Journal of Big Data

RESEARCH

Open Access

Evaluation of recent advances in recommender systems on Arabic content

Mehdi Srifi[1*], Ahmed Oussous[2], Ayoub Ait Lahcen[3] and Salma Mouline[1]

*Correspondence:
mehdi_srifi@um5.ac.ma
[1] LRIT, Associated Unit to CNRST (URAC 29), Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco
Full list of author information is available at the end of the article

Abstract

Various recommender systems (RSs) have been developed over recent years, and many of them have concentrated on English content. Thus, the majority of RSs from the literature were compared on English content. However, the research investigations about RSs when using contents in other languages such as Arabic are minimal. The researchers still neglect the field of Arabic RSs. Therefore, we aim through this study to fill this research gap by leveraging the benefit of recent advances in the English RSs field. Our main goal is to investigate recent RSs in an Arabic context. For that, we firstly selected five state-of-the-art RSs devoted originally to English content, and then we empirically evaluated their performance on Arabic content. As a result of this work, we first build four publicly available large-scale Arabic datasets for recommendation purposes. Second, various text preprocessing techniques have been provided for preparing the constructed datasets. Third, our investigation derived well-argued conclusions about the usage of modern RSs in the Arabic context. The experimental results proved that these systems ensure high performance when applied to Arabic content.

Keywords: Arabic, Recommender systems, Collaborative filtering, Rating prediction, User reviews, Text preprocessing

Introduction

In the era of big data, the volume of information on the web has grown at an unprecedented rate [1]. Due to this exponential growth, web users are flooded with countless information, which is known as the information overload issue [2]. In this case, it becomes laborious for users to find the relevant information that they need. Aiming to guide the users in a personalized way to interesting information, recommender systems (RSs) have become useful and powerful tools for helping to manage the information overload problem [3]. The main goal of RSs is to make predictions about user interest according to their previous interaction data [4, 5]. They have proven the efficiency in facilitating the decision-making process and helping businesses increase benefices in various areas like e-commerce, e-tourism, e-learning, multimedia, social networks, etc. [6].

Collaborative Filtering (CF) techniques, especially, Matrix Factorization (MF) method, represents the most widely adopted and successful paradigm to implement RSs [7]. The mainstream of MF methods is to characterize both users and items utilizing latent features deduced from rating matrix patterns, and then predict missing ratings based on the inner product of the users' and items' feature vectors [8]. However, as a traditional CF method, the MF easily suffers from the rating sparsity issue that usually arises in real-world recommendation platforms in which most users rate only a restricted fraction of items [3]. Furthermore, relying solely on numeric scores makes it hard for these techniques to accurately capture users' preferences [9].

The above two drawbacks have been widely evoked and studied. In fact, to enhance ratings and deal with the rating sparsity issue, several types of auxiliary information have been integrated into MF, namely, tags [10], items' descriptions [11], social relationships [12], and review texts [13]. Among them, reviews remain the most used ones because they contain precious supplementary details that are effective to model users' and items' characteristics; and explain the underlying rationales behind ratings [14]. Consequently, in the last decade, a variety of models have been proposed to use reviews and ratings [5, 14, 15]. All these methods have shown a positive impact of user reviews on the performance of recommendation accuracy [5, 15]. However, most of these existing works have focused only on reviews written in English. From the literature, we notice that fewer works are achieved on other languages such as Arabic, which has become the fourth most spoken language worldwide, and one of the most used languages on the Internet [16]. It also should be noted that the Arabic web users have recently become major consumers of Internet services [17], and therefore they share a large amount of textual content [18]. Thus, it became possible to incorporate such content into RSs to predict users' preferences in an Arabic context.

Motivated by these reasons, we choose to deal with RSs in the Arabic context by exploiting textual reviews written in the Arabic language. Thus, in our study, we carried out experiments with five state-of-the-art RSs from the literature. We notice that the performance of these systems has been demonstrated for English content; however, it has never been proved for content in the Arabic language. Unluckily, there is no available dataset containing Arabic content for assessing RSs. Therefore, in this article, we have focused all our efforts to fill this large need, which is still understudied by the research community. The contribution of our paper can be summarized as follows:

- Building new Arabic datasets for RSs. We collected and translated data from the English Amazon datasets. The main goal is to assess RSs on large-scale Arabic data.
- Providing a text preprocessing scheme for Arabic and English languages.
- Applying the recent RSs in the Arabic context.
- Analyzing and evaluating the performance of various state-of-the-art review-based RSs when using Arabic textual data.
- Demonstrating that modern RSs provide promising results when applied to the Arabic content.
- Shedding light on the large need for RSs devoted to Arabic content and the importance of having additional studies in this direction.

The remainder of the paper is organized as follows: "Research motivation" section explains this work's motivation. The "Related works" section summarizes the related works in this field. The "Methodology" section describes the used methodology for investigating recent RSs in the Arabic context. The "Experiments" section presents the experimental framework and interprets the achieved empirical results. Finally, the "Conclusion and future work" section concludes the paper and provides future work.

Research motivation

After reviewing the RSs literature, we found that only very few RSs have exploited the Arabic content. Among the works that we found are [19, 20], which proposed RSs based on Arabic-language reviews to predict users' preferences on items. Although these works were the precursors in the Arabic RSs field, they suffer from different limitations discussed in the next section. The main ones are:

- In these studies, methods and recent advances in the RSs field are not empirically used or compared.
- The reviews' datasets used in the experimentations of these works are of modest size.
- None of these studies has reported if there is or not a necessity to apply a special scheme for preparing the Arabic reviews before their usage in the RSs.

In view of all those shortcomings as well as the lack of works that apply the recent recommendation approaches in the Arabic context, we decided to conduct this work in order to answer the following research questions:

> ***RQ1****: Is it possible to apply recent RSs in the Arabic context?*
>
> ***RQ2****: If so, does the Arabic content need a particular preprocess step to incorporate it in these RSs?*
>
> ***RQ3****: If so, does the application of these RSs to Arabic content provide good results like when applying them to the original content (English)?*

The aforementioned research questions gave a motivation to do this research work. To the best of our knowledge, no study has been carried out in this direction.

Related works

Over the last decade, many researchers have attempted to enhance rating prediction accuracy by extracting useful information from the text of reviews and incorporating them into the recommendation process [21–27]. HFT [21] and TopicMF [22] fuse latent factors from ratings and latent features from reviews (inferred based on latent Dirichlet allocation model in [21] and non-negative-MF in [22]) adopting a transform function to predict missing ratings. RBLT [23] and ITLFM [24] discover latent features from reviews based on a topic modeling technique. By assuming that the latent features and factors share the same space, they linearly combine them using the MF model for predicting unknown ratings. EFM [25] extracts feature opinion pairs from textual reviews. It uses a phrase level-sentiment analysis for building two matrices, namely, user-feature attention and item-feature quality, which are then merged with the rating matrix to estimate the ratings in the MF approach. ALFM [26] extracts topics from reviews to infer the importance of the target aspect and then incorporates aspect importance into a Latent Factor Model (LFM) for rating prediction. A3NCF [27] extracts features from reviews using topic modeling, then integrates them with embeddings from ratings into network architecture for deriving users' attentions on items. Although all these approaches have been validated to outperform the techniques solely based on rating scores (user-item interactions), they remain limited due to treating review texts as simple bag-of-words that not consider the context of words and consequently lose a lot of semantic information.

To tackle this limitation, several works [28–33] modeling the contextual information from the text of reviews with neural treatments are proposed. In these works, the Convolutional Neural Network (CNN) architecture is exploited to model users and items from their associated reviews. For instance, ConvMF [28] incorporates CNN into MF for rating prediction. In this model, the latent semantic features are derived from review text using CNN, which can consider the words' order and their local context. DeepCoNN [29] utilizes two parallel CNN networks to individually capture the latent semantic representations of users and items by using their reviews. Then, the obtained representations of users and items are concatenated and sent to a Factorization Machine (FM) for estimating unknown ratings. To ameliorate DeepCoNN, TransNet [30] extends it by introducing an extra layer for learning the latent representation of the target user-item review during the training phase and then regularizing the output of this layer based on the learned representation. Another improvement of DeepCoNN consists of PARL [31], which plugged into DeepCoNN a plug-and-play model to enrich target user's preferences exploiting reviews written by similar users, especially in the case when reviews of the target user are incomplete or sparse. CARL [32] learns latent features from reviews by utilizing convolution operations and an attention mechanism, then incorporates them with latent rating embeddings into a FM model to derive missing scores. CARP [33] firstly extracts logic units (each one formed by a user viewpoint and an item aspect) from the users' and items' reviews using a self-attention mechanism incorporated into a convolutional layer; then, based on a novel Routing by Bi-Agreement process, it derives the sentiment representations in user-item level for rating prediction.

As we can notice from the literature review, RSs have attracted research interest in the past few years, but mostly for the English content. Specifically, all the works mentioned above have demonstrated high efficiency and accuracy on various English reviews' datasets for RSs, including Yelp[1] and/or Amazon,[2] and so on. Thus, many high-quality systems are now available for English content. However, very restricted researches were achieved to investigate RSs utilizing Arabic reviews' datasets. The existing studies that we found include [19, 20], which combine Sentiment Analysis (SA) systems with CF-based RSs to predict users' preferences on items. The work [19] uses three datasets containing, respectively, 100 reviews in Algerian dialect, 1000 reviews in Arabic, and 2000 reviews in French and English. It adopts a semi-supervised support vector machine algorithm for determining the polarity score of the reviews (positive/neutral/negative). Finally, to predict missing ratings, it integrates them into user-based CF RS. In [20], an Opinion Corpus for Arabic (OCA) dataset has been used. It contains 500 Arabic-reviews about different movies from different websites. To determine the reviews' polarity score (-1, $+1$), the Singular Value Decomposition (SVM) classifier was adopted. This phase's

[1] https://www.yelp.com/dataset.

[2] http://jmcauley.ucsd.edu/data/amazon/.

output is then combined with numerical ratings into the MF technique for predicting missing ratings. The experimental results of the two studies [19, 20] both shown the positive impact of SA incorporation into RSs. Nevertheless, these surveyed works stay limited because they share different weak points, namely:

- Their reported experiments have been conducted on small product datasets of less than a thousand reviews, and none of them has focused on large Arabic datasets.
- The authors used very simple text analysis techniques to extract information from reviews and did not analyze the impact of the Arabic text preprocessing step on RSs.
- These works use only traditional RSs contrarily to modern and performant systems adopted in studies devoted to English content.
- None of these works has used several Arabic reviews' datasets to evaluate its proposed RS in the Arabic context properly. Thus, a more in-depth empirical study is required.

However, current works are thus not mature yet. Therefore, they do not derive conclusions about the Arabic content's exploitation by recent advances in the RSs field. Aiming to shed light on this situation, in this work, we report an extensive investigation on incorporating the Arabic content into recent recommendation methods.

Methodology

In this section, we present our methodology used for evaluating RSs while exploiting textual reviews written in the Arabic language. It consists of running experiments with five review-based recommender engines to evaluate and compare their performance when varying the language of reviews from the original language (English) to Arabic. Our methodology can be summarized as follows: first, we prepare different English reviews' datasets and generate their equivalent Arabic versions. Second, we pre-processed the constructed datasets. Third, we applied various recommending paradigms for rating prediction in both contexts (English and Arabic). Finally, we measured and compared their accuracy and efficiency. The details of each phase are explained in the following sub-sections. The overall process is illustrated in Fig. 1.

Data preparation and generation

Being aware of the lack of resources related to the Arabic language in the RSs field, we decided to build four publicly available Arabic datasets for RSs. Each of these datasets was constructed by preparing and translating English-language reviews obtained from a specific publicly-available Amazon dataset (see Footnote 2). In fact, each record in these existing English-language datasets contains one user's review on a specific item in the related category. Thus, each record consists of nine fields [34], namely:

- reviewerID: ID (Identifier) of the reviewer in the Amazon platform.
- asin: ID of the item in Amazon websites.
- reviewerName: represents the name of the reviewer.
- helpful: helpfulness score of the review.
- reviewText: contains the review's text.

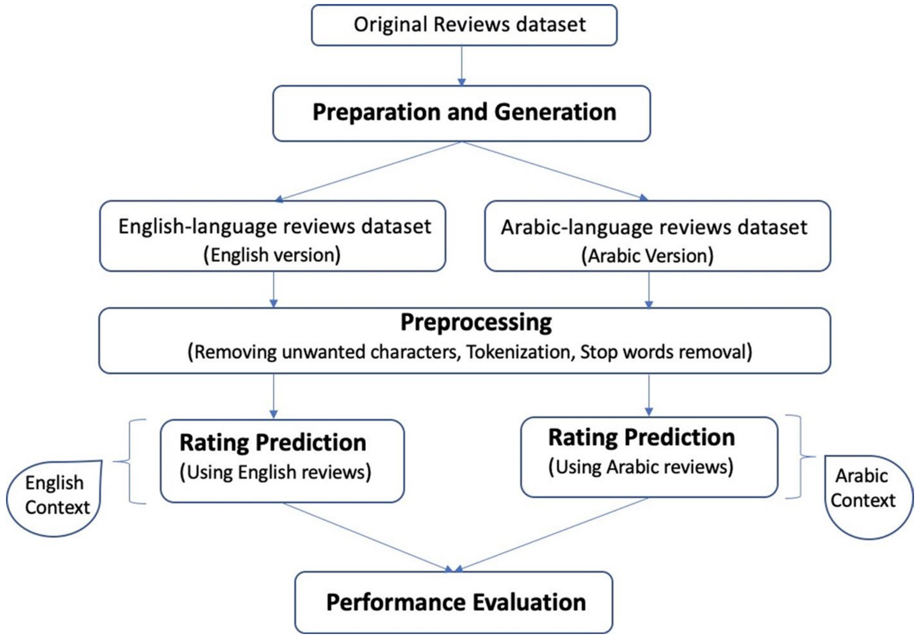

Original Reviews dataset
Preparation and Generation
English-language reviews dataset (English version)
Arabic-language reviews dataset (Arabic Version)
Preprocessing
(Removing unwanted characters, Tokenization, Stop words removal)
English Context
Rating Prediction (Using English reviews)
Rating Prediction (Using Arabic reviews)
Arabic Context
Performance Evaluation

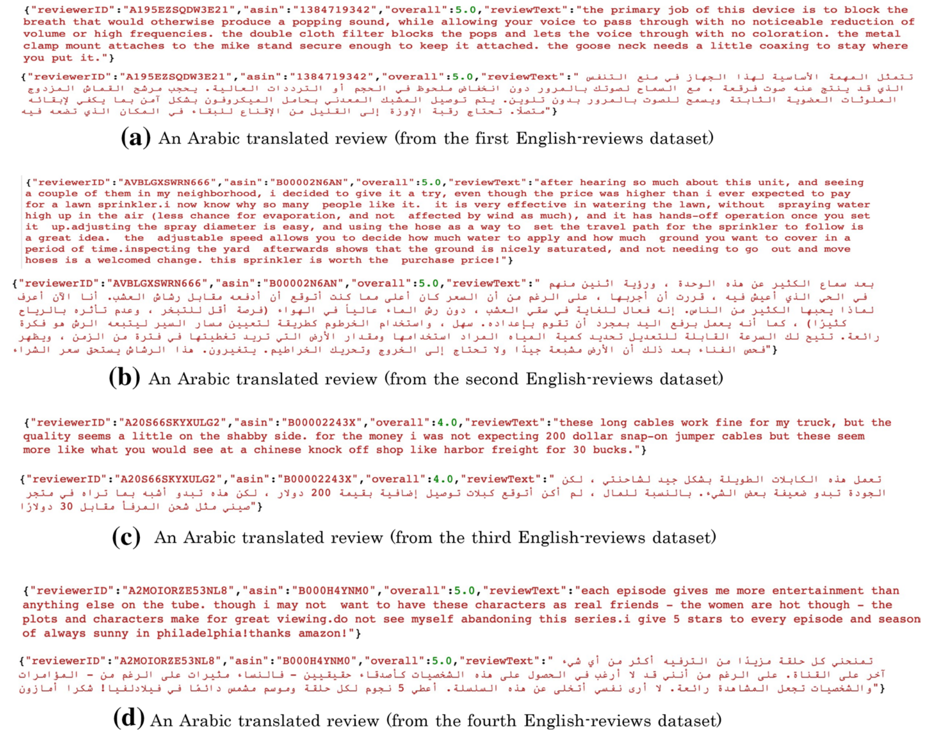
Fig. 1 Steps followed in the proposed methodology

- overall: the rating score of the item.
- summary: contains the review's summary.
- unixReviewTime: the review's time (Unix[3] time).
- reviewTime: the review's time and date (raw).

In order to prepare the datasets used in our work, we implemented a special parser in Python, allowing us to extract reviews from the original English datasets [JavaScript Object Notation (JSON) files] by retaining only the specific fields that we require. In fact, for each review, only the reviewerID, asin, reviewText, and overall were preserved. The remaining other fields were ignored since they are needless for this task. Thus, each record in our datasets contains one review, including the targeted four fields' information.

However, online reviews are usually short informal texts generated by non-experts. They are characterized by using everyday language, grammar and misspelling errors, non-standard vocabulary such as replicated characters, non-formal abbreviations, and so on. Thus, the reviews' cleaning stage is essential to ensure the quality of their content. Its main goal is to clean reviews text from spelling errors and slang words to help the downstream process easily understand and resolve their meanings. To attain it, we proceed as follows: first, all the text data are converted into lower case letters to ensure that the texts are in a uniform format. The application of this task to our reviews' text makes sure that "The" and "the" or "Do" and "do" are treated as the same. Then, we adopted regular expressions to eliminate extra-characters in any sequence that repeated more than

[3] https://en.wikipedia.org/wiki/Unix.



{"reviewerID":"A195EZSQDW3E21","asin":"1384719342","overall":5.0,"reviewText":"the primary job of this device is to block the
breath that would otherwise produce a popping sound, while allowing your voice to pass through with no noticeable reduction of
volume or high frequencies. the double cloth filter blocks the pops and lets the voice through with no coloration. the metal
clamp mount attaches to the mike stand secure enough to keep it attached. the goose neck needs a little coaxing to stay where
you put it."}

{"reviewerID":"A195EZSQDW3E21","asin":"1384719342","overall":5.0,"reviewText":" تتمثل المهمة الأساسية لهذا الجهاز في منع التنفس
الذي قد ينتج عنه صوت فرقعة ، مع السماح لصوتك بالمرور دون انخفاض ملحوظ في الحجم أو الترددات العالية. يحجب مرشح القماش المزدوج
الملوثات العضوية الثابتة ويسمح للصوت بالمرور بدون تلوين. يتم توصيل المشبك المعدني بحامل الميكروفون بشكل آمن بما يكفي لإبقائه
متصلاً. تحتاج رقبة الإوزة إلى القليل من الإقناع للبقاء في المكان الذي تضعه فيه"}

(a) An Arabic translated review (from the first English-reviews dataset)

{"reviewerID":"AVBLGXSWRN666","asin":"B00002N6AN","overall":5.0,"reviewText":"after hearing so much about this unit, and seeing
a couple of them in my neighborhood, i decided to give it a try, even though the price was higher than i ever expected to pay
for a lawn sprinkler.i now know why so many people like it. it is very effective in watering the lawn, without spraying water
high up in the air (less chance for evaporation, and not affected by wind as much), and it has hands-off operation once you set
it up.adjusting the spray diameter is easy, and using the hose as a way to set the travel path for the sprinkler to follow is
a great idea. the adjustable speed allows you to decide how much water to apply and how much ground you want to cover in a
period of time.inspecting the yard afterwards shows that the ground is nicely saturated, and not needing to go out and move
hoses is a welcomed change. this sprinkler is worth the purchase price!"}

{"reviewerID":"AVBLGXSWRN666","asin":"B00002N6AN","overall":5.0,"reviewText":" بعد سماع الكثير عن هذه الوحدة ، ورؤية اثنين منهم
في الحي الذي أعيش فيه ، قررت أن أجربها ، على الرغم من أن السعر كان أعلى مما كنت أتوقع أن أدفعه مقابل رشاش العشب. أنا الآن أعرف
لماذا يحبها الكثير من الناس. إنه فعال للغاية في سقي العشب ، دون رش الماء عالياً في الهواء (فرصة أقل للتبخر ، وعدم تأثره بالرياح
كثيرًا) ، كما أنه يعمل برفع اليد بمجرد أن تقوم بإعداده. سهل ، واستخدام الخرطوم كطريقة لتعيين مسار السير ليتبعه الرش هو فكرة
رائعة. تتيح لك السرعة القابلة للتعديل تحديد كمية المياه المراد استخدامها ومقدار الأرض التي تريد تغطيتها في فترة من الزمن ، ويظهر
فحص الفناء بعد ذلك أن الأرض مشبعة جيدًا ولا تحتاج إلى الخروج وتحريك الخراطيم. يتغيرون. هذا الرشاش يستحق سعر الشراء."}

(b) An Arabic translated review (from the second English-reviews dataset)

{"reviewerID":"A20S66SKYXULG2","asin":"B00002243X","overall":4.0,"reviewText":"these long cables work fine for my truck, but the
quality seems a little on the shabby side. for the money i was not expecting 200 dollar snap-on jumper cables but these seem
more like what you would see at a chinese knock off shop like harbor freight for 30 bucks."}

{"reviewerID":"A20S66SKYXULG2","asin":"B00002243X","overall":4.0,"reviewText":" تعمل هذه الكابلات الطويلة بشكل جيد لشاحنتي ، لكن
الجودة تبدو ضعيفة بعض الشيء. بالنسبة للمال ، لم أكن أتوقع كبلات توصيل إضافية بقيمة 200 دولار ، لكن هذه تبدو أشبه بما تراه في متجر
صيني مثل شحن المرفأ مقابل 30 دولارًا"}

(c) An Arabic translated review (from the third English-reviews dataset)

{"reviewerID":"A2MOIORZE53NL8","asin":"B000H4YNM0","overall":5.0,"reviewText":"each episode gives me more entertainment than
anything else on the tube. though i may not want to have these characters as real friends - the women are hot though - the
plots and characters make for great viewing.do not see myself abandoning this series.i give 5 stars to every episode and season
of always sunny in philadelphia!thanks amazon!"}

{"reviewerID":"A2MOIORZE53NL8","asin":"B000H4YNM0","overall":5.0,"reviewText":" تمنحني كل حلقة مزيدًا من الترفيه أكثر من أي شيء
آخر على القناة. على الرغم من أنني قد لا أرغب في الحصول على هذه الشخصيات كأصدقاء حقيقيين - فالنساء مثيرات على الرغم من - المؤامرات
والشخصيات تجعل المشاهدة رائعة. لا أرى نفسي أتخلى عن هذه السلسلة. أعطي 5 نجوم لكل حلقة وموسم مشمس دائمًا في فيلادلفيا! شكرا أمازون"}

(d) An Arabic translated review (from the fourth English-reviews dataset)

Fig. 2 An example of the obtained translated reviews' texts from the prepared datasets

twice. After that, we addressed the misspelled words in reviews based on a Spelling corrector in Python called Autocorrect.[4] This tool corrects all the detected misspelled terms and replace them with the correct words. At the final stage, we replace the contractions marked by clitic apostrophes with their extended forms. This is achieved by transforming the Wikipedia English contraction-to-expansion list[5] into a python dictionary and then exploited a regular expression for expanding all the existent contractions.

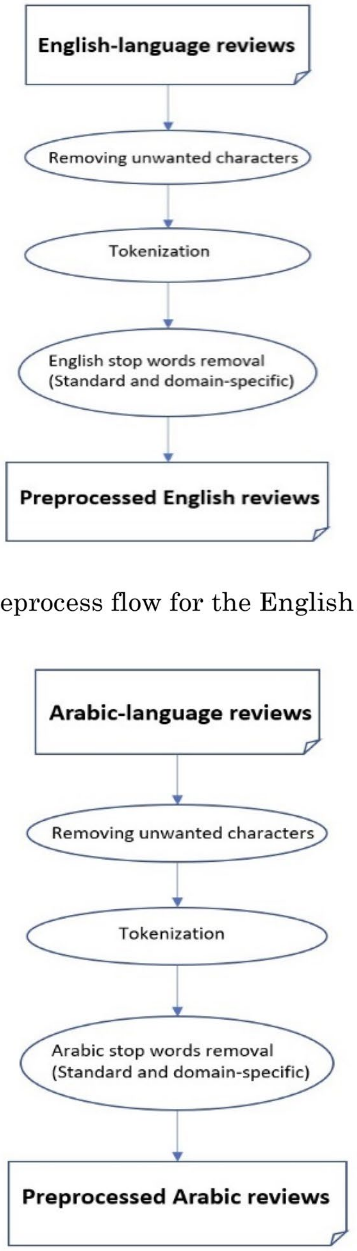
After that, we used a free online Machine Translation (MT) service to generate the Arabic versions of the prepared English datasets. We decided to use the Google TranslateMT tool due to its high efficiency in the translation task. Thus, we implemented a python script using the Python googletrans library[6] to interact with Google Translate API.[7]For each review in each of the collected datasets, we sent the content of its corresponding reviewText field to the API to get its translation into the Arabic language. Figure 2 shows an example of the obtained translated reviews' texts from the prepared English datasets.

[4] https://libraries.io/pypi/autocorrect.

[5] https://en.wikipedia.org/wiki/Wikipedia: List of English contractions.

[6] https://pypi.org/project/googletrans/.

[7] https://translate.google.com.


English-language reviews
Removing unwanted characters
Tokenization
English stop words removal
(Standard and domain-specific)
Preprocessed English reviews
Arabic-language reviews
Removing unwanted characters
Tokenization
Arabic stop words removal
(Standard and domain-specific)
Preprocessed Arabic reviews

(a) Preprocess flow for the English reviews.

(b) Preprocess flow for the Arabic reviews.

Fig. 3 Preprocessing scheme for the used English and Arabic reviews

Reviews preprocessing

Due to the texts of reviews are unstructured, it won't be easy to analyze them directly. Thus, it is recommended to preprocess all the reviews before extracting the feature vectors that will be integrated into RSs. To do this, we created our own text preprocessing scheme, which implies four stages, including unwanted characters removal, tokenization, removing standard stop words, and deleting domain-specific stop words. The details related to each step are discussed below. Figure 3 illustrates the adopted overall preprocessing flow.

Removing unwanted characters

The first stage of the preprocessing module consists of removing unwanted characters. It removes all marks, extra whitespaces, and any other non-alphabetic characters detected in the reviews. This phase is very significant as such instances do not add any hint or value to reviews' content. Hence, removing these characters will help for manipulating and processing text information. Our preprocessing scheme uses Python's re module[8] for removing all the aforementioned unwanted characters from reviews text.

Tokenization

As another step of preprocessing, there is tokenization. This step aims to split the text into a set of tokens (words) based on the punctuation or whitespaces characters. In our text preprocess scheme, this task was applied for each review text to divide it into multiple words utilizing whitespaces characters. There are many libraries to perform tokenization like NLTK,[9] SpaCy,[10], and TextBlob.[11] The NLTK was used in this work. The obtained tokens are exploited for further preprocessing stages.

Standard stop words removal

Stop words like prepositions, pronouns, articles, and conjunctions are employed frequently in reviews. Such words carry less meaning than other keywords, and thus they are not significant for the feature extraction task for RS. Therefore, in our study, a stop words removal step was performed before begin the recommendation process. In the literature, there is no unique universal stop words list used by all text mining tools. For this, in our work, we manually created two lists of stop words. The first one is devoted to the English language, and the second one concerns the Arabic language. In this step, we compare each target token with its corresponding-language stop word list. If it appears in the list, then it is eliminated. For instance, the words such as (have, he, in, is, it, its) are suppressed from the English and Arabic reviews.

Domain-specific stop words removal

While we removed the frequent stop words, it is a crucial practice in text mining to identify unusual words with little discriminant value within a specific field or context. These words are called Domain-specific stop words because they differ from one domain to another. The metric most widely utilized to perform the Domain- specific Stop words removal is "Term Frequency Inverse Document Frequency" (TF-IDF). It computes the relevance of words in a document based on their occurrence's frequency on several documents. Using this metric in our scheme, we may pick out the most pertinent words, allowing us to represent better the reviews related to a target dataset. Thus, we

[8] https://docs.python.org/3/library/re.html.

[9] https://pypi.org/project/nltk/.

[10] https://pypi.org/project/spacy/.

[11] https://pypi.org/project/textblob/.

decided to apply the TF-IDF technique on all datasets by retaining the top 20,000 distinct words (with high TF-IDF scores) as a vocabulary for each small dataset (number of reviews < 15,000) and the top 50,000 distinct words for large datasets (number of reviews > 15,000). All words out of the vocabulary are removed and considered as domain-specific stop words. We further filter out the records containing empty reviews after datasets preprocessing.

The recommender systems used

In our experimentation, the five following RSs were used, namely ALFM, A3NCF, PARL, CARL, and CARP. These RSs are presented below. We refer the readers to read the original papers [26, 27, 31–33] to have more details about these RSs.

ALFM

ALFM is the state-of-the-art recommender model that adopts the Latent Dirichlet Allocation (LDA) paradigm with the Probabilistic MF for rating prediction. This model firstly runs an LDA-based algorithm on reviews' texts to model user's preferences and item's properties in different aspects, thus capturing the importance of aspects for the user and item. It exploits an Aspect-aware Topic Model (ATM) for modeling aspect importance for target user/item as a probability distribution of composite topics, each of which is represented by a set of words from reviews. Then, the output from ATM is combined with ALFM, which associates latent factors with different aspects exploiting the MF approach, such that the model can predict aspect ratings. Finally, the overall rating is obtained by a linear combination of the aspect ratings, which are weighted by the importance of corresponding aspects.

A3NCF

This is the state-of-the-art RS that fuses topic modeling and deep learning. This system firstly adopts LDA to obtain the topics vectors of users and items from textual reviews. Thus, it models users' and items' feature vectors in different topics (the aspects of items that users discuss in reviews) as probability distributions of words that refer to the same topic. For each user and item, their related topics vectors and embedding vectors (from ratings) are fused into an attention neural network to learn their final representation by considering the user's attention weights concerning the different aspects of the target item. The model represents user and item as one-hot encoded vectors and then incorporate them into an embedding layer to get the embedding features for users and items. Finally, an attentive interaction between the user's and item's final representations is feed into fully connected layers (MultiLayer Perceptron) with regression to predict the final ratings.

PARL

PARL is a plug-and-play deep-learning architecture that has been plugged into DeepCoNN, one of the state-of-the-art review-based RS to improve its prediction accuracy upon different user-item pairs. DeepCoNN RS uses two parallel CNNs and a word embedding method for capturing latent representations from the reviews' words associated with the target user and item. The model concatenates the user and item vectors and then transmits it to a regression layer involving the FM method to predict ratings. Although DeepCoNN has shown good effectiveness for rating prediction task, it remains limited. The review sparsity issue is one of the significant limitations when the reviews' texts are short and scarce. In this case, few useful features can be extracted by CNN from the incomplete text data. PARL extracts useful user-item pair-dependent features from user's auxiliary reviews (written by other users with the same rating scores as given by the target user) to alleviate this limitation. Like DeepCoNN, PARL incorporates the extracted auxiliary reviews into CNN layers to transform them into feature vectors. To preserve the useful features for each target user-item pair, PARL incorporates the obtained feature vectors into an abstracting layer involving a highway network and a gated-mechanism. For final ratings, the auxiliary vectors are combined with their corresponding users' vectors in DeepCoNN.

CARL

This is the state-of-the-art RS that learns context-aware representations for each user-item pair based on their characteristics and interactions by exploiting both the textual reviews and the user-item interaction data. CARL consists of two learning components, namely review-based feature learning and interaction-based feature learning. The first one adopts a CNN (using two parallels subnetworks) and an attention mechanism to jointly learn useful latent features for a user-item pair based on the user and item reviews. In this component, CARL also adopts an abstraction layer to obtain pertinent latent features using an average pooling strategy. On the other hand, in the interaction-based component, complementary features are learned for each user and item based on their interaction data. To produce ratings through each module, CARL feeds each component's latent representation into the FM. The final rating score is then computed based on a dynamic fusion strategy that fuses both components' ratings.

CARP

CARP represents the state-of-the-art RS that uses Capsule Network to extract the semantic contextual information from reviews for rating prediction. CARP is based on two modules: viewpoint and aspect extraction and sentiment capsules. The first component adopts a variant of self-attention stacked over a convolutional layer to capture the logic units formed by a given user viewpoint and an item aspect extracted from user and item reviews. In the second component, positive and negative capsules are exploited by a Routing by Bi-Agreement architecture to jointly choose some logic units as the informative ones and produce output vectors that encode their sentiments. Each constructed vector encodes the user's attitudes on a given item in the target sentiment. Besides, the lengths of the vectors suggest the probability of each of these two sentiments. Finally,

Table 1 Statistical details of the datasets

Dataset	#Users	#Items	#Reviews	Sparsity (%)
MI	1429	900	10,261	99.2
Arabic_MI				
PLG	1686	962	13,272	99.1
Arabic_PLG				
Auto	2928	1835	20,473	99.6
Arabic_Auto				
IV	5130	1685	37,126	99.5
Arabic_IV				

to predict a user-item pair's missing rating, the magnitudes and odds in their two corresponding sentiment poles are incorporated into a one-layer highway network and then passed to a rescaled sigmoid function.

Experiments

This section presents some issues about our experiments: first, we introduce the used datasets. Second, we describe the experimental settings and the evaluation metric. Finally, we discuss and interpret the achieved empirical results.

Datasets

To answer the research questions stated in the "Research motivation" section, we run our experiments on several English-language reviews' datasets and their equivalent versions in the Arabic language. Specifically, we used eight datasets: four 5-core datasets, Musical Instruments (MI), Patio Lawn and Garden (PLG), Automotive (Auto), and Instant Video (IV), were chosen from Amazon and contain English-language reviews related to various products. The other four datasets represent the Arabic versions of the first ones. They contain the same reviews but in the Arabic language. We built all datasets based on the preparation and generation process discussed in the "Methodology" section. Table 1 summarizes the statistical details of the used datasets.

Experimental settings

We randomly selected 80% of each dataset as the training set for our evaluations and the remaining 20% as the test set. We trained the adopted models on the training set and evaluated the performance on the test set. To ensure that the testing set reviews are unavailable during the recommending process, such in real-world applications, we utilized the review information only in the training set.

The parameters of all evaluated models (ALFM, A3NCF, PARL, CARL, and CARP) are set as reported in their corresponding papers with the best performance.

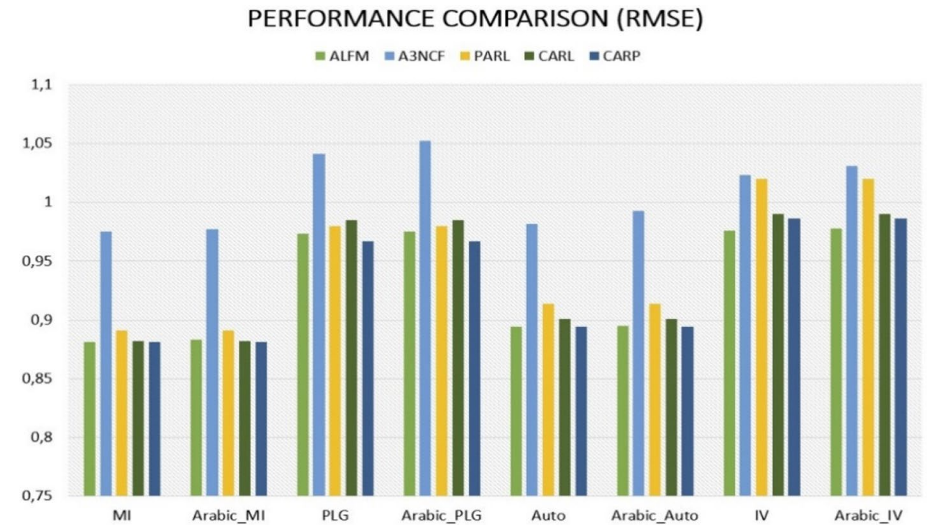


Fig. 4 Comparing the adopted recommendation models in the English and Arabic contexts

For the evaluation metric, we used Root Mean Square Error (RMSE) as a performance metric, which is broadly utilized in several related works for performance evaluation [35–37], formulated as:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{(u,i)\in T}(\widehat{r}_{u,i} - r_{u,i})^2}$$

where T is the total number of data points being tested, $\widehat{r}_{u,i}$ is the predicted rating, and $r_{u,i}$ is the real rating.

Experimental results and discussions

This sub-section presents the realized experiments to address our research questions: ***RQ1***, ***RQ2***, and ***RQ3*** ("Research motivation" section).

The analysis of the attained empirical findings is organized as follows. We started by discussing the applicability of the modern RSs in the Arabic context. Next, we evaluated the effect of applying a preprocessing scheme on the Arabic texts before incorporating them into RSs. We ended up by assessing the performance of the recent RSs in the Arabic context.

(RQ1) Verifying the applicability of recent RSs in the Arabic context

Appendix 1: Table 2 and Fig. 4 show the performances of the five RSs ALFM, A3NCF, PARL, CARL, and CARP in both contexts: English and Arabic. From the results, we can note that for the four Arabic datasets: Arabic_MI, Arabic_PLG, Arabic_Auto, and Arabic_IV,

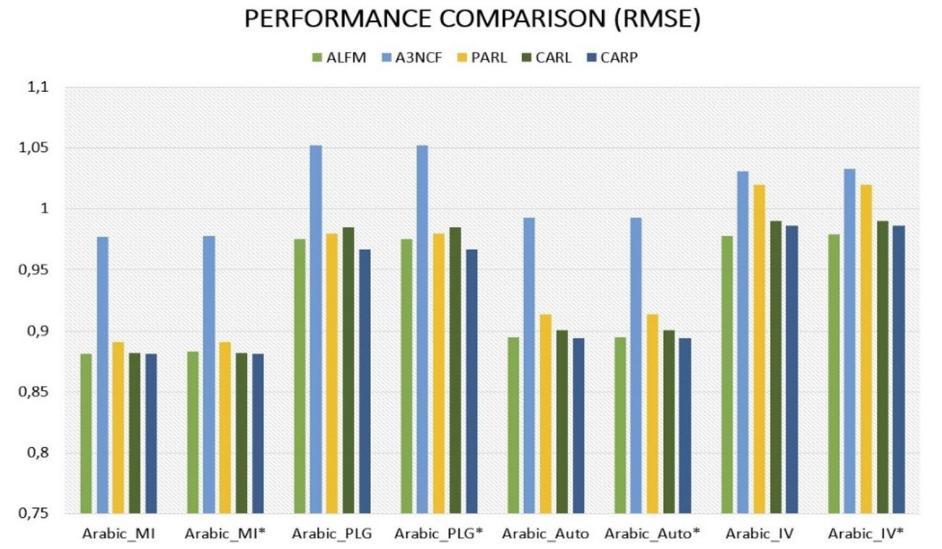


Fig. 5 Impact of the preprocessing phase on RSs when applied to Arabic content

all tested RSs worked well, i.e., overall, they reached good RMSE scores. According to the achieved RMSE scores, we note that each used RS successfully predicted most of the hidden ratings when using Arabic reviews' texts. Therefore, we can conclude from our results that it's possible to apply modern recommender engines to Arabic content.

(RQ2) Analyzing the impact of the preprocessing phase on recent RSs when applied to Arabic content

As done by other researchers in [19, 20], and due to the Arabic language's complex structure [17, 18, 38], we have opted to preprocess the Arabic language reviews before incorporating them into the RSs. However, in this part, we aim to verify if adopting the preprocessing stage is necessary to apply these RSs on the Arabic datasets. To achieve it, we test the five RSs on different Arabic reviews' datasets in both cases: using preprocessing stages and without-using preprocessing tasks. The results are shown in Appendix 1: Table 3 and Fig. 5. By comparing the results achieved from the ALFM, A3NCF, PARL, CARL, and CARP in both cases, we can note that each of these RSs has maintained the same performance on each preprocessed reviews dataset (Arabic MI, Arabic PLG, Arabic Auto, and Arabic IV) and its no-preprocessed version (Arabic_MI*, Arabic_PLG*, Arabic_ Auto* and Arabi_IV*), respectively. Those results confirm that the preprocessing stage is not essential to have better accuracy and performance for the modern RSs when using

Arabic texts. Such systems use effective feature extraction techniques that can capture the relevant information from the texts.

(RQ3) Analyzing the performance of the recent RSs in the Arabic context

In this part, we compare the performance of the RSs (ALFM, A3NCF, PARL, CARL, and CARP) in two contexts: English and Arabic. In particular, we aim to verify if the used recommendation models' differences in performance are statistically significant or not when changing the content's language. The experimental results are shown in Appendix 1: Table 2 and Fig. 4. From the experimental results, we can analyze the performances of the tested RSs. The results show that the RS ALFM has maintained very close performance for each English dataset and its Arabic version. The accuracy decrease on the four datasets' pairs (MI and Arabic_MI, PLG and Arabic_PLG, Auto and Arabic_Auto, IV and Arabic_IV) is 0.21%, 0.21%, 0.11%, and 0.20%, respectively. Similarly, A3NCF has also maintained very nearly accuracy on each English dataset and its Arabic version. The accuracy decrease on the four datasets' pairs is 0.21%, 1.06%, 1.12%, and 0.78%, respectively. Such accuracy differences are insignificant (very minor). We explain these because these RSs use Bag-of-Words (BoW) techniques for review text processing, which are negatively impacted by the noise and irrelevant information introduced within the reviews during the translation phase. The results of this phase depend on the quality of the translation tool.

On the other hand, PARL, CARL, and CARP's performance do not degrade when changing the reviews' text language. Specifically, each of these three RSs has maintained the same rating prediction accuracy (accuracy decrease is 0%) on each English dataset and its Arabic version. We suggest that maintaining the same performance is due to processing the translated reviews by Deep Learning architectures, which help delete the noisy and unimportant information and develop appropriate feature representations for RSs. These results confirm the conclusions obtained in other studies for English content [39], which prove that neural network architectures' modularity allows handling heterogeneous and unstructured text content.

By considering those outputs, we can confirm that ALFM, A3NCF, PARL, CARL, and CARP do not lose performance when changing the application context (English to Arabic). Consequently, from these experimental results, we can conclude that applying the recent RSs to Arabic content provides good rating prediction accuracy as when using English content.

Conclusion and future work

This article aims to provide a comprehensive evaluation of modern RSs when applied to Arabic content. For that, extensive experiments have been conducted using different Arabic reviews' datasets. These datasets were built leveraging the English-language reviews regarding different product categories in the Amazon e-platform. The experiments were performed utilizing five recent RSs. Each of these systems was tested on different-language datasets: original English ones and their corresponding Arabic versions. This study was conducted to achieve three objectives.

The first aim is to verify the applicability of recent RSs to the Arabic content. Our experimental results have shown that the used RSs could perform rating prediction tasks while exploiting Arabic data. The second goal was to determine if there is a necessity to apply a particular preprocessing phase on the Arabic content before incorporating it into RSs. We tested each of the RSs on Arabic reviews within two cases: with and without-preprocessing. The results showed that the preprocessing stage does not impact RSs' performance. The third objective of this work is to evaluate if the performance of these RSs varies when changing the application context from the original one (English) to Arabic. The experimental findings proved that the adopted RSs had maintained the same performance in both contexts.

The results demonstrated the important potential of recent advances in the RSs field when exploiting Arabic content in terms of accuracy and performance. However, it has to be noted that in our experiments, we only used a MT tool to translate reviews' text from English into the Arabic language for building the Arabic data. Thus, as future work, we will try to test on original Arabic language datasets by collecting real reviews from online Arabic stores to assess RSs in the Arabic context better.

This work will encourage the researcher community and serve as a road map to advance the Arabic RSs field. Such advancement would positively impact various sectors, including Arabic: e-commerce, e-learning, e-tourism, e-government, social networks, etc.

Acknowledgements
The authors thank anonymous reviewers for their valuable comments and suggestions.

Authors' contributions
MS took on the main role performed the literature, designed, performed experiments, analyzed data and wrote the paper. AAL and SM supervised the research. AO reviewed the manuscript language and helped edit the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The datasets used in our experiments are available at https://jmcauley.ucsd.edu/data/amazon/ ("Small" subsets).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] LRIT, Associated Unit to CNRST (URAC 29), Faculty of Sciences, Mohammed V University in Rabat, Rabat, Morocco. [2] Informatics Department, Faculty of Sciences and Technologies of Mohammedia (FSTM), Hassan II University, Casablanca, Morocco. [3] Laboratory of Engineering Sciences, National School of Applied Sciences (ENSA), IbnTofail University, Kénitra, Morocco.

Appendix 1

See Tables 2 and 3.

Table 2 Performance comparison on eight datasets in terms of RMSE

Datasets	ALFM	A3NCF	PARL	CARL	CARP	Q% (English vs. Arabic)				
						ALFM (%)	A3NCF (%)	PARL (%)	CARL (%)	CARP (%)
MI	0.881	0.975	0.891	0.882	0.881	0.21	0.21	0	0	0
Arabic_MI	0.883	0.977	0.891	0.882	0.881					
PLG	0.973	1.041	0.980	0.985	0.967	0.21	1.06	0	0	0
Arabic_PLG	0.975	1.052	0.980	0.985	0.967					
Auto	0.894	0.982	0.914	0.901	0.894	0.11	1.12	0	0	0
Arabic_Auto	0.895	0.993	0.914	0.901	0.894					
IV	0.976	1.023	1.020	0.990	0.986	0.20	0.78	0	0	0
Arabic_IV	0.978	1.031	1.020	0.990	0.986					

Q%: accuracy decrease (pair of English-Arabic datasets)

Table 3 Impact of the preprocessing phase on RSs in the Arabic context

Datasets	ALFM	A3NCF	PARL	CARL	CARP
Arabic_MI	0.881	0.977	0.891	0.882	0.881
Arabic_MI*	0.883	0.978	0.891	0.882	0.881
Arabic_PLG	0.975	1.052	0.980	0.985	0.967
Arabic_PLG*	0.975	1.052	0.980	0.985	0.967
Arabic_Auto	0.895	0.993	0.914	0.900	0.894
Arabic_Auto*	0.895	0.993	0.914	0.900	0.894
Arabic_IV	0.978	1.031	1.020	0.990	0.986
Arabic_IV*	0.979	1.033	1.020	0.990	0.986

*No-preprocessed dataset

Received: 15 December 2020 Accepted: 23 January 2021
Published online: 17 February 2021

References

1. Tsai CW, Lai CF, Chao HC, Vasilakos AV. Big data analytics: a survey. J Big Data. 2015;2(1):1–32.
2. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng. 2005;17:734–49.
3. Sun Z, Guo Q, Yang J, Fang H, Guo G, Zhang J, Burke R. Research commentary on recommendations with side information: a survey and research directions. Electron Commer Res Appl. 2019;37:100879.
4. Isinkaye FO, Folajimi YO, Ojokoh BA. Recommendation systems: principles, methods and evaluation. Egypt Inform J. 2015;16(3):261–73.
5. Srifi M, Oussous A, Lahcen AA, Mouline S. Recommender systems based on collaborative filtering using review texts—a survey. Information. 2020;11:317.
6. Lu J, Wu D, Mao M, Wang W, Zhang G. Recommender system application developments: a survey. Decis Support Syst. 2015;74:12–32.
7. Liang Y, Qian T, Yu H. Artan: align reviews with topics in attention network for rating prediction. Neurocomputing. 2020;403:337–47.
8. Han J, Zheng L, Huang H, Xu Y, Philip SY, Zuo W. Deep latent factor model with hierarchical similarity measure for recommender systems. Inf Sci. 2019;503:521–32.
9. Nassar N, Jafar A, Rahhal Y. Multi-criteria collaborative filtering recommender by fusing deep neural network and matrix factorization. J Big Data. 2020;7:1–12.
10. Liang N, Zheng HT, Chen JY, Sangaiah AK, Zhao CZ. Trsdl: tag-aware recommender system based on deep learning–intelligent computing systems. Appl Sci. 2018;8(5):799.
11. Alshammari G, Jorro-Aragoneses JL, Polatidis N, Kapetanakis S, Pimenidis E, Petridis M. A switching multi-level method for the long tail recommendation problem. J Intell Fuzzy Syst. 2019;37(6):7189–98.
12. Lai CH, Chang YC. Document recommendation based on the analysis of group trust and user weightings. J Inf Sci. 2019;45(6):845–62.
13. Chu PM, Mao YS, Lee SJ, Hou CL. Leveraging user comments for recommendation in E-commerce. Appl Sci. 2020;10(7):2540.
14. Chen L, Chen G, Wang F. Recommender systems based on user reviews: the state of the art. User Model User-Adapt Interact. 2015;25:99–154.
15. Srifi M, Hammou BA, Mouline S, Lahcen AA. Collaborative recommender systems based on user-generated reviews: a concise survey. In: 2018 international symposium on advanced electrical and communication technologies (ISAECT). New York: IEEE; 2018. p. 1–6.
16. Stats IW. Top ten languages used in the web. 2020. https://www.internetworldstats.com/stats7.htm.
17. Alharbi A, Taileb M, Kalkatawi M. Deep learning in Arabic sentiment analysis: an overview. J Inf Sci. 2019. https://doi.org/10.1177/0165551519865488.
18. Oussous A, Benjelloun FZ, Lahcen AA, Belfkih S. ASA: a framework for Arabic sentiment analysis. J Inf Sci. 2020;46:544–59.
19. Ziani A, Azizi N, Schwab D, Aldwairi M, Chekkai N, Zenakhra D, Cheriguene S. Recommender system through sentiment analysis. In: 2nd international conference on automatic control, telecommunications and signals. 2017.
20. Harrag F, Al-Salman AS, Alquahtani A. Arabic opinion mining using a hybrid recommender system approach. 2020. arXiv preprint. arXiv:2009.07397.
21. McAuley J, Leskovec J. Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings of the 7th ACM conference on recommender systems. 2013. p. 165–72.
22. Bao Y, Fang H, Zhang J. Topicmf: simultaneously exploiting ratings and reviews for recommendation. In: Proceedings of the AAAI conference on artificial intelligence, citeseer; 2014. pp. 2–8.

23. Tan Y, Zhang M, Liu Y, Ma S. Rating-boosted latent topics: understanding users and items with ratings and reviews. In: IJCAI. 2016. p. 2640–6.
24. Zhang W, Wang J. Integrating topic and latent factors for scalable personalized review-based rating prediction. IEEE Trans Knowl Data Eng. 2016;28:3013–27.
25. Zhang Y, Lai G, Zhang M, Zhang Y, Liu Y, Ma S. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. 2014. p. 83–92.
26. Cheng Z, Ding Y, Zhu L, Kankanhalli M. Aspect-aware latent factor model: rating prediction with ratings and reviews. In: Proceedings of the 2018 world wide web conference. 2018. p. 639–48.
27. Cheng Z, Ding Y, He X, Zhu L, Song X, Kankanhalli MS. A3NCF: an adaptive aspect attention model for rating prediction. In: IJCAI. 2018. p. 3748–54.
28. Kim D, Park C, Oh J, Lee S, Yu H. Convolutional matrix factorization for document context-aware recommendation. In: Proceedings of the 10th ACM conference on recommender systems. 2016. p. 233–40.
29. Zheng L, Noroozi V, Yu PS. Joint deep modeling of users and items using reviews for recommendation. In: Proceedings of the tenth ACM international conference on web search and data mining. 2017. p. 425–34.
30. Catherine R, Cohen W. Transnets: learning to transform for recommendation. In: Proceedings of the eleventh ACM conference on recommender systems. 2017. p. 288–96.
31. Wu L, Quan C, Li C, Ji D. Parl: let strangers speak out what you like. In: Proceedings of the 27th ACM international conference on information and knowledge management. 2018. p. 677–86.
32. Wu L, Quan C, Li C, Wang Q, Zheng B, Luo X. A context-aware user-item representation learning for item recommendation. ACM Trans Inf Syst (TOIS). 2019;37:1–29.
33. Li C, Quan C, Peng L, Qi Y, Deng Y, Wu L. 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. 2019. p. 275–84.
34. AlZu'bi S, Alsmadiv A, Al Qatawneh S, Al-Ayyoub M, Hawashin B, Jararweh Y. A brief analysis of amazon online reviews. In: 2019 sixth international conference on social networks analysis, management and security (SNAMS). New York: IEEE; 2019. p. 555–60.
35. Liao X, Li X, Xu Q, Wu H, Wang Y. Improving ant collaborative filtering on sparsity via dimension reduction. Appl Sci. 2020;10(20):7245.
36. Anwar T, Uma V. CD-SPM: cross-domain book recommendation using sequential pattern mining and rule mining. J King Saud Univ Comput Inf Sci. 2019. https://doi.org/10.1016/j.jksuci.2019.01.012.
37. Hasanzadeh S, Fakhrahmad S, Taheri M. Based recommender systems: a proposed rating prediction scheme using word embedding representation of reviews. Comput J. 2020. https://doi.org/10.1093/comjnl/bxaa044.
38. ALMarwi H, Ghurab M, Al-Baltah I. A hybrid semantic query expansion approach for Arabic information retrieval. J Big Data. 2020;7(1):1–19.
39. Khan ZY, Niu Z, Sandiwarno S, Prince R. Deep learning techniques for rating prediction: a survey of the state-of-the-art. Artif Intell Rev. 2020;54:1–41.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.