Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 107 – 123
© TÜBİTAK
doi:10.3906/elk-1905-114

Integrated topic modeling and sentiment analysis: a review rating prediction approach for recommender systems

Anbazhagan MAHADEVAN[1,*], **Michael AROCK**[2]
[1]Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, India,
[2]Department of Computer Applications, National Institute of Technology, Trichy, India

Received: 20.05.2019 • **Accepted/Published Online:** 16.10.2019 • **Final Version:** 27.01.2020

Abstract: Recommender systems (RSs) are running behind E-commerce websites to recommend items that are likely to be bought by users. Most of the existing RSs are relying on mere star ratings while making recommendations. However, ratings alone cannot help RSs make accurate recommendations, as they cannot properly capture sentiments expressed towards various aspects of the items. The other rich and expressive source of information available that can help make accurate recommendations is user reviews. Because of their voluminous nature, reviews lead to the information overloading problem. Hence, drawing out the user opinion from reviews is a decisive job. Therefore, this paper aims to build a review rating prediction model that simultaneously captures the topics and sentiments present in the reviews which are then used as features for the rating prediction. A new sentiment-enriched and topic-modeling-based review rating prediction technique which can recognize modern review contents is proposed to facilitate this feature. Experimental results show that the proposed model best infers the rating from reviews by harnessing the vital information present in them.

Key words: Recommender systems, topic modeling, latent dirichlet allocation, valence aware dictionary and sentiment reasoner, regression analysis

1. Introduction

With the growing amount of information available on the Internet and substantial rise in the number of Internet users, it becomes important for the E-commerce websites to use recommender systems (RSs) to keep their customers informed of products they are likely to buy. Being a subset of information filtering, RSs must be capable of making accurate recommendations to the users by digging out valuable facts from the enormous amount of information created on the Internet every day [1]. Such RSs are mimicking the natural word of mouth recommendation process, e.g., what item to purchase (Amazon), which movie to watch (Netflix), where to stay during this vacation (TripAdvisor), and whom to send friend request (Facebook). In recent years, RSs also involving supplementary information sources apart from traditional user-item (UI) matrix have become available.

RSs using only star rating are not effective in making accurate recommendations because it is hard to cater to the preferences of two users who gave an item the same rating. For example, user A gave 4 star rating for a phone she purchased because she likes its display and camera quality, user B also gave 4 star rating for the same model of phone because she likes its battery life and processing speed. Therefore, ratings cannot properly capture the views of users. However, reviews are capable of capturing the preferences of users on various aspects

[*]Correspondence: anbazhagan-cse@saranthan.ac.in

of items i.e. how they are finding different features [2]. Moreover, E-commerce users of the present day are very much particular in going through the reviews written by the other users before making a decision to buy an item. Considering the facts discussed, many contemporary RSs use not only ratings given by users but also text reviews authored by users.

Review rating prediction which strives to estimate the numerical rating, given the corresponding text review, has become a popular research area in recent days [1]. There are plenty of techniques available to accomplish this process, e.g., sentiment analysis and topic modeling. Topics present in a review exhibit the thematic structure of that review and play an important role in understanding the user's opinion about different aspects. For example, in a movie RS, a negative sentiment expressed towards the topic "screen play" will have more impact than a negative sentiment expressed towards the topic "running time". If a user expresses a negative sentiment towards "screen play", the review will be more negative than if he/she criticizes the "running time" of the movie. Hence, combining the topics and sentiments is a promising idea for making accurate recommendations.

The goal of the proposed research is to predict ratings using topics and sentiments present in reviews. To achieve the same, a model named sentiment enriched and latent-Dirichlet-allocation-based review rating prediction (SELDAP) is proposed in this paper. The motivation for the idea of combining topics and sentiments is illustrated in Figure 1. First, latent topics available in reviews are taken out. In the meantime, sentiment analysis is performed on reviews to compute their relative sentiment scores with respect to the mined topics. It has been reportedly shown that the present days' reviews contain not only regular texts but also certain new things such as emoticons, acronyms, and slang [3]. Hence, a suitable technique to capture the sentiment in all such newbies is needed. The extracted topics and their sentiments are finally integrated into a supervised machine learning prediction model. This paper has the following two-fold contributions:

1. First, to introduce a new feature named "topic-sentiment" which is nothing but the sentiment specific to topics in order to improve the accuracy of recommendation making process.
2. Secondly, to propose a model named sentiment enriched and latent-Dirichlet-allocation-based Prediction (SELDAP) which uses "topic-sentiment" feature to accurately predict ratings from reviews.

The rest of the sections in this paper are organized as follows: Related works carried out in topic modeling, review rating prediction, and sentiment analysis are presented in Section 2. Section 3 covers the proposed SELDAP model in its entirety. Details on experiments and discussion on results are given in Section 4. Section 5 concludes the present work and provides a direction for future work.

2. Related work

For an E-commerce user, gathering information on what other people think about an item is very important while making a purchase decision. According to a survey of more than 2000 American adults, 81% of Internet users have done online product research at least once, and consumers are interested to pay more for a 5-star rated item than a 4-star rated item [4]. However, there are occasions where a user may give a good star rating for an item due to its overall impression and write a negative review about a particular aspect of that item. Therefore, reviews have become an inevitable source of helpful information. Due to the unprecedented growth of opinion-rich reviews and blogs, users face the following challenges: information overloading, lack of information, difficulty in finding the appropriate information, and so on. As a result, there is a need of assisting users by building a high-quality information-filtering system which serves them what they really need [5].

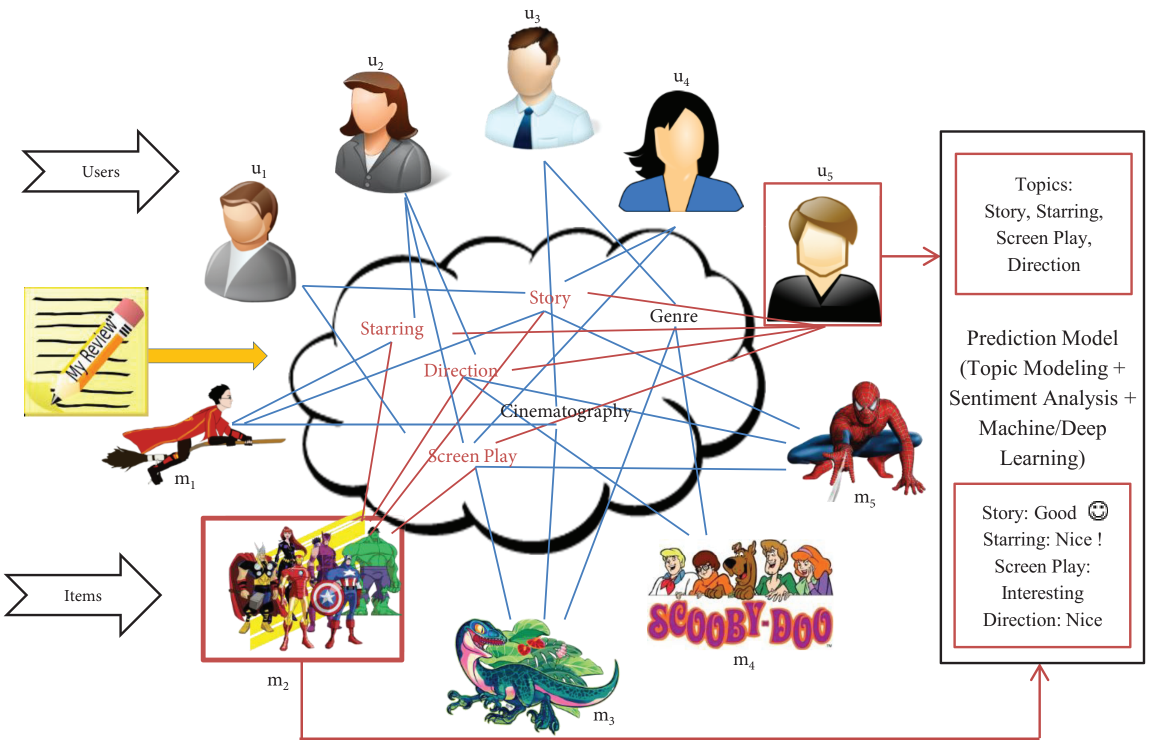


Figure 1. Motivation for the proposed approach.

A generative probabilistic topic modelling technique named latent Dirilecht allocation (LDA) which automatically discovers the hidden topics from discrete document collection was proposed by Blei et al. in [6]. LDA represents the document to be modelled as a finite mixture of certain topics which reflect the thematic structure of that document. The LDA treats the probability distribution of documents over topics as a K-parameter hidden random variable instead of treating the same as a large set of individual parameters. The task of inferring numerical rating from free text reviews is done using LDA and opinion lexicon as baseline sentiment analysis model in [3] and [6]. The experimental results showed that the trained model in combination with sentiment inputs makes accurate predictions of ratings corresponding to reviews.

A new approach to combine three factors such as the user's own social sentiment which reflects his/her preference, interpersonal sentiment which creates an influence among like-minded users, and reputation of the item which reflects user's evaluation was proposed in [7] for improving the prediction accuracy. A probabilistic matrix factorization-based RS model which fully explores user reviews was proposed in [8]. User preferences, social sentiment of the user, and interpersonal influence from reviews are fused into the proposed model to make accurate predictions. Feature extraction methods such as unigram, bigram, trigram, and latent semantic indexing are combined with machine learning algorithms such as logistic regression, naïve Bayes, perceptron, and linear support vector classification to come up with 16 different learning models in [9]. Logistic regression with features extracted using unigram and bigram methods outperformed other combinations.

Features of a product for which users have expressed their opinions are mined from reviews given by those users and are summarized in [10]. This task was performed in three different steps: (i) mining the features quoted in reviews, (ii) detecting the opinions (positive/negative) expressed towards those features, and (iii) summarizing the results. An online LDA algorithm is run in [11] to discover the latent subtopics from free text reviews. It aims to highlight what users really need from a large amount of high dimensional user reviews; the earned topics gave useful insights to users. Opinions are extracted from movie reviews and are used as features for RS to improve the prediction accuracy in [12]. Three methods are used for opinion extraction namely (i) manual clustering, (ii) semiautomatic clustering, and (iii) fully automatic clustering.

A statistical model that combines latent rating dimensions obtained from user ratings using latent factor RS and latent topics extracted from user reviews using LDA is proposed in [13] with an aim to improve the prediction accuracy. The proposed model possesses advantages such as allowing justifying ratings with review texts, reducing the effect of cold start problem, and also helping identify useful reviews. A sentiment analyzer that extracts sentiments about a specific topic from online reviews is presented in [14]. Rather than classifying sentiments of entire document, it identifies all references to the given topic at first and determines sentiments in each of the identified references using natural language processing. An unsupervised model that detects document level sentiment and also extracts mixture of topics from text simultaneously is proposed in [15]. The proposed joint sentiment/topic model is found to be more flexible and to give competitive performance in document level sentiment analysis.

Machine learning techniques such as naïve Bayes, maximum entropy classification, and support vector machines are applied to document sentiment classification in [16] to investigate their effectiveness. It is observed that machine learning techniques outperformed human produced baselines, but they performed poorly for traditional topic-based categorization. According to [16,17] most of the existing sentiment analysis techniques are done at sentence/document level to capture the overall sentiment expressed towards an item. However, understanding the topic/aspect level sentiment is really useful to know the user opinion about a particular feature of the item. The sentiment score of a particular aspect is defined by the weighted sum of sentiment scores of all words in the sentence. Although many such techniques to capture topic level sentiments have been introduced, they do not capture topics and sentiments present in reviews concurrently to a large extent. To overcome this shortcoming, the proposed SELDAP model tries to integrate both topics and their sentiments into a supervised machine learning prediction model.

A novel approach to exploit information present in short-text reviews for word vector representation-based recommendations was proposed in [18]. Two word vectors namely item-vector and user-vector to represent items and users, respectively are built at first. The concatenation of these two vectors with numeric ratings will together constitute a training set for the random forest regression model. An accurate learning-based approach that leverages social data for addressing the cold start problem is proposed in [19]. Linear regression, low-rank parametrization, and randomized SVD are the three key ingredients of this approach. An expected risk minimized matrix approximation method was proposed in [20] with an aim to achieve better trade-off between optimization error and generalization error to reduce the expected risk of approximation models. A novel matrix factorization method with neural network architecture is proposed in [21]. A user-item matrix with explicit rating and nonpreference implicit feedback is constructed at first. The input matrix is then fed to a deep learning architecture for better optimization.

Almost all these researches used only the sentiment analysis techniques which require preprocessing of reviews to remove punctuations, convert uppercase to lowercase and so on before computing the sentiment polarity or sentiment score. As a result, some valid things which have key role in reviews such as emoticons, abbreviations, upper-case letters to emphasize the positivity of a feature, and expressive symbols such as '!' and '?' are removed from the reviews. A rule-based model named valence aware dictionary and sentiment reasoner (VADER) for sentiment analysis was proposed in [3]. A list of lexical features along with their sentiment intensity measures are constructed using the combination of both quantitative and qualitative methods. All these features are constructed considering 5-rules concerning syntactical and grammatical conventions. In order to overcome the aforesaid problem because of ignoring modern review contents, the proposed SELDAP model incorporates VADER sentiment analyzer to compute the relative compound sentiment score of a review with respect to the topics pulled out.

3. The proposed model

The proposed model aims to leverage the useful insights present in reviews in order to improve the efficiency of recommendation making process. To accomplish this, the latent topics which indicate item features and topic-level sentiments i.e. topic-specific sentiments present in the review are extracted using the proposed SELDAP model. The distributions of the topics mentioned above are used alone as well as together with sentiment score to investigate the importance of sentiments in review rating prediction performance. The framework of the proposed SELDAP model is presented in Figure 2. First, the input dataset that has reviews and their associated ratings is collected and made to be in a form which is suitable for further proceedings. Inside the SELDAP model, the topic distribution of the reviews and sentiment scores specific to the latent topics extracted are computed. These two features along with the corresponding star ratings are then used to train a supervised machine learning regression model. Given a review with modern contents, the trained model is now ready to predict the rating corresponding to the input review.

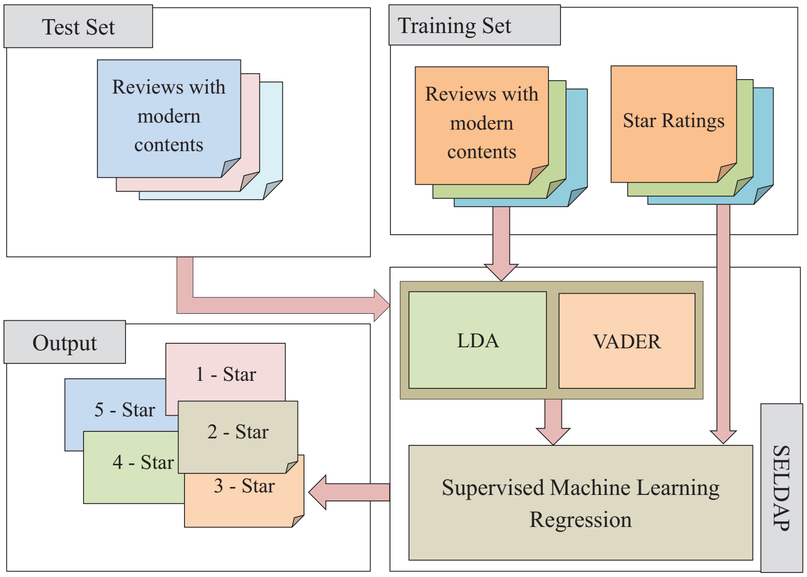

Test Set
Reviews with modern contents
Training Set
Reviews with modern contents
Star Ratings
LDA
VADER
Output
1 - Star
5 - Star
2 - Star
4 - Star
3 - Star
Supervised Machine Learning Regression
SELDAP

Figure 2. The Proposed Framework.

3.1. LDA-based topic modeling

The proposed model uses LDA-based topic modeling to mine the topics/aspects from reviews. LDA is an unsupervised generative probabilistic model for automatically identifying topics present in a text corpus [6]. A repeating pattern of cooccurring words in a text corpus is referred to as a topic. Each document $d \in D$ is associated with a K-dimensional topic distribution θ_d by LDA. The topic distribution θ_{dk} is a probability that the words in document d talks about topic k. Like document d, each topic k is having a word distribution φ_k which is the probability that a particular word is used for the topic k. The word distribution φ_k, topic distribution θ_d, and topic assignment z_{dn} are included in the final model.

The generative process of LDA generates documents given a set of topics. However, we already have a corpus of reviews (documents) and we want LDA to learn the distribution of K topics in each review and the distribution of N words in each topic, then LDA backtracks from documents to identify the topics that would have possibly generated the corpus of reviews. The corpus D is represented as a document-word (DW) matrix which has in its rows the corpus of M reviews $d_1, d_2, \ldots d_M$ and the vocabulary of N words $w_1, w_2, \ldots w_N$ in columns. $DW[i,j]$ gives the occurrence of word w_i in review d_i. Being a matrix factorization technique, LDA transforms this Document-Word matrix into two matrices of lower dimension namely Document-Topic (DT) matrix and topic-word matrix (TW) [22]. Gibbs sampling [23] is used in the proposed model to infer the parameters that probably generated the reviews in the corpus.

$$P(z_i = j|z_{-i}, w_i, d_i, \cdot) \infty \frac{TW_{w_i j} + \beta}{\sum_{w=1}^{W} TW_{wj} + W_\beta} \frac{DT_{w_i j} + \alpha}{\sum_{t=1}^{T} DT_{d_i t} + T_\alpha} \tag{1}$$

where z_i is the topic assigned to the i^{th} token in the whole corpus of reviews, w_i and d_i are the word type of i^{th} token and document containing the i^{th} token, respectively, and z_{-i} is the topic assignments for the remaining topics. Moreover, '$\cdot$' represents the hyper parameters also known as Dirichlet prior such as α and β. Finally, TW and DT stand for topic-word and document-topic matrix, respectively. Using these two matrices as in equation 1, we can infer the posterior estimates of θ and β as follows in Eqs. 2 and 3.

$$\theta_{dj} = \frac{DT_{dj} + \alpha}{\sum_{k=1}^{T} DT_{dk} + T_\alpha} \tag{2}$$

$$\varphi_{ij} = \frac{TW_{ij} + \beta}{\sum_{k=1}^{W} TW_{kj} + W_\beta} \tag{3}$$

where θ_{dj} and φ_{ij} indicate the distribution of topic j in review d and the distribution of word type i for review j. The graphical plate notation of the proposed SELDAP model is shown in Figure 3, where the outer rectangle indicates the corpus of D reviews and inner rectangle indicates the repeated choice of words (w) and topics (z) within a review. The latent and observed variables are represented by empty and shaded circles, respectively. Arrows are for representing the dependencies between latent and Dirichlet parameters.

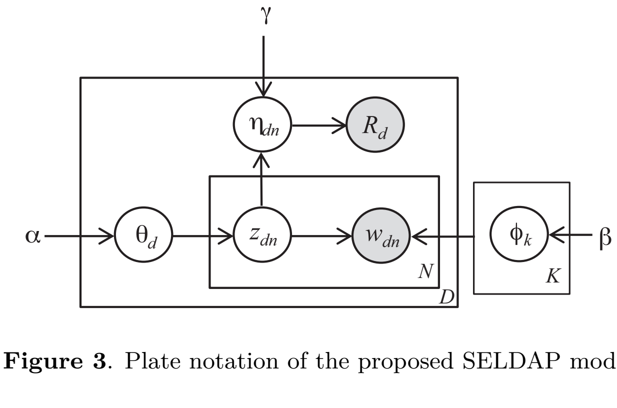


Figure 3. Plate notation of the proposed SELDAP model.

The key terms used in Figure 3 and their descriptions are presented in Table 1. The Dirichlet prior γ refers to the hyper parameters required for sentiment analysis.

3.2. VADER sentiment analysis

Understanding the sentiments of people through the review text authored by them has been a well-researched area in recent years. The motivation behind this process is to know people's views about a specific product. Nowadays, reviews are not only made up of regular texts but also excessive amount of modern contents including emoticons (☺, ☹ etc.), acronyms ("LOL", "OMG" etc.), and slangs ("Nah", "meh", "giggly" etc.). The practice of using these modern content elements makes sentiment analysis a problematic one. Similarly, use of multiple sentiments in a single text and use of figurative language are really tough to comprehend for computers. Examples of the uses of modern content elements in Amazon product dataset are presented in Table 2 (highlighted).

Table 1. Key terms.

Term	Description
D	Review corpus
M	Number of reviews
K	Number of topics
N	Number of words in a review
α, β, γ	Dirichlet priors
θ_d	Topic distribution in review d
φ_k	Word distribution for topic k
z_{dn}	Topic assignment for n^{th} word in review d
w_{dn}	n^{th} word in review d
η_{dn}	Sentiment score for the topic assigned to n^{th} word in review d
R_d	Predicted rating for the review d

Table 2. Amazon sample reviews containing modern elements.

Dataset	Sample Reviews Containing Modern Elements
AIV	OMG! What a twist! I can't wait until the next episode!
ELEC	Saw it and the price was right to experiment..... Nahhhh, won't do that again. Makes the phone look like an afterthought. Just did not look right.
VG	It came on time in great shape!! Looks just like a new game! Looks better than I thought it would!! LOL! I cannot wait to play it!
CPA	These broke when I tried to place them around my ear. They are too weak with thin plastic and too small for the phone. Sorry but I have to say the truth. :(

3.3. Algorithm

This proposed model treats the review rating prediction problem as a supervised regression problem. Topics associated with review texts are extracted by applying LDA-based topic modeling. Due to the fact that the topics extracted from reviews do not have any associated sentiments, sentiment analysis of reviews is also done by using VADER to compute the relative sentiment scores with respect to the topics. The computed sentiment score and distribution of the topics present in reviews are then composed into a training set for the regression-based prediction model. Gradient boosting regression technique is used to build the prediction model in the proposed SELDAP model.

Algorithm: Sentiment-enriched and LDA-based review-rating prediction.

```
1. FOR review in review_corpus DO
     • processed_review = PRE-PROCESSING(review)
     • Put the processed_review, review and associated rating in one of the five groups based on rating
     • Name the groups with labels
       {1_star, 2_star, 3_star, 4_star, 5_star}
   END FOR
2. FOR i in {1, 2, 3, 4, 5} DO
     • i_star_train, i_star_test = Split(i_star) // 80:20 split
   END FOR
3. $all\_train\_set = all\_train\_set \cup i\_star\_train, \forall 1 \le i \le 5$
4. $all\_test\_set = all\_test\_set \cup i\_star\_test, \forall 1 \le i \le 5$
5. FOR processed_review in all_train_set DO
     • Build a dictionary out of the processed_review from previous step
     • Convert the processed_review into bag of word representation $(token\_id, token\_count)$
   END FOR
6. Train an LDA model using the dictionary built and bag of word representation created in step 5 to find
   the topic distribution of each review.
7. FOR review in all_train_set DO
     • Sentiment_Score = VADER(review)
     • Map the Sentiment_Score with latent topics extracted using LDA
     • Store the Sentiment_Score alongside the topic distribution
   END FOR
8. Train the regression-based review rating prediction model using the latent topics and Sentiment_Score
   in all_train_set
9. FOR review in all_test_set DO
     • Apply step 5 & 6 to extract the hidden topics from all_test_set
     • Apply step 7 Calculate the Sentiment_Score present in all_test_set
     • Input the topics and Sentiment_Score to the model trained in step 8 to predict the rating
   END FOR
```

4. Experiments

Experiments are conducted to investigate the performance of the proposed SELDAP model. Three different machine learning regression techniques namely gradient boost, decision tree, and random forest are taken into account for training the review rating prediction model. An empirical review of the relative performances of these regression techniques while working with the proposed model is conducted in order to choose one technique that is most suitable for the proposed SELDAP model.

4.1. Dataset

The experiments are done with the real-world Amazon product dataset [13]. This dataset consists of user ID, product ID, ratings, reviews, review summary, timestamp, and helpfulness for many product categories. Only the subset of the dataset in which all users and items have at least five reviews is chosen for the experimental purpose, as the original dataset is too large in size. As a matter of fact, products/items available on E-commerce websites can be classified into two types, namely (i) search products and (ii) experience products. Search products are easy to evaluate, as feedbacks for this kind of products can be easily collected from the prior users. However, experience products are difficult to evaluate before using them, as they heavily depend on the tastes of different users. Keeping these two types in mind, the experiment done here pertains to two experience products (Amazon instant videos & video games), and two search products (electronics & cell phones and accessories). The summary of the dataset is presented in Table 3.

Table 3. Details of the dataset.

Product category	Length of the longest review	Total number of reviews	Average review length	5-star ratings	4-star ratings	3-star ratings	2-star ratings	1-star ratings
Amazon instant video	18152	37126	515	20890	8446	4187	1885	1718
Electronics	32703	1689188	552	1009026	347041	142257	82139	108725
Video games	32689	231780	1134	120185	54804	28275	13663	14853
Cell phone & accessories	32110	194439	492	108664	39993	21439	11064	13279

4.2. Evaluation metric

Understanding the context is important before choosing an appropriate metric because each machine learning model is trying to solve a problem with a different objective using a different dataset. Root mean square error (RMSE) and mean absolute error (MAE) are the two familiar metrics when the underlying output variable is of type continuous. If a comparison in terms of interpretation point of view is required to be done then MAE is a suitable metric. RMSE penalizes higher errors more than MAE. However, the loss function defined in terms of RMSE is smoothly differentiable and hence RMSE is suitable for many models. MAE which is nothing but the average of the absolute difference between predicted values and observed values is presented in Eq. 4.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|R_{ui} - R'_{ui}\right| \tag{4}$$

RMSE which is nothing but the sample standard deviation between predicted values and observed values is presented in Eq. 5.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{ui} - R'_{ui})^2} \tag{5}$$

In both Eqs. 4 and 5, R_{ui} is the actual rating given by user u for item i, R_{ui} is the predicted rating by the underlying model and n stands for number of observations in the test set. Considering the strengths and weakness of using of MAE and RMSE, both quality metrics are used in the proposed model.

4.3. Baselines

To show the effectiveness of the proposed SELAP model, it is compared to some of the latest baseline models. The baseline models compared in the experiments are presented below in Table 4.

Table 4. State-of-the-art models.

Sl. No.	Model name	Description
1	HFT	Hidden factor and topic model which combine latent rating information and latent review topics [13].
2	ConMF	Convolutional Matrix factorization which uses convolutional neural networks to capture the contextual review information[24].
3	ACF	Advanced collaborative filtering which uses only latent factors obtained from rating information for predicting rating for user u and item i [25].
4	LoCo	Low-rank linear cold-start recommendation system - a special case of linear content based model with different constraints on the learned weights [19].
5	ERMMA	Expected risk minimization for matrix-approximation-based recommender systems which is designed to achieve a better tradeoff between optimization error and generalization error [20].
6	DMF	Deep matrix factorization models for recommender systems - a novel matrix factorization approach with neural network architecture which uses both explicit ratings and implicit feedback [21].
7	R3	Representation of recommendation with short review: an approach to gear up the information available in short reviews by constructing the word vector representations of users and items [18].
8	LDA only	The proposed model without using sentiment score.

4.4. Results and discussions

4.4.1. Choosing the optimal topic size and regression technique

This section presents the quantitative experimental results of the proposed work. First of all, it has been decided to carry out an empirical analysis to choose an optimal value for the topic size (K). Hence, the proposed model is trained with different topic sizes (K=10/20/30/40/50/60). Moreover, for each topic size, the model is trained three times using different regression techniques (gradient boost, a representative for boosting /decision tree, a representative for single regression/random forest, a representative for bagging) to select one best technique which best predicts ratings with the underlying dataset. Each of the eighteen resultant combinations (3 regression techniques × 6 topic sizes) is trained with the aforementioned four product categories.

The resulting RMSEs of all these eighteen models are recorded in Table 5, where the lowest test set RMSE value under all topic sizes (column-wise minimum) in all four product categories are highlighted in bold fashion and also the lowest test set RMSE under all three regression techniques (row-wise minimum) corresponding to four product categories are highlighted with gray background. From this table, column-wise minimum RMSEs are observed twelve times under the topic size $K = 40$, two times under the topic size $K = 30$, and only one time under the topic size K=10. Likewise, row-wise minimum RMSEs are observed five times when gradient boost regression is the regression technique and three times when decision tree is the regression technique. A notable observation with respect to row-wise minimum RMSE value is that for the product category AIV, the same RMSE value of 0.94 is produced by both gradient boosting regression and decision tree regression. This might be due to the nature of the instances available in the product category AIV.

Table 5. Experimental results - RMSE.

Regression technique	Dataset	K = 10	K = 20	K = 30	K = 40	K = 50	K = 60
Gradient boost regression	CPA	1.04	1.03	1.02	**1.01**	1.02	1.02
	AIV	0.95	0.95	**0.94**	**0.94**	0.95	0.96
	ELEC	0.92	0.91	0.91	**0.90**	0.91	0.93
	VG	1.03	1.03	1.03	**1.01**	1.03	1.04
Decision tree	CPA	1.05	1.06	1.05	**1.04**	1.06	1.07
	AIV	**0.94**	0.95	**0.94**	**0.94**	0.96	0.97
	ELEC	0.93	0.93	0.92	**0.91**	0.93	0.93
	VG	1.04	1.05	1.04	**1.03**	1.05	1.06
Random forest	CPA	1.10	1.10	1.08	**1.07**	1.08	1.09
	AIV	1.01	1.00	0.98	**0.97**	0.98	1.00
	ELEC	0.97	0.95	0.95	**0.93**	0.95	0.96
	VG	1.10	1.08	1.08	**1.05**	1.07	1.09

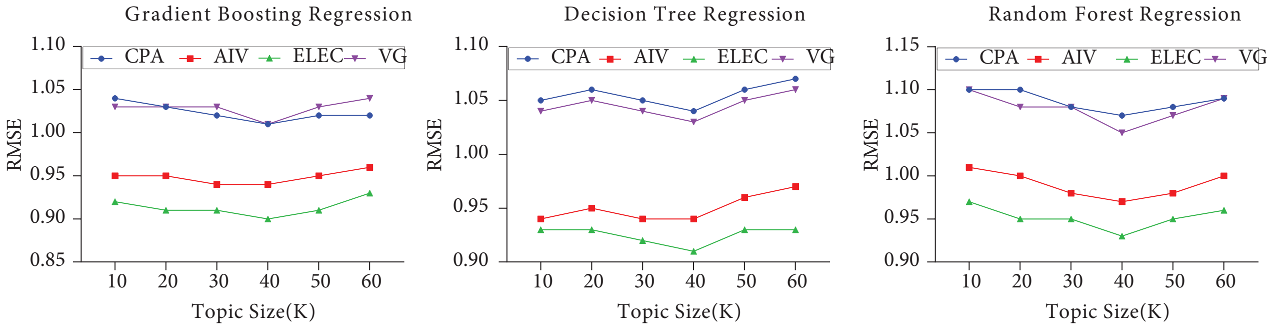
The resulting MAEs of all the eighteen models mentioned earlier are recorded in Table 6, where the lowest test set MAE value under all topic sizes (column-wise minimum) in all four product categories are highlighted in bold fashion and also the lowest test set MAE under all three regression techniques (row-wise minimum) corresponding to four product categories are highlighted with gray background. From this table, column-wise minimum MAEs are observed ten times under the topic size $K = 40$, two times under the topic size $K = 50$. Similarly, row-wise minimum MAEs corresponding to four product categories are observed only when Gradient Boost Regression is used in the underlying prediction model.

Figures 4 and 5 illustrate the impact of different topic sizes on prediction errors such as RMSE and MAE, respectively. It can be observed from Figures 4 and 5 that RMSE and MAE values get decreased as the number of topics are increased from 10 to 40. After the number of topics becomes 50, both RMSE and MAE have gradually started increasing with all three regression techniques and four product categories. It is also observed that most of the time the prediction model with different regression techniques makes prediction with less prediction error when the topic size (K) is 40. Hence, it is decided to choose 40 as the optimal topic size for accurate predictions with less error.

Now, to decide on a suitable machine learning regression technique which best performs review rating prediction, the performance in terms of RMSE and MAE of all three regression techniques with all four product

Table 6. Experimental results - MAE.

Regression Technique	Data Set	K = 10	K = 20	K = 30	K = 40	K = 50	K = 60
Gradient boost regression	CPA	0.76	0.75	0.75	**0.73**	0.75	0.75
	AIV	0.70	0.70	0.70	**0.69**	0.70	0.71
	ELEC	0.61	0.60	0.60	**0.59**	0.61	0.61
	VG	0.75	0.74	0.75	0.74	**0.73**	0.75
Decision tree	CPA	0.80	0.80	0.80	**0.79**	0.81	0.82
	AIV	0.72	0.72	0.72	**0.71**	0.73	0.73
	ELEC	0.65	0.65	0.64	**0.63**	0.65	0.67
	VG	0.79	0.80	0.79	**0.77**	0.78	0.79
Random forest	CPA	0.81	0.80	0.79	**0.78**	0.79	0.80
	AIV	0.75	0.73	0.72	**0.71**	0.72	0.73
	ELEC	0.66	0.64	0.63	**0.62**	0.63	0.65
	VG	0.80	0.79	0.79	0.78	**0.76**	0.78

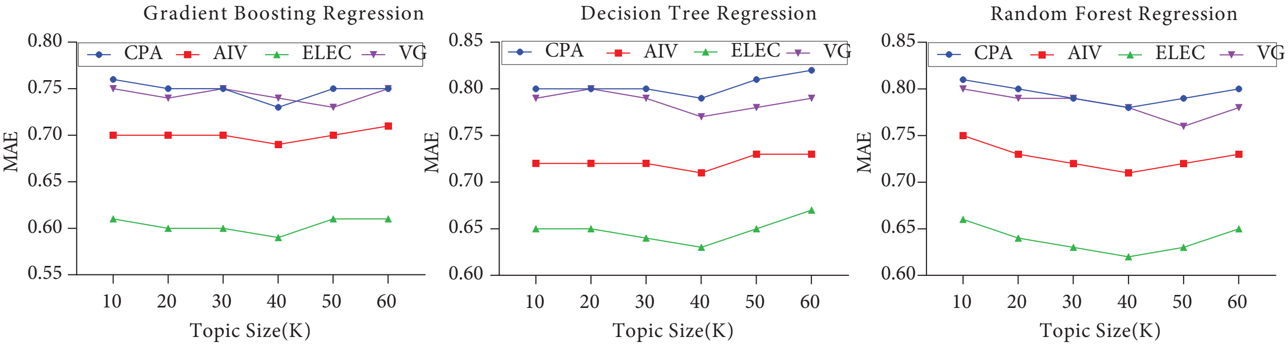


(a) GBR

(b) DTR

(c) RFR

Figure 4. RMSEs for different topic sizes (K).



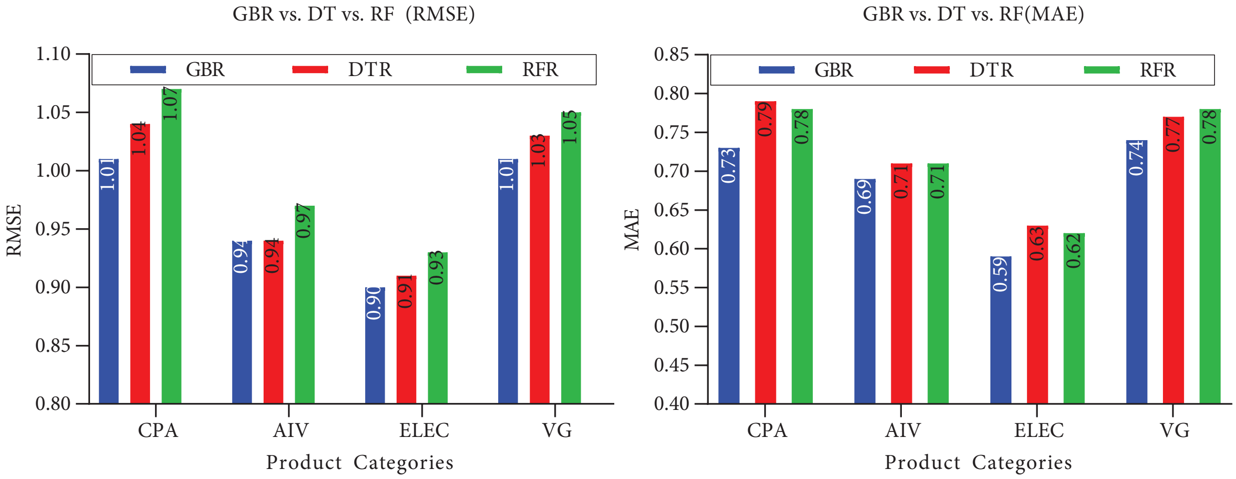
(a) GBR

(b) DTR

(c) RFR

Figure 5. MAEs for different topic sizes (K).

categories when the topic size k = 40 is investigated in Figures 6a and 6b. From Figures 6a and 6b, it is perceived that gradient boost regression technique exhibits relatively good prediction performance with all four product categories compared to the other techniques. Therefore, gradient boosting is chosen as the suitable regression technique to train final prediction model. A special observation from Figure 6a is that for Amazon Instant Video dataset, both gradient boosting and decision tree regression are resulting in the same value of RMSE. Despite this, gradient boosting is chosen as it performs well with all product categories most of the time.



(a) with respect to RMSE

(b) with respect to MAE

Figure 6. Comparison of regression techniques when $K = 40$.

4.4.2. Comparison with baseline models

After finalizing the optimal topic size K and regression technique, the proposed SELDAP model which uses topics size $K = 40$ and Gradient Boosting Regression for prediction model is now compared with the baseline models. All these models are evaluated and the lowest test RMSE and MAE are reported in Figures 7 and 8, respectively. To make a more innate evaluation on how well these models are performing on each product category, the RMSEs of these models are plotted in Figure 7. The figure exhibits a fact that the prediction performance of the proposed model is attributed to the use of latent topics and their associated sentiments extracted from reviews.

As a matter of fact, in all eight baseline models, ACF is the only model that uses latent rating dimensions alone for making predictions. All the other models are using either the latent review dimension alone or in combination with rating dimension. The prediction performance of HFT and ConMF are almost identical in all product categories. Moreover, LoCo, ERMMA, and DMF models are producing higher RMSEs with CPA product category, but for the other three product categories their RMSEs are not that much higher. This is due to the reason that reviews in CPA product category contain heavy use of modern content elements which when removed during preprocessing step higher prediction errors will occur (RMSE). To the best of our knowledge, the proposed SELDAP model is the only model that takes into account ratings, latent topics and topic-specific sentiments obtained from modern review contents to train the prediction model so as to improve the prediction performance.

Similarly, the comparison of the SELDAP model with baseline models in terms of MAE is illustrated in Figure 8. It is inferred from Figure 8 that once again the best performing model over the baseline models is

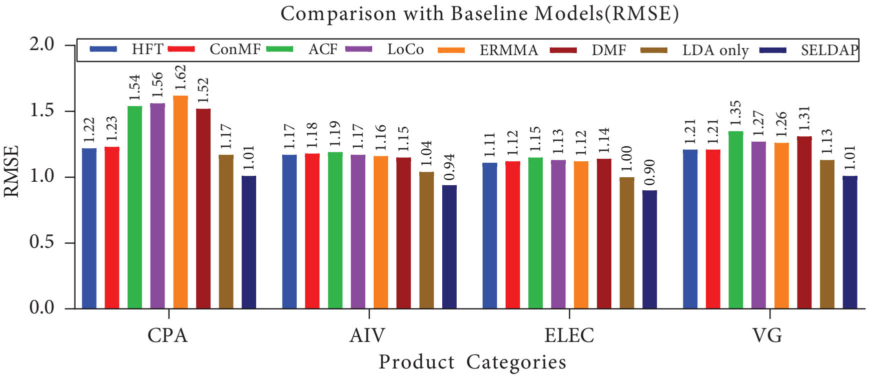


Figure 7. Comparison of SELDAP with Baseline Models in terms of RMSE

SELDAP model. From Figures 7 and 8, it could be understood that the RMSE and MAE values of LDA only model and R3 are almost close to each other and little less compared to the baseline models except SELDAP. The reason for this is that the review information without sentiments cannot accurately capture the user preference and the topic-specific sentiments and latent factor is preferable a preferable combo to reduce the prediction errors. There are models such as HFT that uses both reviews and ratings, but their prediction performance is not the same as SELDAP. This is because the SELDAP model tries to comprehend the importance of each and everything used in the review. The ability of the SELDAP model to process and understand the modern review contents is attributed to the VADER sentiment analysis tool.

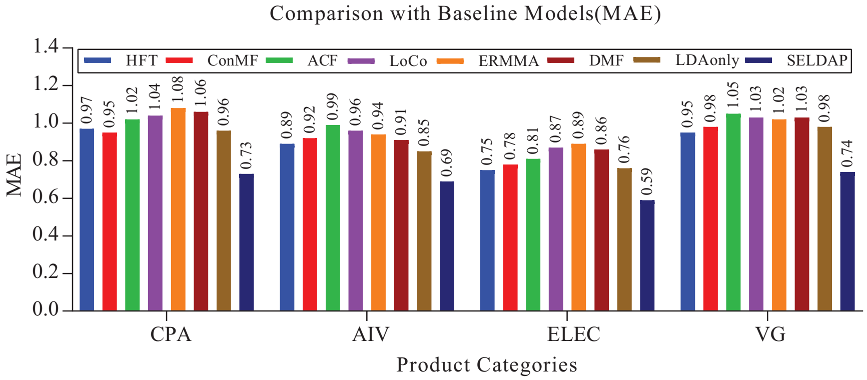


Figure 8. Comparison of SELDAP with Baseline Models in terms of MAE

To give a more clear comparison, the percentage of improvement in the predictiom performance/accuracy of the proposed model over the models compared is presented in Table 7, where ACF an example for rating only model uses only rating information as input and ignores the reviews and models such as HFT and R3 uses both rating and review information but ignores the sentiments present in the reviews. HFT, R3, and SELDAP are examples for review and rating model. Moreover, models namely LoCo, ERMMA, ConMF, and DMF use only review information but not ratings; they are examples for review-only models. The comparison also involves the LDA-only model which means the proposed model with topic distribution and without sentiment score. The performance improvement over ACF clearly shows a point that the use of review information is essential for making accurate recommendations and improvement over HFT, R3, and LDA-only models shows that sentiments present in review are having a strong impact on the accuracy of review rating prediction.

Table 7. Improvement achieved by the SELDAP model over baseline models w.r.t RMSE and MAE.

	RMSE				MAE			
% of improvement over	CPA	AIV	ELEC	VG	CPA	AIV	ELEC	VG
HFT	17.4	19.6	19.0	16.6	24.7	22.5	21.3	22.1
ConMF	17.9	20.2	20.0	16.5	23.2	25.0	24.4	24.5
ACF	34.2	20.9	21.8	25.3	28.4	30.3	27.2	29.5
LoCo	35.1	19.9	20.5	20.6	29.8	28.1	32.2	28.2
ERMMA	37.5	19.2	19.8	19.9	32.4	26.6	33.7	27.5
DMF	33.6	18.3	20.7	22.8	31.1	24.2	31.4	28.2
R3	14.4	11.3	11.8	9.0	25.5	20.7	27.2	26.7
LDA-only	14.0	9.2	10.3	11.0	24.0	18.8	22.4	24.5

From Table 7, it can be perceived that the improvement achieved by the SELDAP model in terms of MAE over the baseline models is relatively greater than the improvement with respect to RMSE. This is because RMSE uses to penalize the higher prediction errors, whereas MAE stands for absolute difference between observed and predicted values. The proposed model achieves 25–35% improvement in RMSE and 27–30% improvement in MAE over the rating-only model ACF. Moreover, 9–37% improvement in RMSE and 23–28% improvement in MAE is achieved by the SELDAP model over review-only models. Finally, there is 9–19% improvement in RMSE and 18–25% improvement in MAE achieved by the proposed model over review and rating models. The reason why the proposed SELDAP outperforms other baseline models is because it makes use of both rating and review information and in particular it captures sentiments present in the modern reviews contents for making recommendations.

From the experimental results and comparison of the performance of the proposed SELDAP model with the baseline models, it is clearly understood that the supplementary/side information such as review text is really useful as well as an expressive source of information to improve the accuracy of predictions made by RS. Numerical ratings alone cannot help make accurate predictions, as they capture only the overall impression of the target item. However, reviews are the right medium for online users to express their feature-specific opinions about items. Moreover, ratings are not useful when it comes to experience products such as movies, restaurants, and video games. Users will not prefer to give their feedback in the form of ratings when the target product is an experience product.

5. Conclusion & future research

This paper investigates the possibilities of incorporating user reviews into RS for making accurate recommendations in order keep the customers of E-commerce websites away from information overloading. To do so a review rating prediction model named Sentiment Enriched and LDA based review rating prediction is proposed in this article to capture the usefulness of review text by processing its modern contents. To enhance the accuracy of review rating prediction task, intensity of the sentiments expressed by users in reviews and topics extracted using LDA are integrated into a supervised machine learning prediction model. The experimental results prove that the proposed model outperforms all the baseline models in terms of making accurate recommendations with less prediction error such as RMSE and MAE. The proposed model also considers the sentiments expressed in modern review contents, which if ignored, will not yield accurate recommendations.

For future research, it has been planned to analyze the effect of using other representation models including nonnegative matrix factorization-based topic modeling technique in place of LDA. In addition, it has also been planned to deal with data imbalance in the datasets by introducing proper resampling techniques.

References

[1] Chen L, Chen G, Wang F. Recommender systems based on user reviews: the state-of-the-art. User Modeling and User-Adapted Interaction 2015; 25 (2): 99-154. doi: 10.1007/s11257-015-9155-5

[2] Baccianella S, Esuli A, Sebastiani F. Multi-facet rating of product reviews. In: European Conference on Information Retrieval; Berlin, Heidelberg; 2009. pp. 461-472. doi: 10.1007/978-3-642-00958-7_41

[3] Hutto CJ, Gilbert E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media; Michigan, USA; 2014. pp.1-20.

[4] Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 2005; 1 (6): 734-749. doi: 10.1109/TKDE.2005.99

[5] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval 2008; 2 (1-2): 1-35. doi: 10.1561/1500000011

[6] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003; 3: 993-1022.

[7] Lei X, Qian X, Zhao G. Rating prediction based on social sentiment from textual reviews. IEEE Transactions on Multimedia 2016; 18 (9): 1910-1921. doi: 10.1109/TMM.2016.2575738

[8] Ma X, Lei X, Zhao G, Qian X. Rating prediction by exploring user's preference and sentiment. Multimedia Tools and Applications 2017; 77 (6): 6425-6444. doi: 10.1007/s11042-017-4550-z

[9] Asghar N. Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362; 2016.

[10] Hu M, Liu B. Mining and summarizing customer reviews. In: Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Seattle, WA, USA; 2004. pp. 168-177.

[11] Huang J, Rogers S, Joo E. Improving restaurants by extracting subtopics from yelp reviews. Social Media Expo 2014; 1: 1-5.

[12] Jakob N, Weber SH, Müller MC, Gurevych I. Beyond the stars: exploiting free-text user reviews to improve the accuracy of movie recommendations. In: 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion; Hong Kong, China; 2009. pp. 57-64. doi:10.1145/1651461.1651473.

[13] McAuley J, Leskovec J. Hidden factors and hidden topics: understanding rating dimensions with review text. In: 7th ACM Conference on Recommender Systems; Hong Kong, China; 2013. pp. 165-172.

[14] Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE International Conference on Data Mining; Melbourne, FL, USA; 2003. pp. 427-434. doi: 10.1109/ICDM.2003.1250949

[15] Chenghua L, Yulan H. Joint sentiment/topic model for sentiment analysis. In: ACM Conference on Information and Knowledge Management; Hong Kong, China; 2009. pp. 375-384. doi: 10.1145/1645953.1646003

[16] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing; Philadelphia, PA, USA; 2002. pp. 79-86. doi: 10.3115/1118693.1118704

[17] Liu B, Zhang L. A survey of opinion mining and sentiment analysis. In: Aggarwal, Charu C, Zhai, ChengXiang (editors). Mining Text Data. Boston, MA: Springer, 2012, pp. 415-463. doi: 10.1007/978-1-4614-3223-4_13

[18] Zheng X, He W, Li L. Distributed representations based collaborative filtering with reviews. Applied Intelligence 2019; 49 (7): 2623-2640. doi: 10.1007/s10489-018-01406-z

[19] Sedhain S, Menon AK, Sanner S, Xie L, Braziunas D. Low-rank linear cold-start recommendation from social data. In: Thirty-First AAAI Conference on Artificial Intelligence; San Francisco, California, USA; 2017. pp. 1502-1508.

[20] Li D, Chen C, Lv Q, Shang L, Chu SM et al. ERMMA: Expected risk minimization for matrix approximation-based recommender systems. In: Thirty-First AAAI Conference on Artificial Intelligence; San Francisco, California, USA; 2017. pp. 1403-1409.

[21] Xue HJ, Dai X, Zhang J, Huang S, Chen J. Deep Matrix Factorization Models for Recommender Systems. In: International Joint Conference on Artificial Intelligence; Macao, China; 2017. pp. 3203-3209. doi: 10.24963/ijcai.2017/447

[22] Mnih A, Salakhutdinov RR. Probabilistic matrix factorization. In: NIPS'07 Proceedings of the 20th International Conference on Neural Information Processing Systems; Vancouver, British Columbia, Canada; 2008. pp. 1257-1264.

[23] Griffiths TL, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences of the United States of America; Washington, USA; 2004. pp. 5228-5235. doi: 10.1073/pnas.0307752101

[24] Kim D, Park C, Oh J, Lee S, Yu H. Convolutional matrix factorization for document context-aware recommendation. In: 10th ACM Conference on Recommender Systems; Boston, Massachusetts, USA; 2016. pp. 233-240. doi: 10.1145/2959100.2959165.

[25] Koren Y, Bell R. Advances in collaborative filtering. In: Ricci F, Rokach L (editors). Recommender Systems Handbook. Boston, MA, USA: Springer, 2015, pp. 77-118. doi: 10.1145/1401890.1401944