EMBEDDED LEARNING FOR LEVERAGING MULTI-ASPECT IN RATING PREDICTION OF PERSONALIZED RECOMMENDATION

Nurkhairizan Khairudin[1], ***Nurfadhlina Mohd Sharef***[1], ***Shahrul Azman Mohd Noah***[2] and ***Norwati Mustapha***[1]

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia
43400 Serdang, Selangor, Malaysia

[2]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia

Email: nurkhairizan@gmail.com[1], nurfadhlina@upm.edu.my[2], shahrul@ukm.edu.my[3], norwati@upm.edu.my[4]

DOI: https://doi.org/10.22452/mjcs.sp2018no1.3

ABSTRACT

Collaborative filtering that relies on overall ratings has been widely accepted due to the ability to generate satisfactory recommendations. However, the most challenging difficulty of this approach is the lack of sufficient ratings or the so-called data sparsity. Moreover, sometimes these ratings alone are not sufficient to precisely understand users' specific behaviours. A user may show his/her overall preferences on an item through the overall ratings but at the same time, they may not satisfy with certain aspects of the item. This situation happened due to the emphasis on aspects that may be different among users and will effect a user's final decisions. Therefore, in this paper, we proposed a model called Neural Network model for Multi-Aspect with Strong Correlation (NNMASC) that utilize the significance of aspect's correlation to enhance the predictive accuracy of personalized recommendation. We integrate the user, item, aspects and overall ratings via embedding features by utilizing the available multi-aspect ratings from hotel reviews dataset. NNMASC adopts a feed-forward neural network with back propagation algorithm to make rating prediction. The experimental result using MAE shows that the proposed model has significantly outperformed the traditional models and the state-of-the-art approaches in terms of prediction accuracy.

Keywords: recommender system; collaborative filtering; multi-aspect; neural network; correlation strength;

1.0 INTRODUCTION

Recommender Systems (RS) are designed to predict user preferences for items by producing a set of relevant recommendations. The ability to produce personalized content to the user has made this kind of system a popular choice since it emerges to deal with the information overload problem, where users are overwhelmed by the expanding of data on the web.

In recent years, a variety of RS implementing various kinds of methods and approaches has been developed. The most popular method is Collaborative Filtering (CF) which relies on user-provided ratings to infer user preferences on items. CF works well and has achieved great success in various kinds of applications. However, the most challenging difficulty of CF is when there is a lack of sufficient ratings [1]. Although the number of items and users reaches hundreds of millions, the overall coverage of items by each user is relatively low. This situation generates an extremely sparse user-item matrix.

Recently, the increasing popularity of commercial website has encouraged users to express their assessment on an item through reviews. These reviews are useful to other users in order to get the idea of the characteristic of an item such as the quality and aspects that they expected. Although the original reviews are in textual form and difficult to be understood by machines; the advanced topic modelling and opinion mining or sentiment analysis can make it possible to process the reviews in order to extract valuable elements that can help to improve the accuracy or recommendation.

Approaches that exploit textual reviews are termed as review-based RS, and based on the existing research; these systems are able to address the rating sparsity issue in several ways. These include creating term-based user

preferences [2][3], generating virtual ratings/rating prediction [4][5] and augmenting the available ratings [6] with additional preference information. Although these studies do employ textual reviews, most of them do not consider how aspects of the review influence the overall rating scores.

Aspect is one of the significant review elements in review-based RS. From the collections of textual reviews, aspects can be extracted accordingly using an opinion mining approach. The main problem with most of the existing aspect-based recommendation methods is that they do not consider user preferences during the learning processes [7]. Therefore it is difficult to precisely predict user interest on the items which they have no experience with. Furthermore, these review-based methods only take a review text or a sentence and analyse its sentiment directly without considering how each available aspect influence the user's rating behaviour.

Recently, Neural Network (NN) has gain popularity due to the advance of its computational capabilities. One of the advantages of using an NN approach is the ability to represent a complex nonlinear relationship between features and independent variables. This is done by choosing a suitable non-linear activation function such as Rectified Linear Unit (ReLU) in order to identify higher polynomial features, interactions and to utilize the availability of multiple optimization algorithms [8]. Therefore, when dealing with multi features (i.e. aspects) the advanced of NN approach can be employed to improve recommendation prediction.

Therefore, in this research, we incorporate multi-aspect ratings into the prediction process by using NN in order to infer the overall rating prediction. The model called NNMASC is proposed to discover the significance of aspects to enhance the predictive accuracy of multi-aspect based recommendation. In this model, we utilize the correlation strength of all aspects to the overall rating using Pearson Correlation (PC). Based on this correlation values, we select the aspect with a strong correlation to be used as an input representing the aspect. After that, we integrate the users, items, overall ratings and strongly correlated aspect rating via embedding features that are used to feed the NN model. We conduct a series of experiments by utilizing the available multi-aspect rating data from TripAdvisor reviews dataset. NNMASC adopts a feed-forward neural network with back propagation method to make a rating prediction. Finally, the prediction accuracy is measured using the Mean Absolute Error (MAE) which is defined in section 4.2. The result is expected to provide more accurate recommendation than the existing single rating-based approaches and state-of-the-art aspect-based recommendation.

We highlight the main contributions of this research as follows: (1) we propose a model that exploits correlation strength of aspects for prediction of ratings in RS, (2) we adopt a feed-forward NN with back propagation model to generate features embedding for the users, items and aspects from the available dataset, and; (3) the prediction performance is improved by integrating the multi-aspect ratings in the NN model as evidenced from the evaluation results.

The remainder of this paper is organized as follows: firstly, reviews of some of the related works are described in section 2. Section 3 then provides an explanation of the proposed recommendation model, the NNMASC. Next, experimental settings and results are presented in Section 4 followed by Section 5 that concludes the findings of this research.

2.0 RELATED WORKS

Typically, RS approaches are classified as Collaborative filtering (CF), Content-based (CB), Knowledge-based (KB) and hybrid [9] depending on the different set of knowledge sources and the algorithmic approach employed. Therefore, this section will elaborate on the popular RS approaches followed by the single and multi-aspect recommendation.

2.1 RS Approaches

The most popular method is CF which relies on user-provided ratings to infer user preferences. For the purpose of personalized recommendation, CF techniques have been proven to be able to achieve great success by assisting the user in making their choices of preference items based on the ratings of other users who share similar interests. Since CF relies on the rating data, it only can produce efficient recommendation when this rating data is sufficient. However, rating data suffer from the data sparsity problem because there are huge numbers of items available but it is impossible for the user to provide ratings for most of the item. As a result, many types of research try to improve the prediction accuracy of CF by proposing various ideas in order to resolve the data sparsity problems. CF can be classified into two categories: the user-based and the item-based approaches [10]. For the user-based, the items preferred by similar

users will be recommended to the current user. Otherwise, for item-based, a user will receive recommendations of items that are similar to those they favoured in the past. The only input of conventional CF method is a matrix of given user-item ratings. CF then typically will generate the output in terms of (a) a prediction indication based on what degree the current user will like or dislike a certain item and (b) a list of n recommended items.

Since CF suffers from the rating sparsity problems, another approach called CB comes into consideration because they rely on the content representations of items instead of users' ratings. CB is used to locate items that have similar content to items the target user liked. Although CB methods are not very sensitive to the data sparsity [11], however, they suffer from other limitations such as diversity and overspecialization [12] as well as invalid recommendations for complicated objects [13]. Therefore, to overcome this limitation and to improve the quality of the recommendations, a hybrid approach where the combination of two or more recommendation approaches has been introduced [14].

Currently, many hybridization approaches have been proposed to be the recent techniques that reduce the sparsity problem. The techniques create hybrid systems with multiple representations including demographic [15] and user-generated content such as tags [16][12] and social relationship information [17][16][18][19] to augment the accuracy of recommendation. However, when the user has limited historical data, these methods are still inadequate. Therefore, other available information must be employed to be sufficiently reliable, complete or representative.

2.2 Single and multi-criteria recommender systems

Traditionally, CF works on a two-dimensional matrix of users and items. Rows and columns represent users and items accordingly. Users explicitly express their preferences through the overall ratings based on the item they experienced. As such, the value of the overall ratings most of the times are influenced by specific aspects of the items in concern. Then, the main task of RS is to recommend appropriate items to the users by estimating the suitability of unseen items based on the given overall rating.

In comparison to single rating recommendation, multi-criteria ratings will have *n* criteria and *one* overall rating. Fig. 1 and Fig. 2 show the different views when we plot single and multi-criteria rating using the same dataset. From Fig. 1, we observe that *user 1* and *user 2* have strong similar preferences for all hotels. However when we consider Fig. 2 with multi-criteria ratings of three aspects (R-Room, L-Location and C-Cleanliness), generally *user1* is seen to be more similar with *user* 3. From this visualization, we can see how the multi-criteria rating can affect the preferences of the items.

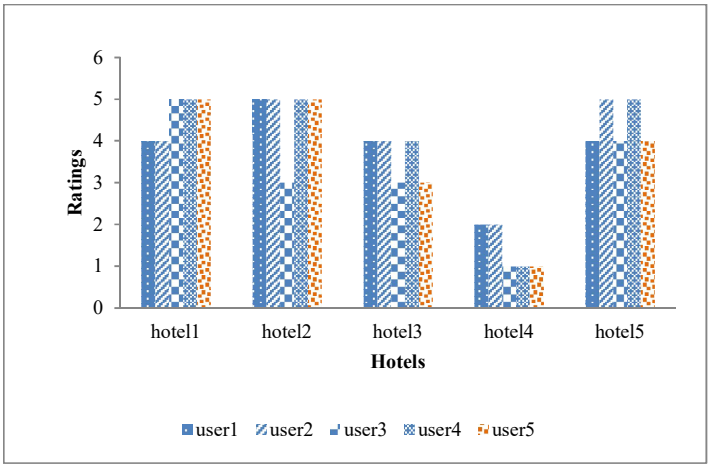


Fig. 1. Single overall rating for the recommendation

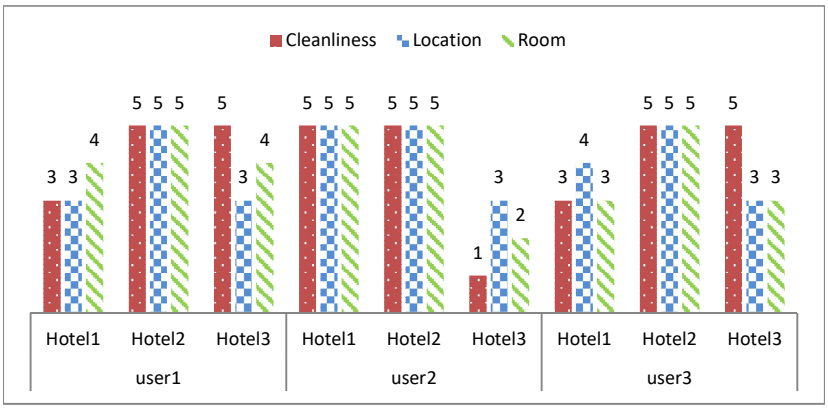


Fig. 2. Multi-aspect rating for the recommendation

When considering only overall rating to make a prediction of unseen items, the user similarity which is the main concern in CF, has a limited performance. This is because the preferred items for a user may depend on more than one criteria that effects the user consideration while rating an item they experienced [20][21]. For that reason, it is better to incorporate other information such as weights on various criteria of the item for capturing user preferences efficiently. This consideration would improve the accuracy of personalized RS [22]. Instead of two-dimensional matrixes of users and items used in CF, the multi-criteria recommendation is represented as three-or-more dimensional matrixes of users, items and criteria.

There are various ways to employ multi-criteria in the recommendation. Musto et al. [23] and Nilashi et al. [24] proposed a multi-criteria recommendation approach to enhance the CF performance. They use aspect ratings as a multi-criteria component in order to provide multi-faceted representation. Kant et al. [25] also proposed a multi-criteria recommendation but using fuzzy Bayesian approach to generate recommendation matrix for each criterion that consists of membership degree. Evaluation has proven the improvement of the multi-criteria recommendation as compared to the single-criteria recommendation. Instead of using all criteria for the recommendation as illustrated by Musto et al. [23], this research will select only one criterion (i.e. aspect) based on the evaluation of the correlation strength so that the dimensionality of the input features can be reduced in order to overcome the computational complexity of NN model.

2.3 Aspects as multi-criteria recommendation

In recent years, a number of recommendation techniques that attempt to take advantage of s multi-criteria preference such as aspects, have been proposed. Aspect is one of the valuable elements in the review that can be used to identify user preferences. Users usually placed different emphasis towards different aspects of the item they experienced. Therefore, the aspect element can be used to improve prediction accuracy by suggesting more accurate recommendation based on users' preferences.

To the best of our knowledge, there are few studies that incorporate user preferences from the reviews into CF. Most of the studies try to enhance the CF performance by integrating Matrix Factorization (MF) with other techniques such as in [26]. Lei et al.[27], Addio et al. [28] and Ma et al. [29] employ review-based methods that analyze sentiment from the reviews directly. Despite this, our main objective is to improve the inference of rating prediction based on multi-aspect rating rather than dealing with the topic modelling and opinion mining or sentiment analysis.

Since the employment of multi-criteria involves more than 2-dimensional matrix, Chen et al. [30] and Yang et al. [31] use Tensor Factorization (TF) to learn user interest from reviews and understand users' preferences by utilizing the multi-aspect ratings. Yang et al. [7] has extended Wang et al. [32] work to employ TF in measuring the aspect weights. Then, TF is used for the second time to predict the unknown overall rating where this tensor is constituted by weighted aspect ratings and overall ratings. Another aspect-based recommendation is also done by Bauman et al. [33] and Qiu et al. [34] where they consider aspects from the reviews to infer the overall ratings. The models mentioned have made great contributions to the modelling of user review information in recommendation tasks. However, there are many spaces available to improve the recommendation performance since the main problem of most existing review-based methods is that they do not consider user preferences in the learning processes.

3.0 METHODOLOGY

This section introduces the proposed recommendation model based on NN for multi-aspect ratings to be validated with the hotel reviews dataset. The general algorithm of the proposed model is to exploit the aspects and overall ratings for CF using NN model.

We start with the pre-processing of raw data to be organized accordingly to the standard data for the proposed model. In order to precisely evaluate the significance of aspects to the overall rating prediction, we select only five aspects with the most available ratings. They are *'cleanliness', 'location', 'room', 'service'* and *'sleep quality'*. In the next step, we implement the proposed model of NN for multi-aspect based on the aspects of correlation evaluation. Then, the selection of aspects is determined by considering the correlation strength as will be described in Section 4.

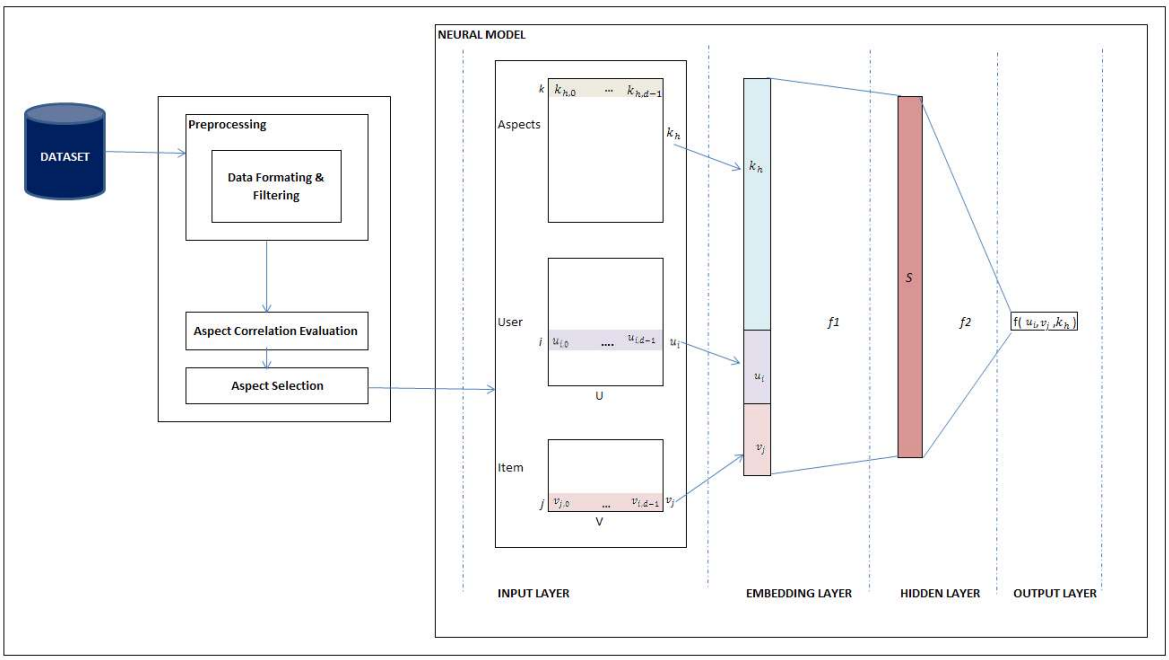


Fig. 3. NNMASC Recommendation Framework

Fig. 3 shows the NNMASC recommendation framework that consists of the following four sections : (a) data preprocessing, (b) aspect correlation evaluation (c) aspect selection and (d) NN multi-aspect model. The purpose of the processing is to remove redundant reviews and construct the available data into standardizing dataset according to the requirement of the implementation environment. After that, PC is calculated to measure the correlation strength of the all aspect ratings to the overall rating. Then, the aspect selection determines the aspect input dimension for the recommendation. Finally, this additional input is fed to the NN model in order to make a rating prediction.

The neural model has four layers consisting of the input layer, the embedding layer, the hidden layer and the output layer. The hidden layer with different dimension is used to encode the input in high-level abstraction. In the output layer, he output of $f(u_i, v_j, k_h)$ in Fig. 3 is calculated as (1):

$$x = (u_i, v_j, k_h)$$
$$S = \sigma(f_2(x)) \quad (1)$$
$$f(x) = \delta(f_1 S)$$

where f_1 and f_2 are the weight matrices for the input layer to the hidden layer and from the hidden layer to output layer respectively. We use the ReLU activation function to learn higher-order features interaction with functions δ and σ are the non-linear activation functions. Using ReLU can lead to faster convergence and avoiding the problem of vanishing gradient[35]. Therefore, ReLU has become the most popular activation function since it has been used in most of the NN models.

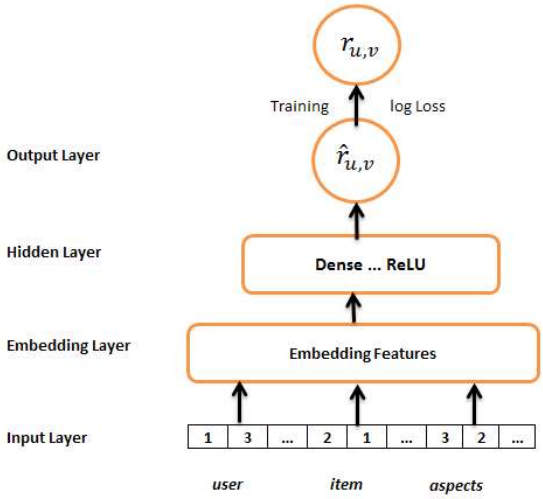


Fig. 4. NN Model for multi-aspect recommendation

As can be seen in Fig. 4, the framework consists of three types of input which are users, items and a set of aspects ratings. $\hat{r}_{u,v}$ indicates the rating prediction, where $r_{u,v}$ is the true ratings available from the dataset. The main objective of the proposed NN for multi-aspect recommendation is to model the interactions between users *u*, items *v* and a set of aspects ratings *k* in order to estimate the overall ratings. To achieve this objective, we design a model that able to learn the function *f* () which can project user *u*, item *v* and a set of aspect *k* to a predicted rating $\hat{r}_{u,v}$ as shown by (2).

$$\hat{r}_{u,v} = f(u_i, v_j, k_h) \tag{2}$$

For the training phase, the training data consist of users, items, overall ratings and a set of five aspects with their ratings. The complete training set is represented by T = $\{< u_i, v_j, r_{u,v,} k_{u,v,h} >\}$, where *u* and *v* represent users and items respectively. The overall rating for user *u* and item *v* is denoted by $r_{u,v}$ and $k_{u,v,h}$ represents the ratings for aspect *k* of user *u* on item *v*.

For the optimization, we used a gradient-based method called *Stochastic Gradient Descent (SGD)* as has been pre-defined in the TensorFlow library.

4.0 EXPERIMENTS AND RESULTS

This section presents the experimental detail of our proposed models. We start with a detailed analysis of the dataset, followed by the experimental settings and the mechanism used to implement the proposed model in order to demonstrate the effectiveness of the proposed model as compared to other benchmark models in an empirical manner.

4.1 Dataset

We demonstrate the efficiency of the proposed model using multi-aspect rating dataset provided by Wang et al. [36]. The preferences information on the hotel dataset was provided by users on eight predefined aspects namely

'cleanliness', *'location'*, *'room'*, *'services'*, *'sleep quality'*, *'value'*, *'business'* and *'check-in'*. An *overall* rating that measures the final users' satisfactory is also available in the dataset. Example of the data is as shown in Table 1. The numbers outside the bracket are the *overall* ratings and the numbers inside the bracket are the ratings for the eight aspects. The value of -1 represent those missing aspect ratings.

Table 1: Multi-aspect rating matrix example

		Hotel			
		h_1	h_2	h_3	h_4
User	u_1	**4** (4,3,4,4,4,4,-1,-1)	**5** (4,5,5,-1,5,4,-1,-1)	**3** (3,3,3,2,3,1,-1,-1)	**3** (3,3,3,2,3,1,-1,-1)
	u_2	**3** (3,3,3,2,3,1,-1,-1)	**4** (4,4,3,4,2,5,-1,-1)	**5** (5,5,5,5,5,1,-1,-1)	**4** (4,4,3,4,2,5,-1,-1)
	u_3	**2** (2,1,3,2,2,1,-1,-1)	**4** (5,4,4,4,4,4,-1,-1)	**5** (2,3,3,2,2,1,-1,-1)	**1** (2,1,1,-1,-1,1,-1,-1)
	u_4	**5** (5,5,5,5,5,1,-1,-1)	**2** (2,3,3,2,2,1,-1,-1)	**2** (2,-1,-1,2,2,1,-1,-1)	**2** (2,2,1,2,1,1,-1,-1)
	u_5	**2** (2,-1,-1,2,2,1,-1,-1)	**4** (4,3,3,4,4,4,-1,-1)	**5** (2,3,3,2,2,1,-1,-1)	**5** (5,5,5,5,5,5,-1,-1)
	u_6	**3** (3,3,3,2,2,1,-1,-1)	**5** (5,5,5,5,5,5,-1,-1)	**4** (4,3,4,4,4,4,-1,-1)	**4** (4,3,3,4,4,4,-1,-1)

During the experiment, we categorized the data into two categories namely exclusive and inclusive sub-datasets. Each sub-dataset contain a number of reviews and each review have a *user id, hotel id*, and *overall rating* with only five predefined aspects ratings which are *'cleanliness', 'room', 'location', 'service'* and *'sleep quality'*. These five aspects were selected because they have the most available ratings. We do not remove any reviews from the dataset in order to remain the high sparsity level for the dataset.

Table 2: Dataset density and sparsity for exclusive and inclusive sub-datasets

	users	items	reviews	density	sparsity
Exclusive					
LARA-9K	8880	100	9090	1.02	**98.98**
LARA-53K	43533	565	53570	0.22	**99.78**
LARA-89K	82195	1371	89060	0.08	**99.92**
LARA-123K	102315	1114	123148	0.11	**99.89**
LARA-300K	199748	1215	306327	0.13	**99.87**
Inclusive					
LARA-17K	17390	220	17390	0.45	**99.55**
LARA-45K	41421	496	45554	0.22	**99.78**
LARA-88K	78421	795	88266	0.14	**99.86**
LARA-141K	116915	1341	141836	0.09	**99.91**
LARA-334K	213547	2495	334646	0.06	**99.94**

For the exclusive sub-dataset, we randomly extract five different and an increasing number of reviews. Each content in exclusive sub-dataset is distinct from the others with a different number on reviews. For the inclusive sub-datasets, the number of reviews also keep on increasing. However, the content of the inclusive sub-datasets is related to each other. For example, LARA-45 actually consists of data from LARA-17K with 17390 reviews and additional reviews data that makes the total reviews to 45554. On the other hand, LARA-88K collectively has the content of all data from the LARA-45 and some additional reviews data that make the total reviews to be 88266, and so on. The summary of the various different distribution of the sub-dataset and dataset density and sparsity level for all of these two sub-datasets are shown in Table 2.

The density and sparsity are measured using (3) and (4) respectively as follows:

$$Density = \frac{N_w}{N_u \,.\, N_v} \,.\, 100\% \tag{3}$$

$$Sparsity = (\,1 - \frac{N_w}{N_u \,.\, N_v}\,) \,.\, 100\% \tag{4}$$

where N_w , N_u and N_v are the number of reviews, users and items respectively.

In order to evaluate the strength of correlations between ratings of the five aspects to the *overall* rating, we conduct a basic statistics where the result is shown in Table 3 and Table 4. Correlation is a statistical technique to determine how strongly pairs of the variable are related. The correlation values range in between -1.0 and 1.0, whereby a correlation of -1.0 indicates a perfectly negative correlation and a correlation of 1.0 indicates a perfect positive correlation. Contrarily, a result of 0 indicates no relationship at all. The most common measure of correlation is the PC that can show a linear relationship between two sets of data as illustrated in equation (5).

$$PC = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}} \tag{5}$$

Therefore, in this research, the strong correlation strength is determined by the maximum value which is the values closest to 1 or -1; meanwhile the weak correlation is indicated by the minimum value which is closest to 0. In Table 3 and 4, bold values indicate a strong relationship, whereas underline values indicate a weak correlation to the overall ratings. As can be seen from the correlation values for the exclusive sub-datasets, the *overall* rating has a strong relationship with the *'service'* aspect for all sub-datasets. However, the weak correlation values are shown in three sets of data for *'location'* aspect and two with '*sleep quality'* and '*location*' aspect.

Table 3: Basic statistical data for overall ratings in the exclusive (a) and inclusive (b) sub-datasets

	LARA-9K	**LARA-53K**	**LARA-89K**	**LARA-123K**	**LARA-300K**
count	9090	53570	89060	123148	306327
mean	4.01	3.97	4.17	3.97	4.07
std	1.21	1.18	1.09	1.19	1.12
min	0	0	0	0	1
25%	3	3	4	3	4
50%	4	4	5	4	4
75%	5	5	5	5	5
max	5	5	5	5	5

(a)

	LARA-17K	**LARA-45K**	**LARA-88K**	**LARA-141K**	**LARA-334K**
count	17390	45554	88266	141836	334646
mean	3.93	3.88	3.93	3.95	3.85
std	1.23	1.23	1.20	1.19	1.24
min	0	0	0	0	0
25%	3	3	3	3	3
50%	4	4	4	4	4
75%	5	5	5	5	5
max	5	5	5	5	5

(b)

Meanwhile, for the inclusive sub-dataset, the correlation values also show that the '*service*' aspect does has a strong relationship with the *overall* rating, but weak correlation values with *'location'* and *'sleep quality'* aspects but for different sub-datasets distribution. Based on this statistic, we choose to execute the correlated aspects with the strongest correlation ('*service'* aspect) value and the weakest correlation ('*sleep quality*' aspect) value to measure and find how the relationship of the correlated aspects to the overall ratings do affect the prediction result. We then implement the same model to the all five aspects to see the different performance between the correlated aspects strength with all five aspects prediction results.

Since data is an important element in the NN model, the different distribution of the dataset will be used in the experiment to identify whether this kind of situations does affect the effectiveness of the prediction model.
\

Table 4: PC values for all aspects of the overall rating for exclusive (a) and inclusive (b) sub-datasets

aspects	**LARA-9K**	**LARA-53K**	**LARA-89K**	**LARA-123K**	**LARA-300K**
cleanliness	0.54	0.47	0.52	0.47	0.47
location	0.26	0.22	0.25	0.21	0.23
room	0.48	0.49	0.38	0.45	0.40
service	**0.61**	**0.57**	**0.67**	**0.54**	**0.56**
sleep quality	0.18	0.16	0.33	0.21	0.25

(a)

aspects	**LARA-17K**	**LARA-45K**	**LARA-88K**	**LARA-141K**	**LARA-334K**
cleanliness	0.49	0.54	0.52	0.50	0.55
location	0.22	0.28	0.24	0.23	0.23
room	0.45	0.47	0.45	0.46	0.50
service	**0.55**	**0.60**	**0.58**	**0.58**	**0.61**
sleep quality	0.18	0.25	0.26	0.22	0.26

(b)

During the performance evaluation, we randomly divided each sub-dataset into training and test sets in the ratio of 80% and 20% respectively. In our case, the division of training and test sets is repeated five times in order to avoid over-specialization. The average performance of the five learned models are then presented. In some existing studies, in order to reduce the error caused by missing aspects rating and to evaluate some existing methods with more effective results, usually, some of the data is removed. The removed data are based on certain conditions such as data for users with less than ten reviews or items with a certain number of reviews. However, in this experiment, we use the original dataset without removing the available data except for redundant reviews. Therefore, as can be seen in Table 2, all exclusive and inclusive sub-datasets have low data density rates with high sparsity rates.

4.2 Evaluation Metrics

Regarding the Pareto Principle, the common and optimal allocation of resources is 80/20. Therefore, for the purpose of evaluation, the dataset is divided into two sections: the training set and the testing set in the ratio of 80:20. In this experiment, we adopted the error metric, the *MAE* which is widely used for predictive accuracy. MAE evaluates the difference between the ratings predicted by the recommender and given by the user. In this metric, r_{uv} and $\hat{r}_{uv}$ represents the true rating value and the predicted rating value respectively, with $|N_t|$ as the number of test sample set. The *MAE* is defined in (6) as follows:

$$MAE = \frac{1}{|N_t|} \sum_{(u,v)\in N_t} |r_{uv} - \hat{r}_{uv}| \tag{6}$$

Generally, in this research, the concern is to precisely predict the ratings rather than shortlisting of interesting items to the users. MAE can measures how much the predicted rating differs from the true rating. A lower value indicates a more accurate and better performance of the predictions.

4.3 Experimental Setting

Experiments are carried out in order to answer the following research questions:

(a) How multi-aspect effect the recommendation performance as compared to a single rating-based recommendation?

(b) Does increasing the number of data (reviews) lead to a better predictive accuracy for the single and multi-aspect RS?

(c) How does the proposed algorithm perform, when compared to the state-of-the-art CF techniques based on TF?

As for the experimental settings, all series of the test samples are implemented in an Intel Core i5 1.6GHz, 4 GB 1600 MHz DDR3 of RAM in Mac OS X Yosemite version 10.10.5 operating system. We implemented all models using one of the machine learning frameworks that can be used to design, build and train NN models called Tensorflow.

We started measuring the recommendation quality of the proposed model in terms of the relationship strength. Firstly, we divided the dataset info five categories with different numbers of reviews. These data contain high sparsity rates as shown in Table 2. Then, we conducted a series of experiment with different strength of the aspect, k where $k = \{1\}$. For the aspect with the strongest and weakest correlation values with the overall ratings, we named it as NNMASC and NNMWC respectively. After that, we executed the experiment to the same model on all five aspects selected as previously described in section 4.1. We named this method NN for Multi-Aspect with All aspects (NNMAA).

In order to evaluate the performance of our proposed model, we conduct a series of experiments with various setting. First, we observe the use MF method compared to NN approach for the single overall ratings only (in this experiment we represent it as MF on Single Rating (MFSR) and NN for Single Rating (NNSR)). Both approaches make predictions base on only the interactions of user latent factors and item latent factors. Then, we compared the results of NNMASC, NNMAWC and NNMAA to show the performance of the NN model for different correlation strength of aspects. Further, the performance of the proposed model is compared with the state of the art aspect-based recommendation conducted using TF[7] for Multi-Aspect, TFMA.

We set the embedding to 30, batch size to 1000 and epoch size to 30. We employ ReLU as the activation function and SGD for optimization in order to perform a parameter update for each training example of NN models. The other hyper-parameters are left as the default value set by the available pre-defined function in TensorFlow library. In order to make a comparison analysis, we learn all models by optimizing the square loss.

4.4 Results and Discussion

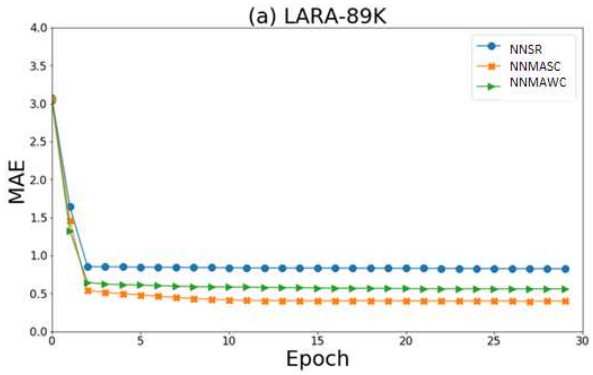
We perform experiments for the proposed NN models and observe the results base on the effect of data sparsity, the correlation strength of aspects, the impact of the amount of data and the number of aspects used for recommendation.

The effect of data sparsity: The analysis of the data sparsity and density for the sub-datasets used in this research has been presented in Table 2 in Section 4.1. Basically, for the inclusive sub-datasets, the sparsity increase with the increment amount of data. However, for the exclusive dataset, random distribution of sparsity is observed because of the ignorance of the relationship between the sub-datasets. Table 5 shows MAE and sparsity values for the exclusive and inclusive sub-datasets, sorted by the sparsity value in ascending order. The result shows the prediction accuracy for each model change differently. For both sub-datasets, NNMAA shows the best prediction accuracy when the data is very sparse. Moreover, in the exclusive sub-dataset, LARA-89K with the maximum sparsity rate shows the best prediction accuracy for all NN models. For the inclusive sub-datasets, where the reviews data is collectively increased, the best performance distributed randomly but still better for sparser data. This result proves that NN models can better adapt to data sparsity especially when more aspects used.

Table 5: MAE and sparsity percentage for exclusive and inclusive sub-datasets

	Sparsity(%)	MFSR	NNSR	NNMAWC	NNMASC	NNMAA
Exclusive						
LARA-9K	**98.98**	4.021	0.884	0.789	0.697	0.534
LARA-53K	**99.78**	3.975	0.867	0.748	0.479	0.431
LARA-300K	**99.87**	4.078	0.825	0.652	0.452	0.387
LARA-123K	**99.89**	**3.973**	0.879	0.722	0.489	0.423
LARA-89K	**99.92**	4.160	**0.824**	**0.561**	**0.400**	**0.349**
Inclusive						
LARA-17K	**99.55**	3.933	0.906	0.866	0.589	0.481
LARA-45K	**99.78**	3.893	0.915	0.758	0.517	0.454
LARA-88K	**99.86**	3.930	0.901	0.720	**0.470**	0.429
LARA-141K	**99.91**	3.950	**0.881**	**0.714**	0.497	0.423
LARA-334K	**99.94**	**3.847**	0.917	0.732	0.498	**0.413**

The effect of the correlation strength of aspects: The PC is used to evaluate the strength of correlation of the five aspects to the overall rating as described in Section 4.1. The closer the correlation values to 1 or -1, the stronger the strength of the aspect correlation to the overall rating. As presented in Table 4 in Section 4.1, both exclusive and inclusive sub-datasets show that the '*service*' aspect has the strongest correlation value. However, the weakest correlations distribute among '*location*' and '*sleep quality'* aspects. For this category of the experiment, NNMACS represent the NN model with strong correlation and NNMWC for weakest correlation value. We also compare the result of NN without aspect involved in the prediction, NNSR. From the previous result shown in Table 5, we choose sub-dataset LARA-89K for inclusive and LARA-141K for exclusive sub-dataset, which have almost similar sparsity with better accuracy. This is to show the effect of correlation strength without any influences of other situation such as sparsity. Fig. 5 (a) and (b) show the performance of NN models base on the correlation strength in terms of MAE for the exclusive and inclusive sub-datasets respectively.



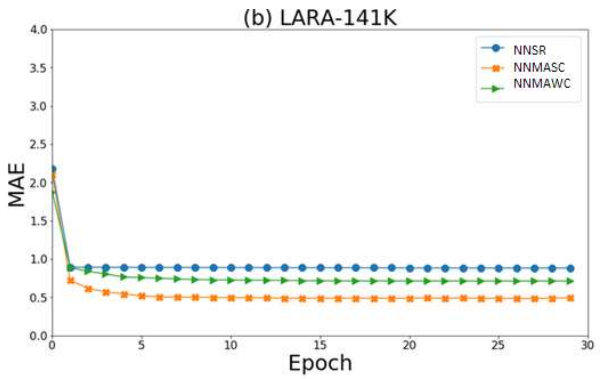


Fig. 5: Performance in MAE base on the correlation strength for exclusive (a) and inclusive (b) sub-datasets

From Fig. 5, all NN models start to converge to the minimum MAE and start stabilized at almost the same epoch for both exclusive and inclusive sub-datasets. NNMASC which has the strongest correlation show the best performance followed by NNMAWC, with the weakest correlation; and NNSR for the single rating-based recommendation. This result presents the significant correlation strength in the aspect-based recommendation. When the right aspect is selected, the dimensionality and complexity of the NN models can be reduced significantly for the efficient recommendation. Furthermore, the utilization of only one aspect of the recommendation process still can produce a good impact even with the weakest correlation aspect. This is shown by the plot pattern in Fig. 5(a) and 5(b) where the different between NNSR and NNMAWC clearly can be seen after the convergence of the minimum MAE.

The impact of the amount of data: There are various distributions and combinations of data have been used in this experiment as stated in Table 2 in Section 4.1. Two groups of the data are separated into exclusive and inclusive sub-datasets. This is to show the prediction performance when the data increase in this two different situation. The Exclusive sub-dataset contains an increasingly different collection of data and the inclusive sub-dataset that collectively increase the amount of data. The detail of this group is discussed in Section 4.1. Fig. 6 (a-e) and Fig. 7 (a-e) show the MAE for exclusive and inclusive sub-dataset respectively with the different amount of data.

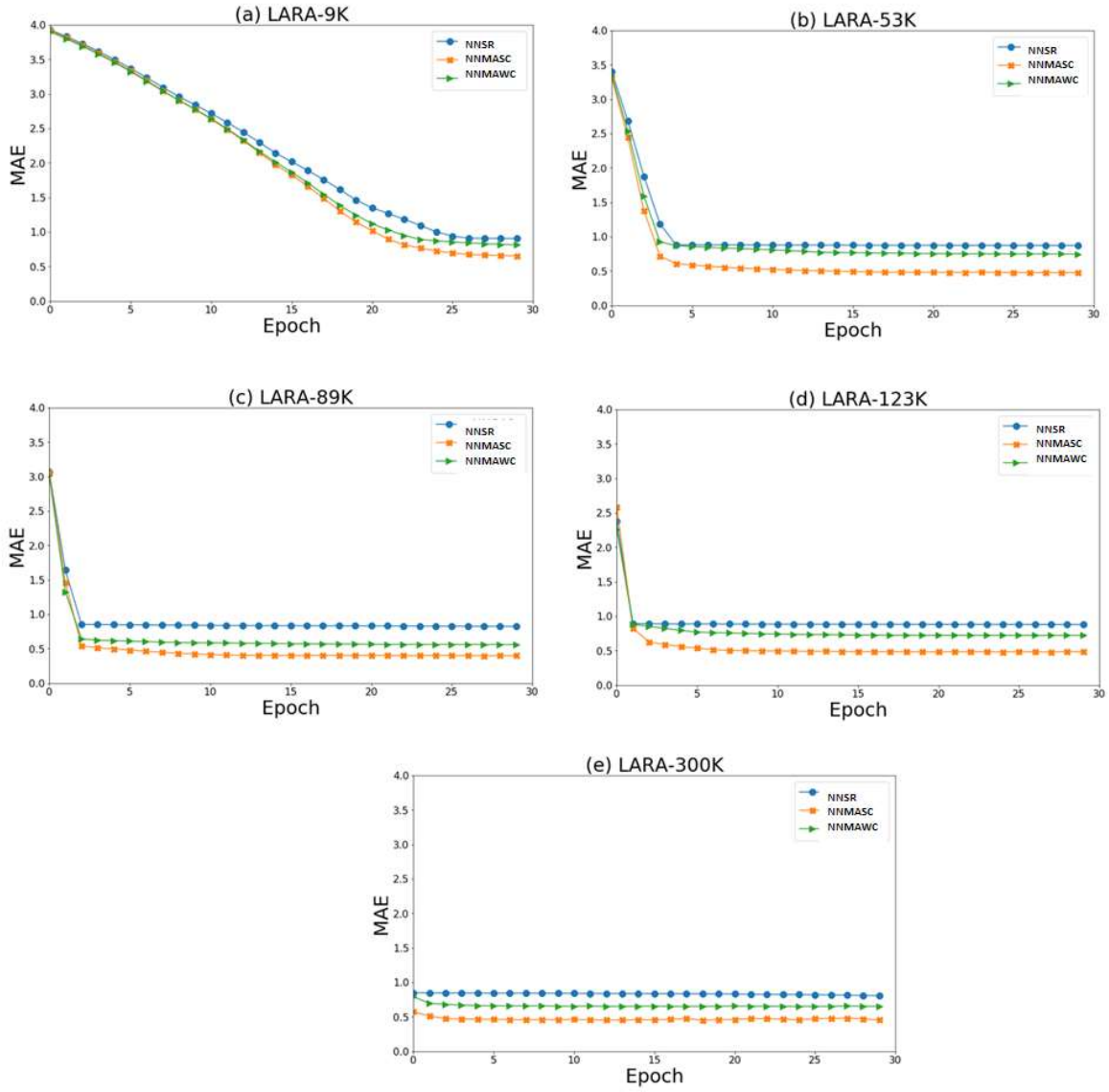


Fig. 6: Performance in MAE for exclusive sub-datasets with increasing amount of data

In order to determine the impact of the amount of data in the prediction quality of NN models, we remain the experimental setting with the previous correlation strength experiment. We just want to investigate the different prediction accuracy when we increase the amount of data in two situations; exclusive and inclusive. From Fig. 6 and Fig. 7, we can observe that the amount of data do effect the convergence of the minimum MAE before the value is stabilized. The result applies to both inclusive and exclusive dataset, where the best MAE is achieved in later epoch for the smaller amount of data. As the amount of data increases, the minimum MAE is converged in an earlier epoch. For example for LARA-300K, the best MAE is achieved at the beginning of the epoch. However, for all NN models, the MAE at the first epoch starts at different values depends on the amount of data. From this result, we observed that the amount of data effects the performance of the prediction efficiency. Furthermore, the result for both exclusive and inclusive dataset does not show significantly differences. Base on the result, we can conclude that as long as more data is used for the training purposes, the prediction performance can be improved regardless of the relationship among the data.

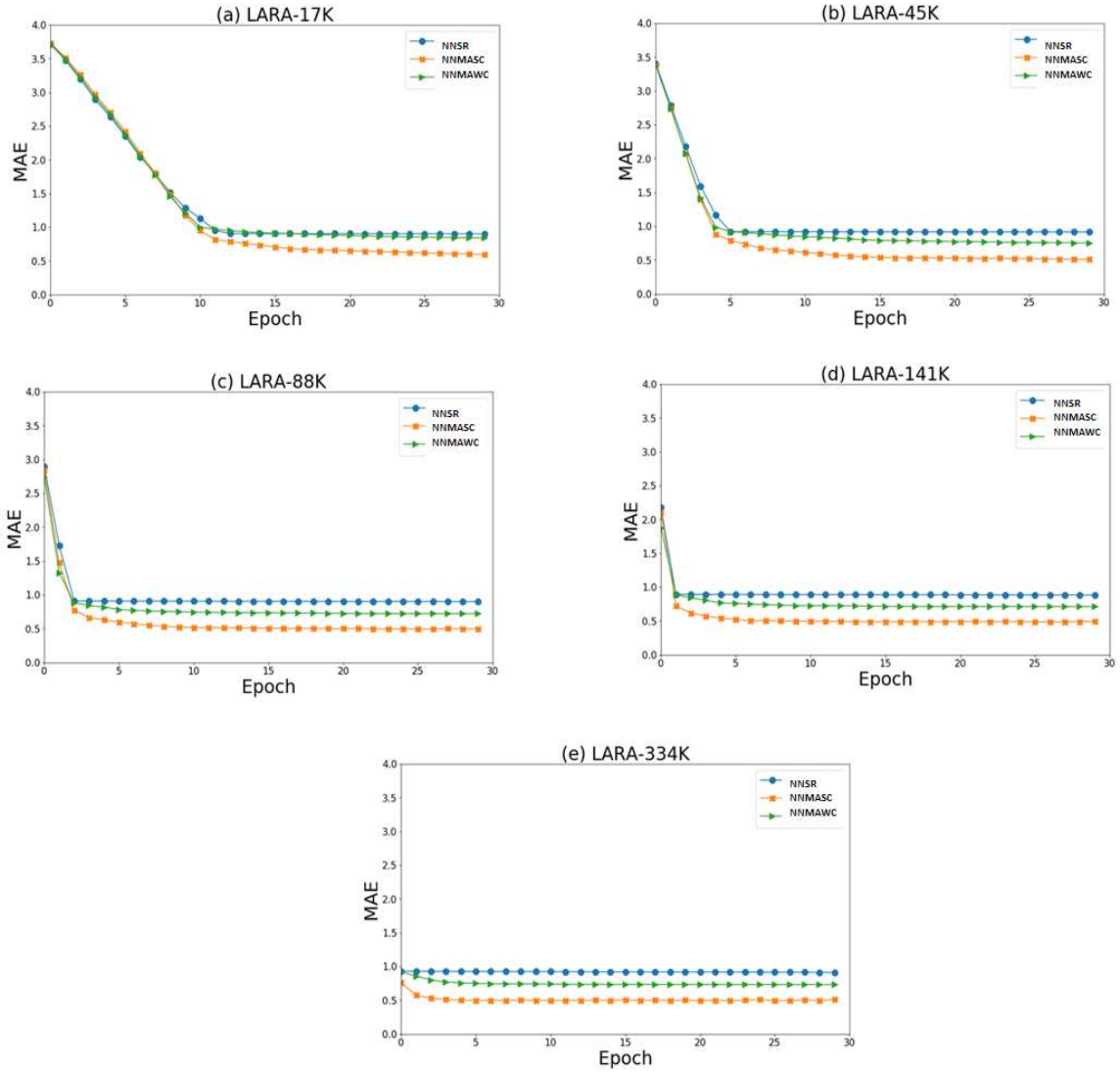


Fig. 7: Performance in MAE for inclusive sub-datasets with increasing amount of data

The effect of the number of aspects: As discussed in Section 4.1, we try to utilize all of the aspects available from the dataset and at the same time, try to reduce the complexity of the NN model. With the correlation strength of aspects to the overall rating, we then select the strongest and weakest aspects in order to investigate the performance of both approaches as compared to NN model that utilized more aspects. Fig. 8 shows the performance of the prediction accuracy when different numbers of aspects are used. Since no significant difference with the result from exclusive and inclusive sub-datasets in term of the amount of data, we choose only one sub-dataset which achieved the best MAE values for all NN models as shown in Table 5. We also compare the result with the popular MF method, MFSR and NN without aspect utilization, NNSR to show the NN models with aspects utilization do improve the prediction accuracy. By summing up all of the experimental results, as shown in Fig. 8, we demonstrate that the utilization of aspects can enhance the prediction accuracy.

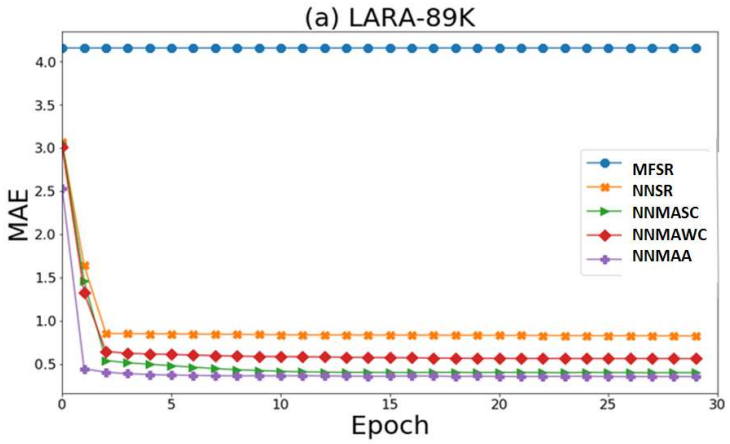


Fig. 8: MAE for the utilization of aspects

The MAE for MFSR is almost static throughout the epoch with no significant changes. Otherwise, for NN based approaches, the MAE start at a high value but then immediately converge to the minimum values after two or three epochs. It shows that, when more aspects are used (NNMAA), the results start to show better performance at the earlier epoch than the others. It is also interesting to observe that as the number of aspect increase, the performance accuracy of NN does not very much improve as compared to the NN model with only one strong correlation aspect (NNMSC). NNMAA utilize all five aspects of the dataset. However, in practice, it creates high dimensionality and complexity in the NN models for the implementation. Therefore, we conclude that the performance of NNMASC is good enough in order to utilized aspect for an aspect-based recommendation for better and efficient implementation.

Finally, we compared the performance of NN models with TF for multi-aspect based rating prediction, TFMA. TF is used to measure the aspect weights and then it is implemented for the second time to infer the overall rating. From the result in Table 6, all the NN models with aspects utilization significantly improved the prediction accuracy.

Table 6: MAE for TF based method and NN based models

	TFMA	NNMAWC	NNMASC	NNMAA
MAE	1.134	0.56	0.40	0.349

5.0 CONCLUSIONS AND FUTURE WORKS

In order to assess the effect of aspect ratings for predicting the overall ratings, we proposed a model based on a neural network called NNMSC which integrate strongly correlated aspect rating into the overall rating prediction model. In this case, we assumed that the overall rating is strongly influenced by the strongly correlated aspect with the overall rating. In order to identify the strongly correlated aspect, we first evaluate the correlation strength by using a PC. Then, the selected aspect is combined with the input of the NN model. From the experimental result, we observed that the proposed model effectively outperformed the baseline method in terms of the prediction accuracy. Based on the observation, some useful findings can be drawn from the experimental results: (a) NN models can overcome the sparsity problem which is one of the main drawbacks of the CF approach; (b) the performance of rating prediction with even only one strongly correlated aspect is shown to improve by employing the NN models, and (c) more aspects are shown to improve the rating prediction but it is computationally intensive and time-consuming because of the higher dimensionality. To date, evaluation was carried out on a single domain. Thus, the near future work is to extend for other domains such as movies and product reviews. Furthermore, another possible research direction is to investigate the performance of other popular NN approaches, such as recursive and convolutional NN, which get an outstanding achievement in other research fields.

REFERENCES

[1] M. Papagelis, D. Plexousakis, and T. Kutsuras, "Alleviating the sparsity problem of collaborative filtering using trust inferences," *Trust Manag.*, vol. 3477, pp. 224–239, 2005.

[2] D. Yu, Y. Mu, and Y. Jin, "Rating prediction using review texts with underlying sentiments," *Inf. Process. Lett.*, vol. 117, pp. 10–18, 2017.

[3] X. Zheng, W. Ding, Z. Lin, and C. Chen, "Topic tensor factorization for recommender system," *Inf. Sci. (Ny).*, vol. 372, pp. 276–293, 2016.

[4] D. H. Alahmadi and X. J. Zeng, "ISTS: Implicit social trust and sentiment based approach to recommender systems," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8840–8849, 2015.

[5] H. Feng and X. Qian, "Mining user-contributed photos for personalized product recommendation," *Neurocomputing*, vol. 129, pp. 409–420, 2014.

[6] J. Peng, Y. Zhai, and J. Qiu, "Learning latent factor from review text and rating for recommendation," *Proc. 2015 7th Int. Conf. Model. Identif. Control. ICMIC 2015*, no. Icmic, pp. 3–8, 2016.

[7] C. Yang, X. Yu, Y. Liu, Y. Nie, and Y. Wang, "Collaborative filtering with weighted opinion aspects," *Neurocomputing*, vol. 210, pp. 185–196, 2016.

[8] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, 1996.

[9] N. Tintarev and J. Masthoff, "Recommender Systems Handbook," *Recomm. Syst. Handb.*, vol. 54, pp. 479–510, 2011.

[10] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *ACM, no. Proc. 10th Int. Conf. World Wide Web*, pp. 285–295, 2001.

[11] P. Pirasteh, D. Hwang, and J. J. Jung, "Exploiting matrix factorization to asymmetric user similarities in recommendation systems," *Knowledge-Based Syst.*, vol. 83, pp. 51–57, 2014.

[12] Y. Zuo, J. Zeng, M. Gong, and L. Jiao, "Tag-aware recommender systems based on deep neural networks," *Neurocomputing*, vol. 204, pp. 51–60, 2016.

[13] G. Lv, C. Hu, and S. Chen, "Research on recommender system based on ontology and genetic algorithm," *Neurocomputing*, vol. 187, pp. 92–97, 2016.

[14] M. Al-Hassan, H. Lu, and J. Lu, "A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system," *Decis. Support Syst.*, vol. 72, pp. 97–109, 2015.

[15] V. Balakrishnan and H. Arabi, "HyPeRM: A hybrid personality-aware recommender for movie," *Malaysian J. Comput. Sci.*, vol. 31, no. 1, pp. 48–62, 2018.

[16] T. Ma, X. Suo, J. Zhou, M. Tang, D. Guan, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "Augmenting matrix factorization technique with the combination of tags and genres," *Phys. A Stat. Mech. its Appl.*, vol. 461, pp. 101–116, 2016.

[17] G. Guo, J. Zhang, and D. Thalmann, "Merging trust in collaborative filtering to alleviate data sparsity and cold start," *Knowledge-Based Syst.*, vol. 57, pp. 57–68, 2014.

[18] Q. Shambour and J. Lu, "An effective recommender system by unifying user and item trust information for B2B applications," *J. Comput. Syst. Sci.*, vol. 81, no. 7, pp. 1110–1126, 2015.

[19] P. Moradi, S. Ahmadian, and F. Akhlaghian, "An effective trust-based recommendation method using a novel graph clustering algorithm," *Phys. A Stat. Mech. its Appl.*, vol. 436, pp. 462–481, 2015.

[20] G. Adomavicius, "New Recommendation Techniques for Multi-Criteria Rating Systems," pp. 1–28.

[21] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, 2005.

[22] D. Jannach, Z. Karakaya, and F. Gedikli, "Accuracy Improvements for Multi-criteria Recommender Systems," *Ec*, vol. 1, no. 212, pp. 674–689, 2012.

[23] C. Musto, M. De Gemmis, and G. Semeraro, "A Multi-criteria Recommender System Exploiting Aspect-based Sentiment Analysis of Users ' Reviews," pp. 321–325, 2017.

[24] M. Nilashi, O. Bin, N. Ithnin, and R. Zakaria, "A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques," 2014.

[25] V. Kant, T. Jhalani, and P. Dwivedi, "Enhanced multi-criteria recommender system based on fuzzy Bayesian approach," *Multimed. Tools Appl.*, vol. 77, no. 10, pp. 12935–12953, 2017.

[26] C. Jiang, R. Duan, H. K. Jain, S. Liu, and K. Liang, "Hybrid collaborative filtering for high-involvement products: A solution to opinion sparsity and dynamics," *Decis. Support Syst.*, vol. 79, pp. 195–208, 2015.

[27] X. Lei, X. Qian, and G. Zhao, "Rating Prediction Based on Social Sentiment from Textual Reviews," *IEEE Trans. Multimed.*, vol. 18, no. 9, pp. 1910–1921, 2016.

[28] R. M. D. Addio, M. G. Manzato, and S. Carlos, "A Sentiment-Based Item Description Approach for kNN Collaborative Filtering," *SAC 2015 Proc. 30th Annu. ACM Symp. Appl. Comput.*, pp. 1060–1065, 2015.

[29] X. Ma, X. Lei, G. Zhao, and X. Qian, "Rating prediction by exploring user's preference and sentiment," *Multimed. Tools Appl.*, vol. 77, no. 6, pp. 6425–6444, 2018.

[30] X. Chen, Z. Qin, Y. Zhang, and T. Xu, "Learning to rank features for recommendation over multiple categories," pp. 305–314, 2016.

[31] C. Yang, X. Yu, Y. Liu, Y. Nie, and Y. Wang, "Neurocomputing Collaborative fi ltering with weighted opinion aspects," vol. 210, pp. 185–196, 2016.

[32] Y. Wang, Y. Liu, and X. Yu, "Collaborative filtering with aspect-based opinion mining: A tensor factorization approach," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1152–1157, 2012.

[33] K. Bauman, B. Liu, and A. Tuzhilin, "Aspect Based Recommendations : Recommending Items with the Most Valuable Aspects Based on User Reviews."

[34] L. Qiu, S. Gao, W. Cheng, and J. Guo, "Aspect-based latent factor model by integrating ratings and reviews for recommender system," *Knowledge-Based Syst.*, vol. 110, pp. 233–243, 2016.

[35] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional Matrix Factorization for Document Context-Aware Recommendation," pp. 233–240, 2016.

[36] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," *17th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, pp. 618–626, 2011.