A Lens into Employee Peer Reviews via Sentiment-Aspect Modeling

Abhinav Maurya*, Leman Akoglu†, Ramayya Krishnan‡
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
*ahmaurya@cmu.edu, †lakoglu@andrew.cmu.edu, ‡rk2x@cmu.edu

Daniel Bay
Uber Inc.
San Francisco, CA
dbay@uber.com

Abstract—**Given a corpus of employee peer reviews from a large corporation where each review is structured into pros and cons, what are the prevalent traits that employees talk about? How can we describe the performance of an employee with just a few sentences, that help us interpret what their work is praised and criticized for? What is the best way to summarize an employee's reviews, while preserving the content and sentiment as well as possible?**

In this work, we study a large collection of corporation-wide employee peer reviews from a technology enterprise. Motivated by the challenges we outline in our analysis of employee review data, our work makes two main contributions in the domain of people analytics: (a) Sentiment-Aspect Model (SAM): we introduce a stylized log-linear model that identifies the hidden aspects and sentiment within an employee peer review corpus, (b) Interpretable Sentiment-Aspect Representations (EMPLOYEE2VEC): using SAM, we produce a vector space embedding for each employee, containing an overall sentiment score per aspect, and (c) Summarization of Employee Peer Reviews (PEERSUM): we summarize an employee's peer reviews with just a few sentences which reflect the most prevalent traits and associated sentiment for the employee as much as possible.

We show that our model SAM can use the structure present in the dataset as supervision to discover meaningful latent traits and sentiment embodied in the reviews. Our employee vector representations EMPLOYEE2VEC provide a compact, interpretable overview of their evaluation. The review summaries extracted by PEERSUM provide text that explains the professional performance of an employee in a succinct and objectively quantifiable way. We also show how to use our techniques for people analytics tasks such as the analysis of thematic differences between departments, regions, and genders.

I. INTRODUCTION

Given a large text collection of employee peer reviews within an enterprise, how can we extract meaningful structure that can help us answer a variety of questions about the employees or the enterprise at large? Specifically, what are the latent traits (or aspects) that employees discuss (praise or criticize) about one another? Which key phrases do they use to talk about certain traits, positively and negatively? How can we identify the aspect-level sentiment with which each employee is reviewed?

Peer reviewing is a commonly used means for evaluating employees in an enterprise, where each employee provides 'free text' feedback for each of their peers [1]. Around 90% of Fortune 1,000 companies have adopted a specific form of peer reviewing known as 360°-reviews [2]. Despite their drawbacks of leniency bias and centrality bias [3], peer evaluations have multiple key advantages. Reviews to an employee are written by people who closely work with them and hence can provide the most actionable feedback. Peer reviews reduce centralized bureaucracy and also allow employees to freely discuss their opinions about one another, without being tied to pre-specified questions, like in a survey.

The unstructured nature of peer reviews, however, comes at a cost for data analysis. The enterprise needs to extract implicit information from text to answer questions like "*What themes or traits are most prevalent across the vast amount of free text reviews?*", "*Are employees mentioning X (e.g., 'communication', 'execution', etc.) as a positive trait?*", "*What are the major themes that top managers are being praised for?*", "*Are there gender disparities in how and with respect to which traits employees are criticized?*", etc. Existing methods for text analytics rely entirely on structured data [4], ignore any sentiment in the peer reviews [5], [6], [7], assume that reviews are labeled with a given set of aspects [8], [9], or that various parts of a review deterministically inherit the overall rating provided by the user along with the review [8].

To effectively address our motivating questions, our work sets out to identify latent structure within the employee peer review corpus from a large ride-sharing enterprise. Latent structure here is two-fold: (a) *Aspects:* hidden traits about which the employees praise or criticize each other, such as work values like 'teamwork' or 'diligence' and (b) *Sentiment:* how positive or negative the feedback is, as evaluations are inherently subjective and convey opinions. Specifically, we propose a structured aspect-sentiment log-linear model that uses review keywords to describe various aspects as well as how positive or negative feedback about those aspects is conveyed by the employees. Moreover, we can summarize an employee review with just a few sentences that best reflect and justify the associated sentiment per aspect for that employee. Figure 1 illustrates an overview of our proposed approach.

We summarize our main contributions to the domain of people analytics as follows:

- **Sentiment-Aspect Model (SAM):** We propose a new multi-aspect sentiment model, called SAM, toward identifying hidden themes and sentiment in a given review corpus of employee peer reviews. Similar to topic models, our model is interpretable; it reveals not only the specific

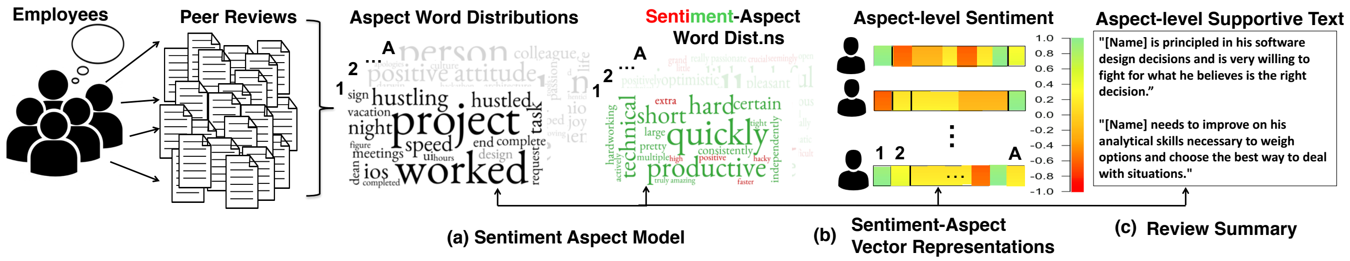

Employees
Peer Reviews
Aspect Word Distributions
Sentiment-Aspect Word Dist.ns
Aspect-level Sentiment
Aspect-level Supportive Text
1
2
...
A
person
colleague
positive
attitude
hustling
hustled
project
night
speed
end complete
design
meetings
worked
ios
task
request
hard
certain
short
quickly
technical
productive
optimistic
pleasant
positive
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
"[Name] is principled in his software design decisions and is very willing to fight for what he believes is the right decision."
"[Name] needs to improve on his analytical skills necessary to weigh options and choose the best way to deal with situations."
(a) Sentiment Aspect Model
(b) Sentiment-Aspect Vector Representations
(c) Review Summary

Fig. 1: Overview of proposed techniques. Given a large review corpus, (a) we find interpretable hidden aspects, and word distributions over aspects and aspect-level sentiment, (b) we construct vector representations of employees capturing overall aspect-level sentiment, and finally (c) we provide aspect-level supportive text summaries, reflective of overall sentiment.

aspect words, but also general sentiment words as well as aspect-specific sentiment words. (§IV)

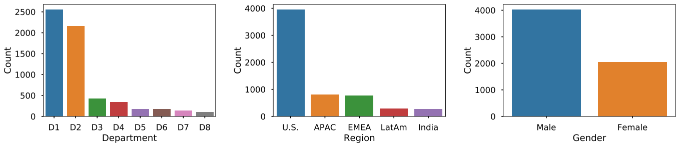
- **Employee Representation** (EMPLOYEE2VEC)**:** We employ "SAM at work" and show how to construct an interpretable vector representation per employee from all their reviews using SAM. EMPLOYEE2VEC is a representation where aspects are the dimensions, and entries reflect the sentiment for each aspect. (§V-A)
- **Peer Review Summarization** (PEERSUM)**:** Based on our model and the derived vector representations of each employee, we also produce an aspect-level text summary. The summary includes a few representative sentences per aspect that are reflective of its overall sentiment. (§V-B)

Our model formulation lends itself to efficient inference. The proposed sentiment-aspect modeling, representation learning, and text summarization tasks all scale linearly with the number of feedbacks in the review corpus. Our proposed techniques enable data scientists in the people analytics teams of large organizations to structure, analyze, and summarize large, unstructured peer reviews corpora. We evaluate the interpretability and utility of our methods through extensive experiments, and demonstrate different ways in which they can reveal new insights. We provide all the source code for this paper at https://bit.ly/2IYQFms.

II. RELATED WORK

Piazza et al. [4] provides a review of data mining applications in people analytics. Data mining of human resources data has typically focused on issues such as predicting employee churn, matching workers to jobs or tasks, assessing worker competencies for jobs, mining career trajectories, etc. To the best of our knowledge, ours is the first attempt in literature to develop statistical data mining techniques for analyzing employee peer reviews in the context of a technology corporation.

Topic models such as LSI [5], PLSI [6], and LDA [7] mainly focus on latent aspect/factor/topic modeling without explicit emphasis on sentiment. Moreover, topics they learn are often not representative of ratable aspects [10]. There are also LDA-extensions that explicitly model or infer topic and sentiment words from text [11], [12], [13], but none of those work can effectively exploit the availability of structured pros-and-cons.



(a) (b) (c)

Fig. 2: Histogram of employees by (a) department, (b) geographical region, and (c) gender.

Aspect-based sentiment models that leverage *a single numeric rating* (e.g., $r_{\text{Hilton}} = 4$) along with review text have been studied later. Specifically, [14], [15] infer (hidden) aspect-level ratings, based on review text and an overall rating using hidden factor models. Most related to our present work are aspect-sentiment models that use text as well as *a set of aspect-level (i.e., multi-aspect) ratings* (e.g., $r^{(room)}_{\text{Hilton}} = 3, r^{(staff)}_{\text{Hilton}} = 4, r^{(location)}_{\text{Hilton}} = 5$), in addition to an overall rating [8], [9]. These models, however, assume that the aspects are *pre-specified* and limited to those explicitly-rated aspects. However, it is a restricting assumption; it is observed that other non-explicit aspects (e.g., breakfast, parking, etc.) are mentioned in (e.g., hotel) reviews in addition to explicit aspects (e.g., room, location, etc.) [8]. Moreover, in these work when text segments (e.g. sentences) are assigned to aspects, they "inherit" the rating of their assigned aspect. In other words, all sentences of a certain aspect are assumed to have the *same* rating, which is also restrictive. In contrast, we find fully-latent rather than pre-defined aspects.

III. DATA AND CHALLENGES

In this work, we study a large collection of peer reviews among employees of a large ride-sharing company. The company goes through a performance evaluation cycle every six months. Evaluations consist of 360°-reviews: each employee writes reviews for their co-workers, manager, subordinates (if any), as well as a self-review [16]. For simplicity, we refer to all the reviews as peer reviews. Our collection contains all reviews from both the mid-year (MY: Jan-June) and end-of-year (EY: July-Dec) 2016 cycles.

The review form used to submit employee reviews consists of 6 free-text boxes, where the reviewer can highlight upto 3

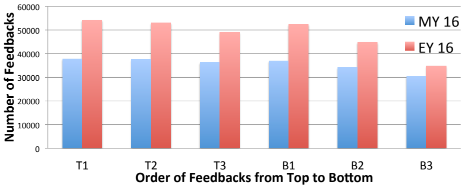
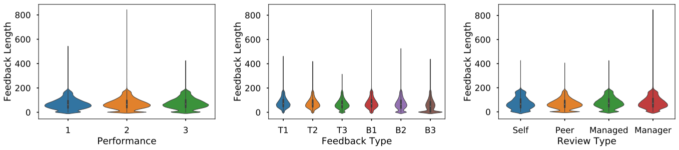


Fig. 3: Frequency of feedbacks by order. Employees tend to give fewer 'bottom' feedbacks.



(a) (b) (c)

Fig. 4: Violinplots of feedback length versus (a) employee performance (1=low, 2=moderate, 3=high), (b) question type (T's and B's), and (c) reviewer.

top and 3 bottom "qualities/traits/things" about the reviewee. T's can be thought as the qualities that the reviewee excels at, and B's as those that they are suggested to focus on improving. In this paper, we call each of these 6 components (3 T's and 3 B's) of a review a *feedback* and the associated T or B tag as the *feedback type*. Figure 2 shows the distribution of employees across eight departments of the corporation (2a), across five main geographical regions (2b), and across genders (2c). Most of the employees work in the US, belong to two prominent departments, and around 66% of all employees are men.

Figure 3 shows the overall frequency of feedbacks split by feedback type, for both MY and EY cycles. Increase in counts in EY is due to new hires, as the company is still growing at a substantial rate. Employees tend to provide notably fewer B2 and B3 (negative) feedbacks. This behavior is evident in both cycles, and even more so in EY. Figure 4 shows violinplots of the distribution of feedback length (in words) against employee performance (4a), feedback type i.e. B's and T's (4b), and reviewer type (4c). Length of feedback is not significantly correlated with employee performance, feedback type, or reviewer type. We therefore shift our focus to analysis of feedback content to discover employee insights.

Each feedback consists of feedback-id, reviewer-id, reviewee-id, feedback type (viz. T1, T2, T3, B1, B2, B3), and feedback text. An individual *review* consists of feedbacks with the same (reviewer-id, reviewee-id) pair. In total, the dataset contains more than half a million ($N = 501,337$) individual pieces of feedback.

Analysis of the employee reviews dataset is challenging for the following reasons:

- **Topic Analysis:** The corpus is difficult to model and analyze using off-the-shelf methods such as Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA). Table I shows prominent keywords from example topics extracted using NMF and LDA. The topics are not semantically coherent, indicating that the assumptions of these methods are violated leading to poor output on the dataset. In other words, our data is not as "topical" as many of the public text datasets.

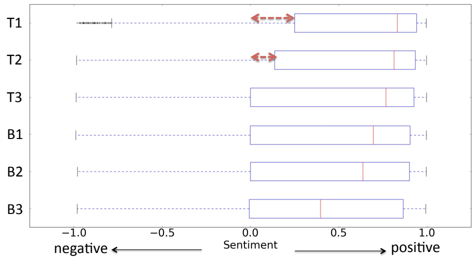


Fig. 5: Sentiment distribution (via boxplots) of feedbacks by order T1…B3. Employees tend to give feedbacks with non-negative sentiment, suggesting constructive criticism, where positivity drops from top to bottom.

- **Sentiment Analysis:** For preliminary analysis, we study the sentiment distribution among the T and B feedbacks. Using the `nltk.sentiment` sentiment analysis toolbox, we compute a sentiment score between $[-1, 1]$ for each feedback. Figure 5 shows the distribution of scores for each feedback group. The average sentiment drops from top to bottom. While fewer in number, B3 feedbacks are significantly less positive, but the sentiment scores still overlap heavily between the six feedback types. Almost all feedbacks have non-negative scores, suggesting mostly constructive criticisms. This general positivity and lack of polarity in sentiment (characterized as leniency and centrality biases respectively in [3]) indicates that analyzing useful distinctions between the T's and B's using an off-the-shelf sentiment analysis tool will perform poorly.

Motivated by these challenges, we develop a stylized log-linear model that jointly models both the aspects and the sentiment associated with the feedbacks, and utilizes the supervision provided by the feedback types - B's and T's - in modeling both general sentiment keywords as well as aspect-specific sentiment keywords.

IV. Proposed Model: SAM

In this work, our goal is to extract hidden information, to structure, summarize, and ultimately better interpret a large, unstructured peer reviews corpus. We address this problem in two steps.

First, we model the corpus with a log-linear **sentiment-aspect model**, called SAM, that identifies latent aspects, as well as word distributions over the aspects and positive/negative sentiment. We further capitalize on our model to transform all the reviews of each employee into an **employee representation** called Employee2Vec, that captures the overall sentiment per aspect. Second, we formulate a **summarization task**, called PeerSum, to explain the overall sentiment in employee embeddings by extracting short, supportive text from their reviews, which builds on SAM and Employee2Vec.

#	topic keywords ordered by importance	#	topic keywords ordered by importance
NMF 1	fc favours fawry faye faze fazed fazer fb fbi fbr	LDA 1	set works encourage fast quickly expectations work hard
NMF 2	team members member rest leader building player	LDA 2	learning want attention line space finish saying engaged pay
NMF 3	work life ethic balance easy pleasure makes efficiently	LDA 3	feedback design end bring provide closely research experience

TABLE I: Top 3 topics from Non-negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) on our dataset. NMF topics sorted by explained variance, and LDA topics by corpus-wide assignment counts. Both methods fit for 100 topics.

A. Model Specification

Formally, a feedback contains words $w \in f_i$ from a vocabulary $\mathcal{V}$ with size $V = |\mathcal{V}|$. The text corpus $\mathcal{T} = (\mathcal{F}, \mathcal{R})$ consists of feedbacks $\mathcal{F} = \{f_1, \ldots, f_N\}$ and their corresponding feedback types $\mathcal{R} = \{r_1, \ldots, r_N\}$, $r_i \in \{T, B\}$ respectively for Top and Bottom. Other than having been organized into top and bottom feedbacks, the reviews contain free text.

By design, each feedback f_i talks about a *single aspect* of the reviewee. However the aspects are *latent*; since individual feedbacks are not tagged or categorized explicitly in the dataset. Formally, each feedback should be associated with one of K latent aspects $\mathcal{A} = \{a_1, \ldots, a_K\}$, where $a(f_i)$, or a_i for short, denotes the (hidden) aspect of feedback f_i.

Besides unearthing the underlying aspects, we aim to learn a model that also captures sentiment. That is, we want a model that can predict the sentiment of a given piece of text. As we will show later, this enables us to summarize reviews while preserving sentiment. Our modeling problem is then stated as follows.

Problem 1 (Sentiment-Aspect Modeling): **Given** a review corpus $\mathcal{T} = (\mathcal{F}, \mathcal{R})$, containing individual feedbacks and their ratings, **Find** a model $\mathbf{M}$, that can (a) discover hidden aspects $\mathcal{A} = \{a_1, \ldots, a_K\}$ in $\mathcal{T}$, (b) map feedbacks $f_i \in \mathcal{F}$ to individual aspects $a_i \in \mathcal{A}$, (c) discover sentiment keywords that describe corpus-wide positive or negative sentiment, and (d) discover aspect-specific sentiment keywords beyond the general sentiment keywords.

To this end, we propose a new sentiment-aspect model called SAM. Specifically, we model the probability that a feedback f discusses a particular aspect a with feedback type r as

$$p(a, r|f) = p^{(\theta)}(a|f) \cdot p^{(\beta,\phi)}(r|a, f) \tag{1}$$
$$= \frac{1}{Z^{(\theta)}} \exp\left\{\sum_{w \in f} \theta_{aw}\right\} \cdot \frac{1}{Z^{(\beta,\phi)}} \exp\left\{\sum_{w \in f} (\beta_{rw} + \phi_{raw})\right\}$$

where we employ two log-linear models for probability estimation, with normalization constants $\frac{1}{Z^{(\theta)}} = \sum_{a' \in \mathcal{A}} \exp(\sum_{w \in f} \theta_{a'w})$ and $\frac{1}{Z^{(\beta,\phi)}} = \sum_{r' \in \{T,B\}} \exp(\sum_{w \in f} (\beta_{r'w} + \phi_{r'aw}))$.

The parameters of our model are $\mathbf{\Theta} = \{\theta, \beta, \phi\}$; where

- θ_{aw} is the 'weight' of word w for aspect a,
- β_{rw} is the 'sentiment-weight' of word w for feedback type r,
- ϕ_{raw} is the aspect-specific 'sentiment-weight' of word w for feedback type r under aspect a;

respectively capturing the aspects, general-sentiment, and aspect-level sentiment.

There are dependencies between feedbacks given or received by the same employee. Such dependencies are hard to capture without relational models. In this work, we make the simplifying assumption of i.i.d. feedbacks, and write the corpus likelihood as

$$L(\mathcal{F}, \mathcal{R}; \mathbf{\Theta}) = p^{\mathbf{\Theta}}(\mathbf{a}, \mathcal{R}|\mathcal{F}) = \prod_{i=1}^{N} p^{(\theta)}(a_i|f_i) \cdot p^{(\beta,\phi)}(r_i|a_i, f_i) \tag{2}$$

where $\mathbf{a}$ depicts the unknown aspect assignments of the feedbacks.

B. Model Inference

We estimate the model parameters from the corpus by maximizing the (regularized) log likelihood. Moreover, the feedback aspects a_i's are latent and need to be inferred. As such, the objective is

$$\hat{\mathbf{\Theta}}, \hat{\mathbf{a}} = \arg\max_{\mathbf{\Theta}, \mathbf{a}} \underbrace{\log p^{\mathbf{\Theta}}(\mathbf{a}, \mathcal{R}|\mathcal{F})}_{\text{corpus log-likelihood}} - \underbrace{R(\mathbf{\Theta})}_{\text{regularizer}}, \tag{3}$$

where

$$R(\mathbf{\Theta}) = \lambda_1(\|\theta\|_2 + \|\beta\|_2 + \|\phi\|_2)$$
$$+ \lambda_2(\log\det(\theta\theta^T) + \log\det(\beta\beta^T) + \log\det(\phi\phi^T)) .$$

In addition to the commonly used l_2 regularizer, $\log\det(\cdot)$ is a determinantal regularizer [17] that promotes diversity in the selection of parameters θ, β, and ϕ.

For the inference of our latent variable model, we use a coordinate ascent approach by alternately optimizing

$$\mathbf{a}^t = \arg\max_{\mathbf{a}} \log p^{\mathbf{\Theta}^t}(\mathbf{a}, \mathcal{R}|\mathcal{F}) \tag{4}$$

$$\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \log p^{\mathbf{\Theta}}(\mathbf{a}^t, \mathcal{R}|\mathcal{F}) - R(\mathbf{\Theta}) \tag{5}$$

at every iteration t, until aspect assignments converge; $\mathbf{a}^t = \mathbf{a}^{t-1}$. Eq. (4) consists of independently maximizing

$$a_i^t = \arg\max_{a \in \mathcal{A}} p^{(\theta^t)}(a|f_i) \cdot p^{(\beta^t,\phi^t)}(r_i|a, f_i)$$

for every feedback f_i. Note that this is a hard-assignment.

On the other hand, parameter estimation in Eq. (5) is concave and can be optimized via gradient ascent. Partial derivatives include:

$$\delta(\theta_{aw}) = \frac{\partial L(\mathbf{\Theta})}{\partial \theta_{aw}} = \sum_{i:a_i=a} \sum_{\substack{w' \in f_i \\ w'=w}} 1 - \sum_{i=1}^{N} p^{(\theta^t)}(a|f_i) \sum_{\substack{w' \in f_i \\ w'=w}} 1 \tag{6}$$

where the first and second terms are respectively the *actual* and the *expected* number of times word w appears in feedbacks with aspect a. A gradient ascent step updates

$$\theta_{aw}^{t+1} \leftarrow \theta_{aw}^{t} + \eta \frac{\delta(\theta_{aw})}{N}$$

where η is the learning rate. Intuitively, the parameter value is increased proportional to its *excess* amount of occurrence over the expectation. That is, the more (less) than expected number of times a word occurs in feedbacks of a certain aspect a, the higher (lower) its 'weight' gets for a. Similar intuition follows for other updates:

$$\delta(\beta_{rw}) = \frac{\partial L(\mathbf{\Theta})}{\partial \beta_{rw}} = \sum_{i:r_i=r} \sum_{\substack{w' \in f_i \\ w'=w}} 1 - \sum_{i=1}^{N} p^{(\beta^t,\phi^t)}(r|a_i, f_i) \sum_{\substack{w' \in f_i \\ w'=w}} 1 \tag{7}$$

$$\delta(\phi_{raw}) = \frac{\partial L(\mathbf{\Theta})}{\partial \phi_{raw}} = \sum_{\substack{i:a_i=a \\ r_i=r}} \sum_{\substack{w' \in f_i \\ w'=w}} 1 - \sum_{i:a_i=a} p^{(\beta^t,\phi^t)}(r|a, f_i) \sum_{\substack{w' \in f_i \\ w'=w}} 1 \tag{8}$$

Notice that our model as given in Eq. (1) is overparameterized, and multiple different combinations of θ, β, ϕ could lead to the same objective value. To enforce uniqueness and prevent oscillations in parameter learning [9], we enforce the absolute sum constraint on θ and β: $|\sum_w \theta_{aw}| = 1 \; \forall a$, and $|\sum_w \beta_{rw}| = 1 \; \forall r$. Such a constraint is easily enforced through projection of θ and β back onto the constraint-compliant polyhedra after each iteration of gradient descent.

Real data is messy, and getting mathematical models to work effectively with such data in practice often requires taking additional steps. To make our model more robust and interpretable, we employ two additional strategies.

By formulation, we expect β and ϕ to put more weight on sentiment words, whereas θ to put more weight on concept words. To capture this intuition, we partition the original vocabulary $\mathcal{V}$ into two: $\mathcal{V}_a$ containing only the nouns and verbs, and $\mathcal{V}_r$ containing the rest of the words (adjectives, adverbs, etc.).

Such a separation not only helps improve interpretability, but also simplifies our model by reducing the number of parameters to be estimated. Overall, we learn $|\mathcal{V}_a|K + 2|\mathcal{V}_r| + 2K|\mathcal{V}_r|$ parameters.

Solving non-concave maximization problems like our objective function in Eq. (3) using alternating optimization may find undesired local optima, depending on the initialization. We employ the following strategies to start on a potentially good initialization.

The parameters of our structured log-linear model can be interpreted as the 'weight's of the features, just like in simpler logit models, where positive (negative) weights increase (decrease) the probability likelihood. Aspects are hidden to us, therefore, we initialize all the aspect-driven parameters θ and ϕ to 0 (neutral), which is the unbiased initialization.

On the other hand, β is to capture general sentiment, independent of specific aspects. To set β, we leverage well-known publicly-available sentiment lexicons. We identify the nouns and verbs in our vocabulary that exist in those lexicons, and assign corresponding lexicon scores as the initial weights.

Each employee belongs to one of 8 departments, and each feedback is assigned to one of 14 coarsely defined "work values." Using this metadata, we initialize each feedback to one of $K = 8 \times 14 = 112$ aspects to discover a diverse set of work-related aspects in a data-driven fashion. In general, external information about the reviewees can be used to initialize aspects for feedbacks in a similar way.

V. SAM at Work

The unstructured nature of peer evaluations make it hard to systematically study, compare and contrast employees (across departments, different levels, etc.). Our next goal is to capitalize on our model to construct vector representations of the employees and use them to summarize an employee's peer reviews succinctly in a few sentences. Unlike various document embeddings, our representations should be interpretable.

Our emphasis is on representing each employee with just a few numbers that have real meaning to a human resources analyst. Similarly, our review summaries should provide an employee a representative sample of peer feedback that closely matches the overall sentiment associated with the employee's peer reviews.

A. Employee Representations via Employee2Vec

Our approach is to identify the dominant aspects that were discussed across all the reviews of a given employee, and estimate an overall sentiment for each such aspect. We construct vector $\mathbf{v}_e \in \mathbb{R}^K$ for each employee e, by computing the expected sentiment per aspect as follows:

$$\mathbf{v}_e[a] = \sum_{f \in \mathcal{F}_e} p^{(\theta)}(a|f) \cdot \left[2 \cdot p^{(\beta,\phi)}(r = Top|a, f) - 1\right], \; \forall a \in \mathcal{A} \tag{9}$$

Notice that we scale $p(r = Top|a, f) \in [0, 1]$ to $[-1, 1]$ to reflect sentiment (the higher the probability, the more positive the feedback is and vice versa), and take a weighted combination of sentiment across all feedbacks, without hard feedback-to-aspect assignment. Finally, we normalize the embeddings by $\frac{\mathbf{v}_e}{|\mathcal{F}_e|}$ as different employees receive different number of reviews and T/B feedbacks.

Intuitively, aspects that do not appear in an employee's reviews receive embedding value 0 since $p(a|f) \approx 0, \forall f \in \mathcal{F}_e$. Moreover, the larger the number of feedbacks that discuss a certain aspect positively or negatively, the larger its absolute value becomes.

As a result, the embeddings provide us with a *unified* representation of the employees, which enables various mainstream tasks such as distribution, cluster and outlier analysis.

B. Peer Review Summarization: PeerSum

Our sentiment-aspect vector representations readily serve as interpretable summaries of the reviews. They quickly route attention to a few dominant aspects among the feedbacks of the employee, and provide an overall sentiment of their evaluations with respect to those aspects.

Utilizing these *numerical* summaries, we aim to provide the employees with *supportive text* summaries. Supportive text is to justify the numerical summaries, i.e., it should be reflective of the estimated aspect-level sentiments. Moreover, it should be short, such that it still serves the purpose of a summary. Given an employee e, all their feedbacks $(\mathcal{F}_e, \mathcal{R}_e)$, vector representation $\mathbf{v}_e$, estimated model parameters $\mathbf{\Theta}$ from SAM, and a budget b; for each dominant aspect a' in $\mathbf{v}_e$, find b sentences $\mathcal{S} \subset \mathcal{F}_e$, $|\mathcal{S}| = b$ such that the expected sentiment of a' with $\mathcal{S}$ well approximates that with $\mathcal{F}_e$

$$p^{(\theta)}(a'|\mathcal{S}) * \left[2 \cdot p^{(\beta,\phi)}(r = Top|a', \mathcal{S}) - 1\right] \approx \mathbf{v}_e[a'] \ . \quad (10)$$

The dominant aspects are those on which the employee received a lot of praise or criticism, i.e., those with large absolute $\mathbf{v}_e[a']$ values. For supportive text, we decide to use individual sentences, as they are the smallest unites of cohesive text information and are easy to present. We also specify a user-defined budget b, which can be tuned interactively depending on the attention span of the user or the information content of the summary sentences.

An advantage of our summarization is its quantifiable and straightforward formulation. We split all the feedbacks of an employee into sentences, score individual sentences by Eq. (10), and choose top b sentences in a greedy fashion to provide the best possible reconstruction accuracy.

VI. Data Analysis and Business Insights

We applied our proposed model SAM and text summarization approach PeerSum to the human resources dataset described in Section III obtained from a technology corporation. We perform a qualitative and quantitative evaluation of SAM in sections VI-A and VI-B. In section VI-C, we use the Employee2Vec employee embeddings to learn which aspects are most associated with important professional demographics such as department, managerial level, and gender. Section VI-D discusses the quality of PeerSum through summariesextracted from the peer reviews of example employees. Finally, section VI-E demonstrates empirical scalability of our proposed techniques.

A. Discovered Aspects and Sentiment

In order to examine the usefulness of SAM, we look at salient aspects (θ), aspect-specific sentiment words (ϕ), and general sentiment words (β) discovered by training the model on peer reviews. In Figure 6, we show five example aspects, with the corresponding aspect-specific sentiment words in Figure 7:

1) First aspect talks about winning litigation and doing the job with an energetic, positive, and infectious outlook.
2) Second aspect discusses many engineering specific subjects such as code reviews, code diffs, design, documentation, features, etc. It is supported by aspect-sentiment phrases that focus on productivity, task readiness, and technical understanding.
3) Third aspect is also related to engineering and design but focuses more on code testing, bugs and their resolution, and overcoming setbacks. Aspect-sentiment phrases related to this aspect are also intuitive, e.g., 'persistent', 'relentless', 'eager', and focus on long term goals.
4) Fourth aspect revolves around teamwork, truly caring, positivity, recharging, and fostering a culture of helpfulness. Phrases such as 'incredibly helpful', 'contagious energy', 'really passionate', etc. form the sentiment associated with this aspect.
5) Fifth aspect talks about an open office environment, enjoyable work culture, and work-life balance. The aspect is powered by similar sentiment phrases such as 'authentic', 'genuine', 'natural', 'comfortable', and 'empower equally'.

In Figure 8, we visualize the general sentiment words in parameter β. The keywords associated with T's i.e. pros are in green and the ones associated with B's i.e. cons are in red. We can see that the parameter captures only general sentiment words such as *best*, *positive*, *willing*, *quickly*, *better* etc. However, the sentiment associated with each aspect is much richer as demonstrated by the variety in the aspect-specific sentiment words. Also, as pointed out in Section III, feedbacks are mostly populated with praise. The most prominent adjective that appears in cons i.e. *better* is indicative of encouragement and suggestions to perform better on the job.

B. SAM *Model Quality*

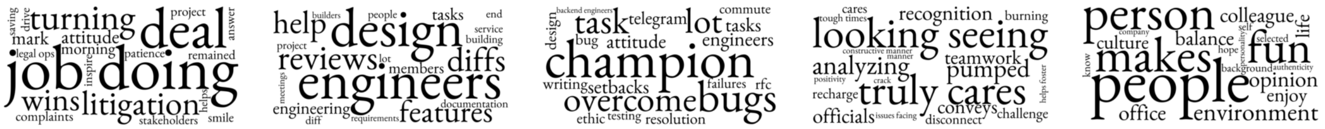
A simple and intuitive way to check for SAM's generalizability is to perform feedback type prediction on *unseen* test documents. To this end, we split the feedbacks into 50% train and 50% test feedbacks, train the model on the former, and predict feedback type i.e. T or B on the latter. We can predict the type of a feedback f given its predicted aspect a by

$$\arg\max_{r} p^{(\beta,\phi)}(r|a, f)$$

We obtain **93.05%** train accuracy and **93.99%** test accuracy on predicting feedback type i.e. distinguishing between B's and T's. Thus, our model learns aspects and aspect-specific sentiment words that help it distinguish effectively between pros and cons of employees on both training and unseen feedbacks.

Evaluation of aspect assignments, on the other hand, is difficult since assignment to aspects is essentially unsupervised and we lack any groundtruth to test aspect prediction on unseen test data. Hence, we examine distributional statistics of aspect assignments across train and test data to confirm that the aspects are meaningfully generalizable.

Figure 9a shows that the perplexity i.e. the negative log-likelihood of heldout validation data decreases with training iteration, indicating that the log-linear probabilistic model we employ learns to fit unseen data from the corpus. Figure 9b compares the average pairwise distance within the cluster versus between clusters using TF-IDF representations of feedbacks. The average intra-cluster distance is significantly lower than the corresponding inter-cluster average by 44.84% for training data and 39.67% for test data. This indicates that feedbacks assigned to the same aspect were far more similar than feedbacks assigned to different aspects, indicating that SAM successfully disentangled useful factors of variation in the unstructured feedbacks.



(a) Aspect I
"job attitude"

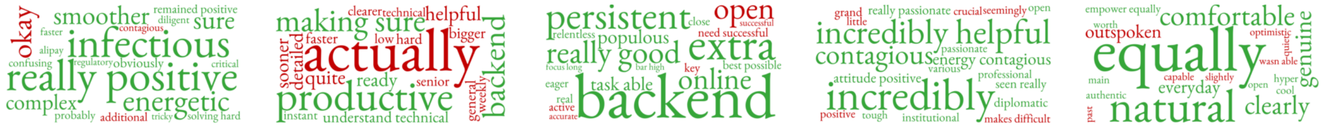
(b) Aspect II
"engineering design"

(c) Aspect III
"engineering testing"

(d) Aspect IV
"team support"

(e) Aspect V
"work environment"

Fig. 6: Example Aspect Wordclouds (θ)



(a) Aspect Sentiment I

(b) Aspect Sentiment II

(c) Aspect Sentiment III

(d) Aspect Sentiment IV

(e) Aspect Sentiment V

Fig. 7: Corresponding Aspect-Sentiment Wordclouds (ϕ) for the Aspects shown in Figure 6 (best in color). Red=negative and green=positive parameters/word weights; word size proportional to corresponding weight's magnitude.

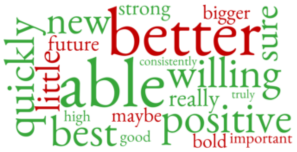
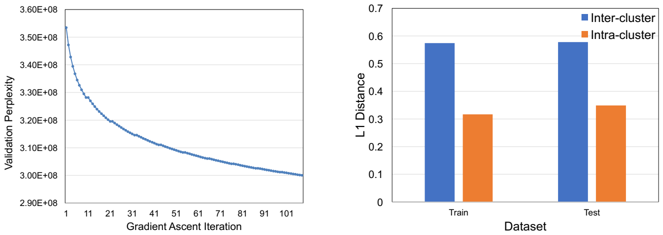


Fig. 8: General Sentiment Wordcloud (β)



(a) SAM Heldout Validation Perplexity

(b) Average inter-cluster vs. intra-cluster pairwise distances

Fig. 9: SAM Quality

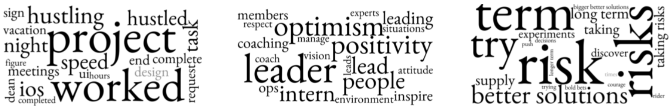
C. *Insights from* EMPLOYEE2VEC

In addition to reflecting the aspect-level sentiment of the employees, we also study how EMPLOYEE2VEC can be used for downstream data analysis tasks. One such way is to investigate which employee characteristics the vector representations are predictive about, and identify specific aspects (features) that capture significant differences among certain employee subpopulations. To this end, we perform the following regressions with sentiment-aspect embedding vectors as input features and certain employee characteristic as binary output response.

Characteristic	Top Distinctive Aspect	Coeff.	p-val
Engineers/Others	"software projects"	28.0721	0.000
Managers/Non-man.	"optimistic leadership"	9.9856	0.001
Males/Females	"risk-taking"	0.5238	0.007

TABLE II: Employee characteristics and aspects that distinguish these characteristics the most in a logit model.

Engineers vs. non-engineers: Since we are analyzing a technology company, it is interesting to identify aspects that distinguish engineers from non-engineers. To this end, we perform logistic regression using the EMPLOYEE2VEC values as input features, a response variable indicating if the employee was an engineer or a non-engineer, and introduce dummy control variables for categorical meta-data such as 'level' and 'gender'. Aspect shown in Figure 10a is found to have a large positive coefficient with statistical significance (see Table II). It places emphasis on terms such as 'project', 'ui', 'ios', as well as 'hustling', '[working] nights' which are indicative of the work done and traits valued in their peers by engineers.



(a) Engineers / Non-engineers

(b) Managers / Non-managers

(c) Male / Female

Fig. 10: Contrastive Aspect Wordclouds Most Distinctive of Certain Employee Subpopulations

Managers vs. non-managers: Another interesting factor that distinguishes employees in the corporate hierarchy is whether they are individual contributors or managers. We perform a similar logistic regression using employee embeddings as inputs and the output indicating if the employee is a manager or not. Figure 10b shows the aspect, sentiment of which are most predictive of managerial status in a statistically significant way. This aspect focuses on 'positivity', 'leadership', 'coaching', '[positive] attitude', and 'expertise'.

Gender: Our dataset did not have gender information available to us as employee meta-data. We inferred it by counting the number of male versus female pronouns in an employee's feedback. If the number of male pronouns was higher than the female ones, we tagged the employee as male, else they were tagged as female. As before, we performed logistic regression using the EMPLOYEE2VEC as input features, and gender as the response variable. Dummies for control variables such as 'level' and 'department' were introduced. The aspect shown in Figure 10c was statistically significant with a positive coefficient. It is about risk-taking and indicates that feedbacks associated with males identified them as more risk-taking than corresponding feedbacks for female employees.

D. *Text Summaries by* PEERSUM

Next we study the text summaries produced by PEERSUM for two example employees with a budget of 3 sentences.

Employee I Summary:

(1) "[Name] can gain more knowledge in business areas and work to be a leader help in coaching others to work on difficult problems."
(2) "I really love working with [Name] and hope to work with him on many more assignments in future."
(3) "He is very knowledgeable not only in [Technology] applications but also in various integrations with 3rd party tools."

The summary here is a mixture of praise and constructive criticism. The first extracted sentence here contains a suggestion for professional improvement, and the next two constitute praise for the employee's work.

Employee II Summary:

(1) "Most people on our team don't really see that side of the world very much."
(2) "He's willing to rollout initiatives even if they are a little more risky as it allows us to test and iterate towards a better experience in the long term and he's confident we'll get there."
(3) "I'd like to see [Name] kickoff some bigger initiatives and projects of his own that don't come out directly of product or sales needs."

In this case, the summary talks about the inter-departmental nature of the employee's job and their ability to take on calculated risks. This is followed by the third sentence in the summary which suggests that the employee should look at the bigger picture and undertake long-term initiatives.

From our detailed examination of the case studies, the extracted sentences above closely capture the overall sentiment associated with the employee's peer reviews effectively.

E. *Scalability*

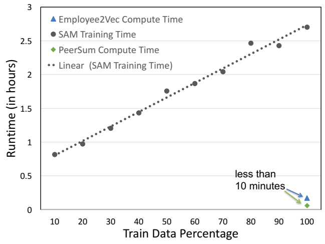


Fig. 11: Model training scalability is linear on dataset size.

In Figure 11, we show the running time to train our model on increasing fractions of the training data ($213,208$ feedbacks in total). We see that SAM's scalability is near-linear on the size of the data, i.e. number of feedback documents. Computing EMPLOYEE2VEC representations as well as generating text summaries using PEERSUM with budget $b = 10$ for most frequent 10 aspects for all the employees take negligible time, less than 10 minutes each on the entire training data.

VII. CONCLUSION

In this work, we modeled and analyzed a large employee peer review corpus from a technology corporation. We introduced a new **sentiment-aspect model** called SAM that can identify multiple latent aspects and associated sentiment. Next, we showed how to "put SAM to work" in order to construct vector representations for employees through EMPLOYEE2VEC, entries of which capture aspect-specific sentiment. We also developed a **peer review summarization** approach called PEERSUM, to extract a few sentences per employee reflective of each aspect's associated sentiment. Through various experiments, we demonstrated the quality and utility of our proposed techniques. We also employed our results on people analytics tasks, to illustrate how they can be used to gain new managerial insights such as revealing gender biases in the workplace and identifying employees with the potential to become good managers.

ACKNOWLEDGMENTS

This research is sponsored by NSF CAREER 1452425 and IIS 1408287. We thank Asna Ansari from Uber Technologies Inc. for helping us in our data analysis. Any conclusions expressed in this material are of the authors and do not necessarily reflect the views, expressed or implied, of the funding parties or our industrial partners.

REFERENCES

[1] Michael Armstrong. *A handbook of human resource management practice*. Kogan Page Publishers, 2006.

[2] David A Waldman, Leanne E Atwater, and David Antonioni. Has 360 degree feedback gone amok? *The Academy of Management Executive*, 12(2):86–94, 1998.

[3] Russell Golman and Sudeep Bhatia. Performance evaluation inflation and compression. *Accounting, Organizations and Society*, 37(8):534–543, 2012.

[4] Franca Piazza and Stefan Strohmeier. Domain-driven data mining in human resource management: A review. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 458–465. IEEE, 2011.

[5] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.*, 41(6):391–407, 1990.

[6] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[7] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[8] Ivan Titov and Ryan T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, pages 308–316, 2008.

[9] Julian J. McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *ICDM*, pages 1020–1025, 2012.

[10] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, 2008.

[11] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW*, pages 171–180, 2007.

[12] Wayne X. Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *EMNLP*, pages 56–65, 2010.

[13] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *ACL*, pages 804–812, 2010.

[14] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *KDD*, pages 618–626. ACM, 2011.

[15] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *KDD*, pages 193–202, 2014.

[16] Leanne Atwater and David Waldman. 360 degree feedback and leadership development. *The Leadership Quarterly*, 9(4):423–426, 1998.

[17] Veronika Rocková, Gemma Moran, and Edward George. Determinantal regularization for ensemble variable selection. In *Artificial Intelligence and Statistics*, pages 1105–1113, 2016.