Effective Rating Prediction using an Attention-based User Review Sentiment Model

Xi Wang[1], Iadh Ounis[2], Craig Macdonald[2]

Unversity of Glasgow, Glasgow, UK
[1]x.wang.6@research.gla.ac.uk
[2]{firstname.secondname}@glasgow.gla.ac.uk

Abstract. We propose a new sentiment information-based attention mechanism that helps to identify user reviews that are more likely to enhance the accuracy of a rating prediction model. We hypothesise that highly polarised reviews (strongly positive or negative) are better indicators of the users' preferences and that this sentiment polarity information helps to identify the usefulness of reviews. Hence, we introduce a novel neural network rating prediction model, called SentiAttn, which includes both the proposed sentiment attention mechanism as well as a global attention mechanism that captures the importance of different parts of the reviews. We show how the concatenation of the positive and negative users' and items' reviews as input to SentiAttn, results in different architectures with various channels. We investigate if we can improve the performance of SentiAttn by fine-tuning different channel setups. We examine the performance of SentiAttn on two well-known datasets from Yelp and Amazon. Our results show that SentiAttn significantly outperforms a classical approach and four state-of-the-art rating prediction models. Moreover, we show the advantages of using the sentiment attention mechanism in the rating prediction task and its effectiveness in addressing the cold-start problem.

1 Introduction

Rating prediction is a classical recommendation task [22], where the recommendation system aims to accurately predict the user rating of an unseen item, so as to better estimate which items to recommend to a user. The predictions are typically based on the existing ratings by users. The rating prediction task remains a challenging and open problem. Indeed, the effectiveness of existing rating prediction-based recommendation systems is still limited, suffering from various types of challenges, including accuracy, data sparsity and the cold-start problem [4, 32]. Therefore, many approaches have been proposed to leverage user reviews [16, 28] – including the sentiment of the reviews [11, 17] – to improve the rating prediction accuracy. Users' reviews can enrich both user and item representations, while sentiment information is often useful for extracting user preferences [11]. However, not all reviews are useful to enhance the rating prediction performance, since they may convey varying actionable information about the users' preferences [1]. Recently, a number of approaches have made use of the attention mechanism to estimate the usefulness of reviews [1, 24]. Attention mechanism focuses on the parts of review content that contribute to the

rating prediction. While these existing approaches demonstrate that the attention mechanism can improve the rating prediction performance, they (i.e. [1, 24]) do not leverage the sentiment information actually captured by the reviews.

Given the effectiveness of sentiment information in extracting user preferences, we hypothesise that sentiment information should also be used in estimating the usefulness of reviews, so as to further improve the rating prediction performance. Indeed, reviews with a clear polarised sentiment (i.e. positive or negative) typically convey richer information about items and are more likely to influence the users' decision making when interacting with the corresponding items [11]. In the literature, several approaches focused on leveraging the sentiment information as an additional feature to address the rating prediction task [6, 28], while ignoring the potential relationship between the sentiment polarity and the usefulness of reviews in users' decision making. In this study, we propose instead to directly leverage the sentiment scores of reviews to address the aforementioned limitation. Inspired by Wang et al. [28], the sentiment score of a review is estimated as the probability of the review having a clear positive or negative polarity as determined by a sentiment classifier. These sentiment scores are then used in a customised attention mechanism to identify informative reviews with rich user preferences. Hence, SentiAttn assumes that reviews with clearly pronounced user preferences are useful for effective rating prediction. In addition, SentiAttn adds another attention mechanism (i.e. global attention [14]) to capture and model the importance of the parts of reviews that are likely to enhance the rating prediction performances. On the other hand, previous works on *neural architecture search* [7, 13] showed that fine-tuning a neural model architecture could have a marked positive impact on the model's performance. To leverage the advantage of fine-tuning the neural models' architectures, in this paper, we propose a strategy where we first concatenate the users' and items' positive and negative reviews as input to SentiAttn, resulting in different SentiAttn architectures with various number of channels (e.g. if we concatenate all reviews for both users and items, then this leads to a single channel-based SentiAttn model). Next, we fine tune the architecture variants of our proposed SentiAttn model with different channel setups on the validation sets of two datasets from Yelp and Amazon, so as to optimise the performances of SentiAttn on different datasets.

Our contributions in this paper are as follows: **(1)** We propose a new sentiment information-based attention mechanism, which weights the usefulness of reviews by their corresponding sentiment scores. These scores reflect the user preferences since they convey a clear sentiment. To the best of our knowledge, this is the first model to directly encode sentiment information for identifying review usefulness in rating prediction; **(2)** We examine the impact of the resulting SentiAttn model architectures using different channels on their rating prediction performances. This examination is conducted by fine tuning the architectures on the validation sets of the Yelp and Amazon datasets; **(3)** We show that SentiAttn achieves a significantly better rating prediction accuracy than one classical (NMF [10]) and four existing state-of-the-art rating prediction models (namely, ConvMF [8], DeepCoNN [32], D-Attn [24] and NARRE [1]) over two datasets; **(4)** We show that SentiAttn is particularly effective in addressing the cold-start problem in comparison to the existing baselines.

2 Related Work

In this section, we briefly discuss two bodies of related work.

Review-based Rating Prediction: Several studies have exploited user reviews to improve rating predictions [27, 29, 32]. Many earlier studies used topic modelling techniques (e.g. Latent Dirichlet Allocation (LDA)) to model user reviews [12, 17]. However, with the emergence of word embedding [18] techniques, it has been shown that rating prediction models based on word embeddings can outperform such topic modelling-based approaches. For example, Zheng et al. [32] proposed a deep learning model that initialised both the user and item matrices with word embeddings before jointly modelling the users and items to make rating predictions. However, not all user reviews provide useful information to enhance the rating prediction performance. With this in mind, some previous studies, e.g. [1, 24], have applied an attention mechanism to identify useful reviews to improve rating predictions. Seo et al. [24] developed two attention mechanisms to learn review usefulness, i.e. local and global attention mechanisms to generate explainable and better-learned review representation latent vectors. Chen et al. [1] initialised user/item latent vectors with review embedding vectors and the corresponding identification information. The authors used a typical attention mechanism to model the latent vectors. However, although the attention mechanism can be effective for modelling the usefulness of reviews, the attention mechanism does not consider the sentiment information of reviews. Sentiment information has been shown to enhance the rating predictions in many studies [15, 26, 31] (we discuss in the remainder of this section). In this paper, unlike prior work, we propose to directly leverage the sentiment information within a customised attention mechanism when addressing the rating prediction task.

Sentiment-enhanced Recommendation: Recently, sentiment-enhanced recommendation approaches have benefited from deep-learning techniques. For example, Wang et al. [28] examined the performance of different state-of-the-art sentiment classification approaches (e.g. CNN [9] and LSTM [5]) to generate review sentiment polarity scores, and then validated the usefulness of sentiment information by replacing user ratings with such sentiment scores for making recommendations. Chen et al. [1] used a convolution operation to convert reviews into latent vectors to represent review sentiment information, thereby enhancing the rating prediction performance. These studies validated the usefulness of using sentiment information to identify user preferences in user reviews. Therefore, we postulate that sentiment information can also be useful for identifying useful reviews. To the best of our knowledge, our proposed SentiAttn model is the first sentiment-enhanced recommendation approach to use sentiment information to weight review usefulness in an attention neural network architecture.

3 The SentiAttn Model

In this section, we first state the rating prediction task and the notations used. Next, we illustrate the motivation of using sentiment information to identify useful reviews and describe our proposed SentiAttn rating prediction model.

Table 1. Review examples with sentiment information.

Positive and High Sentiment Score	
Rating: 5 **Sentiment Score:** 0.9726 **Category:** grocery and gourmet food	**Review 1:** This beverage is so delicious. I would like to order more in the future. I drink it to relax.
Positive but Low Sentiment Score	
Rating: 5 **Sentiment Score:** 0.1783 **Category:** grocery and gourmet food	**Review 2:** My husband insists on making his own yogurt and won't use any other starter. This assures the same consistency month after month.

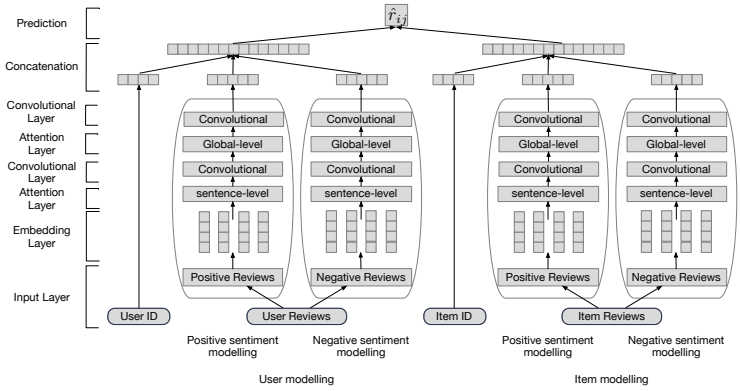


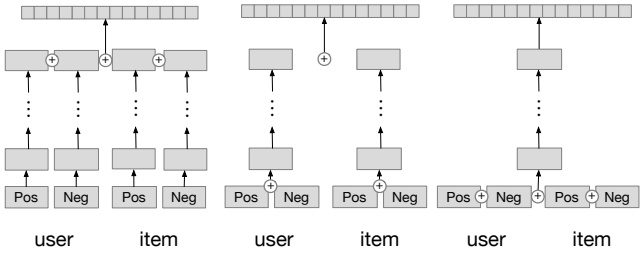
Fig. 1. The architecture of the SentiAttn model

3.1 Task Definition

The rating prediction task aims to predict the ratings of unseen items. Consider a set of users U and items I (of size m and n, respectively). We also have the one-hot embedding vectors E_U and E_I, which map users and items to different randomly initialised vectors. User ratings can be encoded in a rating matrix $R \in \mathbb{R}^{m \times n}$, where entries $r_{u,i} \in R$ represent the previously observed ratings with a range from 1 to 5. In rating prediction, we aim to accurately predict the rating $r_{u,i}$ of an unseen item i for user u. Moreover, each rating $r_{u,i}$ is associated with a textual review $c_{u,i}$. As discussed in Section 1, for each review $c_{u,i}$, we also estimate a corresponding sentiment score $s_{u,i}$, which indicates the probability of the review being polarised, i.e. being strongly positive or strongly negative.

3.2 Review Sentiment Information Analysis

To motivate the use of sentiment information in identifying useful reviews, we provide two illustrative review examples in Table 1. The sentiment score corresponds to the probability of a given review being polarised, as further explained in Section 3.3. These two reviews are both positive and 5 star-rated. However, when we compare these two reviews, Review 1 better conveys the user's preferences, while Review 2 simply describes a personal event, making it hard for the model to capture the user's preferences. Therefore, Review 1 is deemed more useful than Review 2. In particular, the sentiment scores of Reviews 1 and 2 clearly mirror their usefulness difference (Review 1 is scored 0.9726 as being strongly



(a) Four Channels (b) Two Channels (c) One Channel

Fig. 2. Architectures of the original SentiAttn four channels-based model and its variants (i.e. one and two channels-based).

positive while Review 2 is scored 0.1783 only). Therefore, we propose to leverage the relationship between the sentiment scores and the usefulness of reviews in SentiAttn. Our model identifies useful reviews via a novel sentiment information-based attention mechanism to improve the rating prediction performance.

3.3 Model Architecture

To encode the review usefulness information through their sentiment scores, SentiAttn first uses a customised sentiment attention mechanism to embed the review usefulness information. Next, it integrates another global attention mechanism [14] to capture the parts of reviews that are likely to enhance the rating prediction performance. The architecture of SentiAttn (Figure 1) comprises eight layers from the input to the rating prediction layer, described further below:

Input & Embedding Layers: In the input layer, users are represented by the reviews they have posted for items while items are represented by the reviews given by users. In particular, the input layer groups reviews into positive and negative reviews according to their corresponding rating values. If the rating $r_{u,i} \geq 4$, the review $c_{u,i}$ is positive, else, if the rating $r_{u,i} \leq 2$, the review $c_{u,i}$ is negative. A review $c_{u,i}$ with a rating of $r_{u,i} = 3$ or with no provided rating is classified as positive or negative according to a CNN-based binary sentiment classifier (described further in Section 4.3). Therefore, our SentiAttn model can be divided into four parallel networks (i.e. four channels), which model the positive and negative reviews for users and items. The architecture of our SentiAttn model is flexible and can possibly have two additional variants (i.e. one channel or two channels-based). As shown in Figure 2, instead of modelling the polarised reviews for user and items individually, we can concatenate all the reviews for the user or the item, resulting in the two channels-based SentiAttn model variant. Moreover, if we further concatenate all the reviews of the user and the item together, we can obtain the one channel-based SentiAttn model variant. In particular, for each resulting channel, its review modelling pipeline remains the same as each individual channel depicted in Figure 1. It is of note that the one channel-based model variant can only be leveraged by a model that uses a value mapping-based predictor (e.g. the factorisation machine and the multilayer perceptron) and not an interaction-based predictor (e.g. the dot product function), which needs at least two inputs. In this paper, we investigate which SentiAttn

model variant exhibits the best performances on the used datasets. Next, following [1, 32], in the embedding layer, we convert the reviews text into embedding vectors, denoted as X, which are then given as input to the next layer.

Sentiment Attention Layer: In this layer, we customise a sentiment attention mechanism to encode the usefulness of reviews. Our sentiment attention mechanism is inspired by the dot-product attention function [25], which learns the importance (weight) of different embedding vectors. Then, it multiplies the resulting weighting vectors with the initial word embedding vectors to apply the attention mechanism. Unlike the dot-product attention function, our sentiment attention mechanism obtains the weighting vectors from the sentiment scores of the reviews. These sentiment scores can enrich the user's information and might be helpful in addressing the cold start problem. First, the reviews have been labelled as positive or negative in the previous layer. After that, we process these reviews with a given sentiment classifier and obtain the corresponding probabilities of the positive reviews being positive or the negative reviews being negative (denoted as $p_{u,i}$, which naturally ranges from 0 to 1). The corresponding sentiment scores for the positive reviews are $s_{u,i} = p_{u,i}$. Conversely, we use $s_{u,i} = 1 - p_{u,i}$ for the negative reviews. Hence, the sentiment score indicates the probability of a given review being polarised (positive or negative), and a review is deemed more useful if its sentiment score is closer to 1. Next, with a given review embedding vector X, and its sentiment score vector S, the converted vector X' is calculated as $X' = ((SX^T)^T \oplus X)$, where $\oplus$ is a residual connector.

Convolutional Layer Our SentiAttn model applies the convolution operation, as in [24, 32], on the latent vector X' with g neurons to generate feature vectors for the next layer. Each neuron applies the convolution operation to a sliding window t over latent vectors with width T. The convolution operation of neuron e is obtained as follows: $z_e = f(X'_{1:T} * K_e + b_e)$, where $f(.)$ indicates an activation function to filter the output of the convolution operation, $*$ is the convolution operator of neuron e on the corresponding window of vectors and b_e is a bias parameter [8]. After applying the convolution operation, we apply the max pooling function over the output feature vectors, denoted as Z, and obtain the resulting vector o for each neuron (i.e. $o = max(z_1, z_2, ..., z_e^{T-t+1})$). Next, the outputs of the g neurons are concatenated together into the latent vector X_c.

Global Attention Layer: Apart from the proposed sentiment attention layer, we also use the global attention mechanism from [14]. Accordingly, we add the global attention layer to SentiAttn to model the parts of review content that are likely to contribute to enhancing the rating prediction performances. In particular, the global attention mechanism considers all review embeddings as input and calculates the global attention score vector G of the embedding input X_c: $G = \text{SoftMax}(W_g X_c)$. The embedding input X_c is then further weighted by the global attention score vector G as $X_g = (GX_c^T)^T$. After the global attention layer, we add another convolutional layer, which is the same as the one above the sentiment attention layer, to process the review embeddings. We use the outputs from the convolutional layer as the final latent feature vectors for each channel.

Concatenation & Prediction Layer: In the concatenation layer, we concatenate the latent vectors from two groups of inputs: (1) the resulting latent feature vector from the last convolutional layer of the review modelling channels; (2) the one-hot embedding vectors of each user and item. We refer to the concatenated vector as o. Next, in the prediction layer, we use a two-order factorisation ma-

chine [20] as the rating predictor, which is capable of capturing the patterns in data to improve the model's performance [30]. This predictor has also been widely used in the literature to address the rating prediction task [2, 21, 3]. Each predicted rating $\hat{r}_{u,i}$ is calculated as follows: $\hat{r}_{u,i} = w_0 + b_u + b_i + (\sum_{j=1}^{|o|} w_j o_j) + (\sum_{j=1}^{|o|} \sum_{k=j+1}^{|o|} o_j o_k \mathbf{w}_{j,k})$. This equation has five summands: w_0 is the global bias parameter [20]; next, b_u and b_i correspond to the bias parameters for user u and item i, respectively; in the fourth summand, w_j models the weight of the j_{th} variable in o; the final summand models the interactions between pairs of variable vectors o_j and o_k in o, weighted by a factorised parameter $\mathbf{w}_{j,k} \approx \langle \mathbf{v}_j, \mathbf{v}_k \rangle$ as in [20]. SentiAttn is trained by minimising the prediction error between the true rating value $r_{u,i}$ and the predicted rating value $\hat{r}_{u,i}$ with the MSE function.

4 Experimental Setup

We now examine the performance of SentiAttn through experiments on two real-world datasets, in comparison to a number of classical and state-of-the-art rating prediction models. In particular, we address three research questions: **RQ1:** Which architecture variant of the SentiAttn model (based on 1, 2 or 4 channels) performs the best on the two used datasets? **RQ2:** Does SentiAttn outperform other state-of-the-art models in addressing the rating prediction task and how much does it benefit from (i) the proposed sentiment attention mechanism and (ii) the global attention mechanism? **RQ3:** Does SentiAttn outperform the existing baselines when making rating predictions for cold-start users?

4.1 Datasets

To perform our experiments, we use two popular real-world datasets [17, 24]: (i) a Yelp[1] dataset, and (ii) an Amazon Product dataset[2]. The Yelp dataset contains a large number of reviews on venues located in Phoenix, USA. The Amazon dataset contains reviews on products among six categories[3]. The statistics of these two datasets are in Table 2. Following a common setup [1, 32], these two datasets are randomly divided into 80% training, 10% validation and 10% testing sets. Moreover, we follow [6] and denote those users with less than 5 reviews in the training dataset as the cold-start users. Table 2 shows that the Yelp dataset is more sparse (i.e. has a lower density[4]) than the Amazon dataset. This observation suggests that the data sparsity's influence might be amplified in a given model's performance when experimenting with the Yelp dataset. Moreover, as per the positive rating percentages in Table 2, most user reviews are positive in both datasets.

4.2 Baselines and Evaluation Metrics

We compare our SentiAttn model[5] to the following 5 baselines: **(1) NMF** [10]: NMF is a widely used classical baseline, which characterises users and items with

[1] https://kaggle.com/c/yelp-recsys-2013 [2] http://jmcauley.ucsd.edu/data/amazon/ [3] 'amazon instant video', 'automotive', 'grocery and gourmet food', 'musical instruments', 'office products' and 'patio lawn and garden' [4] % Density = #interactions / (#users × #items)) [5] Our source code is available at: https://github.com/wangxieric/SentiAttn.

Table 2. Statistics of datasets.

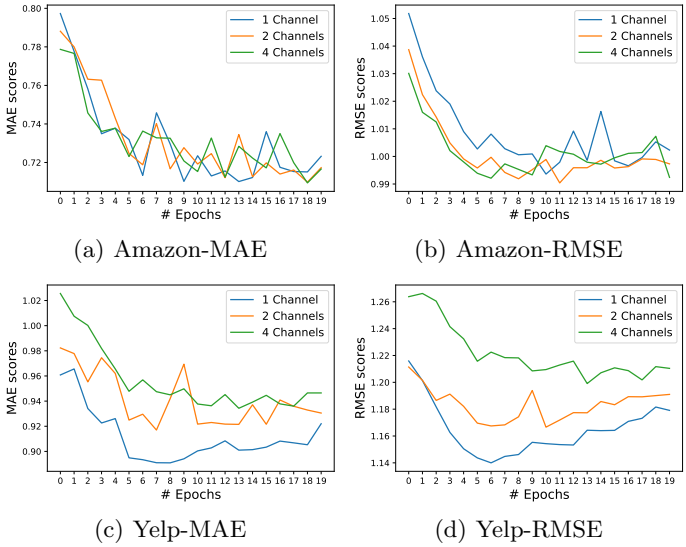
Dataset	#Users	#Cold-start Users	#Items	#Reviews	% Density	% Positive ratings
Yelp	45,981	33,306	11,537	229,907	0.043	67.88
Amazon	26,010	7,874	16,514	285,644	0.066	81.24

their rating pattern-based latent vectors. **(2) ConvMF** [8]: ConvMF extends the latent feature vectors in NMF with the embedding vector of reviews. **(3) DeepCoNN [32]**: DeepCoNN jointly models reviews to characterise users and items with latent vectors. This approach has been widely used as a strong review-based rating prediction model. **(4) D-Attn [24]**: D-Attn is another review-based rating prediction model that includes two global and local attention mechanisms. D-Attn is another review-based rating prediction model. It includes two global and local attention mechanisms, to improve the explainability and rating prediction accuracy of a rating prediction model. **(5) NARRE [1]**: NARRE is a recent state-of-the-art attention-based rating prediction model. It weights reviews by its learned review usefulness scores. These scores are estimated through the use of an attention mechanism. Moreover, we examine the effectiveness of using our proposed sentiment attention mechanism in comparison to three further baselines derived from SentiAttn as follows: One baseline removes both attention layers in the SentiAttn model (denoted by '**Basic**'), while '**+Glb**' and '**+Sent**' add the global attention layer and the sentiment attention layer to the Basic model, respectively. As for the evaluation metrics, we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to measure the performances of SentiAttn and the baselines, which are the commonly used metrics to evaluate rating prediction models [1, 11, 32]. In order to examine the statistical significance of the models' performances, we leverage both the paired t-test, with a significance level of $p < 0.05$, and the post-hoc Tukey Honest Significant Difference (HSD) [23] test at $p < 0.05$ to account for the multiple comparisons with the t-tests[6].

4.3 Model Setting

In the input layer, we use an existing CNN-based binary sentiment classifier to group reviews into positive and negative reviews, which has been shown to have a strong accuracy ($> 95\%$) for sentiment classification [28]. Other sentiment classifiers could have been used, but the investigation of such classifiers is beyond the scope of this paper. For the used CNN-based binary sentiment classifier, we follow the same experimental setup as [28] and train it on 50,000 positive and 50,000 negative review instances that were sampled from a separate dataset, namely the Yelp Challenge Round 12 dataset[7]. Moreover, the classifier provides each review with its probability $p_{u,i}$ of carrying a strong polarised sentiment, so as to generate a sentiment score $s_{u,i}$ in the sentiment attention layer, as explained in Section 3.3. Next, in the embedding layer, we use the pre-trained GloVe [19] word embedding dictionary[8], following [28], and map each word into an embed-

[6] Since RMSE is a non-linear aggregation of squared absolute errors, a significance test cannot be conducted with this metric. [7] https://www.yelp.com/dataset/challenge
[8] We also apply the pre-trained GloVe word embeddings within the baseline approaches, which ensures fair performance comparisons between approaches.



(a) Amazon-MAE (b) Amazon-RMSE

(c) Yelp-MAE (d) Yelp-RMSE

Fig. 3. Validation performances of the SentiAttn variants with different #channels.

ding vector with 100 dimensions. In the convolutional layer, following [24], we set the kernel size to 100 and the activation function to ReLU. In particular, we use the Adam optimiser with a 10^{-4} learning rate. Moreover, it is of note that, to answer RQ1, which investigates the performances of the considered SentiAttn architecture variants, we conduct experiments on the validation sets from the Yelp and Amazon datasets to select the best SentiAttn architecture – thereby mimicking the use of the validation sets for model selection.

5 Results

Next, we report and analyse our obtained results:

Performances of Architecture Variants (RQ1). We investigate which of the SentiAttn model architecture variants leads to the best rating prediction performances. In Figure 3, we report the performances of the three considered architecture variants of the proposed SentiAttn model (namely the one, two and four channels-based SentiAttn architectures). Since we are using the MAE and RMSE error-based evaluation metrics, the lower the metrics' values, the higher is the model's rating prediction performance. First, we compare the performances of the SentiAttn variants on the Amazon dataset in Figure 3(a) and Figure 3(b). For both MAE and RMSE, the three variants show similar trends and performances and are overall comparable. However, on the Yelp dataset, Figure 3(c) and Figure 3(d) show that the SentiAttn model with one-channel consistently outperforms both the original SentiAttn model with four channels and the two channels-based variant. Using the six components of the Amazon dataset corresponding to each of the categories in Footnote 3, we conducted a further analysis to examine

Table 3. Rating prediction accuracy; * denotes a significant difference in MAE with SentiAttn with respect to both the paired t-test and the Tukey HSD test, $p < 0.05$.

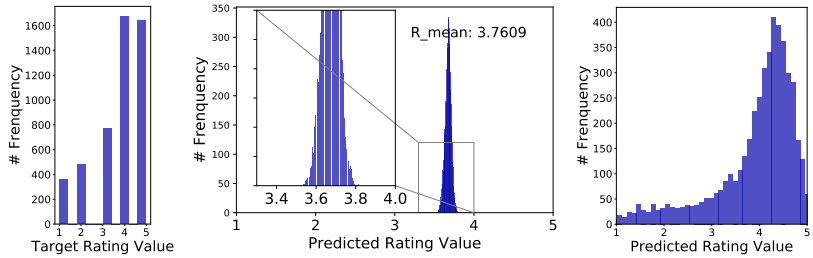
All Users					**Cold-Start Users**				
	Yelp Dataset		**Amazon Dataset**			**Yelp Dataset**		**Amazon Dataset**	
	MAE	RMSE	MAE	RMSE		MAE	RMSE	MAE	RMSE
NMF	0.9866*	1.2630	0.8240*	1.0881	NMF	1.1690*	1.5025	0.9040*	1.1843
ConvMF	0.9748*	1.2329	0.7964*	1.0371	ConvMF	1.0785*	1.3812	0.8565*	1.1154
DeepCoNN	0.9247*	1.1885	0.7233*	0.9929	DeepCoNN	1.0462*	1.3506	0.7882*	1.0749
D-Attn	1.0040*	1.2106	0.8316*	1.0627	D-Attn	1.0154*	1.2394	0.8738*	1.1029
NARRE	0.9163*	1.1781	0.7065*	0.9783	NARRE	1.0289*	1.3481	0.7613*	1.0587
Basic	0.9084*	1.1769	0.7060*	0.9769	Basic	1.0003*	1.3602	0.7451*	1.0520
+Glb	0.8947	1.1734	0.6960	0.9723	+Glb	0.9867	1.3544	0.7253	1.0460
+Sent	0.8932	1.1476	0.6957	0.9685	+Sent	0.9817	1.2408	0.7190	1.0375
SentiAttn	**0.8888**	**1.1463**	**0.6841**	**0.9668**	SentiAttn	**0.9736**	**1.2327**	**0.7090**	**1.0273**

the correlation between a given dataset's statistics and the performances of SentiAttn with different number of channels. The results of our analysis suggest that the higher the density of interactions in a dataset, the better a variant of SentiAttn with a larger number of channels performs[9]. Overall, in answer to RQ1, we conclude that a higher number of channels is preferred for datasets with high density of interactions. As per these results, we select the overall best variant model, namely the one channel-based SentiAttn model for the remaining experiments.

Comparison to the Baselines (RQ2). Table 3 presents the rating prediction errors of both the baseline models and SentiAttn. First, in the obtained results for both the Yelp and Amazon datasets, SentiAttn significantly outperforms the baselines according to both the paired t-test and the Tukey HSD test. In particular, while D-Attn, NARRE and SentiAttn all use an attention mechanism to weight reviews with their estimated usefulness, our SentiAttn model, which relies on a novel sentiment attention and a global attention mechanism returns significantly smaller prediction errors in comparison to competitive baselines on both the Yelp and Amazon datasets. We also evaluate the usefulness of the global attention layer and the proposed sentiment attention layer in SentiAttn by comparing the performances of SentiAttn with the Basic, +Glb, and +Sent models (introduced in Section 4.2). Table 3 shows that SentiAttn significantly (according to both the paired t-test and the Tukey HSD test, $p < 0.05$) outperforms the Basic model on both used datasets, which demonstrates the effectiveness of using the attention mechanisms. Moreover, the results show that the sentiment attention mechanism outperforms the global attention mechanism since it results in lower MAE and RMSE scores (0.8932 vs. 0.8947 (MAE) and 1.1476 vs. 1.1734 (RMSE) for +Sent vs. +Glb in Table 3). In particular, we observe that the sentiment attention mechanism is especially effective in decreasing the variance of the rating prediction errors. Indeed, +Sent outperforms both Basic and +Glb providing lower RMSE scores with wide margins on both used datasets.

To further examine the effectiveness of the proposed sentiment attention mechanism, we conducted further analysis on the results of both datasets. We averaged the sentiment scores of the reviews posted by a given user. We group users into two groups: 'sentiment-polarised' vs. 'sentiment-neutral' users. On the Yelp dataset, the sentiment-polarised users have average review scores > 0.88,

[9] Due to the page limit, we do not include these experimental results in the paper.



(a) Target (b) D-Attn (c) SentiAttn

Fig. 4. Cold-start user rating prediction performance comparison (D-Attn vs. SentiAttn) on the Yelp dataset.

while the sentiment-neutral users have average scores ≤ 0.88[10]. This leads to 25155 sentiment-polarised and 20826 sentiment-neutral users. We would expect the proposed sentiment attention mechanism to mostly benefit the sentiment-polarised users since these users have more reviews that clearly convey their preferences. Next, we compare performances between the global attention '+Glb' and the sentiment attention '+Sent' models. The results on Yelp show that '+Sent' significantly outperforms '+Glb' for 51.6% of the sentiment-polarised users and 43.3% of the sentiment-neutral users (using a paired t-test on users with the MAE metric). Contrastingly, '+Glb' significantly (paired t-test) outperforms '+Sent' for 38.3% of the sentiment-polarised users and 48.1% of the sentiment-neutral users. These results indicate that the proposed sentiment attention mechanism can indeed help the sentiment-polarised users, but does not exhibit better performances than using the global attention mechanism if most of the users' reviews do not contain highly polarised reviews (i.e. sentiment neutral users). We observed similar conclusions on the Amazon dataset. To answer RQ2, we conclude that the obtained results empirically validate the effectiveness of our SentiAttn model in addressing the rating prediction task in comparison to strong baseline models. The results also show the effectiveness of using the sentiment attention mechanism – which weights the review input according to the corresponding review sentiment scores – thereby outperforming the global attention mechanism.

Cold-Start Users (RQ3). We now evaluate the rating prediction performance of SentiAttn on cold-start users. As introduced in Section 4.1, we consider users in the training dataset with less than 5 reviews as cold-start users. Table 3 provides the rating prediction performances of SentiAttn and the various baseline models on both the Yelp and Amazon datasets for cold-start users. The results show that our SentiAttn model obtains a good cold-start performance by significantly outperforming all the strong baseline approaches from the literature on the Yelp and Amazon datasets. Comparing the rating prediction results in Table 3 on the Yelp dataset, we note that as expected from the statistics of this dataset, the rating prediction performances of all models suffer from the cold-start problem. However, the cold-start problem appears to have only a small negative influence on the D-Attn model. To investigate the reasons behind the relative effectiveness of D-Attn in addressing the cold-start problem, we plot the

[10] The threshold (0.88) is the mean value of the reviews' sentiment score distribution.

predicted rating value frequency distribution of the cold-start users on the Yelp dataset using both D-Attn and our SentiAttn model in Figures 4(b) and 4(c), respectively. These distributions are compared with the target rating distribution in Figure 4(a). In Figure 4(b) of the D-Attn model, the predicted rating values shrink between around 3.55 and 3.80, which are all close to the average of the target rating value (i.e. $\overline{R} = 3.7609$). This distribution shows that the performance of D-Attn is less reliable in distinguishing the actual user preferences. On the contrary, in Figure 4(c), the predicted rating value frequency distribution of our SentiAttn model ranges from 0 to 5 and its shape better aligns with the actual rating distribution of the Yelp dataset in Figure 4(a).

We also compare the impact of using two attention mechanisms in addressing the cold-start problem. According to the results in Table 3, our sentiment attention mechanism outperforms the global attention mechanism in improving the rating prediction accuracy of the Basic model (e.g. 1.0003 $\rightarrow$ 0.9817 vs. 1.0003 $\rightarrow$ 0.9867 on the Yelp dataset) and lowers the variances of the rating prediction errors with a wider margin. For example, on the Yelp dataset, the Basic model benefits from using the global attention mechanism and lowers the RMSE score from 1.3602 to 1.3544. However, when applying the sentiment attention mechanism, the RMSE score of the Basic model is decreased from 1.3602 to 1.2408, indicating a higher improvement than when applying the global attention mechanism. Therefore, in answer to RQ3, our SentiAttn model is particularly effective for the cold-start users compared with the five strong baselines from the literature. Our sentiment attention mechanism also shows its usefulness in improving the rating prediction accuracy, especially lowering the variance of the rating prediction errors for cold-start users. In particular, SentiAttn is more reliable than D-Attn in identifying user preferences, as illustrated by the predicted rating distributions.

6 Conclusions

In this paper, we proposed the SentiAttn model, which leverages user reviews as input and deploys a new sentiment attention mechanism. The latter encodes user preferences by initialising the weights of different reviews with their sentiment scores. SentiAttn also integrates a global attention mechanism, which captures the importance of different parts of the review's content. We investigated the effect of using different architecture variants for our SentiAttn model and concluded that a higher number of channels is preferred for datasets with a higher density of interactions. Our results on two real-world datasets showed that SentiAttn significantly and consistently outperformed four existing state-of-the-art rating prediction models. Moreover, we demonstrated the effectiveness of the proposed sentiment attention layer within SentiAttn. We showed that it outperforms the global attention layer in improving the rating prediction accuracy, resulting in a lower variance of the rating prediction errors. Furthermore, we showed that SentiAttn provides a significantly effective rating prediction accuracy and a reliable indication of user preferences for cold-start users. As future work, we plan to consider other review properties (e.g. such as review age) as additional features within SentiAttn to more accurately measure the usefulness of reviews.

References

1. Chen, C., Zhang, M., Liu, Y., Ma, S.: Neural attentional rating regression with review-level explanations. In: Proc. of WWW (2018)
2. Chen, L., Liu, Y., Zheng, Z., Yu, P.: Heterogeneous neural attentive factorization machine for rating prediction. In: Proc. of CIKM (2018)
3. Cheng, C., Xia, F., Zhang, T., King, I., Lyu, M.R.: Gradient boosting factorization machines. In: Proc. of RecSys (2014)
4. Davagdorj, K., Park, K.H., Ryu, K.H.: A collaborative filtering recommendation system for rating prediction. In: Advances in Intelligent Information Hiding and Multimedia Signal Processing, pp. 265–271. Springer (2020)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8) (1997)
6. Hyun, D., Park, C., Yang, M.C., Song, I., Lee, J.T., Yu, H.: Review sentiment-guided scalable deep recommender system. In: Proc. of SIGIR (2018)
7. Jiang, Y., Hu, C., Xiao, T., Zhang, C., Zhu, J.: Improved differentiable architecture search for language modeling and named entity recognition. In: Proc. of EMNLP (2019)
8. Kim, D., Park, C., Oh, J., Lee, S., Yu, H.: Convolutional matrix factorization for document context-aware recommendation. In: Proc. of RecSys (2016)
9. Kim, Y.: Convolutional neural networks for sentence classification. In: Proc. of EMNLP (2014)
10. Koren, Y., Bell, R.M., Volinsky, C.: Matrix factorization techniques for recommender systems. IEEE Computer **42**(8) (2009)
11. Lei, X., Qian, X., Zhao, G.: Rating prediction based on social sentiment from textual reviews. IEEE transactions on multimedia **18**(9), 1910–1921 (2016)
12. Ling, G., Lyu, M.R., King, I.: Ratings meet reviews, a combined approach to recommend. In: Proc. of RecSys (2014)
13. Liu, H., Simonyan, K., Yang, Y.: DARTS: differentiable architecture search. In: Proc. of ICLR (2019)
14. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proc. of EMNLP (2015)
15. Ma, X., Lei, X., Zhao, G., Qian, X.: Rating prediction by exploring user's preference and sentiment. Multimedia Tools and Applications **77**(6) (2018)
16. Manotumruksa, J., Macdonald, C., Ounis, I.: Regularising factorised models for venue recommendation using friends and their comments. In: Proc. of CIKM (2016)
17. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proc. of RecSys (2013)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proc. of NeurIPS (2013)
19. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: Proc. of EMNLP (2014)
20. Rendle, S.: Factorization machines with libfm. ACM Transactions on Intelligent Systems and Technology (TIST) **3**(3) (2012)
21. Rendle, S., Gantner, Z., Freudenthaler, C., Schmidt-Thieme, L.: Fast context-aware recommendations with factorization machines. In: Proc. of SIGIR (2011)
22. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Recommender Systems Handbook. Springer (2011)

23. Sakai, T.: Laboratory experiments in information retrieval. The Information Retrieval Series **40** (2018)
24. Seo, S., Huang, J., Yang, H., Liu, Y.: Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In: Proc. of RecSys (2017)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. of NeurIPS (2017)
26. Wang, H., Fu, Y., Wang, Q., Yin, H., Du, C., Xiong, H.: A location-sentiment-aware recommender system for both home-town and out-of-town users. In: Proc of SIGKDD (2017)
27. Wang, J., Sun, C., Li, S., Wang, J., Si, L., Zhang, M., Liu, X., Zhou, G.: Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning. In: Proc. of EMNLP (2019)
28. Wang, X., Ounis, I., Macdonald, C.: Comparison of sentiment analysis and user ratings in venue recommendation. In: Proc. of ECIR (2019)
29. Wu, C., Wu, F., Qi, T., Ge, S., Huang, Y., Xie, X.: Reviews meet graphs: Enhancing user and item representations for recommendation with hierarchical attentive graph neural network. In: Proc. of EMNLP (2019)
30. Zhang, W., Du, T., Wang, J.: Deep learning over multi-field categorical data. In: Proc. of ECIR (2016)
31. Zhao, K., Cong, G., Yuan, Q., Zhu, K.Q.: Sar: A sentiment-aspect-region model for user preference analysis in geo-tagged reviews. In: Proc. of ICDE (2015)
32. Zheng, L., Noroozi, V., Yu, P.S.: Joint deep modeling of users and items using reviews for recommendation. In: Proc. of WSDM (2017)