An Innovative Framework for Supporting Multi-Criteria Ratings and Reviews over Big Textual Data

Emrul Hasan [1], Chen Ding [1], Alfredo Cuzzocrea [2,3,*] and Islam Belmerabet [2]

[1] *Department of Computer Science, Toronto Metropolitan University, Toronto, Canada*
[2] *iDEA LAB, University of Calabria, Rende, Italy*
[3] *Department of Computer Science, University of Paris City, Paris, France*

Abstract

Nowadays, and thanks to the rise of information technology, recommendation systems have become fundamental tools for e-commerce businesses. Users now are able to provide feedback in form of numerical ratings and comments due to these e-commerce platforms. Recommendation systems are adopted in order to recommend new or unseen items to users depending on those ratings and comments collected previously. In the last years, multi-criteria and multi-aspect based recommendation systems have been a strongly studied topic for the research community in the recommendation field. However, these research works are either driven by ratings or by reviews, and not with both. In this project, we investigate the amenity of further enhancing the overall rating prediction accuracy by integrating numerical multi-criteria ratings along with textual reviews (with multiple aspects). In this paper, we propose a Multi-criteria Rating and Review based Recommendation model (MRRRec), and we show that we can improve the performance by incorporating multi-criteria ratings into multi-aspect ratings extracted from textual reviews. We also demonstrate that our proposed model outperforms a number of state-of-the-art models such as ANR, DeepCoNN, and Deep Multi-criteria Recommendation System in terms of MSE, MAE, precision, recall, and F1-score. We display that, compared to these state-of-the-art models, our model attained an average of 19% and 23.0% lower MSE and MAE respectively and 7.0%, 1.0% and 3.8% higher precision, recall, and F1-score respectively. We furthermore demonstrate how our model performs significantly better with Word2Vec word embedding than the GloVe word embedding methods.

Keywords

Recommendation Systems, Rating Prediction Model, Deep Neural Networks, Multi-Criteria Ratings, Implicit Criteria Rating, Explicit Criteria Ratings, Attention Mechanism, Word Embedding

1. Introduction

Based on historical data, recommendation systems provide users with personalized recommendations. These systems are widely classified into: collaborative filtering, content-based filtering, knowledge-based filtering, and hybrid recommendation systems [1, 2]. The most broadly used methods are Collaborative Filtering (CF) techniques [3]. It makes use of the users' past historical data in order to make recommendations. We can further divide these methods into two main categories i.e., user-based collaborative filtering, and item-based collaborative filtering. User-based collaborative filtering techniques make use of user similarities whereas item-based filtering techniques use item similarities in order to provide a recommendation. Among all CF techniques, the most common ones are based on Matrix factorization (MF) [4], where the user-item rating matrix is decomposed into two

* This research has been made in the context of the Excellence Chair in Big Data Management and Analytics at University of Paris City, Paris, France

SEDB 2023: 31st Symposium on Advanced Database Systems, July 02–05, 2023, Galzignano Terme, Padua, Italy

EMAIL: e1hasan@ryerson.ca (E. Hasan); cding@ryerson.ca (C. Ding); alfredo.cuzzocrea@unical.it (A. Cuzzocrea); ibelmerabet.idealab.unical@gmail.com (I. Belmerabet)
ORCID: 0000-0002-7104-6415 (A. Cuzzocrea); 0009-0003-7878-0991 (I. Belmerabet)

smaller matrices. However, MF suffers from data sparsity problem [5] due to it dependency on the user's explicit ratings, whereas in real-world scenarios, the user-item rating matrix is often a sparse matrix. Moreover, the user's fine-grained preferences on various aspects of the item cannot be captured from a single rating or binary interaction with that item.

In order to mitigate the data sparsity problem in user-item rating prediction, numerous approaches have been proposed. Recommendation with higher accuracy, scalability and explainability [7], [9], [10] can be provided when incorporating textual reviews [6], images [7], and contextual information [8] into MF for example. As, lately, aspect based [11]-[13] and multi-criteria rating [14], [15] based recommendation systems achieved favorable performances in alleviating the data sparsity and cold start problems.

TripAdvisor, the tourism management platform, receives feedback from its customers in three different forms: criteria ratings, textual reviews, and overall ratings. We show an example of a hotel review in Figure 1 where the customer feedback is received on three different categories: ratings on individual criteria (5 stars to value, service, and cleanliness), textual comment in the middle, and the overall rating (5 stars) at the top. Criteria ratings are defined as explicit ratings because the user explicitly provided ratings on individual criteria. We can also extract criteria ratings from textual comments. We define aspects as the terms or phrases or topics that users are concerned about in the review. For example, we consider as aspects the: hotel domain, price, location, value, room size, service, etc. Where each aspect consists of a set of one or more aspect terms. For example, the aspect price includes a set of aspect terms: {expensive, cheap, reasonable . . .}. We can also estimate the associated rating to these aspects [12], [16]. These aspects' ratings are not provided by users explicitly, therefore, they are defined as implicit criteria ratings. It should be noted that these estimated ratings may or may not be similar to the explicit criteria ratings.

In order to predict the overall ratings, multi-criteria rating-based recommendation systems use criteria ratings. Although these methods have succeeded in solving the cold start problem, users' fine-grained opinions expressed through textual comments are still ignored by these methods. In the case of aspect based recommendation systems, users' review texts are used to predict the overall ratings. However, similarly, the users' opinions expressed through explicit criteria ratings are also ignored by these methods. Thus, resulting poor accuracy measure in the overall rating prediction.

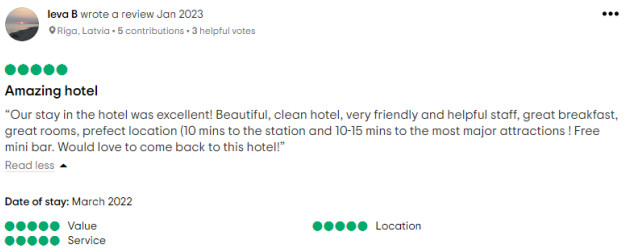


Figure 1: Customer Review on Hotel

We assert that we can improve the performance on the overall rating prediction by integrating implicit criteria ratings into explicit multi-criteria ratings. The publicly accessible TripAdvisor real-world dataset [17], provides users' feedback in form of multi-criteria ratings, textual reviews, and overall ratings. In this paper, we propose an original solution to predict overall ratings based on explicit and implicit multi-criteria ratings. Our model is represented with two parallel paths in order to compute implicit and explicit criteria ratings, and a DNN (Deep Neural Network) at the end to compute the overall ratings.

We describe the main contributions of our work as follows:

1. We have presented an original Multi-criteria Rating and Review based Recommendation model (MRRRec) that computes implicit and explicit criteria ratings in order to predict the overall rating. The implicit criteria rating is computed by the aspect-level user and item representation, and both user and item importance estimation [12]. This paper is to the best of our knowledge the first that makes use of implicit and explicit criteria ratings estimations to compute overall ratings.
2. We have inspected our presented model (MRRRec) with the publicly accessible TripAdvisor dataset and we compared its performance against a number of state-of-the-art baseline methods such

as, DeepCoNN, ANR, and Deep Multi-criteria Recommendation System. We display how our model outperformed all the state-of-the-art methods in terms of MSE, MAE, precision, recall, and F1- score measures.

3. Furthermore, the performance of our model is also compared when using two different word embedding methods i.e., Word2Vec and GloVe word embedding. We also inspected the model with different combinations of hidden layers in the DNN structure in order to attain the best performance.

The rest of the paper is organized as follows. In Section 2, we provide a detailed description of our proposed model. In Section 3 and Section 4 extensive experiments and results are presented respectively. In Section 5, we provide our conclusion and highlight the future works.

2. The Proposed Model - MMRRec

In this Section, we describe our proposed Multi-criteria Review and Rating based Recommendation model (MRRRec). We start by defining the problem statement and describing the components of the model, then we show the related theories. Finally, we demonstrate the effect of different parameters.

2.1. Problem Definition

Considering a corpus C with a set of reviews $V: \{v_1, v_2, v_3, \dots v_w\}$, a set of explicit criteria ratings $E: \{e_1^E, e_2^E, \dots e_w^E\}$ where $X: \{x_1, x_2, x_3, \dots x_{|X|}\}$ is the set of explicit criteria for each review, and a set of overall ratings $R: \{r_1, r_2, r_3, \dots r_n\}$ given by a set of users $S: \{s_1, s_2, s_3, \dots s_{|S|}\}$ to a set of items $M: \{m_1, m_2, m_3, \dots m_{|M|}\}$. In our case, the hotels represent the items. We suppose that each review contains a set of aspects $I: \{i_1, i_2, i_3, \dots i_{|I|}\}$. Therefore, we define the set of implicit criteria ratings as $P: \{p_1^I, p_2^I, p_3^I, \dots p_w^I\}$. It should be noted that X and I may or may not be the same. In Table 1 we summarize the key notations used all over this paper.

Table 1
Definition of Various Notations

Notation	Definition
C	corpus with criteria ratings, textual review and overall ratings.
V	set of reviews
T	set of criteria ratings
w	number of reviews
$\|E\|$	number of explicit criteria ratings for each review
O	a set of overall ratings
S	a set of users
$\|S\|$	number of users
M	a set of items
$\|M\|$	number of items
$\|I\|$	number of implicit criteria for each review
N_S	user documents (set of reviews from users S)
N_M	item documents (set of reviews for items M)
$X_{S,I}$	aspect-level user representation
$Z_{M,I}$	aspect-level item representation
$\alpha_{S,I}$	user aspect importance
$\alpha_{I,K}$	item aspect importance
S_ξ	user id embedding matrix
M_ξ	item id embedding matrix
G_y	weight matrix of hidden layer y
d	hidden layer
Y	number of hidden layers
i_y	bias for hidden layer y
ψ	concatenated user and item ID embedding matrix
F	input features for overall rating prediction DNN

Our primary goal is to predict the overall rating $\hat{O}$ for user-item pairs (S, M). We break up this goal into three sub tasks described as follows:

1. Explicit criteria prediction: we use a deep neural network in order to predict the unknown explicit criteria ratings, E. And we consider user and item IDs as input features for DNN.
2. Implicit criteria prediction: we estimate implicit criteria ratings, P, with aspect-based representation learning for both users and items using an attention-based component. And we use user document N_S and item document N_M as features.
3. Overall rating prediction: we predict the overall ratings, $\hat{O}$, from implicit and explicit criteria ratings using deep neural network.

2.2. Architecture

In Figure 2, we show the detailed architecture of our proposed MRRRec model. Our model is composed of two parallel paths along with a deep neural network at the top. In the first path, on the left side of the network, implicit ratings are predicted based on the reviews. On the other path, explicit criteria ratings are predicted. Finally, the top part of our model is responsible for learning the overall rating using a deep neural network.

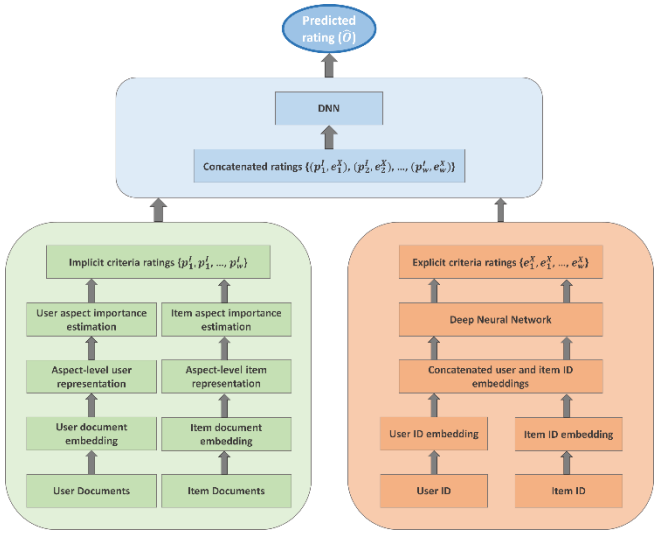


Figure 2: MRRRec Model Architecture

2.3. Implicit Criteria Rating Prediction

We follow the work of Chin et al. [12] in order to predict the implicit criteria ratings. This method utilizes user and item documents as features of the model. User document N_S are represented by a set of reviews written by a set of users S and item document N_M is a set of reviews for a set of items M. First, it transforms item documents into embedding vectors using either pre-trained word embedding GloVe [19] or Word2Vec [20] model. Next, it applies the neural attention mechanism in order to learn the aspect-level user and item representation followed by aspect importance estimation. Considering a set of users, S and a set of items M, a set of implicit criteria ratings R is defined as follows:

$$R = \alpha_{S,I} \cdot \alpha_{M,I} \cdot (X_{S,I}(Z_{M,I})^T) \quad (1)$$

where the set of user aspect importance or a set of users' satisfaction towards a set of I aspects is represented by $\alpha_{S,I}$ and the importance of a set of I aspects for a set of items M is represented by $\alpha_{M,I}$. $X_{S,I}$ and $Z_{M,I}$ represent the aspect-level user and item representations respectively. The paper [12] holds detailed calculations.

2.4. Explicit Criteria Rating Prediction

In this step, we compute a set of explicit criteria ratings T for a set of users S on a set of items M, we implemented this part of our model on basis of the work by Nassar et al. [21] where we use user IDs and item IDs as input features. He et al. [22] suggests that we can represent categorical features such as IDs that do not have logical order as input features. These IDs are embedded and represented as low-dimensional vectors. We initialize embedding vectors with random initialization and adjusted their values during training steps. We concatenate user and item embedding before feeding them to the neural networks. Then, we pass these concatenated embedded feature vectors through a series of hidden layers.

Let the user ID vector S_ξ and item ID vector M_ξ. We concatenate these two vectors as follows:

$$F = Concatenate(S_\xi, M_\xi) \quad (2)$$

where the user ID and item ID embedding matrixes are represented by S_ξ and M_ξ respectively.

The feature vector F represents the input of the DNN where the used activation function is ReLU (Rectified Linear Unit) [23] considering that ReLU is the most efficient activation function.

$$ReLU(f) = max(0, f) \quad (3)$$

We represent mathematically the output of a hidden layer as follows:

$$d_y = ReLU(G_y d_{y-1} + i_y) \quad (4)$$

where weights, bias and hidden layers are represented by G_y , i_y and d_y respectively. We represent the final output of the explicit criteria ratings as follows:

$$T = ReLU(G_Y d_{Y-1} + i_Y) \quad (5)$$

where Y represents the number of layers and G_Y the weight matrix of the final layer Y respectively.

2.5. Overall Rating Prediction

In Figure 3 we show the architecture of the overall rating prediction DNN model. First, we concatenate the explicit criteria ratings T and implicit criteria ratings R. Then we normalize the concatenated ratings due to the DNN's sensitivity to input distribution and scaling [24].

$$F = Concatenate(T, R) \quad (6)$$

where F represents the features for the overall rating prediction DNN. The dimension of the input layer in the DNN is the dimension of F. The output dimension is 1 which is the overall rating. We compute the output of the hidden layers using the Equation (4). Then we predict the overall ratings using the Equation (5) and we define it as follows:

$$\hat{O} = ReLU(G_Y d_{Y-1} + i_Y) \quad (7)$$

where $\hat{O}$ represents the final predicted rating and the input layer is $F = d_0$.

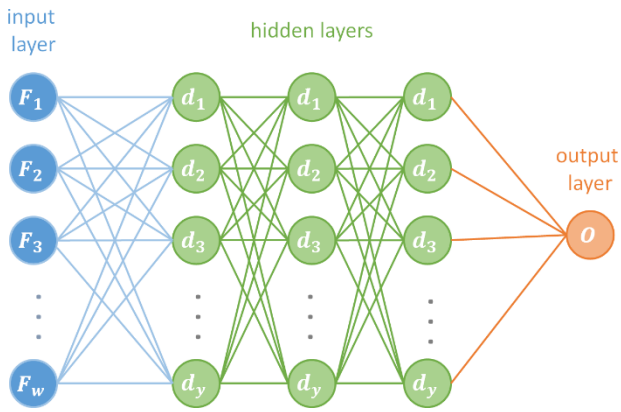
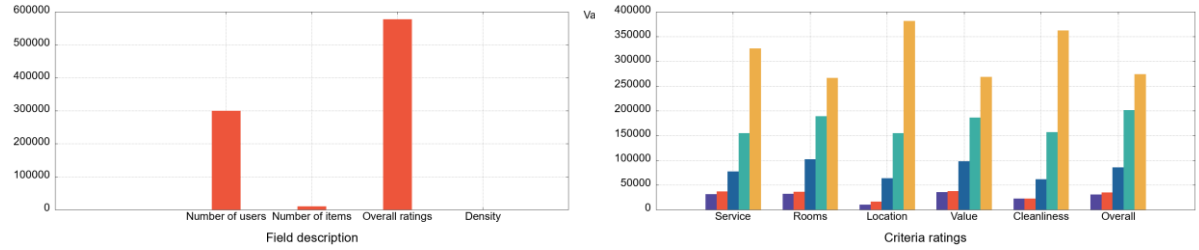


Figure 3: Detailed Architecture of the Overall Rating Prediction DNN

3. Experimental Analysis

We analyze our proposed model's performance with TripAdvisor dataset [17] i.e. a tourism management online platform that offers hotel, flight, restaurant, cruise, and car rental services. This data was collected over 8 years, from 2004 to 2012. This dataset contains multi-criteria ratings, textual reviews, and overall ratings. In which exists, 8 different criteria and associated ratings including value, rooms, location, cleanliness, check in/front desk, service, sleep quality, and business service. Individual criteria ratings and overall ratings are ranging from -1 to 5. Ratings below 1 and criteria with more than 75% missing values are removed.



(*a*) (*b*)

Figure 4: Data Statistics (*a*), and Criteria and Overall Ratings Distribution (*b*)

The remaining data contains 6,260,263 reviews and 5 criteria. Punctuations and next line characters are also removed for data processing purposes. We tokenized all the reviews using the natural language toolkit, NLTK [25] and a vocabulary with most frequent 50,000 words is built. Also, reviews with less than 10 tokens are removed. Each user and item document lengths are truncated to 500 tokens in order to keep consistency with the ANR model [12]. In Figure 4 (*a*), we show detailed statistics about the data, and in Figure 4 (*b*) we show both criteria and overall ratings distribution.

3.1. Baselines

We evaluate the performance of our proposed model MRRRec against the following baselines:

1. ANR [12]: An aspect-based neural recommender where user and item documents are modeled in order to predict ratings. It uses these documents are to learn the aspect based representation, then follows it by using an attention-based component for aspect importance estimation. And finally predicts the ratings.
2. DeepCoNN [22]: Deep Cooperative Neural Network (DeepCoNN) represents one of the state-of-the-art recommendation models where reviews are used for rating prediction. It concatenates user and item representations and utilizes them as inputs to a Factorization Machine (FM) [26] in order to predict overall ratings.
3. Multi-criteria RS by Nassar et al. [12]. A deep learning based multi-criteria recommendation system that predicts overall criteria ratings in two steps. First, it predicts unknown x multi-criteria ratings using a DNN where user and item IDs are used as input features. Then, it uses the DNN model again for learning the relationship between multi-criteria and overall ratings.

3.2. Experimental Settings

We start by randomly splitting the data into training, validation and test sets with a ratio of 80:10:10. Then, we implement the entire model on Google Colab using the configuration, PyTorch: 1.12.1, GPU: NVIDIA Tesla P100, CUDA: 11.2, and RAM: 24 GB. We optimized all the models using Adam optimizer. Next, we evaluate the performances using MSE, MAE, precision, recall and F1-score measures. In the case of ANR, Word2Vec [27] word embedding will be used with a dimension of 300d. We keep all parameters the same as described in this paper. For both ANR and our model MRRRec, we use MSELoss and we train the models for 25 epochs with a batch size of 128.

For DeepCoNN implementation we have used the GloVe word embedding model [19] with a dimension of 100d and we kept the other parameters as specified in the paper. For both ANR and DeepCoNN models, the implementation code is open sourced. However, in the case of deep multi-criteria recommendation system, the code is not open sourced, therefore, we write the code and we set the parameters as specified in the paper. It should be noted that, the model was trained on small datasets crawled from TripAvisor website with the numbers of only 72119, 1850, and 81085 users, items, and ratings respectively. Therefore, we train the model using a larger dataset i.e. shown in Figure 5 having numbers of users, items, and ratings of 299855, 9763, and 577854 respectively.

We use different parameter settings for different components of our model. We describe the set of parameters for optimal performance as follows:

1. Explicit rating prediction: we evaluate the DNN model with different combinations of hidden layers and user and item IDs embedding dimensions. First, the DNN parameters are randomly initialized using a normal distribution with mean of 0 and standard deviation of 0.5. Then, we use for each user and item ID a vector size from 16 to 256. For layer selection, we train the model using six different sets of layers and we obtain the best performance when the combination [256→128→64] is used.
2. Implicit criteria rating prediction: as we use components from ANR to predict implicit criteria ratings, the parameters remain similar. However, we performed experiment using different sets of implicit aspects I while keeping the size of E fixed and changing the size of I, then we train the model with different sizes of I. The size of I is changed because I represents the parameter that determines the number of aspects that will be extracted from the review. Therefore, by changing values of this parameter we are able extract different numbers of aspect ratings.
3. Overall rating prediction: The model's input dimension of the overall rating prediction is defined by the sum of implicit and explicit criteria ratings.

We evaluated the performance using MSE (Mean Squared Error), MAE (Mean Absolute Error), precision, recall and F1-score for both baselines and our proposed model. MSE and MAE are defined mathematically as follows:

$$MSE = \sqrt{\frac{1}{K}\sum_{s,m}^{K}(\hat{O} - O)^2} \tag{8}$$

$$MAE = \frac{1}{K}\sum_{s,m}^{K}|\hat{O} - O| \tag{9}$$

where K represents the total number of ratings, and $\hat{O}$ and O represent the predicted and the known overall rating respectively.

In order to compute the accuracy measure, and based on threshold value of 3.5 we perform a transformation on our predicted ratings to 0 and 1. This transformation is to the goal of deciding whether to recommend an item or not based on the predicted rating. The model makes a recommendation if the rating is higher than 3.5, otherwise it does not. With taking these transformations into consideration, we define this problem as a binary classification problem and we calculate the precision, recall and F1-score measures. Where precision is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

where TP and FP represent the true positive and the false positive respectively. A good classifier scores a precision value of 1. We note that recall is also referred to as measurement of sensitivity or true positive rate, which is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where FN refers to the false negative. We also note that for an ideal classifier, the recall value should be 1.

The F1-score represents the harmonic mean of precision and recall. This measure is the metric where both precision and recall are taken into consideration and it is defined as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{12}$$

We get an F1-score value of 1 when, and only when both precision and recall values are 1.

3.3. Experimental Results

In Figure 5 we show a comparison of performance results between our proposed model MRRRec and the previously described 3 baseline models. Our model outperforms these baselines in all the accuracy measures i.e., MSE, MAE, precision, recall and F1-score, as it scored an $MSE = 0.44043$, $MAE = 0.30011$, $Precision = 0.92477$, $Recall = 0.95104$, and $F1 = 0.93772$. Therefore, our model attained an average of 19% and 23.3% lower MSE and MAE respectively, and 7%, 1.0%, and 3.8% higher precision, recall and F1-score values respectively. Between ANR (aspect-based model) and Nour Nassar et. al. [14] (explicit criteria rating based model), it is noticeable that the multi-criteria rating based model is better in terms of performance. This may be due to the fact that users' opinions on a particular topic are expressed through explicit criteria ratings and that explicitly specified ratings indicate the users' preferences more accurately than the aspect ratings inferred implicitly in textual reviews. Which may be also due to the fact that existing models are unable to estimate the implicit criteria ratings accurately. Compared to the aspect based model ANR, our model attained 7.7%, and 18% lower MSE and MAE, and also 2.6%, 1.7%, and 2.2% higher precision, recall and F1-score values respectively. Similarly, our model MRRRec performed significantly better than DeepCoNN (review-based method) in terms of all the accuracy measures. This shows that adding multi-criteria ratings is important for overall ratings prediction.

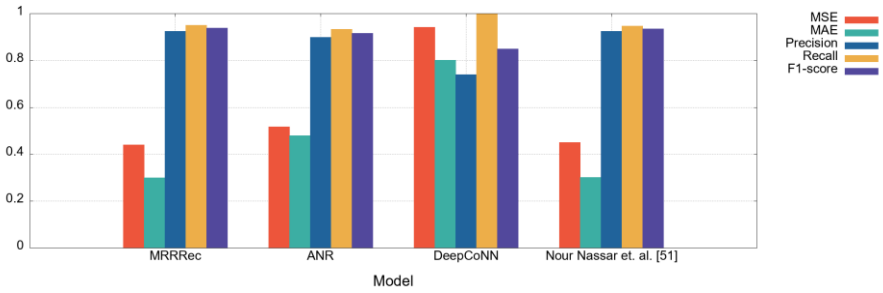


Figure 5: Performance Comparison Against the 3 Baselines

3.4. Ablation Study

In order to analyze the effectiveness of the different components of the model, we perform an ablation study and we show a performance comparison between different network structures in Figure 6.

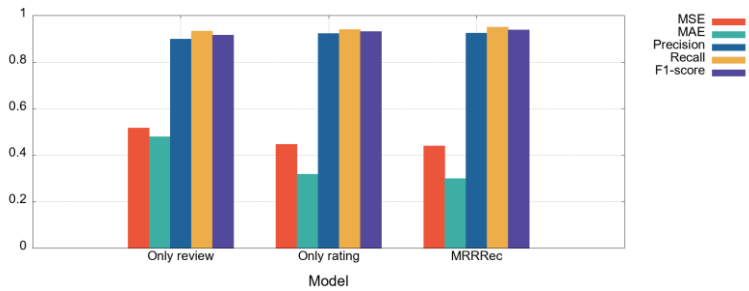


Figure 6: Component Based Performance

1. Only implicit criteria: We obtain higher MSE and MAE scores when we use only review data, than when we use both review and criteria ratings. Which allows this review-only model to score an $MSE = 0.44043$, $MAE = 0.30011$, $Precision = 0.92477$, $Recall = 0.95104$, and $F1 = 0.93772$, therefore, attain 7.7% and 18.0% higher MSE and MAE scores, and 2.7%, 1.6%, and 2.2% lower precision, recall, and F1-score respectively.

2. Only explicit criteria: Similarly to only implicit criteria, the model is unable to extract important hidden aspects from user reviews for rating prediction when only explicit ratings are used. When using only explicit criteria ratings, the model attained 0.6% and 2% higher MSE and MAE scores, and 0.3%, 0.5%, and 1.0% lower precision, recall, and F1-score measures respectively compared to our MRRRec model.

We perform additional experiments using two different word embedding methods Word2Vec and GloVe. Then we report the performance results in Figure 7. Our model performed better with Word2Vec scoring an $MSE = 0.44043$, $MAE = 0.30011$, $Precision = 0.92477$, $Recall = 0.95104$, and $F1 = 0.93772$, therefore, attaining 4.7% and 11% lower MSE and MAE, and 0.1%, 3%, and 1% higher precision, recall, and F1-score values respectively.

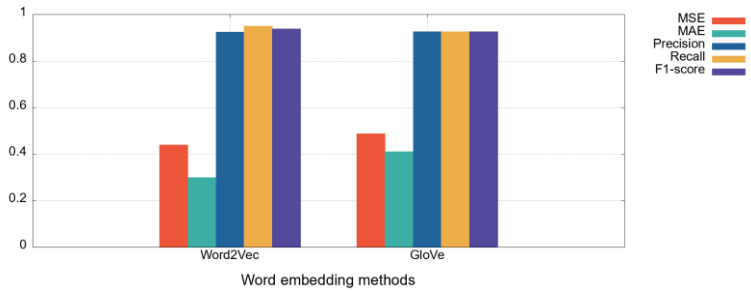


Figure 7: Results of Using Different Word Embedding Methods

3.5. Parameter Sensitivity Analysis

We investigate the effectiveness of using different parameters in the mode. Then, we show the performance results of different layer choices in Figure 8. We can state that making the model complex with too many layers lowers the performance. Similarly, the performance also goes down when the model is too simple and has too few layers. However, the model performs significantly well when we balance it between simplicity and complexity. The best performance is attained when using the hidden layer combination of [256→128 →64], where the model scored an $MSE = 0.44043$,$MAE = 0.30011$,$Precision = 0.92477$,$Recall = 0.95104$, and $F1 = 0.93772$.

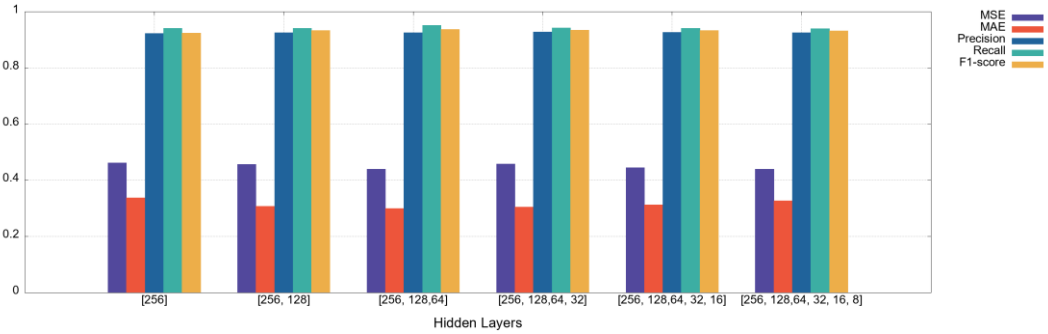


Figure 8: Results of Using Different Hidden Layer Combinations

Similarly, we evaluate the model's performance when using different dimensions of user and item embeddings (explicit criteria rating prediction component), as shown in Figure 9. The model performs poorly when we represent IDs with very low dimensions. On the other hand, the performance improves slowly as we increase the size. The best performance is attained when we use an embedding size of 256 for both user and item embeddings, where the model scored an $MSE = 0.44043$,$MAE = 0.30011$,$Precision = 0.92477$,$Recall = 0.95104$, and $F1 = 0.93772$.

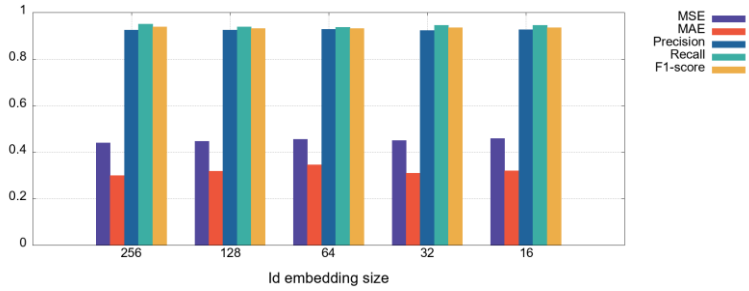


Figure 9: Performance of the Model Using Different Embedding ID Sizes

Furthermore, we inspect the model's performance when using different sizes of implicit criteria. As implicit criteria ratings are estimated from the review and not directly provided by users, we cannot determine how many implicit criteria exist in the review text. Therefore, using fixed explicit criteria, we evaluate the model with different sizes of the set of aspects I. In Figure 10 we report the effect of I, then we show that the higher the value of I is, the higher the value of MSE, and MAE, and the lower the value of precision, recall, and F1-score values are. The best results are attained when using the number $I = 5$ i.e. equal to the number of explicit criteria, where the model scored an $MSE = 0.44043$,$MAE = 0.30011$,$Precision = 0.92477$, $Recall = 0.95104$, and $F1 = 0.93772$.

3.6. Discussion

We discuss in this Section our model's success and failure. Based on the obtained results, it is clear that the model has attained significant improvements in performance when integrating incorporating criteria ratings into explicit criteria ratings. For the overall rating prediction, the existing multi-criteria rating-based recommendation system uses only explicit criteria ratings. On the other hand, the aspect-based recommendation system uses only review for overall rating prediction. The explicit criteria

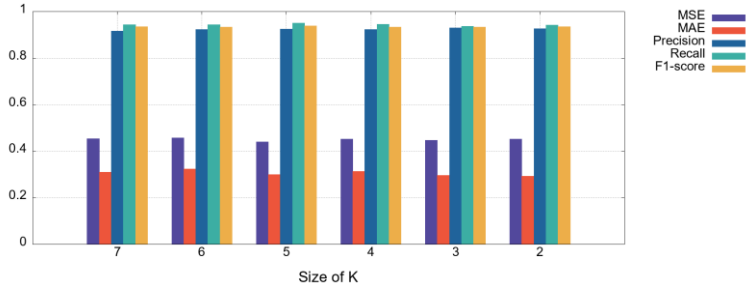


Figure 10: Performance of the Model Using Different Sizes of Implicit Criteria Ratings

rating-based recommendation system is incapable of extracting latent relationships between user-item pairs and the overall ratings. This is due to only explicit criteria ratings being considered, and ignoring the users' latent opinions expressed through textual reviews. Similarly, when considering only reviews, regardless of them being able to extract users' latent preferences from the reviews, they still ignore their explicit preferences on given criteria. With the goal of predicting overall ratings, our model incorporates both implicit and explicit criteria ratings.

Despite the fact that our proposed model attained substantial performance, it is still incapable of extracting explicit criteria ratings from textual reviews. In addition, our current model is unable to determine which implicit criteria ratings are corresponding to which explicit criteria. This is due to the fact that this part of our model (left side of Figure 2) has not been trained with explicit criteria ratings. However, as a future work, we will consider mapping aspect ratings to explicit criteria ratings, which will allow us to predict more accurate overall ratings. Evaluating our model with different datasets can another optimization to make. However, to the best of our knowledge, TripAdvisor is currently the only publicly accessible dataset that offers both of reviews and criteria ratings. Our focus in the future will be on exploring other datasets that provide reviews and criteria ratings, along with creating our own dataset that can potentially serve the research community.

We analyzed our model's performance against the latest state-of-the-art models. However, the latest aspect-based paper does not share the source code [18], and even some shared their codes, a number of files were missing and they didn't provide a proper documentation [16].

4. Conclusions and Future Work

In this work, we have proposed a Multi-criteria Rating and Review based Recommendation System (MRRRec). This model incorporates both explicit and implicit criteria ratings in order to predict overall ratings. It consists of three components, the first component predicts implicit criteria ratings from textual reviews, the second is for unknown explicit criteria rating prediction based on known criteria ratings, and the third for overall rating prediction. In order to predict implicit criteria ratings from review texts, an aspect representation learning and aspect importance estimation is performed. On the other hand, a deep neural network is used for explicit criteria rating prediction where user and item IDs are taken as input features for this network. Finally, to predict overall rating, another deep neural network is used where, the outputs of both implicit and explicit criteria rating models are concatenated and fed as inputs to this final deep neural network. We have shown that our proposed model MRRRec outperforms the three described before state-of-the-art models, namely, ANR, DeepCoNN, and Multi-criteria recommendation system in terms of all MSE, MAE, precision, recall, and F1-score measures. Our model attained an average of 19% and 23% lower MSE and MAE, and 7%, 1%, and 3.8% higher precision, recall, and F1-score respectively.

As for future work, we aim at using state-of-the-art NLP (Natural Language Processing) techniques such as BERT, and NER in order to capture aspects from textual reviews and also for overall rating prediction. On the other hand, integration with *adaptive metaphors* (e.g., [28-30]) is another goal of our future work.

5. References

[1] F. Ricci, L. Rokach, and B. Shapira, "Recommender Systems Handbook", 2nd. ed., *Springer*, 2015, pp. 1–34

[2] K. Falk, "Practical Recommender Systems", *Manning Publications*, 2019

[3] S. K. Raghuwanshi and R. Pateriya, "Collaborative Filtering Techniques in Recommendation Systems", *Data Engineering and Applications 1* (2019) 11–21, Springer

[4] R. B. Yehuda Koren and C. Volinsky, "Matrix Factorization Techniques for Recommendation Systems", *IEEE Computer Society 42.8* (2009) 43–44

[5] M. Grčar, D. Mladenič, B. Fortuna, and M. Grobelnik, "Data Sparsity Issues in the Collaborative Filtering Framework", in: *International Workshop on Knowledge Discovery on the Web*, Springer, 2005, pp. 58–76

[6] W. Zhang and J. Wang, "Integrating Topic and Latent Factors for Scalable Personalized Review-Based Rating Prediction", *IEEE Transactions on Knowledge and Data Engineering 28.11* (2016) 3013–3027

[7] Z. Cheng, X. Chang, L. Zhu, R. C. Kanjirathinkal, and M. Kankanhalli, "MMALFM: Explainable Recommendation by Leveraging Reviews and Images", *ACM Transactions on Information Systems 37.2* (2019) 1–28

[8] Z. Cheng and J. Shen, "Just-For-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation", in: *International Conference on Multimedia Retrieval*, 2014, pp. 185–192

[9] L. Chen, G. Chen, and F. Wang, "Recommender Systems Based on User Reviews: The State Of The Art", *User Modeling and User-Adapted Interaction 25.2* (2015) 99–154

[10] X. He, T. Chen, M.-Y. Kan, and X. Chen, "Trirank: Review-Aware Explainable Recommendation by Modeling Aspects", in: *24th ACM International Conference on Information and Knowledge Management*, 2015, pp. 1661–1670

[11] X. Guan, Z. Cheng, X. He, Y. Zhang, Z. Zhu, Q. Peng, and T.-S. Chua, "Attentive Aspect Modeling for Review-Aware Recommendation", *ACM Transactions on Information Systems 37.3* (2019) 1–27

[12] J. Y. Chin, K. Zhao, S. Joty, and G. Cong, “ANR: Aspect-Based Neural Recommender”, in: *27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 147–156
[13] W. Li and B. Xu, “Aspect-Based Fashion Recommendation with Attention Mechanism”. *IEEE Access. 8* (2018) 141814–141823
[14] N. Nassar, A. Jafar, and Y. Rahhal, “Multi-Criteria Collaborative Filtering Recommender by Fusing Deep Neural Network and Matrix Factorization”, *Journal of Big Data 7.1* (2020) 1–12
[15] Z. Chen, S. Gai, and D. Wang, “Deep Tensor Factorization for Multi-Criteria Recommender Systems”, in: *2019 IEEE International Conference on Big Data (Big Data)* , 2019, pp. 1046–1051
[16] C. Wu, F. Wu, J. Liu, Y. Huang, and X. Xie, “ARP: Aspect-Aware Neural Review Rating Prediction”, in: *28th ACM International Conference on Information and Knowledge Management*, 2019, pp. 2169–2172
[17] TripAdvisor Data. URL: https://notebook.community/melqkiades/yelp/notebooks/TripAdvisor-Datasets
[18] W. Li and B. Xu, “Aspect-Based Recommendation Model for Fashion Merchandising”, in: *Advances in Digital Marketing and eCommerce: Second International Conference*, Springer, 2021, pp. 243–250
[19] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global Vectors for Word Representation”, in: *2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543
[20] K. W. Church, “Word2Vec”, *Natural Language Engineering 23.1* (2017) 155–162
[21] N. Nassar, A. Jafar, and Y. Rahhal, “A Novel Deep Multi-Criteria Collaborative Filtering Model for Recommendation System”, *Knowledge-Based Systems 187* (2020) 104811
[22] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, “Neural Collaborative Filtering”, in: *26th International Conference on World Wide Web*, 2017, pp. 173–182
[23] V. Nair and G. E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines”, in: *27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814
[24] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, in: *International Conference on Machine Learning*. PMLR, 2015, pp. 448–456
[25] E. Loper and S. Bird, “NLTK: The Natural Language Toolkit”, *CoRR cs.CL/0205028*, 2002
[26] S. Rendle, “Factorization Machines”, in: *2010 IEEE International Conference on Data Mining*, IEEE, 2010, pp. 995–1000
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and Their Compositionality”, *Advances in Neural Information Processing Systems 26* (2013) 3111–3119
[28] M. Cannataro, A. Cuzzocrea, and A. Pugliese, “A Probabilistic Approach to Model Adaptive Hypermedia Systems”, in: *WebDyn 2001 International Workshop - ICDT 2001*, 2001, pp. 50–60
[29] M. Cannataro, A. Cuzzocrea, A. Pugliese, “XAHM: an adaptive hypermedia model based on XML”, in: *SEKE 2002 International Conference*, 2002, pp. 627–634
[30] J. Pilault, A. Elhattami, C. J. Pal, “Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data”, in: *ICLR 2021 International Conference*, 2021