sentence to the aspect with the highest matching degree. If aspect with the highest degree of matching is more than one, mark multiple aspects at the same time;
3: Based on this initial aspect annotation, the value $\chi^2$ between each word and aspect in the vocabulary are calculated according to Equ. 1;
4: For each aspect, sort the values of $\chi^2$ from largest to smallest, and add the top $p$ words to the corresponding aspect keyword set $T_i$;
5: Repeat Step2 step4 until the aspect keyword list is no longer expanded or the iteration step limit $p$ reaches the specified value, and finally get the word frequency feature matrix $W_d$ of the word in the vocabulary $V$;
6: Let $W_d$ into the LRR model, calculate the aspect score $S_{di}$ for each aspect $A_i$ in the review text according to Equ. 2, and infer the aspect weight $\alpha_d$ in $d$ according to Equ. 9.
```

2.4 Metrics

For aspect scores and aspect weights, we utilize three different quantitative evaluation indicators: 1) Mean square error Δ^2_{aspect} of aspect score prediction, 2) aspect correlation ρ_{aspect} within reviews, and 3) aspect correlation ρ_{review} between reviews. The definitions of the three quantitative evaluation indicators are as follows:

1) Mean square error Δ^2_{aspect} of aspect score prediction: Suppose S^*_{di} is the true aspect score of aspect A_i and S_{di} is the predicted score. Δ^2_{aspect} measures the difference between the predicted S_{di} and the true S^*_{di}, and we define it as:

$$\Delta^2_{aspect} = \sum_{d=1}^{|D|}\sum_{i=1}^{k}\left(S_{di} - S^*_{di}\right)\Big/\left(k\times|D|\right). \tag{19}$$

2) Aspect correlation ρ_{aspect} within reviews: ρ_{aspect} measures whether the relative order of the score prediction of various aspects of a review text is consistent with the order of the real aspect score, that is, whether the predicted score maintains the preference characteristics of the employee's comments. ρ_{aspect} is defined as follows:

$$\rho_{aspect} = \sum_{d=1}^{|D|}\rho_{S_d,S_d^*}\Big/|D|. \tag{20}$$

ρ_{S_d,S_d^*} represents the Pearson correlation between the two vectors S_d and S_d^*.

3) ρ_{review} measures whether the ranking results of all review texts according to the predicted aspect scores in multiple aspects are consistent with the ranking results based on the real aspect scores. It is defined as:

$$\rho_{review} = \sum_{i=1}^{k}\rho\left(\overrightarrow{S_i}, \overrightarrow{S_i^*}\right)\Big/k, \tag{21}$$

$\overrightarrow{S_i}$ and $\overrightarrow{S_i^*}$ respectively refer to the predicted aspect score vector and the true aspect score vector of aspect A_i in all review texts.

3 DATA AND ASPECT EXTRACTION

We propose the "bootstrapping + LRR" algorithm to the company's sentiment analysis research. The specific process of the research work is shown in **Figure 2**. First, the company's employee review text datasets are distinguished and appropriately preprocessed according to text data and numerical data. We perform aspect-based sentiment analysis on the preprocessed dataset, and finally evaluate the performance of "bootstrapping (semi-supervised) + LRR".

3.1 Data

Figure 3 shows an example of the employee review information on Glassdoor.com. The review information includes numerical data and text data. We usually see that the numerical data includes an overall and star rating for each aspect, it can only express employee emotions in the form of primitive and crude data. Not only does it fail to reflect the emotional differences of employees towards the company, but it also does not dig out the emotion aspects in employees' text comments, that is, it does not reflect the employees' emphasis on different aspects of the company. Therefore, the employee review data are collected from this website in full, and the dataset specifically covers the company name, numerical evaluation data, and text evaluation data. The numerical data are the ratings ranging from 1 to 5 stars, which include the overall rating and the six dimensions of the company: work/life balance, culture and values, diversity and inclusion, senior management, career opportunities, and salary and benefits. It is noted that the aspect of diversity and inclusion appeared after August 2020, so we have not included it. The text data includes pros (positive comments about the company), cons (negative comments), and advice (suggestions for the company).

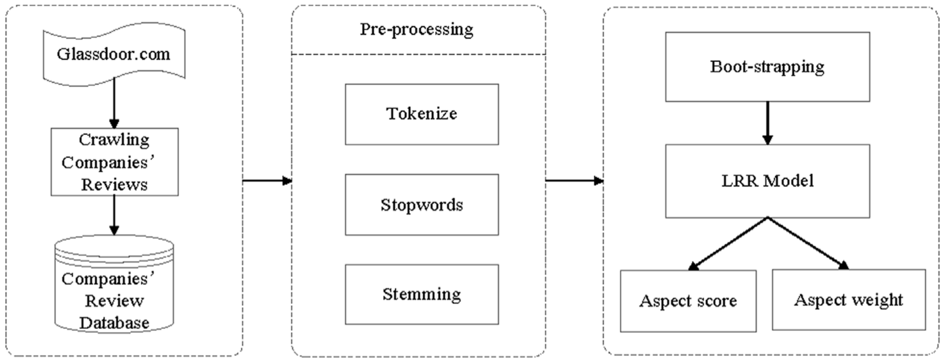
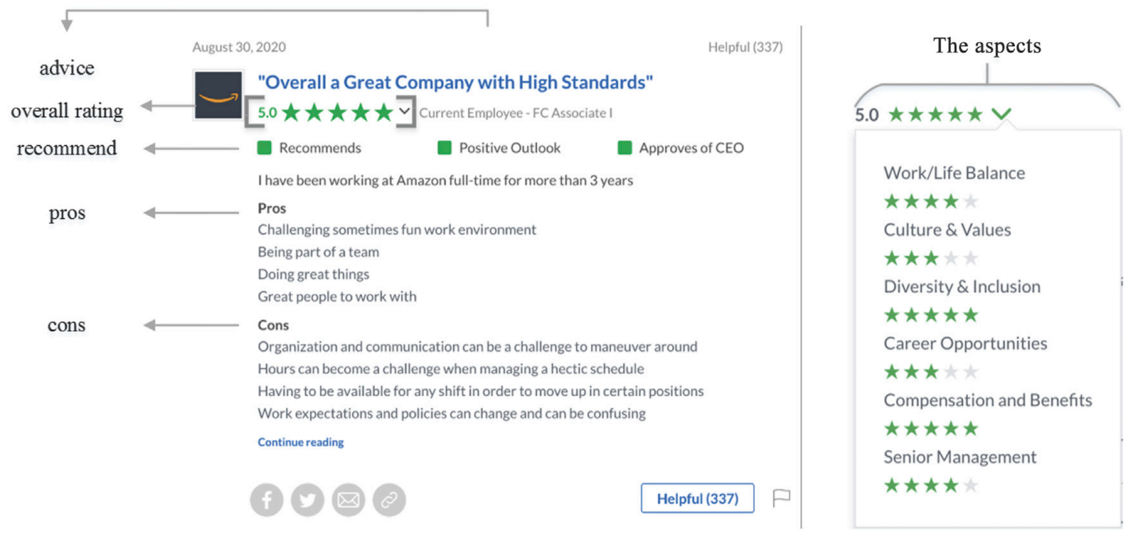


FIGURE 2 | The flow chart of the joint model for extracting latent aspects and their ratings from online employee reviews.



August 30, 2020

Helpful (337)

"Overall a Great Company with High Standards"

5.0 ★★★★★ Current Employee - FC Associate I

🟩 Recommends 🟩 Positive Outlook 🟩 Approves of CEO

I have been working at Amazon full-time for more than 3 years

Pros

Challenging sometimes fun work environment
Being part of a team
Doing great things
Great people to work with

Cons

Organization and communication can be a challenge to maneuver around
Hours can become a challenge when managing a hectic schedule
Having to be available for any shift in order to move up in certain positions
Work expectations and policies can change and can be confusing

Continue reading

5.0 ★★★★★

Work/Life Balance
★★★★☆
Culture & Values
★★★☆☆
Diversity & Inclusion
★★★★★
Career Opportunities
★★★☆☆
Compensation and Benefits
★★★★★
Senior Management
★★★★☆

FIGURE 3 | An example of the employee review information on Glassdoor.com. In the numerical reviews, it is noted that the aspect of diversity and inclusion just appeared after August 2020, so we did include it in this study.

TABLE 1 | Number of reviews per company.

Company	Reviews	Company	Reviews
Allianz	1,290	Fannie Mae	3,342
Amazon	10,350	Ford motor	4,037
Amerisource Bergen	965	General motors	4,626
Apple	9,926	Google	9,848
AT&T	21,103	Honda	867
AXA	1706	JPMorgan	20,236
Bank of China	704	Kroger	8,150
Bank of America	18,820	Marathon petroleum	534
BMW	762	McKesson	3,148
Cardinal health	2,191	Microsoft	10,876
Chevron	2,753	Samsung	4,950
Cigna	2,524	Shell	4,709
Costco wholesale	6,755	Total	743
CVS health	17,738	United health group	10,245
Daimler	405	Verizon	14,306
ExxonMobil	242	Volvo	25,518

We simulate a browser to crawl from Glassdoor.com to the anonymous comment data of the 2020 Fortune 100 (including Amazon, Apple, Microsoft, and other companies) employees of the companies seen in **Table 1**. We select the online comment data of employees who have commented more than 100 times. Finally, the study created data containing anonymous reviews from more than 200,000 employees of various companies. This paper performs preprocessing operations on the noise data obtained as above: 1) Unify the words in the text into lowercase and split the sentence into word sequences. 2) Remove meaningless punctuation marks, stop words, and low-frequency words. 3) Use Porter Stemmer [30] to extract the stem or root of words, and uniformly fill the non-digital data with 0.

3.2 Aspect Extraction

First, we set several initial keywords for each of the five aspects of the company based on the employee's scores for the five

TABLE 2 | Keywords in the five aspects by the bootstrapping algorithm.

Aspects	Five initial aspects of keywords	Keywords in the five aspects by the bootstrapping algorithm
Work/life balance	Schedule, life balance busy minute amusement interest	Weekends lunch work-life week scheduled shift weekend sick life long flexibility balance hour late holidays interest school busy schedules season shifts sheet hard day digital hours set break social personal minute schedule flexible nights days time family amusement
Culture and values	Culture value	Big strong conservative united values clash global office inclusion diversity safety create company deeply diverse value fear innovation toxic inclusive stiff fortune collaborative created gm change positive core mission size corporate culture
Senior management	Senior management staff leader supervisor organization team position	Leader treat advice upper leaders issues listen sales members leadership customers communication department middle staff team store people senior help management service organization position employees job support supervisor levels customer
Career opportunities	Career opportunities	Career mobility internal progression learning networking areas lateral advance network pursue promotional development move grow large explore limited learn develop opportunities room lots expand advancement relocate paths growth movement promotion
Salary and benefits	Benefits bonus commission compensation wage salary pension welfare	Benefits benefits insurance medical bonus decent salary welfare good 401 k discounts dental pension vacation commission reimbursement perks plan wage excellent match pay health tuition great paid compensation competitive hourly minimum

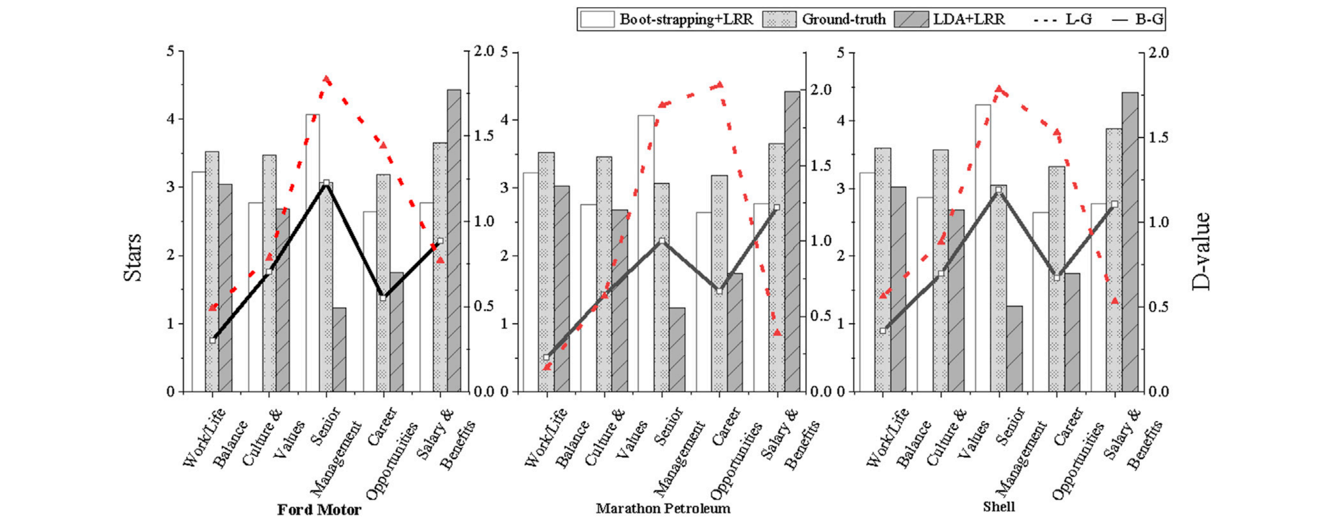


FIGURE 4 | The aspect score of the review text of the employees from the three companies. "B-G" and "L-G" respectively represent the difference value (D-value) between the output value of the two algorithms "bootstrapping + LRR" and "LDA + LRR" and the true value.

aspects of the company, as shown in **Table 2**. Next, the selection threshold of the bootstrapping algorithm is set to $p = 5$, and the iteration step limit is $I = 10$. The five aspects of the obtained keywords are also shown in **Table 2**. The keywords in the five aspects are basically perfected after using the bootstrapping method to extract the text of the company's employee reviews. Each sentence in the review text dataset for each company's review has its own aspect label information. The keywords are expanded from part of the basic words defined manually to the aspect words with a wider coverage.

4 RESULTS

In order to confirm the superior performance of the bootstrapping method in multi-faceted mining of review text, this paper also makes use of the LDA method to conduct an aspect extraction experiment on the employee review text in the same dataset. For the convenience of comparing the performance of the two methods in diverse aspects, this paper will refer to the two methods as the "bootstrapping (semi-supervised) + LRR" method and "LDA (unsupervised) + LRR" method and obtains the corresponding aspect score and aspect weight to compare and analyze the performance of the two methods.

4.1 Assessment of Differences in Aspect Scores

The overall score can judge whether the employee supports their company, but it cannot indicate the employee's emotional differences in different aspects of the company. For the sake of judging the capabilities of the LRR model,

TABLE 3 | The difference in aspect weight of the two employees' published texts.

Reviewer	Overallrating	Work/lifebalance	Culture and values	Seniormanagement	Career opportunities	Salary and benefits
Employee 1	4.0	0.20 (4.0)	0.15 (4.0)	0.13 (4.0)	0.35 (5.0)	0.17 (4.0)
Employee 2	4.0	0.04 (3.0)	0.12 (4.0)	0.13 (4.0)	0.06 (3.0)	0.65 (5.0)

TABLE 4 | Accuracy assessment comparison between the "bootstrapping + LRR" method and "LDA + LRR" method.

Method	$\Delta^2_{\mathrm{aspect}}$	ρ_{aspect}	ρ_{review}
Bootstrapping + LRR	1.098	0.469	0.625
LDA + LRR	1.135	0.331	0.606

this article extracted three companies: Ford Motor, Marathon Petroleum, and Shell, which have the same average overall score (3.5) and different aspect scores. The comparison of the aspect score prediction results obtained by the "bootstrapping + LRR" method and the "LDA + LRR" method with the real value of the aspect score provided in the dataset is shown in **Figure 4**. It can be seen that in the four aspects of work/life balance, culture and values, senior management, and career opportunities, the result of "bootstrapping + LRR" is closer to the true value on the whole.

Different employees attach different importance to all aspects of the same company they have worked for, that is, aspect weights of different comments vary to a certain extent. This paper uses the LRR to model the aspect extraction obtained by the bootstrapping method, so that the aspect weights of different aspects of the review text can be obtained, and the aspect scores of different aspects of the review text can be predicted. **Table 3** shows the real scores of two employees, who work for the same company and give the company an overall score of 4, and different aspect weights of different aspects estimated by the LRR model. While Employee 1 and Employee 2 both gave the company a 4-star overall rating, the two focused on slightly different aspects: Employee 1 values the company's career opportunities the most, while Employee 2 believes that the company's salary and benefits are the most important and superior to other aspects. It can be seen that identifying the weight information of the employees' review texts can reflect the company's employees' preferences on various aspects of the company to better let the company know its own strengths and weaknesses, and help job applicants make better decisions according to the reviews.

4.2 Accuracy Assessment Comparison With Unsupervised LDA

Based on the "bootstrapping + LRR" method and the "LDA + LRR" method, the above three indicators (Δ^2_{aspect}, ρ_{aspect}, and ρ_{review}) are calculated on the experimental results. It can be seen from **Table 4**, in terms of aspect score and aspect weight analysis of employees' comment text, the "bootstrapping + LRR" method adopted in this article is better than the "LDA + LRR" method on the three indicators. A lower value of Δ^2_{aspect} indicates that the "bootstrapping + LRR" method can more accurately predict the aspect score and deviates less from the corresponding real value. A higher ρ_{aspect} indicates that the combination of "bootstrapping" and "LRR" can better distinguish the aspect scores in a review text, and address the trouble of not being able to obtain such information as the relative preferences of different aspects through the overall score. ρ_{review} measures the overall relevance of the review texts of employees in all companies. A higher ρ_{review} value indicates that the sequence of the predicted scores obtained is more consistent with the ranking results based on the real aspect scores.

5 CONCLUSION AND DISCUSSION

More and more employees are free to post their comments on their company on the Internet, but low star ratings do not provide much insight into how employees feel about different aspects of each company. Therefore, we conducted an in-depth study on the potential aspects of opinion information in the employee review text and proposed a joint model (bootstrapping + LRR" method) for extracting latent aspects and their ratings from online employee reviews. Firstly, we used the bootstrapping algorithm to extract aspects of the company's employee review text to complete the aspect identification. Then, based on the overall score and the text review, the LRR model was used to establish the relationship between the known overall score and the unknown aspect score and aspect weight, thus to reveal the employees' emphasis on different aspects of the company and the differences in employees' rating behavior.

The experimental results showed that the combination of bootstrapping and LRR could infer the aspect score and aspect weight of each aspect in the company's employee review text, and its performance was relatively better than the "LDA + LRR" method. Compared with unsupervised learning ("LDA + LRR" method), the proposed semi-supervised learning ("bootstrapping + LRR" method) has a higher degree of reality fit, and completely excavates the implied score in the hidden text evaluation. In addition, the corrected overall score truly shows employees' emotional differences in different aspects of each company, helping to make judgments based on scores more accurate. However, fully supervised learning has not been attempted to determine its effect on the relationship between overall score and aspect scores and aspect weight. Therefore, in future work, we will use fully supervised learning to conduct multi-faceted viewpoint mining on the company's employee review text, and further analyze the application of multi-faceted viewpoint mining in practice. In addition, when modeling the LRR model in this paper, the emotional polarity of the word was used as one of the parameters of the model to generate the overall score. Therefore, in the following research, we can try to combine a sentiment dictionary to optimize parameters to improve the performance of opinion mining.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The datasets used to support the findings of this study are available from the website (https://www.glassdoor.com) or the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Z-MR and XP designed the research; XP and YZ performed the research and analyzed data; Z-MR, YZ, and DW-L wrote the paper.

FUNDING

The work is partially supported by Zhejiang Provincial Natural Science Foundation of China (Grant No. LY21F030019) and National Natural Science Foundation of China (Grant No. 61803137).

REFERENCES

1. Zainuddin N, Selamat A, Ibrahim R. Hybrid Sentiment Classification on Twitter Aspect-Based Sentiment Analysis. *Appl Intelligence* (2018) 48: 1218–32.
2. Ahamed Kabeer NR, Gan KH, Haris E. Domain-specific Aspect-Sentiment Pair Extraction Using Rules and Compound Noun Lexicon for Customer Reviews. *Informatics* (2018) 5. doi:10.3390/informatics5040045
3. Marcacini RM, Rossi RG, Matsuno IP, Rezende SO. Cross-domain Aspect Extraction for Sentiment Analysis: A Transductive Learning Approach. *Decis Support Syst* (2018) 114:70–80. doi:10.1016/j.dss.2018.08.009
4. Marrese-Taylor E, Velásquez JD, Bravo-Marquez F, Matsuo Y. Identifying Customer Preferences about Tourism Products Using an Aspect-Based Opinion Mining Approach. *Proced Comp Sci* (2013) 22:182–91. doi:10.1016/j.procs.2013.09.094
5. Mansour S. Social Media Analysis of User's Responses to Terrorism Using Sentiment Analysis and Text Mining. *Proced Comp Sci* (2018) 140:95–103. doi:10.1016/j.procs.2018.10.297
6. Yadollahi A, Shahraki AG, Zaiane OR. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Comput Surv (Csur)* (2017) 50:1–33.
7. Ren ZM, Du WL, Wen XZ. The Psychological Effects of Digital Companies'employees during the Phase of Covid-19 Pandemic Extracted from Online Employee Reviews. *Sustainability* (2022) 14. doi:10.3390/su14052609
8. Tun Thura Thet TT, Na J-C, Khoo CSG. Aspect-based Sentiment Analysis of Movie Reviews on Discussion Boards. *J Inf Sci* (2010) 36:823–48. doi:10.1177/0165551510388123
9. Liu B. Sentiment Analysis and Opinion Mining. *Synth lectures Hum Lang Tech* (2012) 5:1–167. doi:10.2200/s00416ed1v01y201204hlt016
10. Patil PP, Phansalkar S, Kryssanov VV. Topic Modelling for Aspect-Level Sentiment Analysis. In: AJ Kulkarni, SC Satapathy, T Kang, AH Kashan, editors. Proceedings of the 2nd International Conference on Data Engineering and Communication Technology; Singapore. Springer Singapore (2019). 221–9. doi:10.1007/978-981-13-1610-4_23
11. Poria S, Cambria E, Gelbukh A. Aspect Extraction for Opinion Mining with a Deep Convolutional Neural Network. *Knowledge-Based Syst* (2016) 108:42–9. doi:10.1016/j.knosys.2016.06.009
12. Wang W, Pan SJ, Dahlmeier D, Xiao X. Recursive Neural Conditional Random fields for Aspect-Based Sentiment Analysis. *arXiv preprint arXiv:1603.06679* (2016).doi:10.18653/v1/d16-1059
13. He R, Lee WS, Ng HT, Dahlmeier D. An Unsupervised Neural Attention Model for Aspect Extraction. *Proc 55th Annu Meet Assoc Comput Linguistics* (2017) 1:388–97. Long Papers). doi:10.18653/v1/p17-1036
14. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *J machine Learn Res* (2003) 3:993–1022.
15. Hofmann T. Probabilistic Latent Semantic Indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (1999). 50–7. doi:10.1145/312624.312649
16. Hofmann T (2013) Probabilistic Latent Semantic Analysis. *arXiv preprint arXiv:1301.6705* (2013).
17. Shams M, Baraani-Dastjerdi A. Enriched Lda (Elda): Combination of Latent Dirichlet Allocation with Word Co-occurrence Analysis for Aspect Extraction. *Expert Syst Appl* (2017) 80:136–46. doi:10.1016/j.eswa.2017.02.038
18. Lin C, He Y. Joint Sentiment/topic Model for Sentiment Analysis. In: Proceedings of the 18th ACM conference on Information and knowledge management (2009). 375–84. doi:10.1145/1645953.1646003
19. Moghaddam S, Ester M. Ilda: Interdependent Lda Model for Learning Latent Aspects and Their Ratings from Online Product Reviews. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (2011). 665–74.
20. Wang H, Lu Y, Zhai C. Latent Aspect Rating Analysis on Review Text Data: a Rating Regression Approach. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (2010). 783–92.
21. Wang H, Lu Y, Zhai C. Latent Aspect Rating Analysis without Aspect Keyword Supervision. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (2011). 618–26. doi:10.1145/2020408.2020505
22. Poria S, Cambria E, Ku LW, Gui C, Gelbukh A. A Rule-Based Approach to Aspect Extraction from Product Reviews. In: Proceedings of the second workshop on natural language processing for social media. SocialNLP (2014). 28–37. doi:10.3115/v1/w14-5905
23. Qiu G, Liu B, Bu J, Chen C. Opinion Word Expansion and Target Extraction through Double Propagation. *Comput linguistics* (2011) 37: 9–27. doi:10.1162/coli_a_00034
24. Gindl S, Weichselbraun A, Scharl A. Rule-based Opinion Target and Aspect Extraction to Acquire Affective Knowledge. In: Proceedings of the 22nd International Conference on World Wide Web (2013). 557–64. doi:10.1145/2487788.2487994
25. Su Q, Xu X, Guo H, Guo Z, Wu X, Zhang X, et al. Hidden Sentiment Association in Chinese Web Opinion Mining. In: Proceedings of the 17th

international conference on World Wide Web (2008). 959–68. doi:10.1145/1367497.1367627

26. Rana TA, Cheah Y-N. A Two-fold Rule-Based Model for Aspect Extraction. *Expert Syst Appl* (2017) 89:273–85. doi:10.1016/j.eswa.2017.07.047
27. Lu Y, Zhai C, Sundaresan N. Rated Aspect Summarization of Shor Comments. In: Proceedings of the 18th international conference on World wide web (2009). 131–40. doi:10.1145/1526709.1526728
28. Titov I, McDonald R. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. *proceedings ACL-*. HLT (2008). 08. 308–16.
29. Yang Y, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization. *Icml (Nashville, TN, USA)* (1997) 97:35.
30. Karaa WBA, Gribâa N. Information Retrieval with porter Stemmer: a New Version for English. In: Advances in computational science, engineering and information technology. Springer (2013). 243–54. doi:10.1007/978-3-319-00951-3_24

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ren, Zheng, Du and Pan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.