Jointly Modeling Review Content and Aspect Ratings for Review Rating Prediction

Zhipeng Jin[1,2], Qiudan Li[1], Daniel D. Zeng[1,2,3], YongCheng Zhan[3], Ruoran Liu[1,2], Lei Wang[1], Hongyuan Ma[4]

[1] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2] University of Chinese Academy of Sciences, Beijing, China

[3] Department of Management Information Systems, University of Arizona, Tucson, Arizona, USA

[4] CNCERT/CC, Beijing, China

{jinzhipeng2013, qiudan.li, dajun.zeng, liuruoran2016, l.wang}@ia.ac.cn, yongchengzhan@email.arizona.edu, mahongyuan@foxmail.com

ABSTRACT

Review rating prediction is of much importance for sentiment analysis and business intelligence. Existing methods work well when aspect-opinion pairs can be accurately extracted from review texts and aspect ratings are complete. The challenges of improving prediction accuracy are how to capture the semantics of review content and how to fill in the missing values of aspect ratings. In this paper, we propose a novel review rating prediction method, which improves the prediction accuracy by capturing deep semantics of review content and alleviating data missing problem of aspect ratings. The method firstly learns the latent vector representation of review content using skip-thought vectors, a state-of-the-art deep learning method, then, the missing values of aspect ratings are filled in based on users' history reviewing behaviors, finally, a novel optimization framework is proposed to predict the review rating. Experimental results on two real-world datasets demonstrate the efficacy of the proposed method.

General Terms

Algorithms, Design, Experimentation

Keywords

Review Rating Prediction; Skip-thought Vectors; Aspect Rating; Data Missing

1. INTRODUCTION

With the rapid development of e-commerce, abundant product reviews available on the Web have become a valuable source for consumers and organizations. Given a dataset of customer reviews, rating predictors can learn and predict the rating of a target product by a customer[1]. Horrigan[2] pointed out that consumers are willing to pay from 20% to 99% more to buy a product whose rating is 5-star than 4-star. Companies have more fine-grained requirements for keeping track of product quality. Therefore, review rating prediction is of much importance for sentiment analysis and business intelligence, because it enables the market to estimate how satisfied a customer will be with a product.

Review content and aspect rating play important roles in predicting review ratings[3,4], where the former includes the semantics and users' opinion and sentiment, and the latter reflects users' feelings on more specific aspects of a product. [5] used a linear regression to learn the user's specific aspect preference extracted from the review content, and predicted the overall review rating. [1] improved the accuracy of rating prediction using feature words extracted from customer reviews to reduce the dimension of the feature vector. These methods work well when aspect-opinion pairs can be accurately extracted from review texts and aspect ratings are complete. However, this is often difficult to implement by relying only on natural language processing techniques or other corpora such as WordNet[1]. Furthermore, in a real scenario, aspect ratings are optional for users, the incomplete aspect ratings will reduce prediction accuracy. To solve these problems, in this paper we propose a novel method for review rating prediction.

Recently, deep learning has been a hot research topic and successfully applied to sentiment analysis and natural language processing. The key aspect of deep learning is that it automatically learns features from raw data using a general-purpose learning procedure, instead of designing features by human engineers[6] . The characteristics of requiring very little engineering by hand makes it easily discover interesting patterns from large-scale social media data. [7] proposed a model, called skip-thought vectors, for learning high-quality sentence vectors without a particular supervised task in mind. The model abstracts the skip-gram model of words to the sentence level, and encodes a sentence to predict the sentences around it. Inspired by the good idea, our key idea is to adopt skip-thought vectors to learn the representation of review content, where review contents with similar semantics and sentiment will have similar vector representations. The novel vector representation of review encodes deep semantics and sentiment, thus will help improve the prediction accuracy. Then, we hypothesize that users' history reviewing behaviors of aspect rating can be explored and propose a novel model to alleviate data missing of aspect ratings. Finally, the obtained vector representation of review content and the dense aspect ratings are integrated into a unified optimization framework to predict the overall review rating.

2. METHOD

Figure 1 depicts the proposed method of review rating prediction. It consists of three modules: learning latent vector representation for review content, alleviating the missing values of aspect ratings with users' rating preference and predicting the overall rating in an optimization framework.

Learning latent vector representation for review content

The skip-thoughts [7] is adopted to learn the representation of

SIGIR '16, July 17-21, 2016, Pisa, Italy.

DOI: http://dx.doi.org/10.1145/2911451.2914692

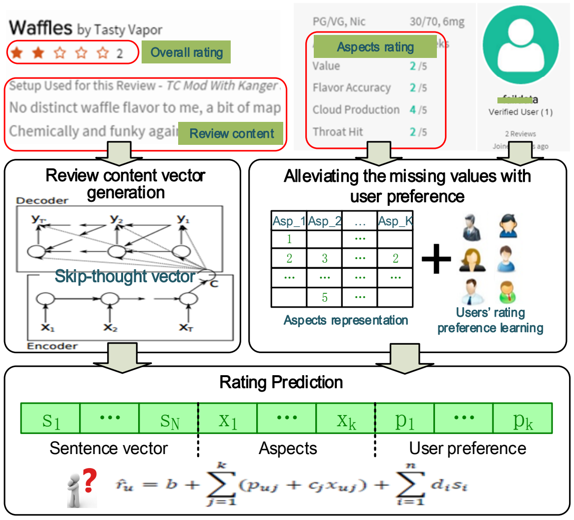

Figure 1. System architecture of the proposed model

review content. The model uses an RNN(Recurrent Neural Network) encoder to map words sequence to a sentence vector and an RNN decoder to generate the surrounding sentences. Given a sentence tuple (y_{i-1}, y_i, y_{i+1}), let w_i^t denote the t-th word in sentence y_i, h_i denote the encoder output of y_i. Then, the model aims to maximize the following objective function:

$$\sum_t logP(w_{i+1}^t | w_{i+1}^{<t}, h_i) + \sum_t logP(w_{i-1}^t | w_{i-1}^{<t}, h_i)$$

Where $P(w_{i+1}^t | w_{i+1}^{<t}, h_i)$, $P(w_{i-1}^t | w_{i-1}^{<t}, h_i)$ denote the probabilities of the t-th word for forward and backward sentences y_{i+1}, y_{i-1} given the previous t-1 words and the encoder representation h_i. The RNN structure can take any length of sentence as input, we thus employ the learnt encoder in [7] as a feature extractor to encode the review contents, then each review content is represented as a n-dimension vector $s \in \mathrm{R}^n$.

Filling in the missing values of aspect ratings The aspect ratings are important features for overall rating. For aspects set $[a_1, a_2, \dots, a_k]$, the normalized aspect ratings of user v_i are represented as$[x_{i1}, x_{i2}, \dots, x_{ik}]$, where k is the number of aspects and x_{i*} is set to be 0 if the aspect rating is missing. The missing values of aspect ratings will result in prediction bias. Generally, random value or mean constant cannot accurately predict the missing values[8]. As shown in[5], users have different preferences on different aspects. Some users are harsh on specific aspect thus give lower ratings while others tend to be tolerant with minor flaws and give relatively higher ratings. Therefore, we propose to learn user's aspect rating preference based on his/her history reviewing behavior to handle the missing values problem. A user preference matrix $\mathrm{P} \in \mathrm{R}^{m \times k}$ is constructed, where m is the number of users, each element p_{ij} denotes the user v_i's preference on aspect a_j. The element of the matrix is automatically learnt during the process of predicting the overall ratings, which is explained in detail in the next step.

Predicting the overall rating Based on the above analysis, the overall rating is highly related to three key factors: semantics of review content , aspect ratings with missing values and user preference, therefore, the overall rating of user v_u for one product is predicted as follows:

$$\hat{r}_u = b + \sum_{j=1}^{k} (p_{uj} + c_j x_{uj}) + \sum_{i=1}^{n} d_i s_i$$

To optimize the model parameters b, p_{uj}, c_j and d_i, we minimize the L2-regularized square error with SGD:

$$\min_{b, p_{uj}, c_j, d_i} \sum (r_u - \hat{r}_u)^2 + \lambda_1 (b^2 + \sum_{j=1}^{k} c_j^2 + \sum_{i=1}^{n} d_j^2) + \lambda_2 \sum_{j=1}^{k} p_{uj}^2$$

For each given rating r_u, prediction error is denoted as $\mathrm{e}_u = r_u - \hat{r}_u$, then the parameters are updated as follows:

- $b \leftarrow b + \gamma(e_u - \lambda_1 b)$
- $p_{uj} \leftarrow p_{uj} + \gamma(e_u - \lambda_2 p_{uj})$ iff $x_{uj} > 0$
- $c_j \leftarrow c_j + \gamma(e_u x_{uj} - \lambda_1 c_j)$
- $d_i \leftarrow d_i + \gamma(e_u s_i - \lambda_1 d_i)$

Where γ is the learning rate, λ_1, λ_2 are regularization parameters. It should be noted that the second rule explains the learning process of user preference matrix. $x_{uj} > 0$ indicates that the specific aspect rating is valid, thus we update p_{uj} to mine the underlying preference.

3. EXPERIMENTS

3.1 Dataset and Baseline Methods

To validate the performance of our review rating prediction model, we build two datasets from product review website JuiceDB[1] and service review website TripAdvisor[2] respectively.

JuiceDB: this website provides a very interesting review service for e-cigarettes, we use API to collect e-liquid reviews from June 2013 to November 2015. The dataset contains 14737 reviews for 4813 e-liquid products, each review is accompanied by an overall rating and a set of four aspect ratings of an e-liquid, each on a scale of 1-5: flavor accuracy, throat hit, value and cloud production. 25% of the reviews have incomplete aspect ratings.

TripAdvisor: this website provides reviews of travel-related content such as tourist attractions, hotels and restaurants. We use API to collect reviews about restaurants in New York from January 2015 to December 2015. The dataset contains 14524 reviews for 117 restaurants. Besides the overall rating, each review has three aspect ratings on a scale of 1-5: food, service, value. 38.3% of the reviews have incomplete aspect ratings.

We randomly partition the data into 75% for training and 25% for testing, repeat this process 5 times and present the averages. For the vector representation of review content, we use a 4800 dimensional vector, according to [7], it is formed by a unidirectional encoder with 2400 dimensions and a bidirectional encoder with another 2400 dimensions.

We compare our method against the following four baselines:

- **PredictMean**: This method simply uses the mean value of the ratings in the training set to predict rating.
- **BoW**(Bag-of-words): This method is a traditional representation method for documents, where unigrams and bigrams are used to represent a document.
- **ParagraphVector[9]**: This method learns the paragraph representations by predicting the surrounding words in contexts sampled from the paragraph. We utilize this unsupervised learning algorithm to represent each review content by a dense vector.
- **Aspects**: this method takes normalized aspect ratings as features to predict the overall rating directly.

The methods of BoW, ParagraphVector and Aspects map each review to different feature spaces. We then predict the

[1] https://www.juicedb.com/

[2] https://www.tripadvisor.com/

review ratings in a unified regression framework with these different feature representations.

3.2 Results and Discussions

To evaluate the accuracy of the method, we adopt mean absolute error (*MAE*) measure and coefficient of determination (R^2) measure:

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}|r_u - \hat{r}_u|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(r_u - \hat{r}_u)^2}{\sum_{i=1}^{M}(r_u - \bar{r_u})^2}$$

Where M is the number of samples, $\bar{r_u}$ denotes the average value of true ratings. MAE reflects the average difference between the predicted rating and the true rating, R^2reflects how much of the total variance is captured by the model. The comparison results of different methods on two datasets are shown below in Table 1 and Table 2 respectively in which Skip is short for Skip-thoughts , Aspects* refers to aspect ratings with user preference.

3.2.1 Review Rating Prediction in JuiceDB

Table 1 shows the performance comparison of different methods in JuiceDB. We empirically set $\gamma, \lambda_1, \lambda_2$ to 0.001, 0.0005, 0.01in this experiment.

Table 1. The comparison results in JuiceDB

	Method	MAE	R^2
Part 1	PredictMean	0.778	0.000
	BoW	0.645	0.322
	ParagraphVector	0.696	0.190
	Aspects	0.659	0.291
	Skip	**0.597**	**0.416**
Part 2	BoW +Aspects	0.616	0.386
	ParagraphVector +Aspects	0.603	0.412
	Skip +Aspects	**0.548**	**0.512**
Part 3	**Skip +Aspects***	**0.517**	**0.555**

- Rate Prediction using only review content or aspect ratings

We first study the prediction effects of different methods using only review content or aspect ratings. It can be observed from part 1 that all these methods perform better than PredictMean which indicates the importance of both review content and aspect ratings. Skip-thoughts model achieves the best performance in terms of both MAE and R^2. Results of Aspects is relatively poor due to the high percentage(25%) of incomplete aspect ratings. Compared with Aspects, better results of Skip-thoughts indicate that accurate semantic representation of review content has more impact than aspect ratings in overall rating prediction.

- Rate Prediction using both review content and aspect ratings

The above experimental results enlighten us to combine both of aspect ratings and review content for overall rating prediction. Results in Part 2 of Table1 demonstrate the effectiveness of the combination and “Skip+Aspects” achieves better performance than the other two. Furthermore, by filling in the missing values of aspect ratings using the learnt user preference matrix, the proposed “Skip+Aspects*” in Part3 that integrates vector representation of review content by skip-thought vectors and aspect ratings achieves the best performance. Our final model can predict a user’s review rating of a product with MAE 0.517 on average. This implies that the deep semantics of review content can be captured by skip-thought vectors, moreover, the learnt user preference matrix based on the users’ history reviewing behaviors is more reliable, thus solving the missing data problem of aspect ratings and improving the performance of overall rating prediction.

3.2.2 Review Rating Prediction in TripAdvisor

Table 2 shows the comparison results of different methods in TripAdvisor. We set $\gamma, \lambda_1, \lambda_2$ to 0.001, 0.0005, 0.01 in this experiment empirically.

- Rate Prediction using only review content or aspect ratings

From Part 1, it can be observed that compared with PredictMean, all the other learning methods achieve significant improvements and Skip still has the best performance in both MAE and R^2. The good performance in both two datasets further indicates that Skip-thoughts model is robust to mine the semantic representations of review contents and predict the ratings.

Table 2. The comparison results in TripAdvisor

	Method	MAE	R^2
Part 1	PredictMean	0.614	0.000
	BoW	0.535	0.165
	ParagraphVector	0.535	0.153
	Aspects	0.588	0.043
	Skip	**0.477**	**0.322**
Part 2	BoW +Aspects	0.531	0.187
	Paragraph Vector +Aspects	0.522	0.191
	Skip +Aspects	**0.473**	**0.344**
Part 3	Skip +Aspects*	0.476	0.308

- Rate Prediction using both review content and aspect ratings

We further combine the review content representation and the aspect ratings to predict the overall rating. From Part 2 and Part 3 in Table 2, we can observe that the three combination ways all achieve better results than methods considering only one factor in Part 1. “Skip+Aspects” obtains the best results in both measures. This further implies that deep semantics of review content and aspects ratings do play important roles in review rating prediction. It can be seen from the results in JuiceDB and TripAdvisor that “Skip+Aspects*” works better than “Skip+Aspects” in the former data set, which has less incomplete aspect ratings. The reason may be that a larger percentage of missing aspect ratings on TripAdvisor makes it difficult for Skip+Aspects* to learn user preference from users’ history reviewing behaviors.

3.2.3 Case Study

The above experiments in JuiceDB and TripAdvisor have demonstrated the effectiveness of our proposed model. Specially, learning semantic representations of review content using skip-thought vectors and filling in missing values of aspect ratings show advantages on improving the accuracy of rating prediction. To give deep insights into the proposed model, we illustrate these two aspects by using intuitive examples in detail.

Firstly, we select some seed reviews and calculate their similar reviews using cosine similarity based on the skip-thought vectors. The results are shown in Table 3 and Table 4. For each seed review, we represent two similar reviews along with their similarities.

In Table 3, we represent one seed review with negative sentiment from JuiceDB. Both the similar reviews are negative and contain negative words like “horrible”, “bad”, “nauseous” which are synonyms to “awful” in the seed. Table 4 presents one positive seed review from TripAdvisor. The similar reviews include similar expressions such as “would definitely return”,“will definitely return”. Besides, the similar reviews contain some positive words like “great”, “good”, “excellent”, “helpful” to express compliments about the experience. The above examples show that without feature engineering, the skip-thought vector is capable of mapping reviews with similar sentiment and semantics into similar vector representations successfully. This verifies the advantage of skip-thought vectors over other representation methods.

Secondly, aspects features are also of importance to the overall review rating prediction. By observing the weights of the aspect features from the trained model, we can have a fine-grained analysis of the aspects. In JuiceDB, aspects are ranked in terms of weights: flavor accuracy, value, throat hit, cloud production. Flavor accuracy and value are the two most important factors when users evaluate one e-liquid product. In TripAdvisor, we rank the aspects in the same way: value, food, service. We can observe that value and food are two factors people care most when they choose a restaurant. Moreover, by incorporating users’ aspect rating preferences, our final model predicts the overall rating more accurately. For example, in the test stage in JuiceDB, a user rates one e-liquid with 3 stars. The “Skip+Aspects” method can predict it with a score of 2.86. But by utilizing the user’s learnt preference vector [0.387 0.305 -1.44 0.385], our final model “Skip+Aspects*” can predict the rating with a score of 2.95 which further illustrate the effectiveness of the preference matrix.

Take JuiceDB for an example, it could serve as an early warning on the use patterns of e-liquid. Policy makers of FDA (Food and Drug Administration) could make use of the predicted ratings and similar sentiments to identify e-liquids that are potentially harmful and addicting. For example, e-liquids that are found to hurt users’ throat, make users feel bad in some way or significantly make users want to use more, could in turn inform the development and implementation of new regulations or laws.

Table 3. The similar reviews mined by Skip-thought vectors in JuiceDB

	review content	similarity
seed review 1	**Awful**. **Just awful**.	
similar reviews	**horrible**. imagine an ashtray with cloves...**just bad**.	0.746
	Pretty **bad**. Can't even vape it without feeling **nauseous**.	0.714

Table 4. The similar reviews mined by Skip-thought vectors in TripAdvisor

	review content	similarity
seed review 2	Really enjoyed our pizza, which was huge. The service was **quick**, **efficient** and really **friendly**. Was very **good** value for money, **will definitely return**.	
similar reviews	We went here for the prix fixe lunch which was very **reasonable** - all the food was **great** and really **good** service - **would definitely return**.	0.839
	A special restaurant for celebrations. Portions are very ample. Quality of food **excellent**. We split an appetizer and salad and thought it was a full order. The staff was **attentive** and extremely **helpful**. **Would definitely return**.	0.832

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel review rating prediction method. Experimental results show that high-quality representation of review content and complete aspect ratings play important roles in improving prediction accuracy. This work is a first step towards learning deep semantics of review content using skip-thought vectors in review rating prediction. The framework can integrate other information such as reviewer’s information, product information, etc. in the future. It will also be very interesting to apply the method to analyze users’ detail experience and how producers and marketers use social media to promote and sell various products or service, which could provide valuable information for regulatory decision-makers.

5. ACKNOWLEDGMENTS

This research is supported by the NNSFC projects (No. 61172106, 91224008, 61402123, 71472175) and Important National Science & Technology Specific Project (No. 2013ZX10004218).

6. REFERENCES

[1] Ochi, M., Okabe, M., and Onai, R., 2011. Rating prediction using feature words extracted from customer reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 1205-1206.

[2] Horrigan, J.A., 2008. Online shopping. *Pew Internet & American Life Project Report*.

[3] Ganu, G., Elhadad, N., and Marian, A., 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content. In *WebDB*. Citeseer, 1-6.

[4] Diao, Q., Qiu, M., Wu, C.-Y., Smola, A.J., Jiang, J., and Wang, C., 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 193-202.

[5] Mukherjee, S., Basu, G., and Joshi, S., 2013. Incorporating author preference in sentiment rating prediction of reviews. In *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 47-48.

[6] Lecun, Y., Bengio, Y., and Hinton, G., 2015. Deep learning. *Nature*. 521, 7553, 436-444. DOI= http://dx.doi.org/10.1038/nature14539.

[7] Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S., 2015. Skip-thought vectors. In *Advances in neural information processing systems*, 3276-3284.

[8] Soley-Bori, M., 2013. *Dealing with missing data: Key assumptions and methods for applied analysis*. Boston University.

[9] Le, Q.V. and Mikolov, T., 2014. Distributed representations of sentences and documents. In *Proceedings of the international conference on machine learning(ICML)*.