Employee Analytics through Sentiment Analysis

Conference Paper · October 2015

DOI: 10.13140/RG.2.1.1623.3688

Employee Analytics through Sentiment Analysis

André Costa, Adriano Veloso
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{andrecosta,adrianov}@dcc.ufmg.br

ABSTRACT

People discuss and talk about the most diverse topics in social media platforms, including about their jobs. This results in a stream of employee-related data, and organizations are increasingly interested in making sense of this data on an ongoing basis in order to assess key factors such as employee engagement, retention and satisfaction. In this paper we propose to estimate such factors from different sentiments that are implicit in employee communications in social media platforms. We introduce sentiment analysis approaches that are based on learning vector representations for employee communications in an unsupervised way, and then these representations are given as input to a state-of-the-art supervised regression algorithm which finally maps text/sentiment to employee factors. We collected a large set of employee communications in social platforms, survey data such as work/life balance, job culture and management, and also official data about retention and salary. Then, we performed a systematic set of experiments using the collected data, and our results show that learning representations leads to better sentiment analysis performance than engineering features based on the standard term-frequency and inverse-document-frequency numbers (i.e., TF-IDF weighting scheme).

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.7 [**Natural Language Processing**]: Sentiment Analysis; J.4 [**Social and Behavioral Sciences**]: Miscellaneous

General Terms

Experimentation, Measurement, Performance

Keywords

Employee Analytics, People Analytics, Sentiment Analysis, Opinion Mining, Social Media, Word2Vec

SBBD'15 Proceedings of the 30th Brazilian Symposium on Databases, October 13-16, 2015, Petrópolis, RJ, Brazil

1. INTRODUCTION

Studies have found that 70% of American workers are not engaged or are actively disengaged in their jobs.[1] There is a direct correlation between employee engagement[2] and the quality of the customer experience, and thus disengaged employees may cost to businesses up to US$550 billion per year. In this paper we propose approaches that capture and analyze employee feedback, and then use them to understand factors that impact employee retention, satisfaction, and productivity. Specifically, our proposed approaches analyse positive and negative evidence and reviews appearing in social media platforms, such as `www.indeed.com` and `www.lovemondays.com.br`, to figure out what employees think about their jobs in order to build predictive models for diverse criteria of interest.

Our proposed approaches to sentiment analysis require: (i) representing text as a fixed-length vector and (ii) using machine learning techniques in order to score sentiments expressed in the text. Textual data (or more specifically job reviews) are represented as features vectors. For instance, a review may be represented as a vector of TF-IDF weights [1]. Then, a state-of-the-art regression or classification algorithm, such as SVR [6] and SVM [8, 9, 3] are used in order to separate sentiments.

Our focus in this paper is in improving text representations for employee analytics (or people analytics) through sentiment analysis. While the TF-IDF scheme has been successfully adopted in many application scenarios, including sentiment analysis [13, 20], its weighting scheme assumes an exact match of words and thus ignores the similarity between synonyms and cannot distinguish polysemy. The accuracy of sentiment analysis largely depends on measuring the semantic similarity of document pairs that are associated with the same sentiments, and thus we assume that an improved semantic representation for documents would lead to more accurate sentiment analysis. Thus, in addition to approaches based on the standard TF-IDF representation, we also evaluate approaches that learn a vector representation for words, such that related words appear next to each other in a common space [16]. The learnt word vectors are then combined to represent a document [10].

Contributions and Findings. In practice, we claim the following benefits and contributions over existing solutions:

[1] `http://www.gallup.com/services/176708/state-american-workplace.aspx`

[2] Employee engagement is the extent to which an employee is enthusiastic about his or her work and committed to furthering the organisation's goals.

- Effective sentiment analysis approaches for employee analytics. Our approaches are based on learning vector representations for reviews posted by employees about their jobs. The representations are chosen in a way such that reviews that share words with similar meanings are placed next to each other in a common metric space.
- We collected a large number of job reviews posted in social platforms, as well as survey data such as work/life balance, management, culture, and also official data about retention and salary. We performed a systematic set of experiments in order to evaluated our proposed sentiment analysis approaches.

In the next section we discuss related work. In Section 3 we present our sentiment analysis approaches for employee analytics. In Section 4 we present our experiments and report results obtained by the proposed approaches. Finally, in Section 5 we conclude our paper.

2. BACKGROUND AND RELATED WORK

Social media is full of opinative content [11]. Analysing such opinative content is a task which is known as sentiment analysis. Approaches for sentiment analysis usually apply machine learning algorithms in order to derive predictive models based on the observed sentiments [18]. Such machine learning algorithms depends heavily on how the training data is represented. Paltoglou and Thelwall [20] systematically explore whether more sophisticated feature weighting schemes from Information Retrieval can enhance sentiment analysis accuracy. They showed that variants of the classic TF-IDF scheme provide significant increases in accuracy, especially when using a sublinear function for term frequency weights and document frequency smoothing.

Delta IDF weighting [13] is a supervised variant of TF-IDF weighting in which the IDF computation is done for each document class and then one value is subtracted from the other. The authors present evidence that this weighting helps with sentiment classification.

It is also common to improve the TF-IDF scheme with latent information, using directed graphical models, notably probabilistic Latent Semantic Indexing [7], as a variant of Latent Semantic Indexing [4] and Latent Dirichlet Allocation [2]. These approaches model each document as mixing proportions for latent components which are viewed as "topics", and a topic is represented as a probabilistic distribution over words to capture their correlations. In [12] the authors present a model to capture semantic similarities among words. The semantic component of their model learns word vectors via an unsupervised probabilistic model of documents.

Most of the existing works have focused on analysing the content of movie or general product reviews [14, 21]. There are also applications to other domains such as debates [23], news [5], and blogs [19]. To the best of our knowledge, this is the first study of sentiment analysis applied to employee analytics.

3. EXTRACTING EMPLOYEE SENTIMENT

The popular TF-IDF [1] scheme represents each employee review using frequency count of words in a basic vocabulary, and normalize the values using inverse review frequency count. Then, TF-IDF vectors are given as input to machine learning algorithms in order to produce predictive models. This is known as the BoW (Bag of Words) approach.

3.1 Learning Review Representations

Despite their popularity, bag-of-words features have two major weaknesses: they lose the ordering of the words in the review and they also ignore semantics of the words. An alternative approach is to learn review representations by capturing semantic and syntactic information of words. Specifically, we represent words by building word projections in a latent space of n dimensions, such that words with similar meanings are placed next to each other in a common $n-$dimensional space metric. There are efficient algorithms to learn word vector representations, and a particularly efficient one follows the skip-gram approach. In this case, a neural network [15, 16] receives as input a word w_i, and the output of the network could be w_{i-1}, w_{i-2}, w_{i+1}, and w_{i+2}. Informally, the task is to predict the context given a specific word w_i. During the process of predicting the context given a word w_i, the network essentially learns a set of weigths, which are then used to represent word w_i. Finally, a review can be represented as vector given as the centroid of the vector representations associated with the words in the review. This approach will referred as Avg-Vec. As with the TF-IDF representation, the Avg-Vec representation can also serve as input to a machine learning algorithm in order to build a predictive model.

A further alternative approach, called Par-Vec, is to learn paragraph vectors [10]. A paragraph vector can also be used to represent a review. The process of learning paragraph vectors is inspired by the process of learning word vector representations. Specifically, paragraph vectors are learnt by concatenating word vectors, and again, using a neural network to predict the next word in the paragraph (or in the review).

3.2 Prediction Models

Once reviews are represented as vectors, either using BoW, Avg-Vec or Par-Vec approaches, they are given as input to a regression or classification algorithm. Specifically, we used the SVR algorithm [6] for regression, and the SVM algorithm [8] for classification. These algorithms follows a supervised learning strategy, and associate patterns in the vector representation of the review and a variable or criterion of interest. Criteria can assume values as salary, retention, management, culture, work/life balance, and others.

4. RESULTS

In this section we present the experimental results for the evaluation of the proposed approaches in terms of predicting employee related variables from textual reviews. In summary, we show that Avg-Vec and Par-Vec approaches achieve similar performance. However, they are significantly superior than the standard BoW approach. All experiments were run on an Intel Core i7 64GB main memory equiped with a Tesla K-40 GPU accelerator with 12GB memory and ultra-fast 288 GB/s throughput.[3] Our algorithms were implemented using Word2Vec[4] for computing word vector representations using skip-gram, and ParagraphVec[5] for computing vector representations of text with arbitrary length. These implementations are compatible with the CUDA interface. We implemented the BoW approach.

4.1 Dataset

We collected employee reviews from two sources, namely: `www.indeed.com.br` (refereed to as Indeed) and `www.lovemondays.com.br` (referred to as LM). We used a total of 11 machines with different IP addresses to efficiently gather large amounts of data. In total we crawled 161,708 reviews (151,349 labeled and 10,359 unlabeled) from Indeed and 20,013 reviews (13,007 labeled and 7,006 unlabeled) from LM. Labeled reviews from Indeed come with ratings (ranging from 0 to 5) based on management, culture, work/life balance, benefits, and career opportunities. An overall rating is also available. Labeled reviews from LM come with the salary of the employee, which are divided into 11 ranges according to the number of minimum wages (ranging from less than a minimum wage to more than 10 minimum wages). Further, we also obtained official data concerning employee retention.

Retention is divided into two groups: companies with low retention (60,158 reviews), and companies with high retention (85,737 reviews). Figure 1 shows the disctribution of the overall rating (Upper), and the distribution of salaries (Bottom). Figure 2 shows the frequency distribution of words, segmented by overall rating on the Indeed dataset (Upper), and segmented by salary on the LM dataset (Bottom).

Figure 3 show wordles that were built reviews of employees associated with different salaries. We can grasp clear differences between the vocabulary used by these employees. The vocabulary also emcompasses names of the companies and the employee position in the company.

4.2 Evaluation Procedure

To evaluate the prediction performance of our approaches, we have used the standard Root Mean Squared Error (RMSE) measure, which gives a summarized measure of the prediction error for regression tasks, and the standard accuracy and F1 measures for classification tasks. We conducted tenfold cross validation using Indeed and LM datasets. Thus, the labeled dataset was arranged into ten folds, including training and test. At each run, nine folds are used as training set, and the remaining fold as test set. The results reported are the average of the ten runs. SVR and SVM parameters used are those that lead to the best results.

4.3 Prediction Performance

Our first experiment is concerned with predicting different objectives, from Indeed dataset, including: management, culture, work/life balance, benefits, and career opportunities. Both Avg-Vec and Par-Vec achieved similar performance numbers, and the BoW approach was the worst performer in the most prediction objectives and for the most word vector representations, as shown in Figure 4.

[3]We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU accelerator for this research.

[4]https://code.google.com/p/word2vec

[5]https://github.com/darshanhegde/ParagraphVec

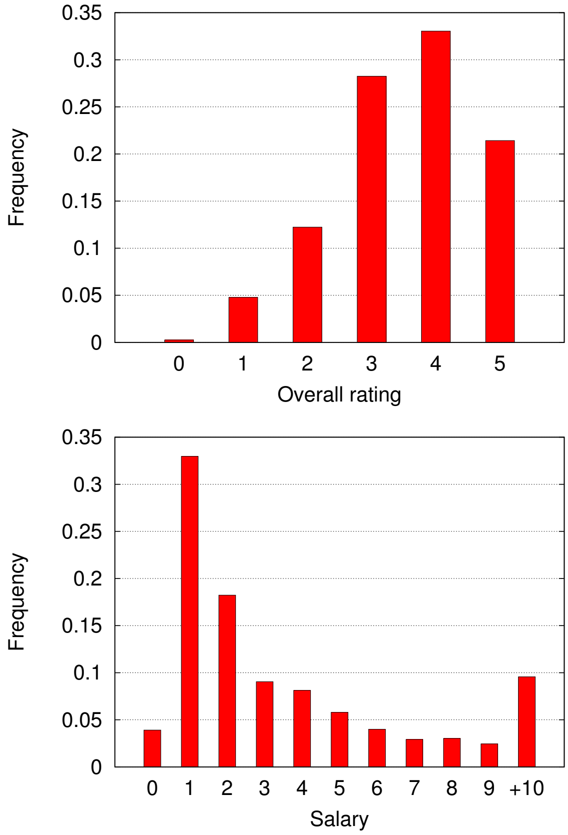

Figure 1: Distribution of ratings and salaries. Upper – Indeed. Bottom – Love Mondays.

The next experiment is concerned with predicting the salary of the employee, from the LM dataset, based on the review he has written. As shown in Figure 5, Avg-Vec provides the best RMSE numbers, followed by Par-Vec, until the representation with 512 features. However, from 1024 features, the performance of Avg-Vec and Par-Vec begins to deteriorate. This happens due to the amount of data available from LM (20,013 reviews). In other words, LM dataset does not have data enough for Avg-Vec and Par-Vec converge in higher dimensions.

According to Mikolov et al., several factors can influence the quality of the word vectors, like amount and quality of the datasets and size of the word vector representations [15, 16, 17]. Therefore, in order to try to improve the performance of Avg-Vec and Par-Vec against BoW, we joined data from Indeed and LM datasets, totaling 181,721 reviews. Figure 6 shows RMSE numbers, confirming that more data improve the results.

Another experiment performed is also concerned with predicting the salary of the employee based on the review he has written. However, for this experiment we varied the amount of unlabeled data available to learn word vector representations. Figure 7 (Upper) shows RMSE numbers by varying the amount of unlabeled data with context (i.e., reviews without the rating). Figure 7 (Bottom), on the other hand, shows RMSE numbers by varying the amount of unlabeled data without context. In this case, text was randomly collected from Wikipedia. Clearly, RMSE numbers decrease as

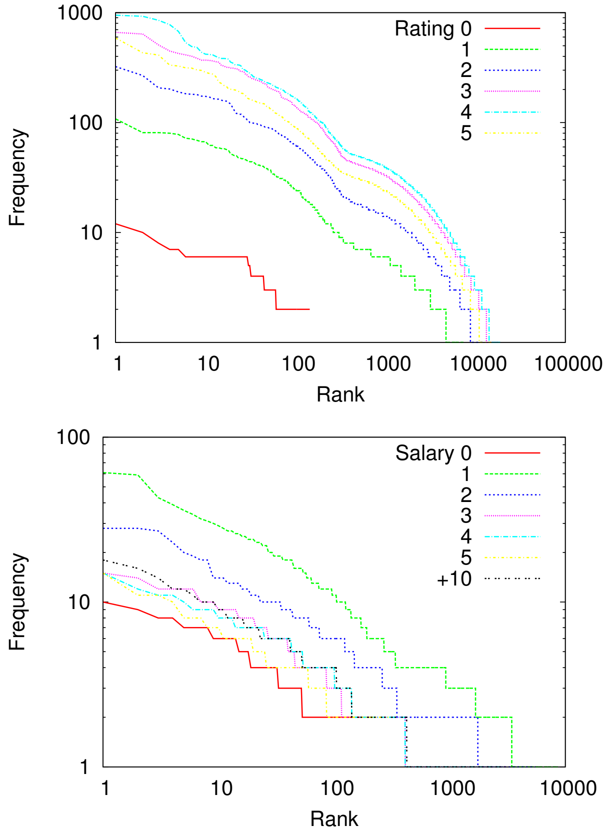

Figure 2: Frequency distribution of words. Upper – Indeed. Bottom – Love Mondays.

more unlabeled data with context is used to learn word vector representations. Unlabeled data without context, however, may be detrimental to learn word representations.

The last experiment is concerned with predicting the employee retention based on the review. In this case, we used the standard accuracy and F1 measures, since the retention can assume only two discrete values: low and high. Again, the best performers were Avg-Vec and Par-Vec, as shown in Figure 8. Specifically, Avg-Vec performed slightly better in terms of F1, while Par-Vec performed better in terms of accuracy. The BoW approach showed the worst results in terms of accuracy, but achieved competitive F1 numbers.

5. SIGNIFICANCE TESTS

We also applied Friedman (non-parametric statistical) and Nemenyi (post-hoc) tests to detect differences between the approaches BoW, Avg-Vec and Par-Vec across results achieved (RMSE, accuracy and F1 values) by experiments performed from all dimensions of vector representations considered [22].

Tables 1, 2, and 3 shows the tests results for different prediction objectives and dimensions. As most Friedman tests indicates significance (p-value < 0.01), it is valid to make multiple comparisons to identify differences between the approaches through the Nemenyi test. Tests with p-value ≥ 0.01 – Retention (F1) [medium-sized vector] and Salary (LM only) [high-dimensional vector] – indicate that the approaches BoW, Avg-Vec and Par-Vec are statistically

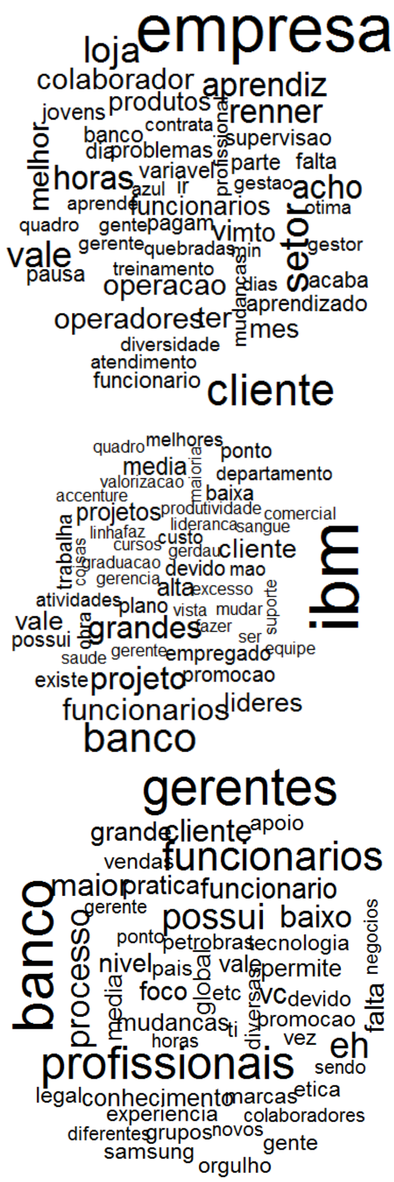

Figure 3: Upper – one minimum wage. Middle – 5-6 minimum wages. Bottom – +10 minimum wages.

similar, not being necessary to apply the Nemenyi test in these cases.

In the following we apply Nemenyi post-hoc tests. As shown in Table 4, we have the post-hoc tests for *low-dimensional* vector representations. The BoW approach differs significantly (p-value < 0.01) to Avg-Vec and Par-Vec approaches in all prediction objectives. Avg-Vec and Par-Vec approaches do not differ (p-value ≥ 0.01) in most cases, i.e., they are statistically similar in these cases. Thus, considering low-dimensional vectors, we can conclude (based on subsection 4.3) that:

- BoW is statistically worse than Avg-Vec and Par-Vec;
- Avg-Vec and Par-Vec approaches diverge (p-value $<$ 0.01) in two prediction criterias – Salary (Indeed+LM) and Retention (F1). In both criterias, Par-Vec is sta-

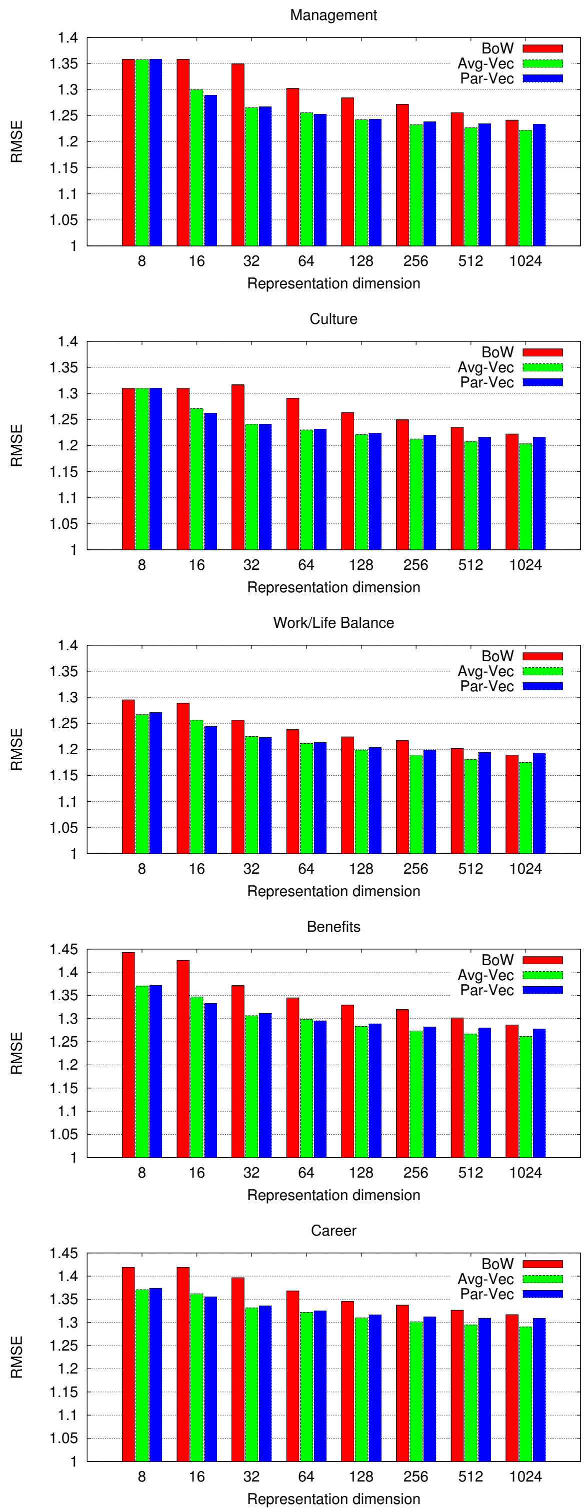

Figure 4: RMSE numbers for different objectives.

tistically better than Avg-Vec.

In Table 5, we have the post-hoc tests for *medium-sized* vector representations. The BoW approach differs significantly (p-value < 0.01) to Avg-Vec and Par-Vec approaches

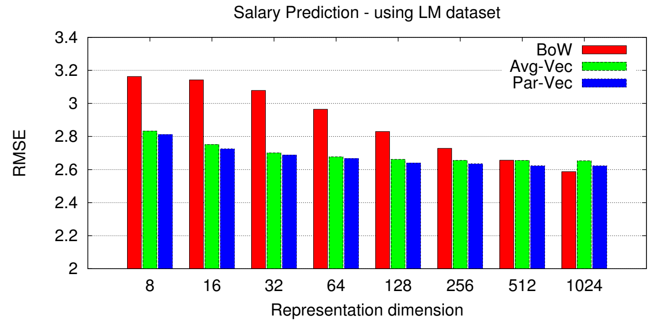

Figure 5: RMSE numbers according to the dimension of the representation (LM dataset)

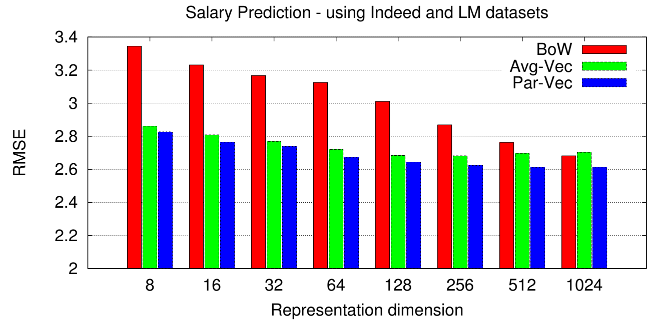

Figure 6: RMSE numbers according to the dimension of the representation (Indeed and LM datasets).

Table 1: Friedman Tests – Low-dimensional vector representations (2^3 to 2^5 features)

Prediction Objectives	p-value
Management	1.87×10^{-9}
Culture	5.93×10^{-6}
Work/Life	1.25×10^{-10}
Benefits	1.48×10^{-10}
Career	1.48×10^{-10}
Salary (LM only)	6.03×10^{-12}
Salary (Indeed + LM)	2.46×10^{-13}
Retention (Accuracy)	5.06×10^{-11}
Retention (F1)	1.39×10^{-12}

in all prediction objectives. Avg-Vec and Par-Vec approaches do not differ (p-value ≥ 0.01) in six criterias, i.e., they are statistically similar in these cases. Thus, considering *medium-sized* vectors, we can conclude (based on subsection 4.3) that:

- BoW is statistically worse than Avg-Vec and Par-Vec;
- Avg-Vec and Par-Vec approaches differ (p-value < 0.01) in three criterias: Work/Life, Career and Salary (Indeed+LM). So, Avg-Vec is significantly superior that Par-Vec in Career and Work/Life criterias; Par-Vec is significantly superior than Avg-Vec considering Salary (Indeed+LM) criteria.

In Table 6, we have the post-hoc tests for *high-dimension*

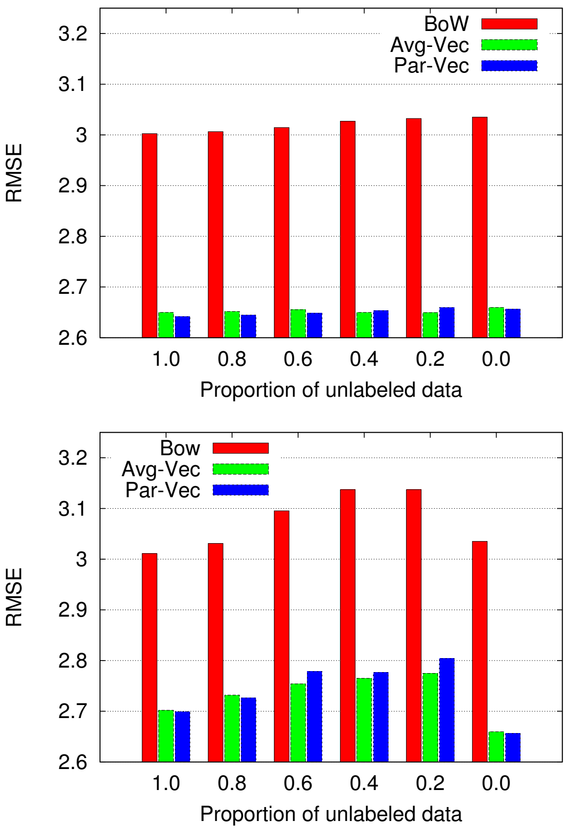

Figure 7: Salary prediction: RMSE numbers according to the proportion of unalabeled data available. Upper – Unlabeled data with context. Bottom – Unlabeled data without context.

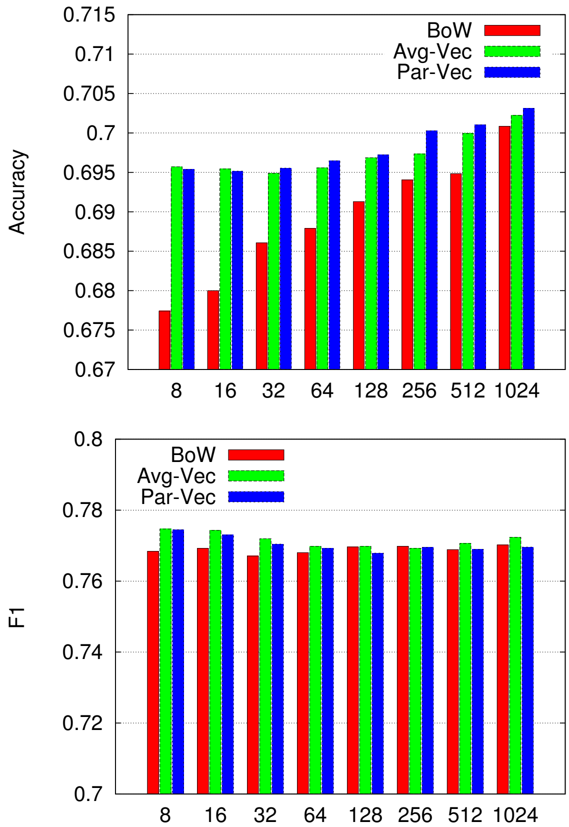

Figure 8: Retention prediction: accuracy and F1 numbers.

Table 2: Friedman Tests – Medium-sized vector representations (2^6 to 2^8 features)

Prediction Objectives	p-value
Management	1.64×10^{-10}
Culture	3.30×10^{-11}
Work/Life	2.46×10^{-13}
Benefits	9.93×10^{-11}
Career	2.46×10^{-13}
Salary (LM only)	1.14×10^{-11}
Salary (Indeed + LM)	2.46×10^{-13}
Retention (Accuracy)	6.65×10^{-11}
Retention (F1)	$\mathbf{8.21 \times 10^{-2}}$

Table 3: Friedman Tests – High-dimensional vector representations (2^9 to 2^{10} features)

Prediction Objectives	p-value
Management	2.06×10^{-9}
Culture	5.33×10^{-9}
Work/Life	2.50×10^{-7}
Benefits	2.06×10^{-9}
Career	2.06×10^{-9}
Salary (LM only)	$\mathbf{1.06 \times 10^{-2}}$
Salary (Indeed + LM)	1.95×10^{-7}
Retention (Accuracy)	2.89×10^{-4}
Retention (F1)	1.76×10^{-5}

vector representations. Thus, considering *medium-sized* vectors, we can conclude (based on subsection 4.3) that:

- BoW approach differs (p-value < 0.01) in most criterias, to Avg-Vec and Par-Vec. So, BoW is statistically worse than Avg-Vec and Par-Vec in these cases;
- in Salary (Indeed+LM) and Retention (Accuracy) criterias, BoW and Avg-Vec are statistically similar;
- in Culture, Work/Life and Retention (F1), BoW and Par-Vec are statistically similar.

- Avg-Vec differs significantly (p-value < 0.01) to Par-Vec in most cases, except Retention (Accuracy). So, Par-Vec is significantly superior that Avg-Vec only in Salary (Indeed+LM); Par-Vec is significantly superior than Avg-Vec in the other criterias.
- in Retention (Accuracy), Avg-Vec and Par-Vec approaches are statistically similar

6. CONCLUSIONS

The focus of this paper is on the analysis of employee opinions and sentiments. Such analysis enables learning effective models that can predict variables such as salary, re-

Table 4: Nemenyi post-hoc tests – Low-dimensional vector representations (2^3 to 2^5)

	BoW/Avg-Vec	BoW/Par-Vec	Avg-Vec/Par-Vec
Management	2.8×10^{-8}	2.8×10^{-6}	$\mathbf{6.8 \times 10^{-1}}$
Culture	5.3×10^{-4}	3.4×10^{-5}	$\mathbf{7.9 \times 10^{-1}}$
Work/Life	1.8×10^{-7}	1.7×10^{-9}	$\mathbf{7.2 \times 10^{-1}}$
Benefits	3.9×10^{-9}	8.5×10^{-8}	$\mathbf{8.6 \times 10^{-1}}$
Career	3.9×10^{-9}	8.5×10^{-8}	$\mathbf{8.6 \times 10^{-1}}$
Salary (LM only)	1.9×10^{-5}	3.8×10^{-12}	$\mathbf{2.7 \times 10^{-2}}$
Salary (Indeed + LM)	1.9×10^{-4}	9.7×10^{-14}	8.8×10^{-4}
Retention (Accuracy)	2.1×10^{-10}	1.0×10^{-6}	$\mathbf{3.3 \times 10^{-1}}$
Retention (F1)	5.8×10^{-13}	6.1×10^{-5}	5.5×10^{-3}

Table 5: Nemenyi post-hoc tests – Medium-sized vector representations (2^6 to 2^8 features)

	BoW/Avg-Vec	BoW/Par-Vec	Avg-Vec/Par-Vec
Management	8.6×10^{-9}	4.0×10^{-8}	$\mathbf{9.6 \times 10^{-1}}$
Culture	5.7×10^{-11}	2.8×10^{-6}	$\mathbf{1.7 \times 10^{-1}}$
Work/Life	9.7×10^{-14}	1.9×10^{-4}	8.8×10^{-4}
Benefits	7.6×10^{-10}	3.6×10^{-7}	$\mathbf{5.6 \times 10^{-1}}$
Career	9.7×10^{-14}	1.9×10^{-4}	8.8×10^{-4}
Salary (LM only)	1.0×10^{-5}	9.4×10^{-12}	$\mathbf{5.3 \times 10^{-2}}$
Salary (Indeed + LM)	1.9×10^{-4}	9.7×10^{-14}	8.8×10^{-4}
Retention (Accuracy)	1.9×10^{-5}	5.7×10^{-11}	$\mathbf{7.2 \times 10^{-2}}$
Retention (F1)	–	–	–

Table 6: Nemenyi post-hoc tests – High-dimensional vector representations (2^9 to 2^{10} features)

	BoW/Avg-Vec	BoW/Par-Vec	Avg-Vec/Par-Vec
Management	7.6×10^{-10}	4.5×10^{-3}	4.5×10^{-3}
Culture	2.1×10^{-9}	$\mathbf{1.2 \times 10^{-2}}$	2.6×10^{-3}
Work/Life	1.3×10^{-6}	$\mathbf{8.0 \times 10^{-1}}$	2.8×10^{-5}
Benefits	7.6×10^{-10}	4.5×10^{-3}	4.5×10^{-3}
Career	7.6×10^{-10}	4.5×10^{-3}	4.5×10^{-3}
Salary (LM only)	–	–	–
Salary (Indeed + LM)	$\mathbf{6.1 \times 10^{-1}}$	5.4×10^{-7}	5.8×10^{-5}
Retention (Accuracy)	$\mathbf{2.0 \times 10^{-2}}$	2.3×10^{-4}	$\mathbf{4.1 \times 10^{-1}}$
Retention (F1)	4.3×10^{-4}	$\mathbf{8.8 \times 10^{-1}}$	5.8×10^{-5}

tention, work/life balance, and career opportunities, based solely on employee reviews. We evaluate the standard TF-IDF bag-of-words approach to represent the reviews, and also more sophisticate approaches based on the skip-gram and paragraph vector represention. We collected employee reviews from social media platforms, as well as official data about salary and retention. We performed an extensive set of experiments and conclude that review representations obtained using skip-gram and paragraph vector approaches offer similar performance. Also, our experiments showed that both approaches are significantly superior than the standard BoW approach.

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England, 2011.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[4] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[5] A. Devitt and K. Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[6] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161, 1996.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[8] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *10th European Conference on Machine Learning*, pages 137–142, 1998.

[9] T. Joachims. Training linear svms in linear time. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217–226, 2006.

[10] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1188–1196, 2014.

[11] B. Liu. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[12] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics*, pages 142–150, 2011.

[13] J. Martineau and T. Finin. Delta TFIDF: an improved feature space for sentiment analysis. In *Proceedings of the 3rd International Conference on Weblogs and Social Media*, 2009.

[14] G. Mesnil, T. Mikolov, M. Ranzato, and Y. Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. In *Proceedings of Workshop at International Conference on Learning Representations*, 2015.

[15] T. Mikolov, K. Chen, G. Conrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*, 2013.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[17] T. Mikolov, W. Yi, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human*

Language Technologies, 2013.

[18] A. Mukherjee and B. Liu. Modeling review comments. In *The 50th Annual Meeting of the Association for Computational Linguistics*, pages 320–329, 2012.

[19] I. Ounis, C. Macdonald, and I. Soboroff. On the TREC blog track. In *Proceedings of the 2nd International Conference on Weblogs and Social Media*, 2008.

[20] G. Paltoglou and M. Thelwall. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, 2010.

[21] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2007.

[22] T. Pohlert. The pairwise multiple comparison of mean ranks package (pmcmr). 2014.

[23] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, 2006.