Document-level Multi-aspect Sentiment Classification by Jointly Modeling Users, Aspects, and Overall Ratings

Junjie Li[1,2], **Haitong Yang**[3] **and Chengqing Zong**[1,2,4]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] School of Computer, Central China Normal University, Wuhan 430079, China
[4] CAS Center for Excellence in Brain Science and Intelligence Technology
{junjie.li, cqzong}@nlpr.ia.ac.cn, htyang@mail.ccnu.edu.cn

Abstract

Document-level multi-aspect sentiment classification aims to predict user's sentiment polarities for different aspects of a product in a review. Existing approaches mainly focus on text information. However, the authors (i.e. users) and overall ratings of reviews are ignored, both of which are proved to be significant on interpreting the sentiments of different aspects in this paper. Therefore, we propose a model called Hierarchical User Aspect Rating Network (HUARN) to consider user preference and overall ratings jointly. Specifically, HUARN adopts a hierarchical architecture to encode word, sentence, and document level information. Then, user attention and aspect attention are introduced into building sentence and document level representation. The document representation is combined with user and overall rating information to predict aspect ratings of a review. Diverse aspects are treated differently and a multi-task framework is adopted. Empirical results on two real-world datasets show that HUARN achieves state-of-the-art performances.

1 Introduction

The ever-increasing popularity of online consumer review platforms, such as Tripadvisor[1] and Yelp[2], has led to large amounts of online reviews that are often too numerous for users to analyze. Consequently, there is a growing need for systems analyzing reviews automatically. Lots of approaches (Xia et al., 2011; Socher et al., 2013; Tang et al., 2015a; Yang et al., 2016) usually focus on determining the overall sentiment rating of a review. Actually, not only does a review express the general attitude of reviewer, but it also conveys fine-grained sentiments towards different aspects of corresponding products. Figure 1 shows an example where *Bob* posts a review about a hotel and gives scores on overall attitude, *location*, *room*, and *service* respectively. The analysis of these aspect ratings could not only benefit mining interested aspects for users, but also help companies better understand the major pros and cons of the product. However, compared with the overall rating, users are less motivated to give aspect ratings. The reviews without aspect ratings are rampant, which are more than 46% in a simple corpus-based statistics[3]. Accordingly, it is really useful to perform document-level multi-aspect sentiment classification, whose goal is to predict ratings for different aspects in a review (Yin et al., 2017).

Multi-task learning (Caruana, 1997; Collobert et al., 2011; Luong et al., 2016) is a straightforward approach for document-level multi-aspect sentiment classification, which shares the input and hidden layers to obtain a document representation as the input of different aspect-specific classifiers. However, the representation fails to capture the differences between aspects. In fact, when we predict the sentiment rating of *service*, the first two sentences in Figure 1 are most helpful and other sentences are auxiliary or even unnecessary for classifying *service*. Therefore, aspect-specific document representation is vital for this task. To this end, Yin et al. (2017) use iterative attention module to mine aspect-specific words and sentences based on a list of aspect keywords and obtain state-of-the-art results.

[1] https://www.tripadvisor.com/
[2] https://www.yelp.com/
[3] The result is computed on 387,805 reviews crawled from https://www.tripadvisor.com/.

Proceedings of the 27th International Conference on Computational Linguistics, pages 925–936
Santa Fe, New Mexico, USA, August 20-26, 2018.

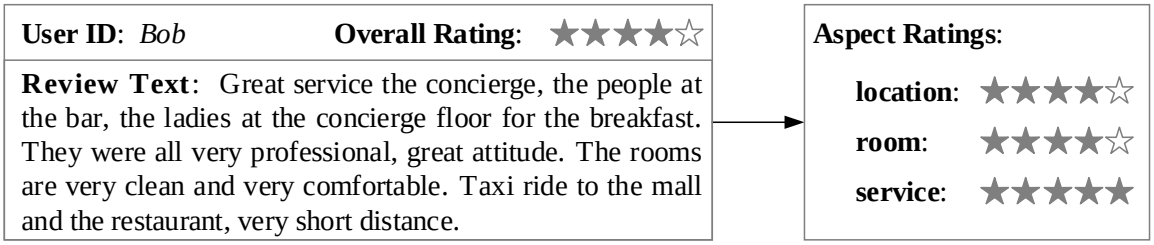


Figure 1: An example of a review. The left part is the review content, the upper right part is the reviewer *Bob* and overall rating of the review and bottom right part is different aspect ratings of the review. We focus on incorporating user preference and overall rating into review content to infer aspect ratings.

Despite the success of methods mentioned above, they typically only use text information. Two kinds of important information are ignored: users and overall ratings of reviews. The results of our statistical analysis are convincing that the two factors have strong correlations with aspect ratings (Section 2). As for users, different users may care about different aspects. When scoring aspects of a hotel, a business traveler may be critical to *service* but lenient with *price* or *room*. Such preference obviously affects the aspect ratings. Actually, many studies (Tang et al., 2015b; Chen et al., 2016; Dou, 2017) have shown that user preference can boost the performance of a related task, document-level sentiment classification that predicts an overall polarity instead of multi-aspect ratings. For our multi-aspect sentiment classification, the overall rating is given, and it can provide prior information to aspect ratings. Usually, the two types of rating are positive correlation. For example in Figure 1, the overall rating is 4 stars and the aspect ratings are all not less than 4 stars.

Inspired by the above analysis, we propose a model called Hierarchical User Aspect Rating Network (HUARN) to consider user preference and overall rating jointly for document-level multi-aspect sentiment classification. Specifically, HUARN utilizes a hierarchical structure to encode word, sentence, to document level information. Then, user and aspect information are embedded as attentions over word-level and sentence-level representation to construct a user-aspect-specific document representation. Based on the document representation, users and overall ratings are combined to express their influences on predicting aspect ratings. Finally, we adopt a multi-task framework to mutually enhance aspect rating prediction between different aspects.

In summary, our main contributions are as follows:

- For document-level multi-aspect sentiment classification, we validate the influences of users and overall ratings in terms of aspect ratings on massive Tripadvisor reviews.
- To the best of our knowledge, this is the first work to incorporate user preference and overall rating into a unified model (HUARN) in this task.
- We conduct experiments on two real-world datasets to verify the effectiveness of HUARN. The experimental results show that HUARN outperforms state-of-the-art methods significantly. The code and data for this paper are available at https://github.com/Junjieli0704/HUARN.

2 Data and Observations

In this section, we first introduce real-world datasets used in our work and present some explorations about the impacts of user preference and overall ratings on aspect ratings.

2.1 Data

We evaluate HUARN on two datasets: TripDMS and TripOUR. They are both crawled from Tripadvisor website and contain seven aspects (*value*, *room*, *location*, *cleanliness*, *check in*, *service*, and *business service*) which are provided by Tripadvisor website. The first dataset is built by Yin et al. (2017). However, there is no available user information in this dataset, thus we create the second one. Statistics

Datasets	#docs	#users	#docs/user	#words/sen	#words/doc
TripOUR	58,632	1,702	34.44	17.80	181.03
TripDMS	29,391	N/A	N/A	18.0	251.7

Table 1: Statistics of our datasets. The rating scale of TripOUR and TripDMS are 1-5.

Datasets	*value*	*room*	*location*	*cleanliness*	*check in*	*service*	*business service*
TripOUR	43,258	41,295	42,354	42,601	1,283	58,449	801
TripDMS	28,778	29,140	23,401	29,184	23,373	28,322	15,939

Table 2: The absolute number of rating of different aspects in TripOUR and TripDMS.

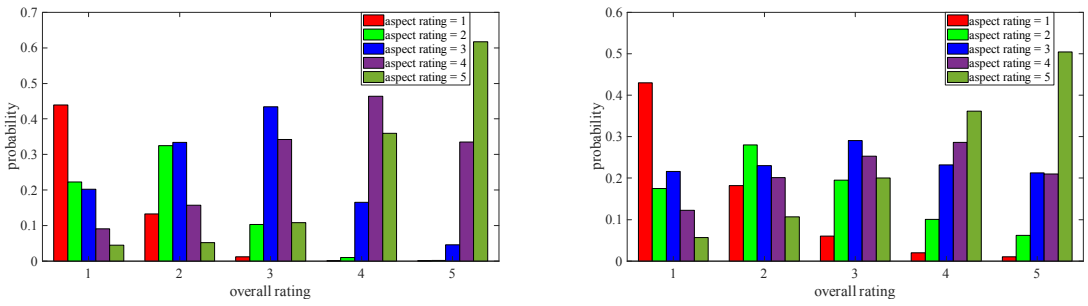
of our datasets are summarized in Table 1. Table 2 presents the absolute number of ratings of these aspects in our datasets.

2.2 Observations

Effects of user preference. Inspired by Tang et al. (2015b), we argue that the influences of users include the following two aspects: (1) *user-rating consistency*: different users have different characteristics in scoring aspects, and aspect ratings from the same user are more consistent than those from different users. (2) *user-text consistency*: different users have different word-using habits to express opinions and texts from the same user are more consistent than those from different users. To verify these consistencies, we conduct hypothesis testing as follows:

First, we construct three vectors $\mathbf{v}_s$, $\mathbf{v}_r$ and $\mathbf{v}_a$ with equal number (l) of elements. $\mathbf{v}_{s_i}$ is obtained by calculating a measurement between two reviews (d_i and d_i^+) posted by the same user, $\mathbf{v}_{r_i}$ is obtained by calculating a measurement between d_i and another random review and $\mathbf{v}_{a_i}$ is a random aspect (such as *service*), where $i \in \{1, 2, ..., l\}$.

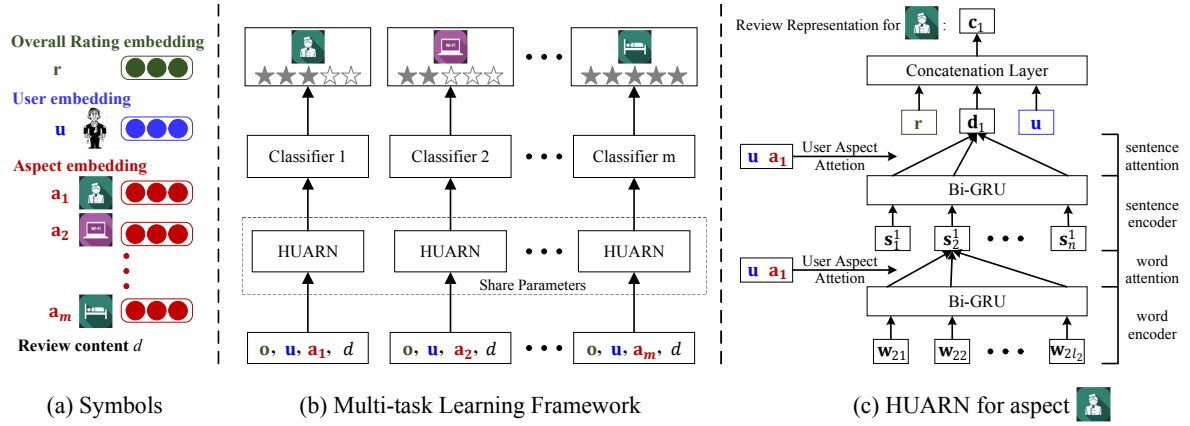
For *user-rating consistency*, the measurement is calculated by $||y - y^+||$ for $\mathbf{v}_s$ or $||y - y^-||$ for $\mathbf{v}_r$, where y, y^+, y^- is aspect rating of review d, d^+, d^- with aspect $\mathbf{v}_{a_i}$ respectively. For *user-text consistency*, the measurement is calculated by the cosine similarity between bag-of-words representation of two reviews. We perform a two-sample t-test on $\mathbf{v}_s$ and $\mathbf{v}_r$. The null hypothesis is that there is no difference between the two vectors, $H_0 : \mathbf{v}_s = \mathbf{v}_r$; the alternative hypothesis is that the difference between reviews with same user is less than with two random reviews , $H_1 : \mathbf{v}_s < \mathbf{v}_r$. The t-test results, p-values, show that there is strong evidence (with the significance level $\alpha = 0.01$) to reject the null hypothesis in *user-rating consistency* test and *user-text consistency* test on TripOUR. In other words, we observe the existence of *user-rating consistency* and *user-text consistency* in TripOUR.



(a) TripOUR (b) TripDMS

Figure 2: Aspect rating distributions for different overall ratings in TripOUR and TripDMS.

Effects of overall ratings. When scoring a product, users may consider multiple aspects of the product. If these aspects could meet the users' requirement, they can give a high overall rating, otherwise, they could give low scores. Therefore, overall rating can partly reflect the user's attitudes to aspects, which



(a) Symbols (b) Multi-task Learning Framework (c) HUARN for aspect

Figure 3: The architecture of HUARN. Left: some embedding symbols used in the figure. Example aspects are *service*, *business service*, and *room*. Middle: Multi-task learning framework for document-level multi-aspect sentiment classification. Right: the architecture of HUARN for aspect *service*.

is called *overall rating prior*. To investigate effects of overall ratings, we compute the aspect rating distributions for different overall ratings from our two datasets and the distributions are shown in Figure 2. We can conclude that high/low overall ratings often result in high/low aspect ratings. For example, when the overall rating is 5 stars, more than 70% aspect ratings are not less than 4 stars in our datasets.

3 Methods

The analysis proves users and overall ratings are significant on interpreting the sentiments of different aspects. Therefore, we introduce these two kinds of information into HUARN and detail the model here. First, we give the formalizations of document-level multi-aspect sentiment classification (Figure 3(a)). Afterwards, we discuss the multi-task learning framework for this task (Figure 3(b)) and how to obtain document semantic representation via Hierarchical Bidirectional Gated Recurrent Unit network. At last, we present user and aspect attention mechanism to construct user-aspect-specific representation and add a concatenating layer to combine user, overall rating and document representation together (Figure 3(c)). The enhanced document representation is used as features for predicting aspect ratings.

3.1 Formalizations

Suppose we have a corpus D about a specific domain (such as "hotel") and m aspects $\{a_1, a_2, ..., a_m\}$ (such as *service* and *room*). Review d is a sample of D with n sentences $\{s_1, s_2, ..., s_n\}$. Sentence s_i consists of l_i words as $\{w_{i1}, w_{i2}, ..., w_{il_i}\}$. The overall rating of review d is r and its author is user u. Document-level multi-aspect sentiment classification aims to predict aspect ratings for these reviews.

3.2 Multi-task Learning Framework

It is natural to model document-level multi-aspect sentiment classification as a multi-task learning. First, we can treat each aspect rating as a classification task. Then, we share document encoder network to obtain document representation and exploits different softmax classifiers to predict ratings of different aspects. The main benefit of the multi-task framework is that it can mutually enhance aspect rating prediction between different aspects.

3.3 Hierarchical Bidirectional Gated Recurrent Network

Since a document is composed of multiple sentences, and a sentence is composed of multiple words, we model the semantics of a document through a hierarchical structure from word-level, sentence-level to document-level. To model the semantic representation of a sentence, we adopt bidirectional GRU (Bi-GRU). Similarly, we also use Bi-GRU to learn document representations.

Given sentence s_i, we embed each word w_{ij} to vector $\mathbf{w}_{ij}$. Then, we use a Bi-GRU to encode contextual information of word w_{ij} into its hidden representation $\mathbf{h}_{ij}$. Hidden states $\{\mathbf{h}_{i1}, \mathbf{h}_{i2}, ...\mathbf{h}_{il_i}\}$ are feed

into an average pooling layer to obtain the sentence representation $\mathbf{s}_i$. In sentence level, we also feed the sentence vectors $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ into Bi-GRU and then obtain the document representation $\mathbf{d}$ similarly.

3.4 Encoding user, aspect, and overall rating

It is obvious that not all words (sentences) contribute equally to the sentence (document) meaning. To consider *user-text consistency* and build an aspect-specific representation, we introduce user attention and aspect attention. Specifically, we employ word (sentence) level user aspect attention to generate sentence (document) representation.

Word-level Attention. We first embed user u and aspect $\{a_k|k \in 1, 2, ..., m\}$ as continuous and real-valued vector $\mathbf{u}$ and $\mathbf{a}_k$. Then, instead of feeding word-level hidden states ($\mathbf{h}_{ij}$) to an average pooling layer, we adopt a user aspect attention mechanism to extract user-aspect-specific words and obtain the sentence representation as follows:

$$\mathbf{m}_{ij} = \tanh(\mathbf{W}_{wh}\mathbf{h}_{ij} + \mathbf{W}_u\mathbf{u} + \mathbf{W}_a\mathbf{a}_k + \mathbf{b}_w) \tag{1}$$

$$\alpha_{ij} = \frac{\exp(\mathbf{v}_w^T\mathbf{m}_{ij})}{\sum_j \exp(\mathbf{v}_w^T\mathbf{m}_{ij})} \tag{2}$$

$$\mathbf{s}_i^k = \sum_j \alpha_{ij}\mathbf{h}_{ij} \tag{3}$$

where $\mathbf{W}_{wh}, \mathbf{W}_{wu}, \mathbf{W}_{wa}$ and $\mathbf{b}_w$ are parameters in the attention layer. α_{ij} measures the importance of the j-th word for user u and aspect a_k and $\mathbf{s}_i^k$ is the representation of sentence s_i for aspect a_k.

Sentence-level Encoder and Attention. After obtaining sentence representation $\mathbf{s}_i^k$ for aspect a_k, we also use a Bi-GRU to encode the sentences and get hidden representation $\mathbf{h}_i^k$ for $\mathbf{s}_i^k$.

When classifying document based on different aspects, different sentences may have different influences. Different users may also pay attention to different sentences. Therefore, in sentence level, we also apply an attention mechanism with user vector $\mathbf{u}$ and aspect vector $\mathbf{a}_k$ in sentence level to select informative sentences to compose user-aspect-specific document representation. The document representation $\mathbf{d}_k$ for aspect a_k is obtained via:

$$\mathbf{t}_i = \tanh(\mathbf{W}_{sh}\mathbf{h}_i^k + \mathbf{W}_{su}\mathbf{u} + \mathbf{W}_{sa}\mathbf{a}_k + \mathbf{b}_s) \tag{4}$$

$$\beta_i = \frac{\exp(\mathbf{v}_s^T\mathbf{t}_i)}{\sum_i \exp(\mathbf{v}_s^T\mathbf{t}_i)} \tag{5}$$

$$\mathbf{d}_k = \sum_i \beta_i\mathbf{h}_i^k \tag{6}$$

where β_i measures the importance of the i-th sentences for user u and aspect a_k.

Concatenation Layer. To explicitly encode *user-rating consistency* and *overall rating prior*, we add a concatenation layer. First, we embed overall rating r as continuous and real-valued vector $\mathbf{r}$ with g_r dimensions. Then, we generate review content representation $\mathbf{c}_k$ by concatenating user embedding $\mathbf{u}$, rating embedding $\mathbf{r}$ and document vector $\mathbf{d}_k$:

$$\mathbf{c}_k = \mathbf{u} \oplus \mathbf{r} \oplus \mathbf{d}_k \tag{7}$$

3.5 Document-level Multi-aspect Sentiment Classification

For each aspect, we obtain review representation $\{\mathbf{c}_k|k \in 1, 2, ..., m\}$. All these representations are high-level representations of the combination of user, aspect, overall rating and document information. It can be used as features for predicting aspect ratings. For aspect a_k, we can use a softmax layer to project $\mathbf{c}_k$ into sentiment distribution $\mathbf{p}(d, k)$ over L classes:

$$\mathbf{p}(d, k) = \text{softmax}(\mathbf{W}_{lk}\mathbf{c}_k + \mathbf{b}_k) \tag{8}$$

where $p_l(d,k)$ is used to represent the predicted probability of sentiment class l for d based on a_k and $\mathbf{W}_{lk}, \mathbf{b}_k$ are parameters of softmax layer for classifying review $\mathbf{c}_k$. Then we define the cross-entropy error between gold sentiment distribution and our model's sentiment distribution as our loss function :

$$L = -\sum_{d \in D} \sum_{k \in \{1,2,\ldots,m\}} \sum_{l=1}^{L} \mathbb{1}\{g_{d,k} = l\} \cdot \log(p_l(d,k)) \tag{9}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function and $g_{d,k}$ represents the ground truth label for review d for aspect a_k.

4 Experiments

In this section, we present data preprocessing and implementation details, all the comparison methods and the empirical results on the task of document-level multi-aspect sentiment classification.

4.1 Data Preprocessing and Implementation Details

We preprocess our datasets as follows: For TripDMS, we use the same splitting method as (Yin et al., 2017). For TripOUR, we tokenize the dataset, split sentences by Stanford CoreNLP (Manning et al., 2014) and randomly split them into training, development, and testing sets with 80/10/10%.

The model hyper-parameters are tuned based on the development sets. For word embeddings, we use the pre-trained word embeddings provided by (Yin et al., 2017), whose embedding size is 200. For user and overall rating embeddings, we initialize them randomly and set their dimensions to 200. For aspect embeddings, we first get aspect keywords[4] from (Yin et al., 2017) and initial aspect embedding by averaging word embeddings of these keywords belong to the aspect. The dimensions of all hidden vectors are set to 150. To avoid model over-fitting, we use dropout with rate of 0.2. All the parameters are trained using Adam (Kingma and Ba, 2014) with a learning rate of 0.001.

4.2 Comparison Methods

We compare HUARN with the following baselines:

Majority is a heuristic baseline method, which assigns the majority sentiment category in the training set to aspect rating in the test dataset.

OverallRatingSame is also a heuristic baseline method, which assigns the overall rating of a review to its aspect ratings.

MajOverallRating splits the training instances into five clusters (per overall rating) and assigns the most frequent rating for the seven aspects per cluster in the test dataset.

SVM and **NBoW** are SVM classifiers with different features. One with unigrams, bigrams as features and another with the mean of word embeddings in a document as features.

CNN (Kim, 2014) performs a convolution operation over a sentence to extract words neighboring features, then gets a fixed-sized representation by a pooling layer.

HAN (Yang et al., 2016) models review in a hierarchical structure and utilizes an attention mechanism to capture important words and sentences, which is only based on text information and achieves state-of-the-art result in predicting overall rating of document.

MHCNN is an extended model of **CNN** with hierarchical architecture and multi-task framework.

MHAN is an extended model of **HAN** with multi-task framework.

DMSCMC (Yin et al., 2017) use iterative attention modules to build up aspect-specific representation for review, and obtain state-of-the-art results in document-level multi-aspect sentiment classification.

HGRUN is the basic form of **HUARN** without user, aspect and overall rating.

HARN is a variant of **HUARN**, which abandons user information from **HUARN**.

4.3 Results

We use Accuracy and Mean Squared Error (MSE) as the evaluation metrics, and the results are shown in Table 3. For heuristic methods, we can see that **Majority** performs very poor because it does not

[4]Sample keywords for *service* are service, food, breakfast, and buffet.

Models	TripOUR		TripDMS	
	Accuracy↑	MSE↓	Accuracy↑	MSE↓
Majority	0.3850	0.954	0.2389†	2.549†
OverallRatingSame	0.3074	1.705	0.2012	3.446
MajOverallRating	0.3487	1.536	0.2414	3.273
SVM	0.4635	1.025	0.3526†	1.963†
NBoW	0.4865	0.912	0.3909†	1.808†
CNN	0.5054	0.752	0.4335†	1.456†
HAN	0.5123	0.705	0.4468†	1.301†
MHCNN	0.5108	0.712	0.4379†	1.398†
MHAN	0.5419	0.629	0.4494†	1.210†
HGRUN	0.5392	0.635	0.4435	1.303
DMSCMC	0.5549	0.583	0.4656†	1.083†
HARN	0.5815*	0.528	**0.4821***	**0.923**
HUARN	**0.6070***	**0.514**	N/A	N/A

Table 3: Document-level multi-aspect sentiment classification on our datasets. Our full model is HUARN. The best performances in **bold**. "†" indicates that the result is reported from (Yin et al., 2017). "*" indicates that the model significantly outperforms DMSCMC. Statistical significance testing has been performed using paired t-test with $p < 0.05$.

capture any text information. **OverallRatingSame** and **MajOverallRating** are also very poor, even though overall rating has strong correction with aspect ratings, it is not enough to decide aspect ratings only based on it.

Compared with **SVM**, **NBoW** achieves higher accuracy by at least 2.3% in both datasets, which shows that embedding features are more effective than unigram and bigram features on these two datasets. When applying more complex neural networks (such as **CNN** and **HAN**), the model can achieve higher accuracy by at least 1.5% in both datasets compared with **NBoW**. Additionally, we observe that the multi-task learning and hierarchical architecture are beneficial for neural networks. Performance on **MHAN** and **MHCNN** are slightly better than **HAN** and **CNN**. Beyond that, we also find attention mechanism is useful. The only difference between **MHAN** and **HGRUN** is that **MHAN** uses attention mechanism to obtain sentence and document representations while **HGRUN** utilizes an average pooling layer, which results in the performance of **MHAN** is better than **HGRUN**. After obtaining aspect-aware representation for the document, **DMSCMC** achieves best results and outperforms other baselines.

Compared to **DMSCMC**, **HARN** achieves improvements of 2.7% and 1.7% on TripOUR and TripDMS respectively, which shows that the incorporation of overall rating and aspect attention helps build up more discriminative representation. Moreover, when incorporating user information, our full model (**HUARN**) can achieve improvements of 5.3% compared with **DMSCMC** on TripOUR[5], which shows user preference can benefit the document-level multi-aspect sentiment classification task.

5 Discussions

In this section, we first give some discussions about the effects of users, aspects and overall ratings on predicting aspect ratings, and then show case study for attention results and visualize user embeddings.

5.1 Effects of Users, Overall Ratings and Aspects

Users, overall ratings and aspects are three kinds of information in HUARN. We present the effects of users, overall ratings, and aspects on document-level multi-aspect sentiment classification in Table 4. From the table, we can observe that: (1) Compared with user-agnostic models (line 1-4), user-aware models (line 5-8) can achieve improvements of 2.2%, 2.5%, 1.2% and 2.5% in TripOUR, which shows

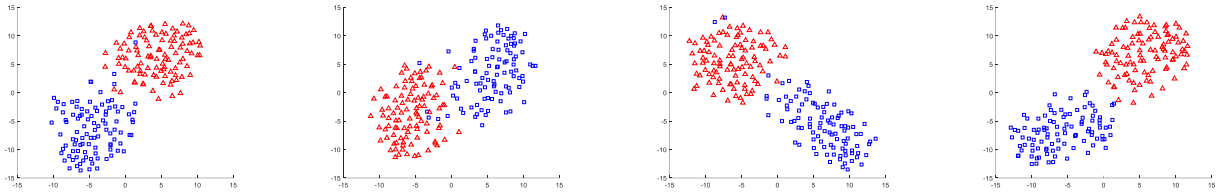
[5]Since there is no user information in TripDMS, we can only compare **DMSCMC** with **HARN**.

No.	Different information			TripOUR		TripDMS	
	User	**OverallRating**	**Aspect**	Accuracy↑	MSE↓	Accuracy↑	MSE↓
1	–	–	–	0.5392	0.635	0.4435	1.303
2	–	–	✓	0.5514	0.599	0.4566	1.256
3	–	✓	–	0.5719	0.555	0.4740	1.093
4	–	✓	✓	0.5815	0.528	0.4821	0.923
5	✓	–	–	0.5640	0.619	N/A	N/A
6	✓	–	✓	0.5764	0.581	N/A	N/A
7	✓	✓	–	0.5839	0.560	N/A	N/A
8	✓	✓	✓	0.6070	0.514	N/A	N/A

Table 4: Effects of user, overall rating and aspect on document-level multi-aspect sentiment classification. Each line represents a variant of HUARN, where "✓" denotes the variant considers the specific information, while "–" denotes not. For example, model in line 1 means HUARN abandons these three kinds of information and degenerates into HGRUN.

that model encoding user information can obtain user-aware document representation and is more suitable for document-level multi-aspect sentiment classification. (2) Compared with models without overall rating information (line 1, 2, 5 and 6), models considering overall ratings (line 3, 4, 7, 8) can obtain 3.2% (3.1%), 3.0% (2.6%), 1.9% (N/A) and 3.1% (N/A) improvements in accuracy in both datasets, which indicates overall rating information is useful for building more discriminative document representation and helpful for predicting aspect ratings. (3) Compared with models without aspect information (line 1, 3, 5 and 7), aspect-based models (line 2, 4, 6, 8) can obtain 1.2% (1.3%), 1.0% (0.8%), 1.2% (N/A) and 2.4% (N/A) improvements in accuracy in both datasets. It shows that aspect information is useful for building aspect-aware document representation and helpful for predicting aspect ratings. (4) After users, overall ratings, and aspects being considered jointly, our model obtains the best performance.

5.2 Visualization of User Embeddings



(a) Service (b) Value (c) Cleanliness (d) Location

Figure 4: t-SNE visualization of user embeddings for different aspects in TripOUR. Blue square and red triangle represent users are "High Score" users and "Low Score" users, respectively.

As different users have different aspect rating preferences and HUARN imports user embedding to consider users, we identify whether such personalized information are encoded in user embedding. To this end, we first rank all users according to their average score in the training set for each aspect. Then the top 100 users are labeled as "High Score" users and bottom 100 users are labeled as "Low Score" users. Due to space limit, here we only show embeddings of users in four aspects (*service*, *value*, *cleanliness*, *location*), which are top frequently scored by all users, in Figure 4. We find "High Score" users and "Low Score" users are separated apparently. The visualization shows that user embedding learned by HUARN can encode personalized traits in scoring different aspects.

5.3 Case Study for Attention Results

To show the ability that HUARN captures user preference and aspect semantic meanings, we take one sentence from TripOUR as example. The content of the sentence is "The food is **good**, but the price is

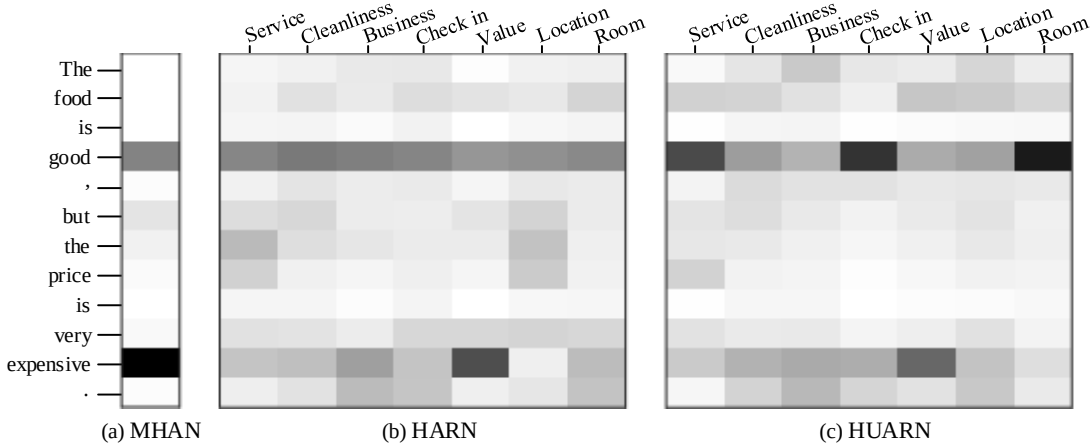


Figure 5: The attention visualization of words. Dark color means higher weight. (a), (b) and (c) show word-level attention weights of **MHAN**, **HARN** and **HUARN**.

very **expensive**", in which "good" is a general sentiment word and can be used to describe many aspects (such as *service*, *room* and so on), while "expensive" is a aspect-specific sentiment word which only applies to describe *value*. We visualize attention weights of the sentence in Figure 5.

From Figure 5(a), we can find that considering word-level attention, **MHAN** distinguishes sentiment words and non-sentiment words, however it is hard to identify the sentiment word is a general one or an aspect-specific one. After adding aspect attention over word-level representations, **HARN** can distinguish these two kinds of sentiment words (Figure 5(b)). The attention weights of "good" for different aspects are very close, while the attention weights of "expensive" for different aspects are different, and the maximum is *value*. When adding user information into word-level attention, **HUARN** can also treat "good" differently. From figure 5(c), we can find attention weights of "good" are different for different aspects, where weights for *service*, *check in* and *room* are higher than weights for other aspects. we check all reviews of the sample review's author and find that he/she often (more than 80%) use "good" to describe *service*, *check in* and *room*.

5.4 Error Analysis

Figure 6: Examples of error cases. GT means ground trouth and P means prediction result of HUARN. Sentences in review text with darker color means higher attention weight for the sentence.

We analyze error cases in the experiments. Some examples of error cases are shown in Figure 6. We can find that HUARN is hard to select important sentences for aspects. For example, sentence "the rooms are large and the bed comfy." and sentence "however the carpet should have been replaced long time." in the first sample in Figure 6 are all important to decide the rating of aspect *room*. However, HUARN pays more attention to the former sentence when predicting rating of *room* and obtains the wrong result.

Based on the literature study, we find that Yin et al. (2017) uses iterative attention models to build up aspect-specific representation for review. It may alleviate this problem. We leave how to encode user and overall rating information into DMSCMC as our future work.

6 Related Work

Multi-aspect sentiment classification is an extensively studied task in sentiment analysis (Pang and Lee, 2008; Liu, 2012). Lu et al. (2011) propose Segmented Topic Model to model document and extract features, then exploit support vector regression to predict aspect ratings based on these features. McAuley et al. (2012) add a dependency term in final multi-class SVM objective to consider the correction between aspects. Many other studies (Titov and McDonald, 2008; Wang et al., 2010; Wang et al., 2011; Diao et al., 2014; Pappas and Popescu-Belis, 2014; Pontiki et al., 2016; Toh and Su, 2016) solve multi-aspect sentiment classification as a subproblem by utilizing heuristic based methods or topic models. However, these approaches often rely on strict assumptions about words and sentences, for example, word syntax has been used to distinguish aspect word or sentiment word, or appending an specific aspect to a sentence. Another related problem is called aspect-level sentiment classification (Pontiki et al., 2014; Dong et al., 2014; Wang et al., 2016; Tang et al., 2016; Schouten and Frasincar, 2016). Wang et al. (2016) and Tang et al. (2016) employ attention-based LSTM and deep memory network for aspect-level sentiment classification, respectively. However, the task is sentence level. Document-level sentiment classification (Li and Zong, 2008; Li et al., 2010; Li et al., 2013; Xia et al., 2015; Yang et al., 2016) is also a related research field because we can treat single aspect sentiment classification as an individual document classification task. However, they did not consider multiple aspects in a document.

In addition to these methods, the work of Yin et al. (2017) is the most related to ours, which focuses on using iterative attention mechanism to build discriminative aspect-aware representation to perform document-level multi-aspect sentiment classification. However, it ignores the influences of users and overall ratings on aspect ratings. Actually, many studies (Tang et al., 2015b; Tang et al., 2015c; Chen et al., 2016; Li et al., 2016; Dou, 2017) have shown considering user preference can boost the performance of Document-level Sentiment Classification. Partially inspired by these approaches, we propose HUARN to consider users, overall ratings and aspects jointly into document-level multi-aspect sentiment classification. Compared with these user-aware approaches, HUARN has some differences: (1) They do not consider aspects. (2) Although Chen et al. (2016) and Dou (2017) embedding user to consider user-text consistency to perform sentiment classification, they ignore user-rating consistency. (3) Tang et al. (2015b; 2015c) embed user in a matrix and build user-specific representation by a convolutional neural network structure. However, it is hard to train with limited reviews for user matrix. Our motivation is that (1) Aspect information is very useful for selecting informative words and sentences and building up aspect-specific representation for document-level multi-aspect sentiment classification, therefore, we add aspect attention into our model. (2) Compared with user-text consistency, user-rating consistency describes the correlation between users and ratings more directly. (3) Embedding user in a vector and using attention mechanism to build user-specific representation is an effective way to consider users. User embedding is enough to encode the relation between user and rating in The most important is that it is easy to train.

7 Conclusion and Future work

In this paper, we present Hierarchical User Aspect Rating Network (HUARN) to incorporate user preference and overall rating into document-level multi-aspect sentiment classification. HUARN encodes different kinds of information (word, sentence and document) into a hierarchical structure. To consider user preference and overall rating, HUARN introduces user information as attention over word-level representation and sentence-level representation, and then generates review representation by combining user, overall rating and document information. Extensive experiments show that our model outperforms state-of-the-art methods significantly. In the future, we will study how to encode user and overall rating information into DMSCMC.

Acknowledgements

We thank Xiaomian Kang and Yang Zhao for valuable discussions. We also thank the anonymous reviewers for their suggestions. The research work descried in this paper has been supported by the Natural Science Foundation of China under Grant No. 61333018 and 61673380.

References

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of EMNLP*.

Ronan Collobert, Jason Weston, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(1):2493–2537.

Qiming Diao, Minghui Qiu, Chao Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of KDD*, pages 193–202.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of ACL*, pages 49–54.

Zi-Yi Dou. 2017. Capturing user and product information for document level sentiment analysis with deep memory network. In *Proceedings of EMNLP*, pages 532–537, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Shoushan Li and Chengqing Zong. 2008. Multi-domain sentiment classification. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers*, pages 257–260.

Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 414–423.

Shoushan Li, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. 2013. Active learning for cross-domain sentiment classification. In *Proceedings of IJCAI*, pages 2127–2133.

Junjie Li, Haitong Yang, and Chengqing Zong. 2016. Sentiment classification of social media text considering user attributes. In *Natural Language Understanding and Intelligent Applications - 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2-6, 2016, Proceedings*, pages 583–594.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *Proceedings of ICDM Workshops*, pages 81–88.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of ICLR*, San Juan, Puerto Rico, May.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Julian Mcauley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of ICDM*, pages 1020–1025.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of EMNLP*, pages 455–466. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of International Workshop on Semantic Evaluation at*, pages 27–35.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Kim Schouten and Flavius Frasincar. 2016. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813–830.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*.

Duyu Tang, Bing Qin, and Ting Liu. 2015a. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of EMNLP*, pages 1422–1432, Lisbon, Portugal, September. Association for Computational Linguistics.

Duyu Tang, Bing Qin, and Ting Liu. 2015b. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of ACL*, pages 1014–1023, July.

Duyu Tang, Bing Qin, Yuekui Yang, and Yuekui Yang. 2015c. User modeling with neural network for review rating prediction. In *Proceedings of IJCAI*, pages 1340–1346.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. *arXiv preprint arXiv:1605.08900*.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL*, pages 308–316. Association for Computational Linguistics.

Zhiqiang Toh and Jian Su. 2016. Nlangp at semeval-2016 task 5: Improving aspect based sentiment analysis using neural network features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 282–288. Association for Computational Linguistics.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of KDD*, pages 783–792.

Hongning Wang, Yue Lu, and Cheng Xiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of KDD*, pages 618–626.

Yequan Wang, Minlie Huang, Li Zhao, and Xiaoyan Zhu. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of EMNLP*.

Rui Xia, Chengqing Zong, and Shoushan Li. 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138–1152.

Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li. 2015. Dual sentiment analysis: Considering two sides of one review. *IEEE Transactions on Knowledge and Data Engineering*, 27(8):2120–2133.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J Smola, and Eduard H Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*, pages 1480–1489.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of EMNLP*, pages 2044–2054. Association for Computational Linguistics.