Multilingual Review-aware Deep Recommender System via Aspect-based Sentiment Analysis

PENG LIU, LEMEI ZHANG, and JON ATLE GULLA, Department of Computer Science, Norwegian University of Science and Technology, Norway

With the dramatic expansion of international markets, consumers write reviews in different languages, which poses a new challenge for Recommender Systems (RSs) dealing with this increasing amount of multilingual information. Recent studies that leverage deep-learning techniques for review-aware RSs have demonstrated their effectiveness in modelling fine-grained user-item interactions through the aspects of reviews. However, most of these models can neither take full advantage of the contextual information from multilingual reviews nor discriminate the inherent ambiguity of words originated from the user's different tendency in writing. To this end, we propose a novel Multilingual Review-aware Deep Recommendation Model (MrRec) for rating prediction tasks. MrRec mainly consists of two parts: (1) Multilingual aspect-based sentiment analysis module (MABSA), which aims to jointly extract aligned aspects and their associated sentiments in different languages simultaneously with only requiring overall review ratings. (2) Multilingual recommendation module that learns aspect importances of both the user and item with considering different contributions of multiple languages and estimates aspect utility via a dual interactive attention mechanism integrated with aspect-specific sentiments from MABSA. Finally, overall ratings can be inferred by a prediction layer adopting the aspect utility value and aspect importance as inputs. Extensive experimental results on nine real-world datasets demonstrate the superior performance and interpretability of our model.

CCS Concepts: • **Information systems** → **Recommender systems**; **Personalization**; • **Computing methodologies** → **Natural language processing**;

Additional Key Words and Phrases: Recommender systems, deep learning, multilingual aspect-based sentiment analysis, neural attention, co-attention

ACM Reference format:
Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2021. Multilingual Review-aware Deep Recommender System via Aspect-based Sentiment Analysis. *ACM Trans. Inf. Syst.* 39, 2, Article 15 (January 2021), 33 pages.
https://doi.org/10.1145/3432049

1 INTRODUCTION

Many e-commerce websites, such as Amazon and Yelp, allow users to naturally write reviews along with a numerical rating to express opinions and share experiences toward their purchased items. These reviews are usually in the form of free text and play the role of carriers that reveal the reasons why users like or dislike the items or services they concern. For example, a review may

This work is supported by the Research Council of Norway under Grant No. 245469.
Authors' address: P. Liu, L. Zhang, and J. A. Gulla, Department of Computer Science, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway; emails: {peng.liu, lemei.zhang, jon.atle.gulla}@ntnu.no.

1046-8188/2021/01-ART15 $15.00
https://doi.org/10.1145/3432049

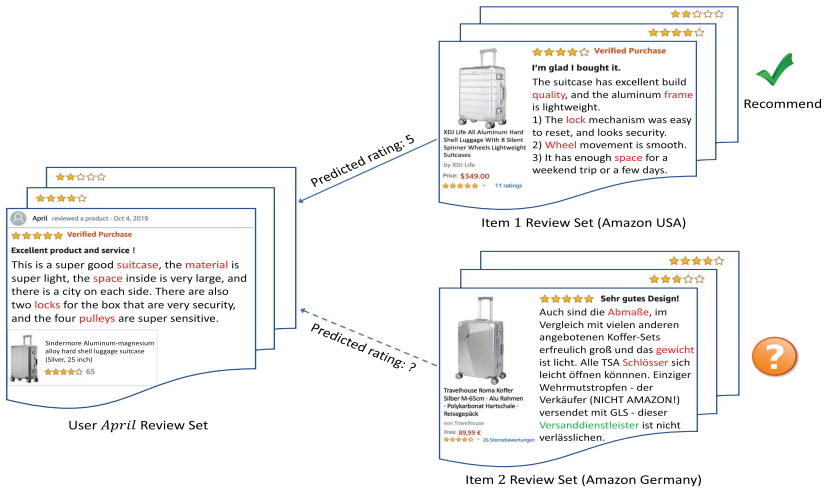


Fig. 1. A schematic to show multilingual scenarios for RSs. Note that the red words represent aspects with positive sentiment and the green words represent aspects with negative sentiment.

include the user's opinions on the various aspects of an item (e.g., its price, performance, quality, etc.), which are of high reference values for other users to make purchasing decisions. Therefore, in recent years, many recommender systems (RSs) [1–5] have been developed by exploiting the semantic information covered in reviews to model a fine-grained user preference and alleviate the data sparsity problem for enhancing personalized recommendations.

Previous works on review-aware RSs are mainly devoted to the monolingual scenario. However, with the growth of the Web and the expansion of the international market, consumers write reviews in different languages, and e-commerce is becoming more and more multilingual. Only addressing monolingual reviews lead to missing a lot of useful information existing in other languages. Indeed, it has been estimated that more than half of the world's population is bilingual, and nearly 45% of the websites provide content in a language different from English [6]. Besides, statistics of Amazon European market[1] show that almost 63% of users on average are non-English speakers, and Amazon provides services with different languages apart from English according to the users' geolocation. Facing the abundance of multilingual information, RSs need to evolve to effectively deal with the challenge of recommending interesting items with their review languages different from that the users adopted to express their preferences. As far as we know, this problem is very prevalent for most e-commerce platforms (e.g., Amazon and Booking) but has never been explored before.

To have a deep insight into the problem of multilingual review-based recommendation for e-commerce, Figure 1 illustrates two different simplified recommendation scenarios the users often encounter when shopping on Amazon. *April* is an American user who usually buys suitcase on Amazon. When she is shopping at home in America, traditional review-based RSs could easily suggest $item1$ to *April*, since the item features contained in its reviews match well with the user preference on different aspects expressed in her reviews. However, when she is travelling or studying abroad in Germany, it would be difficult for such RSs to provide a satisfying recommendation (e.g., $item2$) only according to the English reviews in her purchased history, because most reviews of $item2$ are written in German. Such scenarios can also be easily found on other e-commerce like

[1] https://orangeklik.com/optimize-listings-amazon-europe/.

Foursquare, Booking, and TripAdvisor. This clearly motivates the need for efficient and effective recommendation techniques that cross the boundaries of languages.

So far, there have been few studies on multilingual recommendation in the literature. Existing methods [7–9] attempt to build language-independent user/item profiles by leveraging the concepts contained in external knowledge sources, such as Wikipedia and MultiWordNet. However, they are not suitable for our task due to inability to model fine-grained user-item interactions. Recently, empowered by continuous real-valued vector representations and semantic composition over contextual information, deep-learning-based methods have demonstrated their effectiveness in modelling user's fine-grained preferences to specific item features through the aspects extracted from reviews. The attention mechanism is mainly adopted in these works to automatically learn the aspect importances/weights for different user-item pairs. Guan et al. [3] propose an attentive aspect-based recommendation model that effectively captures the interactions between aspects extracted from reviews for rating perdition tasks. Chin et al. [4] propose to use a neural architecture incorporated with a co-attention mechanism to perform aspect-based representation learning for both users and items and estimate aspect-level importance in an end-to-end fashion.

Despite their state-of-the-art performance, they still suffer from the following limitations: (1) Most methods fail to handle multilingual reviews embodied with significant contextual information, especially when only a few reviews are provided in the monolingual scenario [10]. (2) The users tend to exhibit different criteria when writing reviews, which leads to inherent ambiguity among words, and thus it is difficult for such approaches to precisely capture the user's intent. (3) Most existing methods neglect long-tail items when performing recommendations, which are crucial to gain the diversity of RSs and thereby improve the users' satisfaction. (4) The majority of above-mentioned algorithms take as inputs the concatenation of all the word representations from every associated review, which makes the size of inputs considerably large, and therefore are impractical in the real-world applications.

In this article, to track the above limitations, we propose a novel Multilingual Review-aware Deep Recommendation Model (MrRec), which incorporates the aligned aspects and aspect-specific sentiments in different language reviews for rating prediction and interpretation. Specifically, MrRec consists of two parts: multilingual aspect-based sentiment analysis (MABSA) and multilingual recommendation module (MRM). In the first part, we utilize an unsupervised aspect-based autoencoder to learn a set of language-independent aspect embeddings. Then Multiple Instance Learning (MIL) framework integrated with hierarchical attention mechanism is designed to predict the aspect-specific sentiment distributions of review sentences, and learn aspect-aware sentence representations guided by the overall ratings. Note that the overall ratings serve both as a proxy of sentiment labels of reviews and as a bridge among languages. MIL framework, originated from the work of Reference [11], offers a viable and natural solution for learning in a weakly supervised setting by taking into account the overall opinions of user's reviews. However, most recent works [12, 13] with MIL framework perform sentiment analysis at the sentence level, assuming that an entire section of text/review expresses one sentiment toward one entity, which is not always true. Thus, in our work, we extend MIL on aspect level that allows for multiple opinions toward multiple aspects or entities in a sentence. Instead of learning from manually labelled aspect opinions, which are not always available and demand time-consuming tasks especially in the multilingual scenario, our model only requires document level supervision and learns to judge the sentiment of aspects related to each review sentence introspectively.

In the second part, a multilingual recommendation module is developed to infer the overall rating through a prediction layer with its input of the aspect utilities estimated by a dual interactive attention mechanism, and the corresponding aspect importances of both the user and item considering the different contributions of multiple languages. Many recent researches propose to use

dual attention mechanism in recommendation tasks [14–16]. For instance, in Reference [14], the authors propose to use the dual local and global attention that leverages local layer to learn user's preferences or item properties, and global layer to capture the semantic meaning of the whole review text. The authors of Reference [15] propose to use two dual Graph Attention Networks (GATs) that one dual GAT is used to capture the user's social influence and homophily, while another is to model the item's static and dynamic attributes. Differently, our dual interactive attention mechanism pays attention to a finer-grained aspect-level for the user and item sides. One attention net focuses on the most relevant items the target user previously rated with regard to the candidate item, which takes into account item properties from the item side and long-tail items. Meanwhile, another attention net aims to search for candidate item with potential aspects assessed by other users in accordance with the taste of the target user on the same aspects, which takes into account the preferences of the target user. From these two perspectives, our model enables the balance of the recommendation accuracy and diversity at the same time. We applied our model to several real-world datasets, and experimental results demonstrate the promising and reasonable performance of our approach.

In summary, our contributions are as follows:

- To the best of our knowledge, this is the first study that leverages multilingual reviews as potential resources to improve the interpretability and diversity of recommendation tasks in e-commerce. We also explore the possibility that deep-learning techniques can be adopted to model language-independent user/item profiles in a fine-grained scale.
- We are the first to introduce MIL framework for multilingual aspect-based sentiment analysis, which uses freely available multilingual word embeddings and only requires light supervision (user-provided ratings). It is demonstrated that the overall ratings can serve as the surrogate sentiment labels and bridges to address language barriers.
- We design a novel dual interactive attention mechanism that considers both popular and long-tail items for effectively modelling the fine-grained user-item interactions, as well as balancing between recommendation accuracy and diversity.
- Extensive experiments are conducted on nine datasets from Amazon and Goodreads to verify the effectiveness and efficiency of our model. The results show that MrRec not only outperforms state-of-the-art baselines but also interprets the recommendation results in great detail.

The remainder of the article is organized as follows. Section 2 introduces the related work. In Section 3, we present our MrRec model in detail. We describe the datasets, experimental settings, and the state-of-the-art methods we use in Section 4, as well as experimental results and analysis. Finally, we present the conclusions and future work in Section 5.

2 RELATED WORK

In this section, we briefly review several key areas that are highly related to our work: (1) Review-aware Recommender Systems, (2) Multilingual Recommender Systems, as well as (3) Multilingual Aspect-based Sentiment Analysis.

2.1 Review-aware Recommender Systems

In the past few years, textual reviews were exploited by many researchers for improving the performance as well as enhancing the interpretability of recommendations [17–22].

To extract meaningful features from reviews, some methods concatenate all the reviews belonging to a user (or item) as a user (or item) document, and then employ convolutional neural networks (CNNs) to learn the latent user and item representations. Examples include DeepCoNN

[23], TransNets [24], and D-Attn [14]. Though these methods have been shown to provide good predictive performance, the learned low-dimensional latent representations fail to capture the fine-grained information on the user preference.

In earlier times, aspect-based recommender systems were proposed by leveraging topic models to extract latent semantic topics/aspects from reviews and learn multi-faceted user preferences, for instance, JMARS [25] and FLAME [26]. The recently proposed ALFM [1] integrated aspect importance of a user toward an item estimated by an aspect-aware topic model (ATM) into rating predictions. Despite effectiveness, topics extracted from these topic modelling methods are probabilistic distributions over independent words or phrases, and thus contextual information of words are neglected during the training process. In addition, short reviews make topic model related approaches more difficult to estimate the topic distributions [27]. An alternative type of aspect-based recommendations, such as EFM [28], LRPPM [29], and SULM [2], rely on external NLP tools [30] to extract aspects and sentiments from reviews. Similarly, TriRank [30] adopted the extracted aspects to construct the user-item-aspect tripartite graph for recommendations. Besides the fact that they are not self-contained, such methods largely depend on the performance of the external toolkit.

More recently, there has been a trend of applying deep-learning techniques into aspect-based recommendations. A^3NCF [31] leveraged neural attention layers to capture users' varied interests toward aspects that are defined as a combination of topic vector and embedding vector. AARM [3] modelled the user-item interactions between synonymous and similar aspects to tackle with data sparsity problem, and utilized a neural attention mechanism to consider user, item and aspect information simultaneously. ANR [4] proposed to use a neural architecture incorporated with a co-attention mechanism to perform aspect-based representation learning for both users and items and estimate aspect-level importance in an end-to-end fashion. However, none of the above methods has considered sentiment polarities toward aspects for different users and items such that it cannot explain to what extent a user likes or dislikes an item on various aspects. Very recently, Li et al. [5] proposed a capsule network-based model, namely, CARP, which was capable of reasoning the rating behaviour by discovering the informative logic unit embracing a pair of a viewpoint held by a user and an aspect of an item, and extracting the corresponding sentiments for rating prediction tasks. Despite the interpretability improvements to some extent, this method fails to enhance the diversity of recommendations as it neglects long-tail items. Furthermore, the considerably large inputs of word embeddings render the system less efficient. Apart from these, all of the above-mentioned methods did not consider the multilingual scenario, which is one of the key contributions in our work.

2.2 Multilingual Recommender Systems

Though there have been some studies on multilingual recommendation domain, this topic is still not fully investigated in the literature.

Traditional collaborative filtering is inherently multilingual, since it does not rely on content information of items but solely on the user's rating patterns. However, it encounters cold start issues when there is a rapid turnover of the recommended items. The work of Reference [32] required users trust that is not always easy to obtain, as crucial information to overcome the gap between multiple languages. In Reference [33], the authors proposed an LDA-based cross-lingual keyword recommendation method that can model both English and Japanese simultaneously. However, the problems lie in its inability to process more than two languages simultaneously and provide fine-grained recommendations. Some research works exploited well-known thesauri such as MultiWordNet [7, 8] and Wikipedia [9] to build language-independent user/item profiles for recommendation tasks. Narducci et al. [6] built concept-based representation of items by exploiting two knowledge sources, namely, Wikipedia and BabelNet, in the multilingual

recommendation. These works mainly rely on the use of ontologies and large corpora like Wikipedia, which are the key factors to determine the recommendation performance. However, they fail to consider fine-grained user preferences and sentiment information.

Specifically, in this article, we present a novel approach for multilingual recommendations that can provide fine-grained user and item modelling based on the multilingual aspect extraction and aspect-specific sentiment analysis. The vocabularies in different languages are embedded into the same space such that synonyms and similar words project closely. Meanwhile, the contributions of multiple languages to specific user/item are learned through a neural attention mechanism.

2.3 Multilingual Aspect-based Sentiment Analysis

There are only a handful of researches dealing with fine-grained level (i.e., topic or aspect level) sentiment analysis on multiple languages. One of the difficulties at topic/aspect-level is that the sentiments attach to specific groupings of words, and if these words are mistranslated or their sentiments are incorrectly inferred, there is no way to predict them correctly. Some studies adopt statistical machine translation (SMT) to overcome language barriers [34–36]. However, such approaches assume there is a high-quality machine translation system available for each language pair, which is not always true for under-resourced languages. Barnes et al. [37] compared several types of bilingual word embeddings and machine translation techniques for cross-lingual aspect-based sentiment classification. They show that distributional vector representations are more promising and produce results that are comparable to simple SMT baselines but still require more research.

The cross-lingual topic model provides a potential solution to help the aspect-level sentiment classification in a target language by transferring knowledge from a source language. Boyd-Graber et al. [38] developed the MUltilingual TOpic (MUTO) model to exploit matching across languages on term level to detect multilingual latent topics from unaligned texts. Zhang et al. [39] incorporated soft bilingual dictionary-based constraints into Probabilistic Latent Semantic Analysis (PLSA) so that it could extract shared latent topics in text data of different languages. However, these models do not consider sentiment factors and thus cannot help cross-lingual sentiment analysis. Some studies [40–42] tried to model aspects and sentiments in a unified framework for cross-lingual sentiment analysis. There are mainly two major drawbacks of these approaches. First, they are unable to capture the contextual information of words, which has been proven crucial to preserve topic coherence. Second, parameter-adjusting might be an onerous task when training these models, since they have too many parameters.

Different from the methods mentioned previously, in this work, we design a multiple instance learning framework integrated with hierarchical attention mechanism for multilingual aspect-based sentiment analysis without external resources. Our model first predicts sentiments over K aspects at the sentence-level and subsequently combines predictions up the document hierarchy.

3 THE PROPOSED MODEL

In this section, we elaborate the proposed MrRec, which aims to predict overall ratings based on captured multilingual user-item interactions in a fine-grained scale integrated with aspects and aspect-specific sentiments. First, we present the problem setting followed by the overview of our MrRec model. Then, we describe in detail the multilingual aspect-based sentiment analysis and the multilingual recommendation module for overall rating predictions. The notations used to describe our MrRec model are summarized in Table 1.

Table 1. Notations Used in the Article

Symbol	Description
$\mathcal{U}, \mathcal{I}, \mathcal{R}$	The set of users, items and ratings
$\mathcal{D}$	The set of reviews
$\mathcal{L}$	The set of languages
$\mathcal{A}$	The set of aspects
$\mathcal{F}$	The set of document representations
$\mathcal{P}$	The set of document-level sentiment distributions
$\mathcal{V}_n$	The set of negative samples in a minibatch
K	The number of aspects for items such as price, screen, battery, etc.
C	The number of classes separating the sentiment polarity score
L	The number of languages
N_f	The number of CNN filters
M_l	The number of reviews in language l
N_w	The number of words in the sentence
$r_{u,i}$	The rating rated by user u on item i
$\hat{r}_{u,i}$	The predicted rating of user u on item i
$d_{u,i}, l_{u,i}$	The review and language written and used by user u on item i
$\boldsymbol{\delta}_{u/i}$	K-dimensional vector with each element representing the importance degree of aspects of u/i with respect to i/u
$y_{u,i}^{(a_k)}$	Aspect utility representing user u's satisfaction with aspect a_k of item i
$\boldsymbol{A} \in \mathbb{R}^{K \times d}$	Language-independent aspect embedding matrix
$\boldsymbol{p}_{s,a_k}^{sen}$	C-dimensional vector, aspect sentiment distribution of sentence s on aspect a_k
$\boldsymbol{z}_{s,a_k}$	Aspect-specific sentence representation of sentence s on aspect a_k
$\boldsymbol{F}_{a_k}^{l}$	Document representation in language l on aspect a_k
$\boldsymbol{h}_i \in \mathbb{R}^d$	Multilingual word representation
$\boldsymbol{v}_s$	Sentence embedding of s
$\boldsymbol{p}_s^{asp}$	K-dimensional vector with each element p_{s,a_j} representing the possibility that sentence s belongs to aspect a_j
$\boldsymbol{r}'_i \in \mathbb{R}^d$	Aspect-based word embedding
$\boldsymbol{p}_s^{sen}$	C-dimensional vector representing the sentence-level sentiment distribution
$\boldsymbol{z}_s$	The representation of sentence s
$\boldsymbol{f}_{u/i,m,a_k}^{l}$	Document representation of the m-th review in language l on aspect a_k for user u/ item i
$\boldsymbol{p}_{d,a_k}^{sen}$	C-dimensional vector denoting the aspect sentiment distribution of review d on aspect a_k
$\hat{\boldsymbol{f}}_{u,a_k}^{l,t}$	Feature map obtained by the t-th filter on $\boldsymbol{F}_{u,a_k}^{l}$
$\boldsymbol{s}_{u,a_k}^{l}$	Language embedding on aspect a_k of user u
$\boldsymbol{w}^l$	Language-level contextual vector learned through training process
$\boldsymbol{u}_{a_k}/\boldsymbol{i}_{a_k}$	User/Item representation on aspect a_k
$\boldsymbol{U}_u/\boldsymbol{I}_i$	User/Item representation matrix on all aspects
$\boldsymbol{E}_u$	Affinity matrix whose element represents the similarity between the corresponding user and item pair representations on aspects
ω	C-dimensional sentiment polarity vector whose element denotes the sentiment score in $[-1,1]$
$polarity(d)^{a_k}$	Document-level sentiment polarity on aspect a_k
$\boldsymbol{W}_f$	Projection matrix that maps document-level representations and user representation into the same space
$r_{u \to i/i \to u,a_k}$	The aspect utility of user u/item i w.r.t. item i/user u on aspect a_k

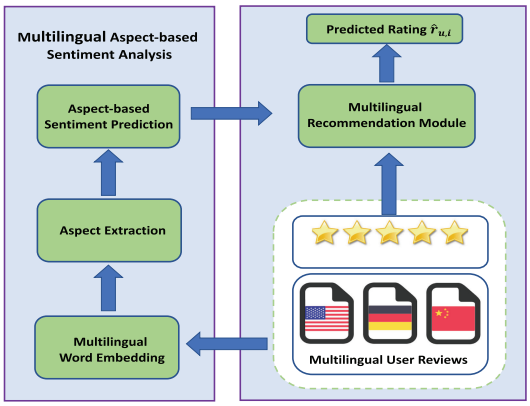


Fig. 2. The proposed MrRec framework for rating prediction tasks.

3.1 Problem Setting

Considering a set of ratings $\mathcal{R}$ accompanied by a set of reviews $\mathcal{D}$, for item set $\mathcal{I}$ and user set $\mathcal{U}$, each user-item interaction can be represented as a tuple $(u, i, r_{u,i}, d_{u,i}, l_{u,i})$, where $r_{u,i}$ is a numerical rating that can be seen as the overall sentiment the user u toward the item i, $d_{u,i}$ denotes the review text written by the user u on different aspects $a \in \mathcal{A}$ toward item i, and $l_{u,i} \in \mathcal{L}$ is the language used by u on i. In this article, we only consider the cases that all the items are from the same category, and we assume that these items share the same set of K aspects $\mathcal{A}$. The primary goal is to predict the unknown ratings of items that the users have not reviewed yet. Before introducing our method, we would like to clarify the necessary concepts being used in our article.

- **Overall rating:** An overall rating rated by user u on item i denoted as $r_{u,i}$ is a integer ranging from 1 to 5 stars. In our article, we set $r_{u,i}$ as a real value within $[1, 5]$ for easy computation.
- **Aspect:** It is a high-level semantic concept denoting the attribute of items the users commented on in reviews. An aspect set $\mathcal{A} = \{a_1, \ldots, a_K\}$ includes K aspects like *price, screen, battery* and *performance* for the mobile phone domain.
- **Aspect utility:** It is denoted as $y_{u,i}^{(a_k)} \in [-1, 1]$ representing the user u's satisfaction with aspect a_k of a given item i. Aspect utility can be derived by aspect sentiment polarities with -1 being the most dissatisfied and 1 being the most satisfied with aspect a_k.
- **Aspect importance:** For user u on item i, the aspect importance is represented by a K-dimensional vector $\boldsymbol{\delta}_u = (\delta_{u,1}, \ldots, \delta_{u,K})$, where the jth dimension $\delta_{u,j} \in [0, 1]$ indicates the importance degree of aspect a_j of u with respect to i. Similarly, for item i on user u, the aspect importance vector is $\boldsymbol{\delta}_i = (\delta_{i,1}, \ldots, \delta_{i,K})$, and $\delta_{i,k}$ indicates the importance degree of aspect a_k of i with respect to u.

3.2 Overview of MrRec Architecture

Figure 2 shows the overall architecture of our model, which consists of two components responsible for aspects extraction as well as aspect-specific sentiment analysis, and overall rating prediction.

Specifically, we feed the review set $\mathcal{D}$, its corresponding ratings $\mathcal{R}$ and languages $\mathcal{L}$ as the inputs to the MABSA module. Note that all inputs are from training split rather than validation or testing split. The training reviews are firstly transformed into a matrix $\boldsymbol{D} \in \mathbb{R}^{n \times d}$ via a multilingual embedding layer, which maps each word from the language vocabulary $\mathcal{V}$ to its corresponding

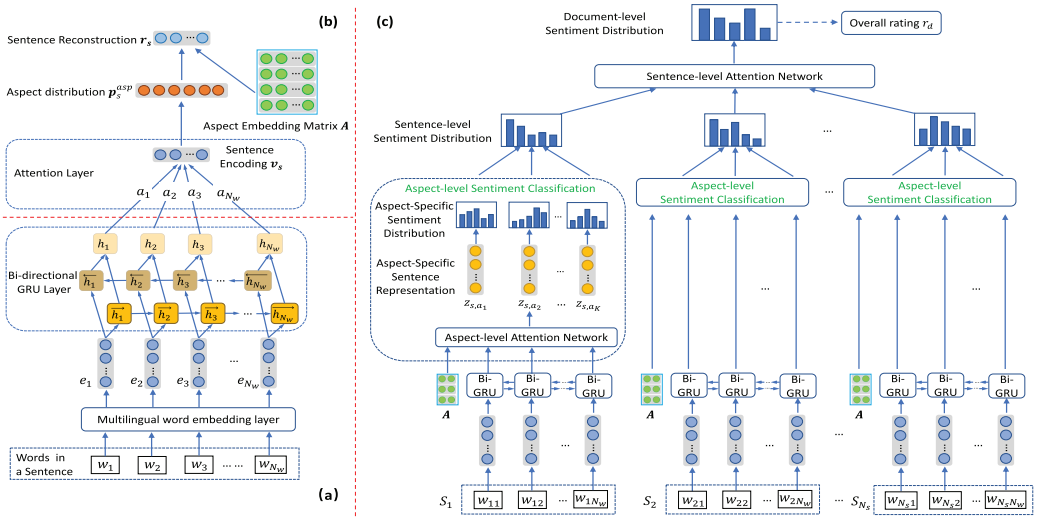


Fig. 3. Multilingual Aspect-based Sentiment Analysis Module. Panel (a) is the multilingual word embedding part that takes a sequence of words as input and outputs the learned multilingual word embeddings incorporated with the words' contextual information. Panel (b) is the aspect extraction part that learns aspect embedding matrix $\boldsymbol{A}$ in an unsupervised manner with the output of panel (a). Panel (c) is the aspect-based sentiment prediction part, which takes the output of panels (a) and (b) as input and learns document-level sentiment distribution with considering the overall ratings.

d-dimensional vector initialized with pre-trained multilingual word embeddings for better semantic representations of user/item documents. n is the number of words in the reviews. Then the embedding matrix $\boldsymbol{D}$ will be used to derive a set of language-independent aspect embedding matrix $\boldsymbol{A} \in \mathbb{R}^{K \times d}$ through multilingual aspect extraction component. After that, aspect-based sentiment prediction part will take $\boldsymbol{A}$ as input and generates aspect sentiment distribution over C classes $\boldsymbol{p}^{sen}_{s,a_k} = (p^{(1)}_{sen,s,a_k}, \ldots, p^{(C)}_{sen,s,a_k},), 1 \leq k \leq K$, and aspect-specific sentence representations $\boldsymbol{z}_{s,a_k}, 1 \leq k \leq K$.

In the second component, the inputs are document representations and document-level sentiment distributions of different aspects achieved through a weighted sum of the outputs from MABSA. Then the document representation set $\mathcal{F} = \{\boldsymbol{F}^l_{a_k} | 1 \leq k \leq K, 1 \leq l \leq L\}$ and document-level sentiment distribution set $\mathcal{P} = \{\boldsymbol{p}^{sen}_{d,a_k} | 1 \leq k \leq K, d \in \mathcal{D}\}$ are fed into MRM along with $\mathcal{R}$. $\boldsymbol{F}^l_{a_k} = (\boldsymbol{f}^l_{1,a_k}, \ldots, \boldsymbol{f}^l_{M_l,a_k})$, where M_l is the total number of reviews in language l, $\boldsymbol{f}^l_{m,a_k}$ is the real-value vector of document representation. The output of MRM is the predicted rating $\hat{r}_{u,i}$ of user u on item i.

3.3 Multilingual Aspect-based Sentiment Analysis Module

The architecture of MABSA module is depicted in Figure 3. The module is basically composed of three parts: (a) multilingual word embedding, (b) aspect extraction, and (c) aspect-based sentiment prediction.

3.3.1 Multilingual Word Embedding. For a given review $d_{u,i} \in \mathcal{D}$, suppose there are N_s sentences in $d_{u,i}$, and the jth sentence is composed by a sequence of words $\{w_{j1}, \ldots, w_{jN_w}\}$, where N_w is the total number of words in the sentence. For each word, we first use the multilingual

word embeddings[2] [43] to represent the word in the multilingual embedding vector space with its representation denoted as $\boldsymbol{e} \in \mathbb{R}^{d_e}$. We then adopt a bidirectional GRU [44] on $\boldsymbol{e}$ by summarizing information from both directions for word, and thus contextual information can be incorporated. Then the final word representation $\boldsymbol{h} \in \mathbb{R}^d$ can be derived through the concatenation of hidden states from both directions:

$$\boldsymbol{h} = [\overrightarrow{GRU}(\boldsymbol{e}); \overleftarrow{GRU}(\boldsymbol{e})]. \tag{1}$$

3.3.2 Aspect Extraction. Our work builds on the basis of the research of Reference [27], which is an analogous autoencoder called Attention-based Aspect Extraction (ABAE) model that learns aspect embedding matrix $\boldsymbol{A} \in \mathbb{R}^{K \times d}$ with K aspects identified by each row by minimizing the reconstruction error.

Given the word embedding $[\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{N_w}]$ of sentence s, the sentence encoding $\boldsymbol{v_s}$ is computed as the weighted average of word embeddings using an attention encoder:

$$\boldsymbol{v}_s = \sum_{i=1}^{N_w} \mu_i \cdot \boldsymbol{h}_i, \tag{2}$$

$$\mu_i = softmax(\boldsymbol{h}_i^T \cdot \boldsymbol{M_a} \cdot \boldsymbol{v}'_s), \tag{3}$$

where $\boldsymbol{v}'_s$ is simply the average of all word embeddings, μ_i is the attention weight on the ith word, and $\boldsymbol{M}_a \in \mathbb{R}^{d \times d}$ is an attention matrix that needs to be learned. The sentence embedding $\boldsymbol{v}_s$ is then fed into a softmax classifier to obtain a probability distribution over K aspects,

$$\boldsymbol{p}_s^{asp} = softmax(\boldsymbol{W}_a \cdot \boldsymbol{v}_s + \boldsymbol{b}_a), \tag{4}$$

where $\boldsymbol{W}_a \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_a \in \mathbb{R}^d$ are weights and bias. $\boldsymbol{p}_s^{asp} = (p_{s,a_1}, \ldots, p_{s,a_K})$ is a K-dimensional vector with each element $p_{s,a_j}, j \in [1, K]$ representing the possibility that sentence s belongs to aspect a_j. The reconstruction of the sentence s is a linear combination of aspects $\boldsymbol{A}$:

$$\boldsymbol{r}_s = \boldsymbol{A}^T \cdot \boldsymbol{p}_s^{asp}. \tag{5}$$

The model is trained by minimizing the reconstruction loss $\mathcal{L}_r = \sum_{s \in \mathcal{D}} max(0, 1 - \boldsymbol{r}_s \cdot \boldsymbol{v}_s + \boldsymbol{r}_s \cdot \boldsymbol{v}_h) + \lambda ||\tilde{\boldsymbol{A}} \cdot \tilde{\boldsymbol{A}}^T - \boldsymbol{I}||$, where $\tilde{\boldsymbol{A}}$ is $\boldsymbol{A}$ nomalized along each row, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{v}_h = argmin_{t \in \mathcal{V}_n} t \cdot \boldsymbol{v}_s$ represents the hardest one in a set of negative samples $\mathcal{V}_n$ in a minibatch.

Different from ABAE, we only focus on the hardest negative samples of different languages for computational efficiency [45]. When training on examples from different languages consecutively, it is difficult to learn a shared space that works well across languages. It is because only a subset of parameters is adjusted when training on each language, which may bias the model away to other languages. To avoid such issue, we follow the work of Reference [46] and sample parallel sentences from different language pairs in a cyclic fashion at each training iteration. Specifically, during each iteration, the number of samples per language is equal to the mini-batch size divided by L. We randomly re-select samples to pad the vacancies for those languages that have fewer reviews.

Note that in Equation (3), ABAE adopts word embedding $\boldsymbol{e}_i$ as input rather than $\boldsymbol{h}_i$, which makes the model originally a neural topic model. It is assumed that the sentence is composed with a bag of independent words, and thus the surrounding context among words are neglected when computing the global context of the sentence, $\boldsymbol{v}'_s$. By using the bidirectional GRU on each word embedding $\boldsymbol{e}_i$, we can summarize the information of the whole sentence centred around word w_i.

[2] https://fasttext.cc/docs/en/aligned-vectors.html.

3.3.3 Aspect-based Sentiment Prediction. Given multilingual word embeddings $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{N_w})$ from Equation (1), aspect matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K)$ and aspect distribution $\boldsymbol{p}_s^{asp}$ as inputs, for sentence s, aspect-based sentiment prediction module will output the document-level sentiment distribution $\boldsymbol{p}_d^{sen}$ on review d.

The idea of this module is based on *Multiple Instance Learning (MIL)* framework [12, 13], which deals with the problems where labels (document-level sentiment polarities in our case) are associated with groups of instance or bags (sentences), while instance labels are unseen. In our scenario, we assume that the sentiment distribution of document (overall rating) is composed as the weighted sum of the sentiments of each segment (sentence), which are the linear combinations of sentiment polarities of their associated aspects. To the best of our knowledge, we are the first that applies MIL framework to multilingual sentiment analysis.

The architecture of our module is shown in Figure 3(c). Particularly, we propose an aspect-level attention mechanism to fuse the information of aspects to the representations of target sentences,

$$\boldsymbol{r}_i' = \boldsymbol{W}_e \cdot [\boldsymbol{h}_i; \boldsymbol{a}_j], \tag{6}$$

$$\alpha_i = softmax(\boldsymbol{h}_c^T \cdot tanh(\boldsymbol{W}_c \cdot [\boldsymbol{h}_i; \boldsymbol{r}_i'])), \tag{7}$$

where $\boldsymbol{r}_i' \in \mathbb{R}^d$ can be seen as aspect-based word embedding, and α_i represents the importance of the ith word in sentence s. $\boldsymbol{W}_e \in \mathbb{R}^{d \times 2d}$ and $\boldsymbol{W}_c \in \mathbb{R}^{d_c \times 2d}$ are weight matrices. $\boldsymbol{h}_c \in \mathbb{R}^{d_c}$ is a learnable parameter. Then, the aspect-aware sentence representation can be achieved by weighted summation of all word embeddings in the sentence:

$$\boldsymbol{z}_{s,a_j} = \sum_{i=1}^{N_w} \alpha_i \cdot \boldsymbol{h}_i. \tag{8}$$

The sentence representation $\boldsymbol{z}_{s,a_j}$ is fed into a softmax layer to predict the aspect-specific sentiment distribution on sentence s with respect to aspect a_j:

$$\boldsymbol{p}_{s,a_j}^{sen} = softmax(\boldsymbol{W}_s \cdot \boldsymbol{z}_{s,a_j} + \boldsymbol{b}_s), \tag{9}$$

where $\boldsymbol{W}_s$ and $\boldsymbol{b}_s$ are the parameters. $\boldsymbol{p}_{s,a_j}^{sen}$ is a real-valued vector $(p_{sen,s,a_j}^{(1)}, \ldots, p_{sen,s,a_j}^{(C)})$ with 1 and C representing the most negative and most positive polarity score, respectively. For instance, supposing a 5-class scenario, C represents 5 classes and $p_{sen,s,a_j}^{(k)}, k \in [1, C]$ denotes the probability that the polarity score equals to k of sentence s with respect to aspect a_j. Thus, the sentence-level sentiment distribution can be calculated as

$$\boldsymbol{p}_s^{sen} = \sum_{j=1}^{K} p_{s,a_j} \cdot \boldsymbol{p}_{s,a_j}^{sen}. \tag{10}$$

Each element $p_{sen,s}^{(k)}, k \in [1, C]$ of $\boldsymbol{p}_s^{sen}$ represents the probability that the polarity score is equal to k of sentence s. After that, the sentence representation on all aspects can be achieved by

$$\boldsymbol{z}_s = \sum_{j=1}^{K} p_{s,a_j} \cdot \boldsymbol{z}_{s,a_j}. \tag{11}$$

Similarly, to capture the context around the target sentence s, we feed $\boldsymbol{z}_s$ to the bi-directional GRU layer $\boldsymbol{h}_s = [\overrightarrow{GRU}(\boldsymbol{z}_s); \overleftarrow{GRU}(\boldsymbol{z}_s)]$. To learn different contributions of sentences in a review, we adopt a sentence-level attention network defined as follows:

$$\beta_s = softmax(\boldsymbol{h}_r^T \cdot tanh(\boldsymbol{W}_r \cdot \boldsymbol{h}_s + \boldsymbol{b}_r)), \tag{12}$$

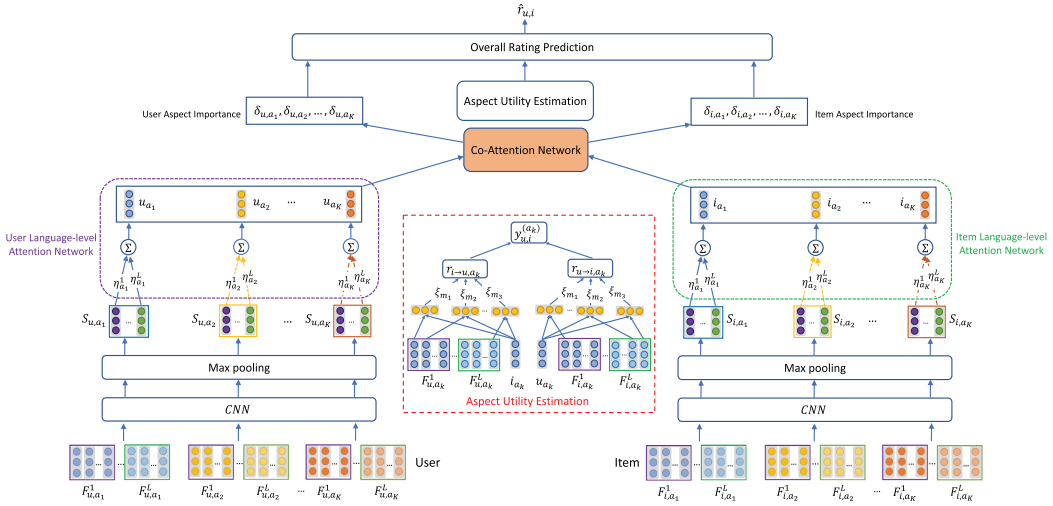


Fig. 4. Multilingual Recommendation Module.

where $\boldsymbol{h}_r \in \mathbb{R}^{d_r}$, $\boldsymbol{W}_r \in \mathbb{R}^{d_r \times d}$, and $\boldsymbol{b}_r \in \mathbb{R}^{d_r}$ are learnable parameters. Finally, we obtain the document-level sentiment distribution as the weighted sum of sentence distributions:

$$\boldsymbol{p}_d = \sum_{s=1}^{N_s} \beta_s \cdot \boldsymbol{p}_s^{sen}, \tag{13}$$

where N_s is the number of sentences in review d.

The aspect-based sentiment prediction is trained end-to-end on all training reviews guided by the overall ratings accompanied with reviews. We use the negative log-likelihood as the objective function:

$$\mathcal{L}_s = -\sum_{d \in \mathcal{D}} log p_d^{(r_d)}, \tag{14}$$

where $r_d \in [1, C]$ is the polarity score of review d.

3.4 Multilingual Recommendation Module

The overall architecture of MRM is depicted in Figure 4. Before we delve into the detail of MRM module, several pivotal intuitions need to be highlighted that we aim to capture through the module.

- **Intuition 1:** Not all languages are of equal importance in review sets for the user and item modelling.
- **Intuition 2:** The importance of the same aspect could be entirely different for different users, which makes it possible that different users have different requirements for the same aspect of an item. Besides, different users may focus on different aspects of the same item.

Based on the above intuitions, the basic idea of MRM is that, given the output from MABSA module and the review sets for target user u as well as candidate item i, to predict the overall rating of user u toward the item i, we first learn user/item representation of each language on different aspects. Then language-level attention network is proposed to learn language importance for user/item on different aspects, and user/item representation on each aspect can be derived with respect to different languages (**Intuition 1**). After that, co-attention network and aspect utility

estimation will be performed in a parallel way to derive the user and item aspect importance for each user-item pair, as well as user's satisfaction toward candidate item on different aspects, respectively (**Intuition 2**). Finally, the overall rating of the target user toward the candidate item can be inferred via a prediction layer with considering the two perspectives.

Specifically, given the review set $\mathcal{F}_u = \{\boldsymbol{F}^l_{u,a_k} | 1 \le k \le K, 1 \le l \le L\}$ written by user u and the review set $\mathcal{F}_i = \{\boldsymbol{F}^l_{i,a_j} | 1 \le j \le K, 1 \le l \le L\}$ written for item i, as input to multilingual recommendation module (MRM). $\boldsymbol{F}^l_{u/i,a_k} = (\boldsymbol{f}^l_{u/i,1,a_k}, \ldots, \boldsymbol{f}^l_{u/i,M_l,a_k})$, where $\boldsymbol{f}^l_{u/i,m,a_k} \in \mathbb{R}^d$ denotes the document representation of the mth review in language l on aspect a_k for user u or item i, and M_l is the total number of reviews in language l. To obtain it, we first learn sentence representation incorporated with contextual fusion using bi-directional GRU with input from Equation (8): $\boldsymbol{h}_{s,a_k} = [\overrightarrow{GRU}(\boldsymbol{z}_{s,a_k}); \overleftarrow{GRU}(\boldsymbol{z}_{s,a_k})]$. Then the importance of sentence s on aspect a_k can be calculated as

$$\beta'_{s,a_k} = softmax(\boldsymbol{h}_t^T \cdot tanh(\boldsymbol{W}_t \cdot \boldsymbol{h}_{s,a_k} + \boldsymbol{b}_t)), \tag{15}$$

where $\boldsymbol{h}_t \in \mathbb{R}^{d_t}$, $\boldsymbol{W}_t \in \mathbb{R}^{d_t \times d}$ and $\boldsymbol{b}_t \in \mathbb{R}^{d_t}$ are learnable parameters. The document representation can be achieved by the weighted sum of sentence representations. Likewise, document-level sentiment distribution on the aspect can also be derived through a weighted sum of aspect sentiment distributions:

$$\boldsymbol{f}^l_{u/i,M_l,a_k} = \sum_{s=1}^{N_s} \beta'_{s,a_k} \cdot \boldsymbol{h}_{s,a_k}, \quad \boldsymbol{p}^{sen}_{d,a_k} = \sum_{s=1}^{N_s} \beta'_{s,a_k} \cdot \boldsymbol{p}^{sen}_{s,a_k}. \tag{16}$$

Since the modelling process for users and items are identical, we focus on illustrating the process for a given user.

3.4.1 Language-specific Aspect-based User Representation. First, the user review set $\boldsymbol{F}^l_{u,a_k}$ is grouped by different languages and aspects, which is fed into MRM as input. To capture the semantic features of reviews, we employ a CNN network to perform convolution operations on each $\boldsymbol{F}^l_{u,a_k}$ matrix with N_f filters. Since we do not consider the orders of reviews for users and items, we set the window size to 1 to extract features from each review independently. Specifically, for review $\boldsymbol{f}^l_{u,j,a_k}$, we perform: $\hat{f}^{l,t}_{u,j,a_k} = \sigma(\boldsymbol{W}_t * \boldsymbol{f}^l_{u,j,a_k} + b_t)$, where $*$ is the convolution operator, $\boldsymbol{W}_t$ is the tth convolution filter, $b_t \in \mathbb{R}$ is a bias term, and σ is a non-linear function, i.e., ReLU. By applying the tth filter on the $\boldsymbol{F}^l_{u,a_k}$ matrix, we obtain a feature map represented as $\hat{\boldsymbol{f}}^{l,t}_{u,a_k} = (\hat{f}^{l,t}_{u,1,a_k}, \ldots, \hat{f}^{l,t}_{u,M_l,a_k})$. Then max-pooling is applied to find the most important feature on the subset of reviews $s^{l,t}_{u,a_k} = max(\hat{\boldsymbol{f}}^{l,t}_{u,a_k})$. After performing on all filters, we obtain the vector $\boldsymbol{s}^l_{u,a_k} = (s^{l,1}_{u,a_k}, \ldots, s^{l,N_f}_{u,a_k}) \in \mathbb{R}^{N_f}$, which can be seen as the language-specific representation of user u on aspect a_k. The outputs from max-pooling layer that represent the same aspect a_k are concatenated to form a matrix $\boldsymbol{S}_{u,a_k} = (\boldsymbol{s}^1_{u,a_k}, \ldots, \boldsymbol{s}^L_{u,a_k}) \in \mathbb{R}^{L \times N_f}$.

3.4.2 Language-level Attention Network. We argue that not all languages are of equal importance to the user. For instance, if a user u's primary language is French and s/he also writes reviews in English, French should be more important than English in most cases. In other words, French contributes more than English in learning user representation. Note that when we refer to *primary language*, we mean the language that is the most informative one for the user u. Therefore, inspired by the related research of self-attention network [47], we propose a language-level attention network.

Indicatively, we measure the importance of the language as the similarity of $\boldsymbol{s}^l_{u,a_k}$ with a language-level context vector $\boldsymbol{w}^l$ and get a normalized importance weight $\eta^l_{a_k}$ through a softmax

function. The context vector $\boldsymbol{w}^l$ can be seen as a high-level representation of a fixed query, "which is the most informative language" over the languages adopted by the user u,

$$\eta^l_{a_k} = softmax((\boldsymbol{w}^l)^T \cdot \boldsymbol{s}^l_{u,a_k}), \tag{17}$$

where $\boldsymbol{w}^l \in \mathbb{R}^{N_f}$ is randomly initialized and learned through model training process. Then a weighted combination of language-specific user representations on aspect a_k is considered as the representation of user u on aspect a_k:

$$\boldsymbol{u}_{a_k} = \sum_{l=1}^{L} \eta^l_{a_k} \cdot \boldsymbol{s}^l_{u,a_k}. \tag{18}$$

The representation of user u on all aspects are denoted as $\boldsymbol{U}_u = (\boldsymbol{u}_{a_1}, \ldots, \boldsymbol{u}_{a_K})$. Similarly, we learn language importance on item i's review set and obtain the item representation matrix denoted as $\boldsymbol{I}_i = (\boldsymbol{i}_{a_1}, \ldots, \boldsymbol{i}_{a_K})$.

3.4.3 Co-attention Network. The self-attention mechanism focuses on the "static" features of users or items rather than the features of user-item interactions, and thus is suboptimal to learn the importance among aspects of user u taken specific item i into account, and vice versa. Therefore, following the work of References [48–50], we propose to learn the aspect importance of user u or item i in a joint manner.

To incorporate item i as context when calculating the aspect importance of user u, we need to know how user u and item i matches on certain aspects:

$$\boldsymbol{E}_u = \sigma(\boldsymbol{U}_u \cdot \boldsymbol{W}_e \cdot \boldsymbol{I}_i^T), \tag{19}$$

where $\boldsymbol{W}_e \in \mathbb{R}^{N_f \times N_f}$ is a learnable parameter, and each entry of $\boldsymbol{E}_u \in \mathbb{R}^{K \times K}$ represents the similarity between the corresponding user and item pair representations on aspects. Next, the aspect-level importance of user u w.r.t. item i can be learned as

$$\boldsymbol{H}_u = \sigma(\boldsymbol{U}_u \cdot \boldsymbol{W}_u + \boldsymbol{E}_u(\boldsymbol{I}_i \cdot \boldsymbol{W}_i)), \boldsymbol{\delta}_u = softmax(\boldsymbol{H}_u \cdot \boldsymbol{v}_u), \tag{20}$$

where $\boldsymbol{W}_u, \boldsymbol{W}_i \in \mathbb{R}^{N_f \times d_f}$, and $\boldsymbol{v}_u \in \mathbb{R}^{d_f}$ are learnable parameters. $\boldsymbol{\delta}_u = (\delta_{u,a_1}, \ldots, \delta_{u,a_K})$ is a K-dimensional vector with each element representing the importance of the corresponding aspect for user u. Likewise, the aspect importance of item i can be derived as $\boldsymbol{\delta}_i = (\delta_{i,a_1}, \ldots, \delta_{i,a_K})$.

3.4.4 Aspect Utility Estimation. When calculating user u's satisfaction with each aspect a_k of item i, for the improvement of recommendation diversity, we need to consider not only the utilities of other users that rated item i on aspect a_k but also the user u's individual utilities assigned by user u to items that are similar to the item i on aspect a_k even though the items are less popular (long-tail items). Hence, a dual interactive attention mechanism is designed to learn the aspect-level ratings of user u on item i and vice versa.

Given the aspect-specific sentiment distribution on document d w.r.t aspect a_k, $\boldsymbol{p}^{sen}_{d,a_k} = (p^{(1)}_{sen,d,a_k}, \ldots, p^{(C)}_{sen,d,a_k})$, and aspect-level document representations $\{\boldsymbol{f}^l_{u/i,m,a_k} \mid 1 \le m \le M_{u/i}, 1 \le k \le K\}$, to estimate the aspect utility of user u on item i $r_{u \to i,a_k}$, and the aspect utility of item i w.r.t. user u $r_{i \to u,a_k}$, we first define a real-valued sentiment polarity vector $\boldsymbol{\omega} = (\omega^{(1)}, \ldots, \omega^{(C)})$, where $\omega^{(c)} \in [-1, 1]$ represents a weight assigned according to the discrete uniform distribution so that $\omega^{(c+1)} - \omega^{(c)} = \frac{2}{C-1}$. For instance, the sentiment polarity vector of a five-class scenario would be $\omega = (-1, -0.5, 0, 0.5, 1)$. Thus, the document-level sentiment polarity on aspect a_k can be calculated as

$$polarity(d)^{a_k} = \sum_{c \in [1,C]} p^{(c)}_{sen,d,a_k} \cdot \omega^{(c)}. \tag{21}$$

Next, to find how the attribute of the candidate item i on aspect a_k characterized by other users, matches the user u's requirement on the same aspect, we define the element-wise product of user representation on aspect a_k and document-level representation of review m_i w.r.t. item i and aspect a_k,

$$\phi(u, m_i) = \boldsymbol{u}_{a_k} \odot (\boldsymbol{W}_f \cdot \boldsymbol{f}_{i,m_i,a_k}), \tag{22}$$

where $\boldsymbol{f}_{i,m_i,a_k} \in \mathcal{F}_i$ denotes the document representation of review m_i that is trained to characterize the attribute of item i on aspect a_k. $\boldsymbol{W}_f \in \mathbb{R}^{N_f \times d}$ is the projection matrix used to map document-level representations and user representation to the same space. The contribution of review m_i to user u can be learned by a softmax layer:

$$\xi_{m_i} = softmax(\boldsymbol{W}_{att}^T \cdot \phi(u, m_i)), \tag{23}$$

where $\boldsymbol{W}_{att} \in \mathbb{R}^{N_f}$ is a learnable parameter. The larger the value of ξ_{m_i} is, the more the review matches closely to the user u's taste on aspect a_k. Then, we can obtain the aspect utility of user u to candidate item i on aspect a_k:

$$r_{u \to i, a_k} = \sum_{m_i=1}^{|\mathcal{F}_i|} \xi_{m_i} \cdot polarity(d_{m_i})^{a_k}. \tag{24}$$

Similarly, the aspect utility of item i w.r.t. user u can be calculated as: $r_{i \to u, a_k} = \sum_{m_u=1}^{|\mathcal{F}_u|} \xi_{m_u} \cdot polarity(d_{m_u})^{a_k}$, where $|\mathcal{F}_u|$ and $|\mathcal{F}_i|$ are total number of reviews in user u's set and item i's set.

To learn user u's overall satisfaction with item i on the aspect a_k, a regression layer is stacked to the concatenation of these two aspect-level ratings:

$$y_{u,i}^{(a_k)} = \boldsymbol{W}_y \cdot \begin{bmatrix} r_{u \to i, a_k} \\ r_{i \to u, a_k} \end{bmatrix}. \tag{25}$$

3.4.5 Overall Rating Prediction. The overall rating for user-item pair can be predicted via a prediction layer with the combination of the user's satisfaction $y_{u,i}^{(a_k)}$ and the aspect importance $\delta_{u,a_k}, \delta_{i,a_k}$ as inputs:

$$\hat{r}_{u,i} = \sigma_C \left(\sum_{k=1}^{K} \delta_{u,a_k} \cdot \delta_{i,a_k} \cdot y_{u,i}^{(a_k)} \right) + b_u + b_i + b, \tag{26}$$

where b_u, b_i, and b are user, item, and global bias, respectively. $\sigma_C(x) = 1 + \frac{C-1}{1+exp(-tan(\frac{\pi}{2}x))}$ is a variant of sigmoid function, producing the value within the range of $[1, C]$. Note that since $x \in [-1, 1]$ needs to be mapped to the range of $[1, C]$, we first map x to radian space, which is then prolonged to $[-\frac{\pi}{2}, \frac{\pi}{2}]$. $tan(\cdot)$ function is adopted to project x to the range of $[-\infty, \infty]$. Finally, the variant of the sigmoid function can be used to achieve the goal. The model parameters can be learned through backpropagation with the standard Mean-squared Error (MSE) as the loss function. The three submodules of MrRec (aspect extraction, aspect-specific sentiment analysis, and multilingual recommendation) need to be learned separately. The performance of each submodule implicitly relies on the outputs from the previous one. Thus, we adopt a pre-trained multilingual word embedding to improve the performance. To train the first two submodules, we uniformly mix the training set with different languages. To deal with the overfitting issue in deep-learning models, we apply the dropout technique with parameter ρ, and L_2 regularization term to the objective function.

3.5 Complexity Analysis

In MrRec, the training process consists of two parts, i.e., MABSA and MRM. The time complexity of the whole model mainly depends on the dimensionality of embedding vectors and the size of training data. The complexity of MABSA consists of aspect extraction module and sentiment prediction module with the complexity of $O(|\mathcal{D}_{train}|KN_sN_w^2d^2)$ and $O(|\mathcal{D}_{train}|KN_s^2N_w^2d^2)$. $|\mathcal{D}_{train}|$ represents the number of reviews in the training set, K represents the number of aspects, N_s is the number of sentences in one review, N_w is the number of words in the sentence, and d is the dimension of word embeddings. In total, the complexity of MABSA is $O(|\mathcal{D}_{train}|KN_w^2N_s^2d^2)$. For MRM, the complexity depends on the aspect importance in Equation (20) and the user's overall satisfaction in Equation (25) with the complexity of $O(K^2N_f^2)$ and $O((|\mathcal{F}_u|^2 + |\mathcal{F}_i|^2)(N_fd + N_s^2d^2))$. In total, the complexity of MRM is $O(|\mathcal{R}_{train}|K((|\mathcal{F}_u|^2 + |\mathcal{F}_i|^2)(N_fd + N_s^2d^2) + K^2N_f^2))$, where $|\mathcal{R}_{train}|$ represents the number of ratings in the training set, $|\mathcal{F}_u|$ and $|\mathcal{F}_i|$ represent the number of reviews written by the user and the number of reviews written for the item, N_f is the number of filters. Since $|\mathcal{R}_{train}| = |\mathcal{D}_{train}|$, the overall complexity of MrRec is $O(|\mathcal{R}_{train}|K((|\mathcal{F}_u|^2 + |\mathcal{F}_i|^2)(N_fd + N_s^2d^2) + K^2N_f^2 + N_w^2N_s^2d^2))$. In practice, the MABSA part can be trained offline in advance, which further improves the efficiency of our model.

4 EXPERIMENTS

To evaluate our proposed MrRec model, we designed and conducted extensive experiments to answer the following six research questions.

- RQ1: How does our MABSA module perform as compared with state-of-the-art multilingual aspect-based sentiment analysis methods? (Section 4.2)
- RQ2: How does MrRec perform in terms of effectiveness and efficiency as compared with state-of-the-art recommendation algorithms? (Section 4.3)
- RQ3: How do the hyper-parameters, such as the dimension of Bi-GRU output d and the number of aspects K, affect the performance of our model? (Section 4.4)
- RQ4: How does MrRec perform in handling the cold-start issue? (Section 4.5)
- RQ5: How different components in our model contribute to the overall performance? (Section 4.6)
- RQ6: How can MrRec interpret the recommendation results? (Section 4.7)

In what follows, we first describe the experimental settings, and then we answer the above six research questions.

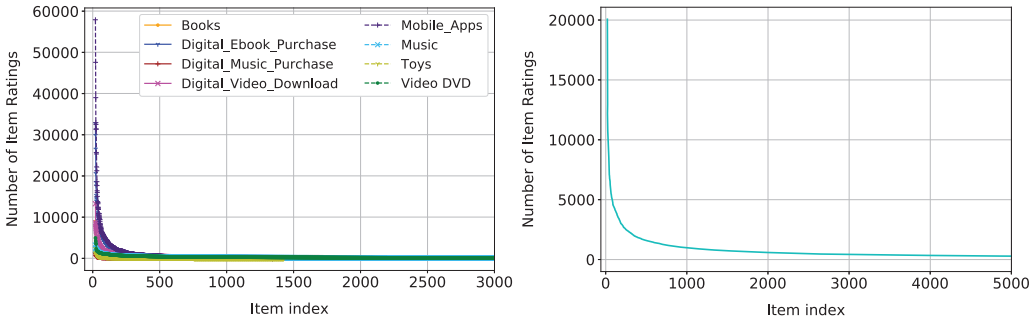
4.1 Experimental Settings

4.1.1 Datasets. We evaluate our proposed model on rating predictions against several state-of-the-art baselines with real-world datasets freely available online. Specifically, we use nine datasets from two sources: Amazon Customer Reviews[3] and Book Reviews.[4] The datasets cover 11 languages: Afrikaans (AF), English (EN), German (DE), French (FR), Catalan (CA), Spanish (ES), Italian (IT), Norwegian (NO), Romanian (RO), Slovenian (SL), and Tagalog (TL). For Amazon Customer Reviews dataset, eight datasets from different domains are used (i.e., Books, Digital Ebook Purchase, Digital Music Purchase, Digital Video Download, Mobile Apps, Music, Toys, and Video DVD). The other is from the Book Reviews dataset. Note that we determine not to apply the *k-core* settings [51, 52] over these datasets, whereby there are at least k ratings/reviews for each user and item, as

[3] https://s3.amazonaws.com/amazon-reviews-pds/readme.html.
[4] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/reviews?authuser=0.

Table 2. Statistics of the Datasets for Evaluating the Recommendation Task

Datasets	*Books*	*Digital Ebook Purchase*	*Digital Music Purchase*	*Digital Video Download*	*Mobile Apps*	*Music*	*Toys*	*Video DVD*	*Goodreads*
# Users	847,499	1,118,718	125,381	758,052	1,161,439	753,598	96,819	1,038,981	440,817
# Items	26,642	5,392	16,310	18,674	1,327	28,540	1,408	37,365	1,901,485
# Interactions/Reviews	1,165,926	1,534,618	159,320	1,078,790	1,709,289	1,318,337	108,547	1,908,260	14,668,579
# Multilingual users	3,017	5,890	528	12,164	12,775	6,324	232	10,682	84,669
# Multilingual items	8,180	2,699	2,518	6,345	1,326	16,632	926	15,536	147,025
# Multilingual interactions	1,033,626	1,493,101	87,749	999,787	1,709,263	1,261,610	105,448	1,627,160	11,478,423
Avg. # words/ review	115.0 *(σ=184.3)*	49.7 *(σ=86.1)*	53.2 *(σ=97.6)*	34.6 *(σ=62.7)*	29.9 *(σ=35.0)*	114.3 *(σ=179.5)*	46.8 *(σ=76.5)*	101.7 *(σ=184.8)*	129.3 *(σ=181.1)*
Avg. # sentences/ review	7.066 *(σ=10.223)*	3.842 *(σ=5.091)*	4.107 *(σ=5.899)*	3.084 *(σ=3.723)*	2.741 *(σ=2.393)*	7.436 *(σ=11.240)*	3.604 *(σ=4.254)*	6.527 *(σ=9.903)*	8.415 *(σ=11.179)*
Avg. # reviews/ user	1.376 *(σ=3.810)*	1.372 *(σ=1.498)*	1.271 *(σ=1.133)*	1.423 *(σ=1.480)*	1.472 *(σ=1.638)*	1.749 *(σ=4.612)*	1.121 *(σ=0.553)*	1.837 *(σ=5.179)*	33.276 *(σ=114.7)*
Avg. # reviews/ item	43.76 *(σ=174.9)*	284.61 *(σ=1247.2)*	9.77 *(σ=31.5)*	57.77 *(σ=290.0)*	1288.09 *(σ=3702.8)*	46.19 *(σ=116.9)*	77.09 *(σ=144.2)*	51.07 *(σ=125.6)*	7.714 *(σ=70.4)*
Density	0.005%	0.025%	0.008%	0.008%	0.111%	0.006%	0.080%	0.005%	0.002%



(a) Amazon Dataset (b) Goodreads Dataset

Fig. 5. The popularity distribution of items in the experimental datasets.

it trivializes the problem of data sparsity, which is inevitable in real-world recommendations. The basic statistics are summarized in Table 2. Besides, we also plot the popularity distribution of item set on two dataset sources in Figure 5, from which we can see a substantial amount of long-tail items that need to be considered when providing recommendations.

To evaluate the performance of our multilingual aspect-based sentiment prediction module, we adopt Trip-MAML[5] dataset, which consists of TripAdvisor hotel reviews in English, Italian, and Spanish. Besides, we also produce a multilingual dataset that incorporates English and French reviews on restaurant domain to test our module. Specifically, we adopt English restaurant reviews[6] follow the work of Reference [53], and French restaurant reviews[7] from Reference [54], which are then combined to form a multilingual datasets denoted as Restaurant Reviews. The statistics of the datasets are presented in Table 3. For both datasets, each review comes with an overall rating on a discrete ordinal scale from 1 to 5 "stars." The datasets are annotated at sentence-level with 3-values sentiment labels including Positive, Negative and Neutral/Mixed. Each sentence is manually anno-

[5]http://hlt.isti.cnr.it/trip-maml/.
[6]http://dilab.korea.ac.kr/jmts/jmtsdataset.zip.
[7]http://metashare.ilsp.gr:8080/repository/search/?q=semeval+2016.

Table 3. Statistics of the Datasets for Evaluating the Aspect-based Sentiment Analysis Task

Datasets	**Trip-MAML**			**Restaurant Reviews**	
	EN	*ES*	*IT*	*EN*	*FR*
# Reviews	442	500	500	652	455
# Sentences	5,799	2,620	2,593	3,418	2,427
# Opinions	5,587	3,416	3,602	3,742	3,484
# Positive opinions	3,344	2,402	2,484	2,278	1,605
# Negative opinions	1,377	792	651	855	1,646
# Neutral opinions	866	222	467	609	233

tated according to 12 recurrent aspects, i.e., *Rooms, Cleanliness, Value, Service, Location, Check-in, Business, Food, Building, Sleep Quality, Other* as well as *NotRelated*, and 7 recurrent aspects, i.e., *Restaurant, Food, Service, Ambience, Price, Location*, as well as *Miscellaneous*, for Trip-MAML and Restaurant Reviews, respectively.

As for preprocessing, we perform the following steps: (1) set maximum length of raw documents to 300; (2) split documents into sentences that are then tokenized into words, and the words are further converted into lowercases; (3) shorten the words with redundant characters into their canonical forms (e.g., cooooool is converted to cool); (4) remove URLs and HTML tags such as $
$; (5) remove the duplicates and records with empty or invalid content. Furthermore, we convert all rating ranges in all datasets to $[1, 5]$, and therefore the C is set to 5. For each dataset, we randomly split the training and testing set according to the ratio of 80:20. Moreover, 10% reviews in the training set are left out as a validation set for hyper-parameter selection. Note that for records in the testing set, at least one interaction for each user or item is included in the training set, and otherwise will be moved from the testing set to the training set. To make the experiments repeatable, we make the pre-processed datasets publicly available.[8]

4.1.2 Evaluation Metrics. Performance of rating prediction tasks is evaluated on the testing set via MSE, which is widely adopted in the recommendation domain.

Despite the importance on measuring the recommendation performance of MSE, user experience can be greatly enhanced if the systems provide diverse recommendations. To evaluate the diversity of our proposed method, we first generate top-N recommendation list $L(N)$ to the target user according to $\hat{r}_{u,i}$ in descending order. More advanced ranking algorithms are out of the scope in this article. These N items should present various characteristics in terms of, i.e., aspects. Then the following metrics are utilized in this article as measurements:

Intra-list Similarity. This metric proposed by Reference [55] assesses diversity on an individual level. The rationale behind this metric is that each user prefers recommendations from various categories. Assuming i and j are two different items in the recommendation list, the similarity between i and j can be measured via binary similarity calculated upon the training set, which is defined as

$$Sim(i, j) = \frac{\#users\ that\ click\ both\ i\ and\ j}{\sqrt{\#users\ that\ click\ i} \cdot \sqrt{\#users\ that\ click\ j}}. \tag{27}$$

[8] https://drive.google.com/file/d/15XSiPVSwjPCdl1SIIAJf6uIgPRJzPKua/view?usp=sharing.

Thus, the intra-list similarity (ILS) can be defined as

$$ILS = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{(i,j) \in L(N), R(i) < R(j)} Sim(i,j). \tag{28}$$

The lower the ILS value is, the more diverse the recommender system is.

Novelty. The novelty of a recommender system evaluates the likelihood of a recommender system to give recommendations to the user that they are not aware of, or that they have not seen before. The definition of novelty is varied in publications according to its context and purpose. In this article, we apply the population-oriented item novelty evaluation metric introduced in Reference [56] as expected popularity complement (EPC) to measure the ability of our recommender system to recommend items from long-tail. Its definition is as follows:

$$EPC = \frac{\sum_{u \in \mathcal{U}} \sum_{r=1}^{N} \frac{rel(u,i_r) * (1 - pop(i_r))}{log_2(r+1)}}{\sum_{u \in \mathcal{U}} \sum_{r=1}^{N} \frac{rel(u,i_r)}{log_2(r+1)}}, \tag{29}$$

where i_r denotes the item ranked to the rth place in the recommendation list. $rel(u, i_r)$ is a binary function with values of 1 or 0 representing if the user u rated the item i_r or not, respectively. The popularity $pop(i)$ is calculated based on the times the item has been rated in training set, and it can be defined as

$$pop(i) = \frac{|rate(i)|}{max_{j \in \mathcal{I}} |rate(j)|}, \tag{30}$$

where $rate(i)$ denotes the number of ratings of item i and the denominator is the maximum number of ratings obtained by an item in item set. It is desirable for a recommender system to have a high EPC value when it not only recommends items from long-tail but also ranks them highly in the recommendation list.

Besides, we adopt the precision (P), recall (R), and F1-score as evaluation metrics for multilingual aspect-based sentiment analysis.

4.1.3 Baselines. To validate the performance of our proposed model in recommendation tasks, we compare with the following approaches:

- **MF [57]:** It characterizes users and items by vectors with implicit feedbacks inferred from item rating patterns.
- **NAIS: [58]** It learns the importance of user's historical clicking items via a neural attention network, which is then integrated into the item-based collaborative filtering for rating prediction.
- **Tran-D-Attn [14]:** It models user preferences and item characteristics by CNNs with dual local and global attention mechanism for review rating prediction.
- **Tran-ALFM [1]:** It is an aspect-based recommender system with aspect discovered by an aspect-aware topic model on review texts. A weighted matrix is introduced to associate latent factors with aspects by using MF approach to predict ratings.
- **Tran-ANR [4]:** It performs aspect-based representation learning to model both user preferences and item properties. The neural co-attention mechanism is introduced to learn the aspect-level user and item importance.
- **Tran-CARP [5]:** The model predicts ratings based on sentiment-aware representations of user-item interactions, which are learned via a novel Routing by Bi-Agreement mechanism.
- **CL-Babelfy [6]:** This is a content-based recommender system aiming to generate cross-lingual recommendations using knowledge-based strategies to build the bond between different languages. In Reference [6], the authors extracted concepts from Wikipedia

Table 4. Comparison of the Methods

Methods	Tasks		Characteristics							
	Sentiment Analysis	Rating Prediction	Latent Factor Model	Neural Model	Topic Model	Cross-lingual	Translation-based	Aspect-based	Attention-based	Sentiment-aware
CLJAS	✓	–	–	–	✓	✓	–	✓	–	✓
Tran-AT-LSTM	✓	–	–	✓	–	–	✓	✓	✓	✓
Tran-CAN	✓	–	–	✓	–	–	✓	✓	✓	✓
MF	–	✓	✓	–	–	✓	–	–	–	–
NAIS	–	✓	✓	✓	–	✓	–	–	✓	–
Tran-D-Attn	–	✓	–	✓	–	–	✓	–	✓	–
Tran-ALFM	–	✓	✓	–	✓	–	✓	✓	–	–
Tran-ANR	–	✓	–	✓	–	–	✓	✓	✓	–
Tran-CARP	–	✓	–	✓	–	–	✓	✓	✓	✓
CL-Babelfy	–	✓	–	–	–	✓	–	–	–	–
MrRec	✓	✓	–	✓	–	✓	–	✓	✓	✓

Fields without the related method are marked with a hyphen.

or BabelNet. Here, we adopt BabelNet,[9] since it can lead to better recommendation performance on the two multilingual datasets.

We evaluate our multilingual aspect-based sentiment analysis module with the following comparative approaches:

- **CLJAS [41]:** It jointly performs aspect-specific sentiment analysis of two languages simultaneously by incorporating sentiment parameter into a cross-lingual topic model.
- **Tran-AT-LSTM [59]:** The attention mechanism is adopted in LSTM to generate the sentence representation. The aspect embedding is used to compute the attention weights.
- **Tran-CAN: [60]** It introduces sparse and orthogonal regularizations when performing aspect-specific sentiment analysis to learn sentiment distributions on the sentence level. Orthogonal regularization is designed especially for reviews with non-overlapping aspect-specific sentiments, which are unknown in two review datasets. Thus, we only adopt sparse regularization for testing.

The comparison of our MrRec and the baseline methods is listed in Table 4. Note that for monolingual baselines such as D-Attn, ALFM, ANR, CARP, AT-LSTM, and CAN, we translate all the reviews from other languages to English using Google Translate,[10] and adopt the prefix **Tran-** as an indicator.

4.1.4 Implementation Details. We implement MrRec with TensorFlow.[11] We initialize our multilingual word embeddings by using the aligned word vectors pre-trained with fastText,[12] while the word embeddings used in the translation baselines for the English language were initialized

[9] https://babelnet.org/.
[10] https://translate.google.com/.
[11] https://www.tensorflow.org/.
[12] We also adopt MUSE from https://github.com/facebookresearch/MUSE, an alternative multilingual pre-trained word embeddings for our experiments by selecting reviews in languages existing in MUSE, but achieved comparative recommendation performance. However, fastText provides pre-trained word embeddings in more languages than MUSE does, and therefore we choose to report results with fastText pre-trained word embeddings as inputs.

by Glove[13] [61]. We also initialize the aspect embedding matrix A with the centroids of clusters resulting from running k-means on word embeddings. The orthogonality penalty weight λ is set to 0.9. We experimented with different numbers of aspects ranging from $[2, 8]$ for all datasets and no major difference was shown with the results. For a fair comparison with other aspect-based baselines, we set the number of aspects K to 5 for Tran-ALFM, Tran-ANR and our model. We also set the number of aspects M for Tran-CARP as 5. The dimension of hidden state output from bi-directional GRU is set to 150. The number of hidden units for each direction is 75. The number of convolution filters N_f is set to 50 for MRM. The number of latent factors d_c, d_r, d_t, and d_f are set to 300, 300, 300, and 100, respectively. MrRec is trained with Adam optimizer, because Adam uses adaptive learning rates for parameters with different update frequencies and converges faster than vanilla stochastic gradient descent. We test the initial learning rate of [0.0001, 0.001, 0.01]. For the coefficient of L2 regularization, [0.0, 0.0001, 0.01, 0.1] is tested. To prevent overfitting, the dropout rate ρ is set to 0.7. The batch size is set as 200 for the Book Reviews datasets, while others are set to 100. The model is trained for a maximum of 300 epochs with early stopping, which means that the training will stop if the performance on validation set does not improve in 10 epochs. The final performances are reported after five runs with the average test results.

A detailed list of parameter settings for both recommendation baselines and sentiment analysis baselines are included in Table 5. For recommendation baseline methods, we adopt the optimization strategies reported in their papers to tune the hyper-parameters. We tune the number of latent factors h for MF, which is selected from $\{5, 10, 15, 20, 25\}$. We tune the dimension of feature vectors m for CL-Babelfy, which is selected from $\{5, 10, 15, 20\}$. We tune the embedding size k, smoothing exponent β, attention factor a for NAIS, where k, a are selected from $\{8, 16, 32, 64\}$ and β from $\{0.2, 0.4, 0.6, 0.8\}$. We tune the dimension of embedding d, window size w, number of filters n_{l-att} and n_{g-att}, filter length w_f, number of hidden factors K_1 and K_2 for Tran-D-Atten, where d is selected from $\{50, 100, 150, 200, 300\}$, w and w_f are selected from $\{2, 3, 4, 5, 6\}$, n_{l-att} is selected from $\{100, 150, 200, 250\}$, n_{g-att} is selected from $\{60, 80, 100, 120\}$, K_1 is selected from $\{300, 400, 500, 600\}$, and K_2 is selected from $\{30, 40, 50, 60\}$. We tune the number of latent factors h for Tran-ALFM, which we select from $\{5, 10, 15, 20, 25\}$. For Tran-ANR, we tune d, c, which stand for the dimension of word embeddings and the width of local context window, respectively. d is selected from $\{100, 150, 200, 300, 400\}$ and c is selected from $\{2, 3, 4, 5\}$. We also tune the number of latent factors h_1, h_2, which are selected from $\{10, 15, 20, 30, 40, 50, 60\}$. We tune the dimension of word embeddings d, the number of latent factors h, the number of filters n and window size c for Tran-CARP, which are selected from $\{100, 150, 200, 300, 400\}$, $\{25, 50, 100, 150\}$, $\{30, 40, 50, 60, 70\}$, and $\{2, 3, 4, 5\}$, respectively.

As for multilingual aspect-based sentiment analysis baselines, the aspect embedding matrix and parameters are initialized by sampling from a uniform distribution $U(-\sigma, \sigma)$, $\sigma = 0.01$ in the AT-LSTM and CAN models. The number of aspects K for Tran-CAN and Tran-AT-LSTM, T for CLJAS are set to 12 and 7 for Trip-MAML and Restaurant Reviews datasets, respectively. We tune the regularization parameter λ and the dimension of word embedding d for Tran-CAN and Tran-AT-LSTM, which are selected from $\{0.001, 0.01, 0.1, 0.15, 0.2\}$ and $\{100, 150, 200, 300, 400\}$, respectively. For CLJAS, we set $\alpha_z = 50/T$, $\beta_w = 0.1$, $\mu = 0.001$, and $\gamma_{tgt}^{tgt} = 0.001$ on both datasets and select γ_{src}^{tgt} from $\{0.001, 0.01, 0.1, 0.15, 0.2\}$.

4.2 Evaluation on Aspect Extraction and Sentiment Prediction (RQ1)

In this section, we conduct experiments to verify if the models are able to extract aspects and predict associated sentiments in different languages simultaneously. Given a review sentence, our

[13] https://nlp.stanford.edu/projects/glove/.

Table 5. Tuned Parameter Values of Different Methods on Different Datasets

Datasets	MF [57]	CL-Babelfy [6]	NAIS [58]			Tran-D-Attn [14]						
	h	m	k	β	a	d	w	n_{l-att}	w_f	n_{g-att}	K_1	K_2
Books	25	15	32	0.6	32	150	5	200	3	100	500	50
Digital Ebook Purchase	25	10	16	0.6	16	100	5	200	3	80	500	40
Digital Music Purchase	25	10	32	0.8	32	150	5	200	3	100	500	50
Digital Video Download	25	10	32	0.6	32	150	4	200	3	100	400	50
Mobile Apps	20	10	16	0.4	16	100	4	150	4	100	400	50
Music	25	15	32	0.6	32	150	5	200	3	100	500	50
Toys	20	10	16	0.8	16	150	5	150	3	80	400	40
Video DVD	25	15	16	0.4	16	100	6	200	3	100	500	50
Goodreads	25	15	16	0.6	16	150	6	200	3	100	500	50
	Tran-ALFM [1]		Tran-ANR [4]					Tran-CARP [5]				
	K	h	d	K	c	h_1	h_2	d	h	n	c	M
Books	5	25	300	5	3	10	50	300	50	50	3	5
Digital Ebook Purchase	5	20	300	5	3	10	50	150	25	50	3	5
Digital Music Purchase	5	25	300	5	3	15	60	300	50	60	3	5
Digital Video Download	5	25	300	5	3	15	60	300	25	50	3	5
Mobile Apps	5	15	300	5	4	15	60	150	25	50	4	5
Music	5	25	300	5	3	10	50	300	25	50	3	5
Toys	5	25	200	5	3	10	50	200	25	60	3	5
Video DVD	5	20	300	5	3	10	50	200	25	50	3	5
Goodreads	5	20	300	5	3	10	50	200	50	50	3	5
	Tran-CAN [60]		Tran-AT-LSTM [59]					CLJAS [41]				
	λ	d	K	λ	d	K	T	α_z	β_w	μ	γ_{tgt}^{tgt}	γ_{src}^{tgt}
Trip-MAML	0.1	300	12	0.001	300	12	12	4.16	0.1	0.001	0.001	0.01
Restaurant Reviews	0.15	300	7	0.01	300	7	7	7.14	0.1	0.001	0.001	0.1

MABSA module assigns one or more inferred aspect labels that correspond to the learned weights higher than a threshold τ[14] according to Equation (4). A summary of the results of the baselines and our MABSA module over the two datasets w.r.t. aspect extraction and sentiment prediction are reported in Table 6.[15] Several observations can be made as follows:

First, the values of all evaluation metrics on Trip-MAML dataset are generally higher than that on Restaurant Reviews dataset. It is probably because we have more training samples in Trip-MAML dataset, which gives the model more opportunities to fit the data well during training. Second, we can observe that Tran-AT-LSTM consistently performs worst of all methods, since the attention mechanism may scatter the distribution of weights across the whole sentence and thus may introduce noisy words or opinion words from other aspects. Besides, machine translation, to some extent, is unable to take into account the divergence in the expression of sentiments across different languages. Moreover, the performance gain of CLJAS baseline compared with Tran-AT-LSTM mainly benefits from the knowledge transferred from the source language, and therefore can capture more statistics characteristics.

[14]We set $\tau = 0.2$ for it achieves the best performance after experiments.
[15]Note that the values of P/R/F1 reported are the average over five runs, and thus the F1-score cannot be computed directly from corresponding P/R values.

Table 6. Comparison Results of the MABSA Part with the Baseline Methods in Terms of Precision, Recall, and F1 Score

Model	**Trip-MAML**						**Restaurant Reviews**					
	Aspect Extraction			**Sentiment Prediction**			**Aspect Extraction**			**Sentiment Prediction**		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Tran-AT-LSTM	0.725	0.702	0.712	0.691	0.716	0.703	0.636	0.582	0.609	0.614	0.592	0.602
CLJAS	0.773	0.685	0.728	0.722	0.781	0.751	0.682	0.577	0.624	0.631	0.602	0.617
Tran-CAN	0.854	**0.882**	**0.867**	0.793	0.779	0.786	0.791	**0.814**	**0.803**	0.737	**0.723**	0.731
MABSA	**0.876***	0.843	0.859	**0.837***	**0.786***	**0.812***	**0.802***	0.761	0.783	**0.758***	0.714	**0.735***

The best results are highlighted in boldface. "*" indicates the improvements are statistically significant for p-value < 0.01 with paired t-test.

Our model outperforms Tran-AT-LSTM and CLJAS on both datasets for aspect extraction and sentiment prediction tasks. This may be because: (1) the usage of bi-directional GRU helps to incorporate contextual information into word embeddings, while CLJAS captures the words co-occurrence based on the assumption of independence of each word in sentences. (2) Different from Tran-AT-LSTM, we fuse aspect information into word representations when learning attention weights, which to some extent, concentrates the importance on more meaningful words. (3) The utilization of the pre-trained multilingual word embeddings that project languages into a shared space also contributes to the performance improvement. Furthermore, the experimental results show that compared with supervised model Tran-CAN, MABSA can obtain comparable performance on aspect extraction tasks, which have convincingly validated the effectiveness of MABSA in extracting aspects. The reason for the relatively low precision and high recall of Tran-CAN compared with MABSA is probably that the attention weights learned by Trans-CAN distribute evenly on words for several cases without explicit aspect terms appearing in the sentences, which leads to more predicted aspect categories than ground truth. It is interesting to note that our module outperforms Tran-CAN on sentiment analysis tasks, which is probably attributed to the hierarchical attention mechanism (including aspect-level and sentence-level attention nets) and aspect fusion that can learn the most indicative sentiment words associated with each aspect in both overlapping and non-overlapping muli-aspect sentences, while Tran-CAN only adopts sparse regularization term that is inadequate to extract sentiment words of non-overlapping aspects.

4.3 Recommendation Performance Evaluation (RQ2)

4.3.1 Recommendation Effectiveness. Table 7 shows the performance comparison of our MrRec model with state-of-the-art methods on the same test dataset. The table is separated in three blocks showing the results on metric MSE, ILS, and EPC, respectively. From Table 7, we can make the following observations:

For the first block, it is not surprising that MF, which depends solely on user-item interactions for rating prediction, consistently yields worst MSE among all approaches on all datasets, which we believe validates the importance of contextual information in reviews. Though NAIS only adopts item IDs rather than textual reviews as inputs, it outperforms review-based method, CL-Babelfy, which we believe is probably credited to the powerful representation learning capacity of neural models. Among all translation-based approaches, Tran-D-Attn model perform worse than others, which is because the model does not consider aspect-level features when modelling users and items and thus cannot capture the fine-grained characteristics of users/items. Whereas Tran-CARP achieves the lowest MSE among all translation-based baselines over all datasets, which shows that the aspects and aspect-specific sentiments derived from textual reviews play crucial roles in

Table 7. Comparison Results of the MRM Part with the Baseline Methods in Terms of Mean-square Error (MSE), Intra-list Similarity (ILS), and Novelty (EPC)

Measures	*Methods*	*Books*	*Digital Ebook Purchase*	*Digital Music Purchase*	*Digital Video Download*	*Mobile Apps*	*Music*	*Toys*	*Video DVD*	*Goodreads*
MSE	*MF*	2.487	2.279	2.583	2.426	2.067	2.361	2.503	2.184	2.091
	CL-Babelfy	2.361	2.183	2.504	2.337	1.972	2.254	2.396	2.068	1.995
	NAIS	2.082	1.969	2.211	2.049	1.746	2.015	2.152	1.833	1.771
	Tran-D-Attn	1.935	1.823	2.086	1.902	1.627	1.885	2.011	1.726	1.654
	Tran-ALFM	1.746	1.628	1.893	1.731	1.433	1.692	1.825	1.542	1.467
	Tran-ANR	1.633	1.527	1.781	1.618	1.326	1.587	1.714	1.425	1.348
	Tran-CARP	1.481	1.366	**1.635**	1.464	1.169	1.437	1.556	**1.232**	**1.183**
	MrRec	**1.307***	**1.253***	1.682	**1.288***	**1.036***	**1.269***	**1.392***	1.243	1.189
	Improvement (%)	11.75∼47.45	8.27∼45.02	−2.87∼34.88	12.02∼46.91	11.38∼49.88	11.69∼46.25	10.54∼44.39	−0.89∼43.09	−0.51∼43.14
ILS	*MF*	8.756	7.427	10.683	8.894	7.048	8.347	9.530	7.795	7.052
	CL-Babelfy	9.877	8.352	11.565	10.032	7.917	9.233	11.158	8.937	7.426
	NAIS	10.453	8.781	11.672	10.684	7.885	9.931	11.797	9.732	8.278
	Tran-D-Attn	11.283	9.693	12.462	11.531	8.392	12.846	12.135	9.524	8.732
	Tran-ALFM	12.732	10.662	14.959	13.263	9.531	12.182	13.677	12.151	9.041
	Tran-ANR	12.345	10.236	13.501	14.025	8.825	11.473	12.972	10.604	9.727
	Tran-CARP	13.527	11.379	14.153	12.672	10.257	10.746	14.579	11.372	10.562
	MrRec	**8.283***	**7.264***	**9.565***	**8.667***	**6.371***	**7.716***	**9.372***	**7.386***	**6.558***
	Improvement (%)	5.40∼38.77	2.19∼36.16	10.47∼36.06	2.62∼38.20	9.61∼37.89	7.56∼39.93	1.66∼35.72	5.25∼39.21	7.01∼37.91
EPC	*MF*	0.653	0.598	0.621	0.633	0.586	0.639	0.591	0.672	0.694
	CL-Babelfy	0.673	0.616	0.641	0.670	0.592	0.672	0.604	0.684	0.709
	NAIS	0.693	0.621	0.675	0.683	0.605	0.687	0.612	0.706	0.723
	Tran-D-Attn	0.712	0.636	0.702	0.708	0.663	0.714	0.635	0.753	0.766
	Tran-ALFM	0.748	0.719	0.713	0.750	0.648	0.742	0.661	0.826	0.825
	Tran-ANR	0.775	0.708	0.748	0.739	0.611	0.781	0.674	0.772	0.841
	Tran-CARP	0. 794	0.687	0.741	0.763	0.681	0.766	0.692	0.831	0.846
	MrRec	**0.842***	**0.775***	**0.798***	**0.817***	**0.732***	**0.826***	**0.767***	**0.859***	**0.873***
	Improvement (%)	6.05∼28.94	7.79∼29.60	6.68∼28.50	7.08∼29.07	7.49∼24.91	5.76∼29.26	10.84∼29.78	3.37∼27.83	3.19∼25.79

The best results are highlighted in boldface. "*" indicates the improvements are statistically significant for p-value < 0.01 with paired t-test.

improving recommendation performance. Though both Tran-ANR and Tran-ALFM attempt to utilize aspects in their architectures, Tran-ANR outperforms Tran-ALFM, whose major drawback is that the proposed model leverages topic model to learn the statistical features of words in reviews, which neglects the contextual information around the word. Our model shows comparable performance compared with Tran-CARP and even shows superior performance on most datasets that have more multilingual reviews. We believe this benefits from the language attention mechanism that can learn the different contributions of reviews in multiple languages and multilingual word embeddings that jointly mine semantic information with textual reviews written in various languages.

Diversity and novelty are measured by ISL and EPC with the results displayed in the rest two blocks, from which we can observe that our MrRec exhibits the dominating performance among all methods across nine datasets. We argue that this is attributed to the aspect utility estimation mechanism, which takes into consideration both users with similar preferences to the target user

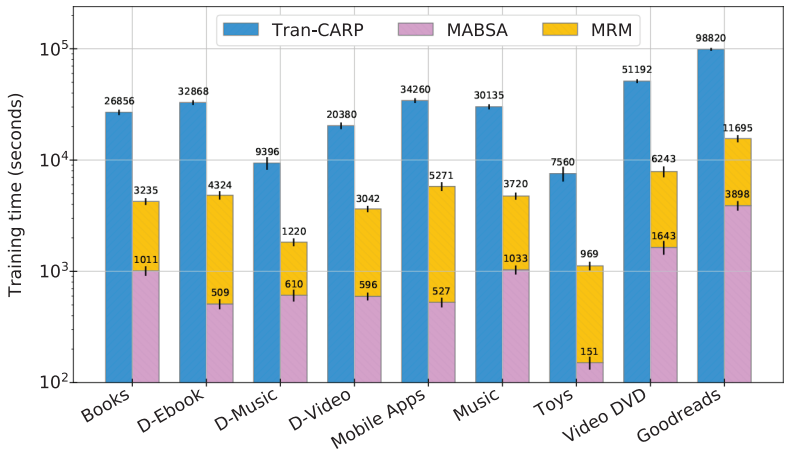
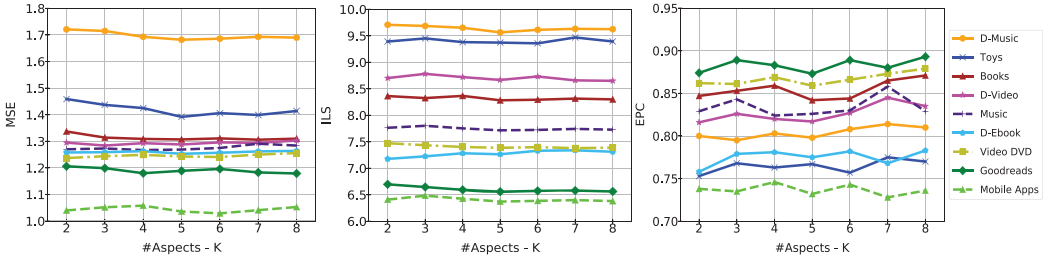


Fig. 6. Runtime comparison (seconds) for training model on all datasets. D-Ebook, D-Music, and D-Video are short for Digital Ebook Purchase, Digital Music Purchase, and Digital Video Download.

and items in user's historical records similar to the candidate item. There is no dominating winner among neural network baseline methods, but they outperform CL-Babelfy and MF on EPC, which is because they focus more on historical user preferences and thus tend to recommend items similar to the items user clicked before rather than the popular items. Because of the same reason, they neglect the diversification on candidate items of the user potential interests, and therefore perform worse than CL-Babelfy and MF on ILS.

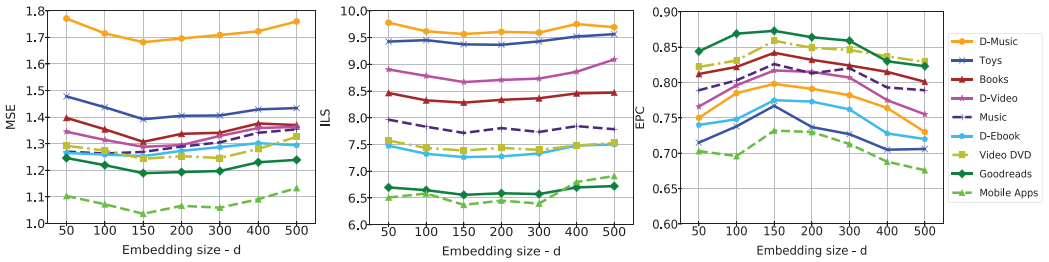
4.3.2 Recommendation Efficiency. Figure 6 illustrates the log scale training time comparisons between MrRec and Tran-CARP, the best performance with MSE among all baselines. Experiments are conducted on GPU machines of Nvidia GeForce GTX TITAN X. Compared with MrRec, whose time complexity is $O(|\mathcal{R}_{train}|K((|\mathcal{F}_u|^2 + |\mathcal{F}_i|^2)(N_f d + N_s^2 d^2) + K^2 N_f^2 + N_w^2 N_s^2 d^2))$ analyzed in Section 3.5, the complexity of Tran-CARP is $O(|R_{train}|K(KN_f N_s^2 N_w^2(|\mathcal{F}_i|^2 + |\mathcal{F}_u|^2)(d + N_f) + Kd^2 + K^5))$. We can derive that the computational cost of MrRec is lower than Tran-CARP, since the sentiment analysis and recommendation procedures of Tran-CARP are coupled together to work in an end-to-end fashion, which means they need to be trained on both user and item sides for each training sample. In contrast, the sentiment analysis module of MrRec only performs once in a separate offline phase. Besides, compared with other review-based methods, i.e., Tran-CARP, which usually feed the reviews with embedding vectors of all words into model, our inputs of MRM are achieved through aspect-based representations of all sentences, and thus the size of input for a specific user/item is changed from $O(|\mathcal{F}_{u/i}|N_w d)$ to $O(|\mathcal{F}_{u/i}|N_s K d)$. From Table 2, we can see $N_s \ll N_w$. Actually in practice, $N_s \times K$ is usually smaller than N_w. Therefore, our MRM module can further accelerate the training efficiency. In Figure 6, similar observations can be found for the two models. As MrRec is composed of two steps, we report the training time for each of them with different colours, and the whole training time for the model is the summation of them. It can be seen that MrRec trains 5 to 6.5 times faster than Tran-CARP over all datasets,[16] which benefits from the reduction of input size and decomposition of training tasks.

[16]It is worth mentioning that the actual training time of MrRec is shorter, because step 1 (MABSA) can be performed offline in advance.



(a) Mean Square Error (b) Intra-list Similarity (c) Novelty

Fig. 7. Effect of the number of aspects.



(a) Mean Square Error (b) Intra-list Similarity (c) Novelty

Fig. 8. Effect of the Bi-GRU output dimension.

4.4 Effects of the Hyper-Parameters (RQ3)

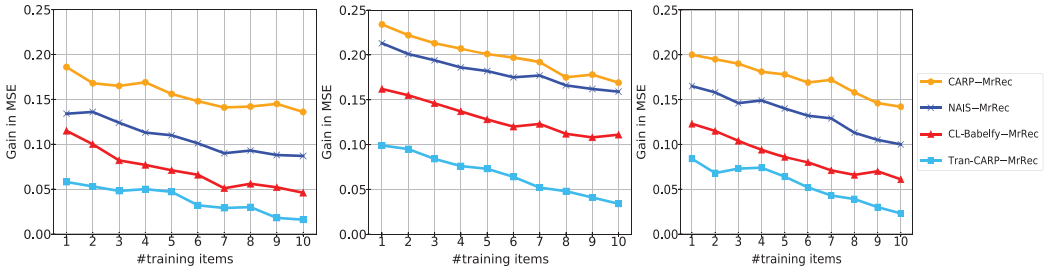
In this section, we analyze the influence of embedding size d and the number of aspects K on the final performance of MrRec. We optimize one parameter with another one fixed to see how performance will change accordingly.

The empirical results displayed in Figure 7 indicate the effect of varying the number of aspects K from 2 to 8 for our model w.r.t. MSE, ILS, and EPC across nine datasets. We can observe that though the optimal value of K varies across different datasets, the overall trends are relatively stable. The comparatively good performance can be achieved with five aspects. We hypothesize that adjusting the number of aspects can only influence the granularity of modelling the textual reviews. As such, varying K within a reasonable range has little impact on the recommendation performance.

Figure 8 illustrates the effect of varying the embedding size d from 50 to 500 across multiple datasets on three metrics. As can be observed, the performance keeps improving with d ranging from 50 to 150 on most datasets. The highest performance appears with d set around 150 and remains relatively stable before d equals to 300. However, the results show the turbulent trends when d is higher than 300, which indicates that further use of larger embedding size does not show significant improvement. Thus, we set $d = 150$ in our experiments.

4.5 Cold Start Evaluation (RQ4)

For monolingual scenario, cold start refers to users/items with limited ratings, which makes it difficult to provide satisfactory recommendations for monolingual recommendation models. MF method can easily lead to cold start issue, since there are only few user-item interactions available. In contrast, review-based methods can alleviate the problem, since reviews contain rich contextual information on users' preference and item characteristics. However, we argue that such problem



(a) Digital Music Purchase (b) Mobile Apps (c) Goodreads

Fig. 9. Performance on the cold start problem.

can further be alleviated by introducing resources from other languages, i.e., textual reviews written in different languages. To verify this assumption and demonstrate the capability of our model in dealing with multilingual user-item interactions, in this section, we conduct experiments on multilingual datasets with our MrRec model and different baselines, i.e., Tran-CARP, NAIS, and CL-Babelfy. We also compare our model with the original version of Tran-CARP, namely, CARP, to test that to what extent translation can help to improve multilingual recommendations.

The experiments are conducted on the selected three of nine datasets, Digital Music Purchase, Mobile Apps, and Goodreads with the lowest, highest, and middle ratio of multilingual interactions, respectively. As preprocessing, we first filter out monolingual user-item interactions, and then split the datasets into training, validation, and testing set based on the number of ratings in each dataset. We also remove users from the testing set who have no rating in the training set. We evaluate the performance of users who have the number of ratings from 1 to 10 in the training set. Figure 9 shows the **Gain in MSE** grouped by the number of user ratings. Gain in MSE can be calculated by the average MSE of baselines minus that of our model, i.e., CARP-MrRec. As can be seen, similar trends can be found across all datasets. Our MrRec model consistently outperforms other baselines on three datasets, since the differences are all positive values. In particular, Tran-CARP substantially improves the rating prediction accuracy compared with CARP, which we believe verifies the importance and benefit of leveraging multilingual reviews for recommendations. Besides, our MrRec model beats the other baselines that integrate multilingual resources. This is attributed to the fact that our model is more effective in exploiting and modelling textual reviews in different languages.

4.6 Ablation Study (RQ5)

In this section, we perform an ablation study to analyze how different components in our proposed model contribute to the overall performance. The experiments are conducted among variants of MrRec and the complete model (denoted as "baseline") with hyper-parameter settings as stated in Section 4.1.4. We incorporate the following variants:

- **Random Word Embeddings (RWE):** Instead of using pre-trained multilingual word embeddings as inputs to our model, we train our convolutional model on word embeddings initialized randomly from a uniform distribution. The word embeddings are part of the trainable parameters of the network in this model.
- **Without Bi-directional GRU Layer (Without Bi-GRU):** To show the effect of adopting Bi-directional GRU Layer to the word representations, we choose to remove the Bi-directional GRU Layer to test its effectiveness in the MABSA module.

Table 8. Comparison of the Model Variants for the Mobile Apps and Goodreads Datasets

Setup	**Mobile Apps**			**Goodreads**		
	MSE	*ILS*	*EPC*	*MSE*	*ILS*	*EPC*
Baseline	1.036	6.371	0.732	1.189	6.558	0.873
RWE	1.432	6.946	0.701	1.508	7.161	0.821
Without Bi-GRU	1.115	6.696	0.712	1.257	6.959	0.837
ULI	1.125	6.789	0.705	1.263	7.020	0.818
UAI	1.143	6.748	0.708	1.278	6.985	0.831
Without AUE	1.229	7.638	0.633	1.307	7.860	0.763
Without ALI	**1.519**	**8.263**	**0.627**	**1.605**	**8.496**	**0.752**

The worst and second-worst results are highlighted in boldface and underlined, respectively.

- **Without Aspect-level Interactions (Without ALI):** We forgo co-attention network and aspect utility estimation component in our model. Instead, we apply a fully-connected layer upon the concatenation of $\boldsymbol{u}_{a_k}/\boldsymbol{i}_{a_k}$ on all aspects to learn the user/item representation. Similar to D-Attn, the user and item representations are then adopted to derive the overall rating.
- **Uniform Language Importance (ULI):** We view each language as equal importance. Specifically, $\eta^l_{a_k}$ is set to $1/L$ in Equation (18).
- **Without Aspect Utility Estimation (Without AUE):** We remove the aspect utility estimation component and use aspect-based user/item representation to predict the overall ratings.
- **Uniform Aspect Importance (UAI):** In Equation (20), each δ_{u,a_k} is replaced with $1/K$ to verify the importance of co-attention network.

The results are shown in Table 8 for the ***Mobile Apps*** and ***Goodreads datasets***. As shown in the table, we can observe that the lack of aspect-level interaction component can lead to large performance degradation on both datasets over three metrics, which is probably because only aspect-level user/item modelling cannot precisely reflect user's satisfaction or requirement on an item, and user's preference toward an item can be comprehensive consequences of different aspects from user-item pair rather than one side. The performance deteriorates second on ILS and EPC when our model is without aspect utility estimation component, which is attributed to the dual interactive attention mechanism of AUE that models the fine-grained user-item interactions with considering both the similarities of candidate item w.r.t. the items the target user previously consumed from item side, and the similarities of target user's preferences w.r.t. other users' attitudes toward the candidate item from user side. Besides, we find that the pre-trained multilingual word embedding provides a crucial starting point for multiple language integration and, consequently affects the overall rating prediction. Finally, excluding either Bi-directional GRU, language attention network, or co-attention network can cause the degradation of recommendation performance to different degrees, which highlights the need of capturing the contextual information of words, adaptively integrating multiple language information, as well as dynamically estimating the user and item aspect importance for each user-item pair in improving the rating prediction, system's diversity and novelty.

4.7 Interpretability Visualization (RQ6)

In this article, a user's preference on an item can be decomposed into the user's preference on different aspects with considering the importance of those aspects from both user and item sides,

Table 9. Top Ten Words of Each Aspect in English (EN) and French (FR) for a User (50989966) from Video DVD Dataset

Film		Style		Time		Character		Value	
EN	*FR*	*EN*	*FR*	*EN*	*FR*	*EN*	*FR*	*EN*	*FR*
$(\eta^{en}_{a_1}: 0.352)$	$(\eta^{fr}_{a_1}: 0.648)$	$(\eta^{en}_{a_2}: 0.365)$	$(\eta^{fr}_{a_2}: 0.635)$	$(\eta^{en}_{a_3}: 0.419)$	$(\eta^{fr}_{a_3}: 0.581)$	$(\eta^{en}_{a_4}: 0.337)$	$(\eta^{fr}_{a_4}: 0.663)$	$(\eta^{en}_{a_5}: 0.473)$	$(\eta^{fr}_{a_5}: 0.527)$
story	*téléfilm* $(TVmovie)^\star$	*fiction*	*genre (kind)*	*august*	*année (year)*	*artists*	*associée (partner)*	*sale*	*affaires (business)*
theatre	*épisodes (episodes)*	*documentary*	*comédie (comedy)*	*october*	*mars (March)*	*referee*	*épouse (wife)*	*profit*	*score (score)*
movie	*éditions (editions)*	*comedy*	*fiction (fiction)*	*months*	*septembre (September)*	*individuals*	*police (police)*	*free*	*dollars (dollars)*
episode	*personnages (characters)*	*historical*	*musical (musical)*	*medieval*	*vie (life)*	*children*	*artistes (artists)*	*money*	*champion (champion)*
actors	*scénariste (scriptwriter)*	*album*	*historique (historical)*	*hours*	*présentent (present)*	*man*	*juifs (Jews)*	*million*	*libre (free)*
actress	*actrice (actress)*	*musical*	*documentaire (documentary)*	*life*	*actuel (current)*	*director*	*chanteuse (singer)*	*industry*	*millions (million)*
character	*studio (studio)*	*philosophy*	*exposition (exhibition)*	*diff*	*évoluant (evolving)*	*chief*	*chiffre (figure)*	*economic*	*moins (minus)*
description	*fin (end)*	*military*	*action (action)*	*throughout*	*quand (when)*	*winner*	*infanterie (infantry)*	*material*	*meilleurs (best)*
families	*séries (series)*	*social*	*images (images)*	*further*	*finale (final)*	*brothers*	*parisien (Parisian)*	*commercial*	*haute (high)*
sports	*écrivain (writer)*	*criminal*	*sociale (social)*	*november*	*parfois (sometimes)*	*citizens*	*filles (girls)*	*trade*	*mesure (measure)*

Each column is corresponding to an aspect attached with an "interpretation" label. $\eta^{l}_{a_k}$ denotes the contribution of language l on aspect a_k for the target user.

* The English translations are shown in brackets.

as well as the sentiment utilities exhibiting from the aspects discovered based on multilingual textual reviews. The learned aspects for the user can be expressed with their representative words, which are found by looking at the nearest words from his/her reviews in the embedding space using cosine as the similarity metric. Specifically, the cosine similarity is calculated between the aspect representation $\boldsymbol{a}_k$ from Equation (5), and the word representation $\boldsymbol{h}_i$ from Equation (1): $sim(k, i) = cos(\boldsymbol{a}_k, \boldsymbol{h}_i)$. The higher value of $sim(k, i)$ is desirable for the word w_i belonging to the kth aspect. The top 10 aspect words in each language of user u from *Video DVD* dataset are shown in Table 9. The contributions of different languages, i.e., English and French user u adopted in total, are listed under the name of each aspect. For instance, $\eta^{en}_{a_1}: 0.352$ represents the contribution of English for aspect a_1 is less than that of French. As shown in Table 9, the five aspects can be semantically interpreted to *Film*, *Style*, *Time*, *Character*, and *Value*. The top aspect words of candidate items can also be achieved in the same way, but here we only illustrate on the user side. Then in Table 10, we demonstrate how to interpret the high and low ratings the user u giving to items on the same dataset. From the table, we can see the aspect importance $\boldsymbol{\delta}_u$ for user and $\boldsymbol{\delta}_i$ for item from Equation (20), as well as aspect utility $y^{a_k}_{u,i}$ from Equation (25) w.r.t. "item 1" and "item 2." As can be observed, the user pays more attention to *Character* and *Film* aspects on both items. Similarly, "item 1" and "item 2" put more importance on *Character* and *Film*. However, the user is more satisfied with *Character* and *Film* on "item 1" than that on "item 2." As a result, according to Equation (26), the overall rating of "item 1" should be apparently higher than that of "item 2," which is 4 to 2, respectively. From the illustration, we can see that our model could capture to what extent the user likes or dislikes an item on an aspect and interpret the recommendation results at a fine level of granularity.

Table 10. Interpretation for Why the "User 50989966" Rated "Item1" and "Item 2" with 4 and 2, Respectively, from Video DVD Dataset

Aspects	*Film*	*Style*	*Time*	*Character*	*Value*
Importance for User (1)	0.214	0.097	0.014	0.653	0.022
Importance for Item (1)	0.103	0.015	0.007	0.861	0.014
Aspect Utility (1)	0.871	0.526	0.145	0.922	−0.263
Importance for User (2)	0.203	0.018	0.003	0.712	0.064
Importance for Item (2)	0.129	0.073	0.006	0.784	0.008
Aspect Utility (2)	−0.576	−0.691	0.193	−0.879	−0.154

5 CONCLUSION

In this article, we have proposed for the first time a multilingual review-aware deep recommendation model (MrRec) for overall rating prediction and item recommendation. The model requires neither external translation tools nor knowledge bases to analyze multilingual reviews. Particularly, instead of labelled datasets, MrRec extracts aspects and analyzes aspect-specific sentiments requiring merely overall ratings, which are leveraged as user sentiments to remove the possible ambiguity contained in the textual reviews. Besides, our model is able to estimate aspect importance for each user-item pair by utilizing co-attention network on the learned aspect-based user/item representations with considering the different contributions of multiple languages. Furthermore, user satisfaction is embodied by the aspect utility derived from a dual interactive attention mechanism with considering both users with similar preferences to the target user and items with similar properties to the candidate item on aspect level. Finally, the overall rating is predicted by adopting a prediction layer on the combination of learned aspect utility and aspect importance. We have compared the MrRec with state-of-the-art baselines on nine real-world datasets, and experimental results demonstrate the effectiveness and efficiency of our model on recommendation accuracy, as well as recommendation diversity, especially for cold start users/items in the monolingual scenario but with extra reviews written in other languages.

Since this article is our initial step to explore the way to improve diversity and novelty in multilingual recommendation tasks, some limitations still exist in the process of doing experiments. First, since we do not consider geolocation information when providing recommendations to users, the recommendations generated by our model incorporate items with reviews in all languages from world-wide websites, while cannot be effectively differentiated according to the real-time location of the target user. Thus, our next step will concentrate on fusing the geolocation factor into the model training procedure. Besides, a few error cases in our experiments show that the sentiment attention weights distribute evenly on nearly all words, because the sentence does not contain any explicit sentiment words or expresses special sentiments such as sarcasm. Therefore, our future work will focus on detecting special sentiments in sentences and thereby integrating them into recommendation tasks.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proceedings of the World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 639–648.

[2] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect-based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 717–725.

[3] Xinyu Guan, Zhiyong Cheng, Xiangnan He, Yongfeng Zhang, Zhibo Zhu, Qinke Peng, and Tat-Seng Chua. 2019. Attentive aspect modeling for review-aware recommendation. *ACM Trans. Info. Syst.* 37, 3 (2019), 28.

[4] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. 2018. ANR: Aspect-based neural recommender. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 147–156.

[5] Chenliang Li, Cong Quan, Li Peng, Yunwei Qi, Yuming Deng, and Libing Wu. 2019. A capsule network for recommendation and explaining what you like and dislike. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 275–284.

[6] Fedelucio Narducci, Pierpaolo Basile, Cataldo Musto, Pasquale Lops, Annalina Caputo, Marco de Gemmis, Leo Iaquinta, and Giovanni Semeraro. 2016. Concept-based item representations for a cross-lingual content-based recommendation process. *Info. Sci.* 374 (2016), 15–31.

[7] Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Marco De Gemmis, Pierpaolo Basile, and Giovanni Semeraro. 2010. Mars: A multilanguage recommender system. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. ACM, 24–31.

[8] Bernardo Magnini and Carlo Strapparava. 2001. Improving user modelling with content-based techniques. In *Proceedings of the International Conference on User Modeling*. Springer, 74–83.

[9] Sebastian Schmidt, Philipp Scholl, Christoph Rensing, and Ralf Steinmetz. 2011. Cross-lingual recommendations in a resource-based learning scenario. In *Proceedings of the European Conference on Technology Enhanced Learning*. Springer, 356–369.

[10] Libing Wu, Cong Quan, Chenliang Li, and Donghong Ji. 2018. Parl: Let strangers speak out what you like. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 677–686.

[11] Jim Keeler and David E. Rumelhart. 1992. A self-organizing integrated segmentation and recognition neural net. In *Advances in Neural Information Processing Systems*. MIT Press, 496–503.

[12] Nikolaos Pappas and Andrei Popescu-Belis. 2017. Explicit document modeling through weighted multiple-instance learning. *J. Artific. Intell. Res.* 58 (2017), 591–626.

[13] Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Trans. Assoc. Comput. Linguist.* 6 (2018), 17–31.

[14] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the 11th ACM Conference on Recommender Systems*. ACM, 297–305.

[15] Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Peng He, Paul Weng, Han Gao, and Guihai Chen. 2019. Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In *Proceedings of the World Wide Web Conference*. 2091–2102.

[16] Donghua Liu, Jing Li, Bo Du, Jun Chang, and Rong Gao. 2019. Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 344–352.

[17] Yang Bao, Hui Fang, and Jie Zhang. 2014. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

[18] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *Proceedings of the World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1583–1592.

[19] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 233–240.

[20] Guang Ling, Michael R. Lyu, and Irwin King. 2014. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender Systems*. ACM, 105–112.

[21] Dongmin Hyun, Chanyoung Park, Min-Chul Yang, Ilhyeon Song, Jung-Tae Lee, and Hwanjo Yu. 2018. Review sentiment-guided scalable deep recommender system. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 965–968.

[22] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Trans. Info. Syst.* 37, 2 (2019), 1–28.

[23] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. ACM, 425–434.

[24] Rose Catherine and William Cohen. 2017. Transnets: Learning to transform for recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*. ACM, 288–296.

[25] 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 the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 193–202.

[26] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect-based opinion mining and collaborative filtering. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*. ACM, 199–208.

[27] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 388–397.

[28] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 83–92.

[29] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to rank features for recommendation over multiple categories. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 305–314.

[30] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. 1661–1670.

[31] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan Kankanhalli. 2018. A^3NCF: An adaptive aspect attention model for rating prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. International Joint Conferences on Artificial Intelligence Organization, 3748–3754. DOI: http://dx.doi.org/10.24963/ijcai.2018/521

[32] Carmen Martinez-Cruz, Carlos Porcel, Juan Bernabé-Moreno, and Enrique Herrera-Viedma. 2015. A model to represent users trust in recommender systems using ontologies and fuzzy linguistic modeling. *Info. Sci.* 311 (2015), 102–118.

[33] Atsuhiro Takasu. 2010. Cross-lingual keyword recommendation using latent topics. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*. ACM, 52–56.

[34] Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained SMT. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 781–787.

[35] Roman Klinger and Philipp Cimiano. 2015. Instance selection improves cross-lingual model training for fine-grained sentiment analysis. In *Proceedings of the 19th Conference on Computational Natural Language Learning*. 153–163.

[36] Mariana S. C. Almeida, Cláudia Pinto, Helena Figueira, Pedro Mendes, and André F. T. Martins. 2015. Aligning opinions: Cross-lingual opinion mining with dependencies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 408–418.

[37] Jeremy Barnes, Patrik Lambert, and Toni Badia. 2016. Exploring distributional representations and machine translation for aspect-based cross-lingual sentiment classification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*. 1613–1623.

[38] Jordan Boyd-Graber and David M. Blei. 2009. Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 75–82.

[39] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. 2010. Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1128–1137.

[40] Jordan Boyd-Graber and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 45–55.

[41] Zheng Lin, Xiaolong Jin, Xueke Xu, Weiping Wang, Xueqi Cheng, and Yuanzhuo Wang. 2014. A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1089–1098.

[42] Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Xueqi Cheng, Weiping Wang, and Dan Meng. 2015. An unsupervised cross-lingual topic model framework for sentiment classification. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 24, 3 (2015), 432–444.

[43] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. Retrieved from https://arxiv.org/abs/1804.07745.

[44] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Retrieved from https://arxiv.org/abs/1409.0473.

[45] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*. 201–216.

[46] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. Retrieved from https://arxiv.org/abs/1601.01073.

[47] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[48] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems*. MIT Press, 289–297.

[49] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2309–2318.

[50] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, and Xiangyang Luo. 2019. A context-aware user-item representation learning for item recommendation. *ACM Trans. Info. Syst.* 37, 2 (2019), 1–29.

[51] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*. 811–820.

[52] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2015. Learning hierarchical representation model for nextbasket recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 403–412.

[53] Md Hijbul Alam, Woo-Jong Ryu, and SangKeun Lee. 2016. Joint multi-grain topic sentiment: Modeling semantic aspects for online reviews. *Info. Sci.* 339 (2016), 206–223.

[54] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq et al. 2016. Semeval-2016 task 5: Aspect-based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*.

[55] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*. 22–32.

[56] Katja Niemann and Martin Wolpers. 2013. A new collaborative filtering approach for increasing the aggregate diversity of recommender systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 955–963.

[57] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[58] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. 2018. Nais: Neural attentive item similarity model for recommendation. *IEEE Trans. Knowl. Data Eng.* 30, 12 (2018), 2354–2366.

[59] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 606–615.

[60] Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2018. CAN: Constrained attention networks for multi-aspect sentiment analysis. Retrieved from https://arxiv.org/abs/1812.10735.

[61] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

Received February 2020; revised October 2020; accepted October 2020