Information Sciences 625 (2023) 738–756

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Hybrid recommendation by incorporating the sentiment of product reviews

Mehdi Elahi [a,b,*,1], Danial Khosh Kholgh [c], Mohammad Sina Kiarostami [c], Mourad Oussalah [c], Sorush Saghari [a,b,c]

[a] University of Bergen, Bergen, Norway
[b] NHH Norwegian School of Economics, Bergen, Norway
[c] University of Oulu, Faculty of ITEE, Oulu, Finland

ARTICLE INFO

*Article history:*
Received 22 December 2021
Received in revised form 4 January 2023
Accepted 5 January 2023
Available online 10 January 2023

*Keywords:*
Recommender systems
Sentiment analysis
Content-based recommendation

ABSTRACT

Hybrid recommender systems utilize advanced algorithms capable of learning heterogeneous sources of data and generating personalized recommendations for users. The data can range from user preferences (e.g., ratings or reviews) to item content (e.g., description or category).

Prior studies in the field of recommender systems have primarily relied on "ratings" as the user feedback, when building user profiles or evaluating the quality of the recommendation. While ratings are informative, they may still fail to represent a comprehensive picture of actual user preferences. In contrast, there are other types of feedback data that differently or complementarily represent users and their preferences, including the reviews and the sentiments encapsulated within them. Such data can reveal important parts of a user's profile that are not necessarily correlated with user ratings, and hence, they potentially reflect a different side of the user's profile.

In this paper, we propose a novel form of hybrid recommender system, capable of analyzing the reviews and extracting their sentiments that are incorporated into the recommendation process. We used advanced algorithms to generate recommendations for users capable of incorporating additional data, such as the review sentiment. We conducted analyses and showed that sentiments of user reviews are not always highly correlated with the ratings (e.g., in music domain). This might mean that sentiment can be indicative of a different aspect of user preferences and can be used as an alternative signal of user feedback. Hence, we have used both ratings and sentiments of reviews when evaluating our proposed hybrid recommender system. We selected two common datasets for the evaluation, Amazon Digital Music and Amazon Video Games, and showed the superior performance of the proposed hybrid recommender system compared to different baselines. The comparison were made in two evaluation scenarios, namely, when the ratings were considered the user feedback and when sentiments of the review were considered the user feedback.



* Corresponding author.
*E-mail addresses:* Mehdi.elahi@uib.no (M. Elahi), danial.khoshkholgh@oulu.fi (D. Khosh Kholgh), Mohammad.Kiarostami@oulu.fi (M.S. Kiarostami), Mourad.Oussalah@oulu.fi (M. Oussalah), sorush9saghari@gmail.com (S. Saghari).

[1] The co-authors have made equal contributions to this work.

https://doi.org/10.1016/j.ins.2023.01.051
0020-0255/

## 1. Introduction

### 1.1. What is a recommender system?

A recommender (or recommendation) system is a digital tool that uses a series of algorithms, data analysis, and possibly artificial intelligence (AI) for information filtering and to generate relevant recommendations to the user. Such a recommendation can be either personalized or generic.

A personalized recommendation subsumes incorporation of the user's profile and any related side information. The quality of the recommendation is directly related to the quantity and quality of the data available as part of the input. For example, in classical recommendation approaches, the more data collected and utilized about a target user, and similar like-minded users, the better and more personalized the recommendations can be [1].

Today, recommender systems are present in most streaming services in social networks, e-commerce and even in app stores that we use on a daily basis. Applications of recommender systems are visible on popular sites such as Netflix, where movies are organized according to genre, age, popularity, previous watch, and navigation, among others [2]. Historically, Amazon was the pioneer in implementing item-based recommendation through customization of the shopping experience, presentation of similar product content, and the popularity of items

### 1.2. How are recommender systems regarded in e-commerce?

In the age of digitalization and social media, a recommender system is seen as an ideal solution for achieving mass customization in the growing e-commerce industry. These systems can be seen as effective tools for dealing with exponential growth in the number of catalog items available in online stores, as well as the increasing number of attributes that can be exploited by the users when making a purchase decision. In this regard, recommender systems contribute to the customization of consumer experiences via the presentation of the available products, summarizing product content and community opinion, fair comparison with alternative choices, among others. This was also motivated by the growing impact of the one-to-one marketing strategy in several key industrial sectors , in which consumers' preferences should be fully accommodated. Similar is the development of ad-targeting, which treats a consumer as both an individual and a member of a market group, which provides golden opportunities for recommender systems research, in order to gain share in the lucrative advertising market. This enables search engines and advertising companies to suggest advertisements that are dynamically displayed on a user's navigation screen according to that consumer's behavior. The promotion of coupons and other financial incentives in exchange for completion of medium - to large-scale online surveys has been adopted by many advertising companies as a way to gain access to valuable user profiles that enhance the efficiency of the corresponding recommender system. This motivates global companies such as Google, Amazon, and Netflix to develop enhanced recommender systems to maintain sustainable competitive advantages in e-commerce. Overall, recommender systems enhance e-commerce activity in at least three distinct ways [3]:

- Browsers into buyers: This explores the user's click and browsing history, even without any purchases, to recommend similar products with a better cost-to-quality ratio, increasing the likelihood of the product being purchased.
- Cross-sell: This exploits the content of the user's shopping cart to recommend additional products that either bear some similarity or are complementary to the items already in the shopping cart.
- Loyalty: This builds on the user's purchase history from a given supplier to create a value-added relationship between the site and the customer, referred to as loyalty programs, which provides some extra benefits and services to the customer. This relationship can be further exploited to learn the user's interests and behaviors, which in turn, will be used to develop appropriate recommendations.

### 1.3. How sentiment analysis contributes to recommendations and e-commerce?

Sentiment Analysis (SA) is an automated process that uses AI and natural language processing (NLP) tools to analyze textual documents and identify sentiments and opinion [4]. Through SA approaches, we can explore the polarity of users' textual inputs and correlate such findings with purchase decisions or any other actions [5]. This can also reveal the root of a positive or negative sentiment. Negative reviews allow business entities to discover hidden cues, potential weaknesses, users' tastes and preferences, and rationality that needs to be addressed to gain customer satisfaction and expand market scope.

Traditionally, many recommender systems have focused on utilizing a user's ratings as the key characteristic of that user's profile to build and evaluate the recommendations made for them. It is trivial that such a criterion is not comprehensive enough to fully capture the dynamics and complexity of users' preferences. For instance, two distinct products can be assigned a full rating of five out of five, even if the detailed review content contains contrasting views (see examples provided in Table 1). Therefore, leveraging rating scores by sentiment polarity is expected to increase the accuracy of the under-

**Table 1**
Examples of inconsistencies between the average ratings provided by users and the average of the extracted sentiment within those reviews. Note: the range of the rating is [1,5] and the range for the sentiment is [0,1], where 0 indicates a (very) negative sentiment, and 1 indicates the (very) positive sentiment.

| Dataset | Item ID | Review Text | rating [1–5] | sentiment [0–1] |
|---|---|---|---|---|
| **Amazon Digital Music** | 1098 | Again I like the older | positive (5) | negative (0) |
| | 5673 | Love the song | negative (1) | positive (1) |
| **Amazon Video Games** | 4092 | I LOST IT | positive (5) | negative (0) |
| | 2734 | it's ok | negative (1) | positive (1) |

lying recommender system. This also expects that it will either implicitly or explicitly provide useful insights about users that could permit a business entity to enhance its loyalty program and use profile information for cross-selling opportunities.

### *1.4. Challenges with comprehending sentiment with recommender systems.*

In essence, the challenges in using SA outcomes are ultimately linked to the inherent ambiguity of human language and the complexity of various linguistic modifiers, composition, and context-associated dilemmas. This can consequently yield a large discrepancy between the automatically induced sentiment polarity and the human interpretation of that sentiment polarity. For instance, the word "suck" can bear positive polarity as in "The vacuum sucks well." and a negative polarity in "This product sucks." Similarly, the sentence "The phone is amazingly bad" would be treated as neutral because it contains one positive and one negative word, while it indicates a clear negative orientation. This is again exemplified in the Amazon dataset shown in Table 1, where one can identify inconsistencies between the automatically generated sentiment polarity using a standard lexicon-based approach and the rating (See Section 2.3 for a more comprehensive review of SA).

In this study, we used a model based on BERT (bidirectional encoder representations from transformers) [2], which is a state-of-the-art machine learning model developed primarily for natural language processing tasks. The model provides encoding for each phrase/sentence/paragraph of textual input, learned from the more than 2 billion items in the Google news corpus, which expects us to encapsulate, to a large extent, the contextual information and homonym information. This does not provide a universal solution to the broad challenges of SA tasks but it at least narrows the scope.

### *1.5. Assumptions*

The assumptions made in this study are summarized in the following.

*Assumption 1*: If a user makes a given item recommendation, this subsumes that the user has purchased the item.

This assumption permits us to link our recommendation-based study with business and e-commerce contexts.

*Assumption 2*: Textual inputs are assumed to be inputted by genuine users.

This assumption discards the recently acknowledged importance of *bots* and robots in generating user-like reviews in a manner meant to influence potential customers' perceptions or achieve another malicious goal.

*Assumption 3*: Users are independent, so the influence among the users, if any, is discarded.

This assumption refutes the importance of common knowledge in biasing user's opinions and feedback concerning specific items.

### *1.6. Research questions*

This paper addresses two main research questions:

- *RQ*1: To what extent are sentiment in reviews written by users correlated with the ratings provided by these users, considering different application domains (e.g., digital music and video games)?
- *RQ*2: To what extent can the sentiments of reviews be effectively utilized as a substitute for the ratings of the users in the evaluation process of recommender systems?

We conducted a set of experiments using two popular datasets, i.e., Amazon Digital Music and Amazon Video Games to address these questions. For RQ1, we computed, in different application domains, the correlation between sentiments in reviews, which were extracted using different techniques, and the ratings of the users.

To address RQ2, we first compared the performances of several well-known recommendation algorithms (including YoutubeRanker and DeepFM) when only ratings are used to generate recommendations and when both ratings and sentiments are used for recommendation. We then chose the top-performing algorithms and compared them in two evaluation scenarios: when user ratings are considered the feedback data or when sentiments are considered the feedback data. In the evaluation, we used additional algorithms (including DeepCoNN, D-Attn, and ESCOFILT), and compared their precision-recall area under curve (PR-AUC) metrics.

[2] The availability of Colab Google cloud infrastructure makes it easy to instantaneously access BERT encoding vectors for any textual input.

### 1.7. Contributions

In summary, the main contributions of this paper are as follows:

- We have proposed and evaluated a novel hybrid recommender system that explicitly accounts for the user's textual feedback and sentiment to exploit both in the recommendation process. In particular, BERT embeddings followed by dimensionality reduction (using PCA) was applied to generate a new encoding vector that summarized the content of the textual feedback. This encoding layer is integrated into two state-of-the-art models-YoutubeRanker and DeepFM-to yield a new, deep-learning based, sentiment-aware recommender system.
- We compared the proposed model with the other state-of-the-art sentiment-aware recommender systems (DeepCoNN, D-attn and ESCOFILT) and demonstrated the feasibility, technical soundness and stable performances of the proposed model.
- We showed that ratings are not necessarily correlated with the sentiment of reviews, in different applications domains, and hence, exploiting both types of user feedback can help to obtain a more comprehensive picture of user profiles.
- We demonstrated that sentiment score can be used as an alternative to user rating in the evaluation of recommender systems.
- We evaluated our proposed hybrid recommender system, in two application scenarios (digital music and video games) and shown that utilizing sentiment in the recommendation process can substantially improve the quality of the recommendation in both scenarios.
- We also showed that, when providing ratings and reviews, music listeners are more likely to express different behaviors than are video game players.

### 1.8. Paper organization

The remainder of this paper is organized as follows. Section 2 summarizes the state-of-the-art literature. The details of the method and dataset used in the present study are described in Section 3. Section 4 emphasizes the results of our experiments and provides further discussions. Finally, concluding statements and prospective work are drawn in Section 5.

## 2. Related work

### 2.1. Recommender systems

Common approaches utilized in recommender systems can be grouped into three categories: *Content-Based Filtering (CBF)*, Collaborative Filtering (CF), and *Hybrid* approaches, according to the type of data available as input and the purpose and mechanism of the recommendation [6,7]. Content-based recommender systems use the item content (e.g., item description and category) and user's profile learned from user's interactions, feedback, purchase history, or inputted preferences to generate recommendations. Therefore, a user receives recommendations of items similar to those that have previously been of interest. This assumes the existence of some model of the user's preferences and/or a history of their interactions with the recommender system to build the user profile, along with a relatively good description of the item's attributes.

CF-based recommender systems can be either user-based or item-based [8]. User-based collaborative filtering identifies other users that bear similarity to the given user and recommends what they have liked. Item-based recommender system suggest items that are most similar to items the user has liked in the past. This assumes that users' preferences regarding items can be inferred collaboratively from other users' preferences. Hybrid recommender systems take advantage of any kind of item and user information that can be extracted or inferred from other sources, regardless of the scope of the item description and the user's profile, by combining two or more recommendation strategies to benefit from their complementary advantages, such as by using a mix of CBF and CF.

### 2.2. Sentiment analysis classes

Sentiment-based systems can be categorized into two large classes: lexicon-based and machine learning based [9]. In the former, sentiment classification is performed using a dictionary of terms, such as those found in SentiWordNet and WordNet, that assign a positive or negative orientation to selected words. The document sentiment is then calculated by summing up the orientation of its individual word components. Word sentiment orientation can also be made dependent on the context, as in the case of corpus-based semantic analysis. Machine learning approaches view sentiment as a classification problem and therefore, make use of the rich panorama of machine learning techniques, provided that an appropriate large-scale training corpus is available to be applied to a wide variety of contextual situations [10]. In this case, one distinguishes the state-of-the-art BERT architectures [11].

Conversely, given the fact that a review may constitute a mixture of positive and negative opinions, SA can be conducted at various levels: word, sentence, and document. Word-level analysis bears similarity with the aforementioned lexicon-based approach and accordingly determines the sentiment orientation of an opinion word or phrase [12]. Sentence-level

analysis acknowledges the existence of "subjective" sentences in the document that can be used as a basis for assessing the sentiment polarity. Such subjective sentences can be identified separately through a machine learning approach or reconstructed using the characteristics of its contained tokens.

Finally, the document or review-level analysis determines the dominant polarity by aggregating the individual sentiment orientations of the document's constituent sentences or phrases. Various refinements of such categorization have also been investigated with regards to the structure of the review document. For instance, in a phrase-level analysis [13], the goal was extracting the sentiment polarity of each feature attribute for which a user expresses an opinion. A major issue in focusing on only the document-level analysis is the fact that not all sentences expressing opinion are necessarily subjective, which intuitively entails some gap in the coverage of the overall sentence aggregation. Therefore, sentence-level analysis is often viewed as accurate enough to grasp the fine-grained variation of sentiment in the review. This, in particular, led to the development of aspect-based semantic analysis [14].

### 2.3. Sentiment analysis in recommender systems

The idea of including SA outcomes in recommender systems is not fully new and has been pursued since the emergence of early works on textual SA tools. Hence, the introduction of SA, can be seen as a way to resolve some common deficiencies of recommender systems [15]. A notable example of such deficiencies is the sparsity of data. This is part of the bigger problem called *cold start*, which occurs when a new item or a new user is added to the system without sufficient data available to learn associated profile(s) [16].

Recent works have focused on exploiting advanced techniques like *Deep Learning* to model sentiment data and incorporate it into the recommendation process [17].

Incorporating sentiment in recommender systems can be performed from various perspectives. First, sentiment polarity can be considered a fully independent piece of knowledge that can be combined with other recommender system attributes. For instance, in the case of CF, this is done via some convex combination rule, as in [18]. Second, SA can be used as a tool to validate or even refute the outcome of the recommender system, as in the work of Preethi et al. [19].

Third, SA can be used as a way of eliciting some relevant information regarding either an item description or users' attitudes, affinities, tastes, and preferences. Such preference elicitation can, therefore, contribute to building an enhanced item description or user profile that will boost the performance of collaborative or content-based recommendation. This approach was pursued in [20]. This view is also supported from a topic modeling perspective, in which sentiment and a user's interests are introduced as topic distributions according to the review content, as in [21]. Similarly, from an item description perspective, one should note the growing body of research in aspect semantic analysis, which can boost the identification of relevant item attributes or categorizations. This approach has been used in a number of prior works on recommender systems, such as EFM [22]. Alternatively, approaches such as JMARS [23] have used an integrated framework where each aspect is represented as a distribution over the words of the dictionary, in the same spirit as latent Dirichlet topic modeling.

Fourth, sentiment review can be considered part of contextual information that can therefore be integrated into the newly emerging area of *Context-Aware Recommender Systems (CARS)* wherein many theoretical premises have been put forward as indicated by review papers [24]. Also noteworthy in this context is the proposed ALFM [25], which includes an Aspect-aware Topic Model (ATM) that models each aspect as a multinomial distribution over the same set of $K$ latent topics, each of which is defined as a multinomial distribution over the vocabulary.

Fifth, SA can be used to enhance the explainability of a recommender system. For instance, Zhang et al. [22] proposed an Explicit Factor Model (EFM) to generate an explainable recommendation by extracting explicit product features and user opinions through phrase-level SA on reviews. Similarly, Pugoy and Kao [26] proposed a novel model that extends CF by utilizing review data. The model is called Extractive Summarization-based Collaborative Filtering (ESCOFILT) and it is capable of using BERT and clustering methods to add an explanatory module to a CF recommender system. We can also cast in this category the broad variety of works where sentiment information is considered part of a user's feedback and, therefore, a variety of approaches are used to build relevant actions, supporting explainability through visualization.

In terms of industry applications of these models, we shall mention several notable e-commerce systems where such formalisms have been successfully implemented. Lei et al. [27] proposed combining the three sentiment factors of (i) user sentiment similarity, (ii) interpersonal sentiment influence, and (iii) item reputation similarity in order to improve the rating predictions of recommender systems. Peleja et al. [28] suggested integrating unrated reviews and movie ratings in their proposed recommender system for TV on the web, in which the reviews of unrated movies are exploited by inferring ratings from other reviews included in the exiting user-ratings matrix. Wang et al. [29] proposed a movie recommendation framework based on a hybrid recommendation model and SA on the Spark platform, in which the sentiment-enhanced recommendation framework is obtained by combining CF and CBF methods. Review texts were represented using vector space model (VSM) and term frequency-inverse document frequency (TF-IDF), and sentiment polarity was inferred through lexicons.

In the stock market, the active research in the field of integrating sentiment in recommender systems is acknowledged to help investors in their business decisions to provide a better summative information and investment suggestions [30]. This research advocates the existence of a positive correlation between stock ranking and stock recommendations. This is motivated, for instance, by the fact that Value Lin and Dartboard had the ability to predict future investment profit [31]. Furthermore, a correlation of stock recommendations with market movement and trading volume, known as publicity effect or announcement effect is well acknowledged in business studies [7].

In the tourism industry, there is a growing number of online recommender systems that aim to attract tourists or provide guidelines [32]. Mastrocaronte [33] is an onboard recommender system for drivers that utilize knowledge-based approaches to recommend attractions, restaurants, hotels, and nearby fuel stations, among others.

Table 2 summarizes some key related works regarding the use of SA in recommender systems. Overall, the preceding testifies to the rich panorama of sentiment polarity in recommender systems. Although a detailed exploration of this taxonomy is outside the scope of this paper, this study espouses the view that SA can be considered an independent piece of evidence whose outcome can be incorporated as additional input data for the recommendation process. It is worth noting that, to the best of our knowledge, the majority of the prior research has primarily focused on proposing novel methods and algorithms capable of extracting, processing, and utilizing sentiment data in recommender systems. While this line of research is important, our work goes beyond that and proposes utilizing user sentiment data as alternative user feedback to not only improve the quality of the recommendation but also to obtain the true opinion of the user when interacting with and assessing the recommendation. More particularly, we propose the incorporation of user sentiment in the "evaluation" process of the recommender systems as a potential alternative. While traditional types of user feedback (such as ratings) are still informative, they do not necessarily project the entire opinion of users about the recommendations. To our knowledge, very limited works have attempted this type of integration of sentiment into the recommendation process.

## 3. Methodology

In this section, we provide details of the research methodology employed in this study. A complete overview and flow of the methodology is illustrated in Fig. 1.

### 3.1. Dataset

We used two datasets popular in the relevant research communities: Amazon Digital Music and Amazon Video Games. Both of these datasets are part of the Amazon Review Data [35], an updated version of an Amazon Product dataset. [3] The original version was published in 2014 and contained reviews provided by nearly 142.8 million users together with the product metadata, collected from May 1996 to July 2014. The updated version included more data, including the ratings, texts, views, and votes from nearly 233.1 million users, collected from May 1996 to October 2018. The dataset contains product category, price, brand, color, size, package type, and even image features. The Digital Music subset contains 1,584,082 reviews provided to 456,392 products. The Video Games subset contains 2,565,349 reviews of 84,893 products. Each of these datasets includes a five-core version, datasets where each user and item have at least five reviews. For the digital music dataset, the five-core version includes 169,781 reviews, while the five-core version of the video game dataset contains 497,577 reviews. We initially cleaned these datasets by removing all data points with missing review texts prior to extracting corresponding sentiment polarity.

### 3.2. Overall method and experimental setup

After cleaning and preprocessing, we randomly split the dataset into two subsets with 80% and 20% of the whole dataset for training and testing purposes, respectively. This procedure was conducted based on the reviews, which is a common approach that ensures all users and items have similar odds of appearing in testing and training subsets. We applied a standard NLP pipeline for data cleaning and pre-processing that includes removing posts with missing data, uncommon characters or symbols, links, URLs, lower casing and tokenization. The effect of preprocessing is vital to subsequent steps, such as providing the textual data as input to the BERT model. A highly noisy input would result in an inappropriate encoding vector that would likely deteriorate the overall performance of the model.

Next, sentence embeddings were extracted from review texts using the BERT [11] and Python Transformers library [36]. The BERT-base-uncased model was used in extracting the last layer representations. This has resulted in a vector size of 768 for each review. To reduce the dimensionality in the experiment, we employed the PCA algorithm to reduce this vector's size to 10. The latter is motivated by noticing that, in our case, the first 10 principal components (PCs) accounted for more than 99% of data variance. Hence, we safely assumed that we could represent the data space using the first 10 components.

As shown in Fig. 1, we considered two categories of recommender systems, CF models and hybrid models.

The computed vectors were primarily representative of the reviews, and hence, they were aggregated (i.e., averaged) over the total to build vectors representing items and users. For instance, if item A has received five reviews from users, the vectors of each of those reviews aggregated to construct a vector representing item A. The same process was conducted for the users.

Furthermore, the same BERT-base-uncased model was used to perform SA on review texts. This was achieved by appending a classification head to BERT (i.e, a linear layer on top of the pooled embeddings) for fine-tuning the SA model. The model was fine-tuned using the IMDb Large Movie Dataset [37], which includes over 50,000 textual reviews with bipolar sentiment labels. Afterward, the fine-tuned model was used to extract sentiments for our target dataset (i.e., Amazon reviews dataset),

[3] https://jmcauley.ucsd.edu/data/amazon/

**Table 2**
Summary of related work in sentiment-based recommender systems.

| Reference | Theory or Concept | Recommender Approach |
|---|---|---|
| [26] | Extractive summarization with BERT | Extended collaborative filtering |
| [27] | Using item reputation quantification | Extended matrix factorization |
| [22] | Exploiting phrase-level sentiment | Explicit factor model |
| [18] | Combining different data with sentiment | Hybridized collaborative filtering |
| [19] | Sentiment to validate the recommendation | Deep learning |
| [34] | Item or user profile eliciting | Machine learning |
| [22] | Using aspect sentiment | Data fusion and statistics |
| [21] | Combining sentiment and topics | Probabilistic matrix factorization |

keeping in mind that these extracted sentiments represented the reviews and not the items and users, and hence, the same aggregation process was performed for the corresponding users and items. Finally, these aggregated sentiments were used both as training features for considered recommender models and also as a basis for calculating some of the evaluation metrics later in the experiments.

### 3.3. Recommendation algorithms

In this section, we describe a set of state-of-the-art recommendation algorithms adopted in this work, namely, YoutubeRanker, DeepFM, DeepCoNN, D-Attn, and ESCOFILT. We further considered two popular algorithms as more classical baselines, i.e., Item-based CF and alternating least square (ALS). The choice of these algorithms was motivated by their widespread popularity (e.g., ALS), which means they had already been implemented in e-commerce applications [38], and their excellent performance, due to the use of advanced models (e.g., YoutubeRanker). We either adopted an implementation of these algorithms in popular libraries, such as *LibRec* library [39], or used their public repositories. A brief description of the algorithms are provided in the following.

#### 3.3.1. Item-based collaborative filtering

Traditionally, CF has been one of the most popular recommendation approaches. Item-to-item CF (Item-CF) is a well-known variant of this approach that was proposed in the early stages of recommender systems research [40]. This approach tries to match the rating history of users and suggest items that are similar to those the user has rated highly in the past.

This approach uses a similarity matrix built among all items, using a metric based on the rating vectors of items. We have utilized *cosine* similarity as a popular metric proposed and adopted in many prior studies [1,41]. The similarity is computed using the following equation:

$$Similarity(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A}\cdot\vec{B}}{\|\vec{A}\| \times \|\vec{B}\|} \tag{1}$$

where $A$ and $B$ are the rating vector of two different items.

#### 3.3.2. Alternating least square

Alternating Least Square (ALS) is another popular approach, commonly utilized to generate personalized recommendations. This approach can model users and items with a vector of size $K$ (i.e., the dimensions). The unobserved ratings of the users are then predicted by minimizing the least-squares error of the observed rating [42]. Hence, ALS tries to minimize an objective function by finding an optimal $X$ and $Y$ (i.e., the user and item vectors). The optimization goal is formulated as follows:

$$\min_{X,Y} \sum_{r_{ui} observed} (r_{ui} - x_u^T y_i)^2 + \lambda\left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2\right) \tag{2}$$

where $x_u$ is a $K$ dimensional vector of user $u$ with its features. The dimension of the vector can be fixed with a relatively small number, such as 10. Similar to the user's vector, $y_i$ is a $K$ dimensional vector of each item $i$. ALS aims to predict $r_{ui}$, the rating that is predicted to be given by the user $u$ to the item $i$ exploiting their vectors. In our experiments, we set the number of epochs (i.e., the number of iterations the optimization algorithm performs) to 50. Moreover, the hyper-parameter $\alpha$ is set to 5. This hyper-parameter governs the baseline confidence in the observed ratings.

#### 3.3.3. YoutubeRanker

The YoutubeRanker algorithm is one of the advanced deep learning-based approaches for recommendations that was originally proposed by Google [43]. YoutubeRanker can effectively learn various types of data and build a model on top of them, exploiting neural layers with several depths. These layers can efficiently model the interactions among the many features elicited from the users and items that are eventually utilized in the recommendation process. YoutubeRanker uses a modified logistic regression to learn and model features, enabling the recommendations to be generated based on multiclass

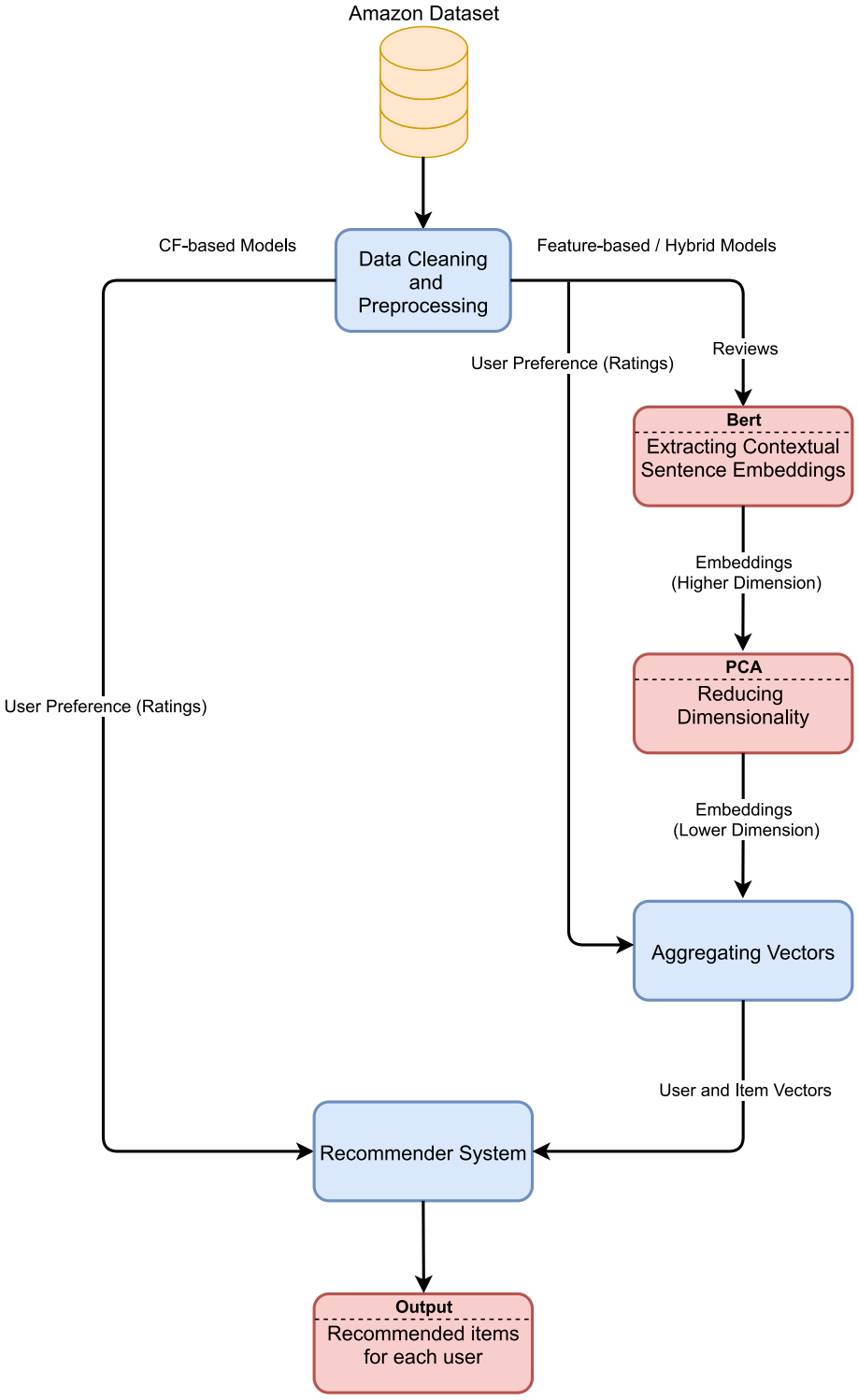



**Fig. 1.** An overview of our approach.

classification. This approach designates a specific item at the time $t$ from among millions of items in a corpus $V$, based on the interactions of the particular user $U$ in the context $C$, as shown in the following formula:

$$P(w_t = i|U, C) = \frac{e^{v_i u}}{\sum_{j \in V} e^{v_j u}}, \quad u, v_j \in \mathbf{R}^{\mathbf{N}} \tag{3}$$

where $u$ and $v_j$ represent a high-dimensional user profile and the item embeddings, respectively. In our experiment, we set the number of epochs to 20 to prevent the over-fitting problem. Furthermore, we set the *learning rate* to 0.001 thus controlling the rate at which the model's parameters are updated, and finally chose 0.4 for the *dropout rate*, which is the proportion of neurons that are dropped during each training to avoid potential over-fitting. The ultimate neural network architecture is [32] and the *embedding size* is 64, meaning the size of the embedding vector extracted from features and fed to the models as input.

### 3.3.4. DeepFM

The DeepFM algorithm is another deep learning based approach that adopts an end-to-end learning mechanism. DeepFM takes advantages of the combination of two different classes of algorithms: deep neural networks (DNNs) and factorization machines (FMs). The benefit of using DeepFM in a recommendation process is that it does not require expertise in feature engineering aside from raw features due to the shared input in distinct parts of this approach [44]. DeepFM adopts the following equation in its core for optimization:

$$\hat{y} = sigmoid(y_{FM} + y_{DNN}) \tag{4}$$

where $y_{FM}$ is the FM module's and $y_{DNN}$ is the DNN module's output, respectively. In our implementation, we set the number of epochs to 10 to prevent the over-fitting problem on the Digital Music dataset. Also, we chose 0.001 as the learning rate of the method and 0.4 for the dropout rate. The neural network architecture is [32,16], and the embedding size is 32. The parameters were fine-tuned enabling the algorithm to achieve its best performance.

### 3.3.5. DeepCoNN

Deep Cooperative Neural Networks (DeepCoNN) [45] use an architecture that more closely resembles our general approach for feature-based models. It consists of two convolutional neural networks, one responsible for capturing user behaviors and the other for modeling items based on reviews creating a user–item pair. These features are extracted through word2vec [46] pre-trained word embedding. Two models have separate weights and architectures, which make use of convolutional and max-pooling layers commonly found in CNNs, but they are jointly used for prediction with a shared, fully-connected layer on top.

In the original paper, authors trained their model on a number of datasets, including the Amazon Review dataset. In our work, we used the code made publicly available by the authors to train it on both of our sub-datasets and later compared the results using the proposed metrics.

### 3.3.6. D-Attn

A similar model used for item recommendation is Dual-Attention (D-Attn)[47], which incorporates a dual approach for capturing user preferences and item properties with both local and global attention mechanisms [48]. Similar to DeepCoNN, separate networks are employed for user and item, with each of them, in turn, having disjoint data-propagation paths for local and global features. The features from local and global attentions for each network are extracted from GloVe embeddings[49], pass through convolutional and pooling layers, get concatenated at the end of the processing stream, and finally, go through a fully connected layer. Lastly, the output of the model is calculated by the dot product of user and item vectors. Similar to our work, the authors in [47] trained and tested their model on a subset of the Amazon Review dataset.

### 3.3.7. ESCOFILT

Extractive Summarization-based Collaborative Filtering (ESCOFILT)[26] is one of the more recent contributions using NLP techniques to further enhance recommendations and is considered one of the state-of-the-art approaches at this time. The model employs BERT embedding and use multilayer perceptrons (MLP) to respectively learn sentence embeddings, extractive representation explanations and user–item interactions from user reviews. Their work somewhat resembles ours in the sense that they also exploit an attention-based model to obtain rich features from the review texts. However, unlike our approach, which additionally extracts sentiments from each review as an explicit feature, they instead implicitly use reviews by building representations for users and items based on text, called, extractive summaries. Such representations contribute to improving the prediction accuracy of CF. The extractive summaries were obtained using BERT Extractive Summarizer [50] with a $BERT_{large}$ model. Then, K-means clustering was used to aggregate multiple sentence embeddings and combine them into a single item or user representation. Lastly, these representations were fed into an MLP to predict the user ratings of items.

### 3.3.8. Naive ideal recommender

Finally, for the sake of comparison, we considered the best possible recommendation as an ultimate baseline. *Ideal* recommender simply selects and recommends to a target user those items for which the user has provided the highest ratings

(i.e., 4 to 5 out of 5) in the past. Then, if there is still room for more recommendations, the items that received (by overall average) the highest ratings from all users are recommended.

*3.3.9. Metrics*

We employed the following two metrics to analyze the performance of the models.

*Global Hit Rate*: First, the items that had high (average) sentiments and received at least 10 reviews from users were selected as global top items. The global top items were not selected in a personalized manner (i.e., per user). Instead, they were selected across the entire dataset (i.e., globally). We considered that an average sentiment computed from all reviews of an item and was equal to 1.0 to be high-level sentiment. Then, *hit rate* [%] for recommendations for each user was computed according to the following formula:

$$GlobalHitRate = 100 \times \frac{number\ \ of\ \ hits}{number\ \ of\ \ recommended\ \ items} \tag{5}$$

where *hit* is a global top item that appeared in the recommendation list of a user. The hit rate scores computed for each user were then averaged over all users to compute the *global* hit rate score. It is worth noting that, in the Amazon Digital Music dataset, 3,870 items were identified as global top items, whereas for the Amazon Video Games dataset, only 1,924 items were identified as global top items.

*Precision:* First, for each user, the items that received reviews with high sentiment levels were considered personal top items for that specific user. Hence, the personal top items were selected in a personalized manner (i.e., per user), meaning that the top items for each user is different from the top items of the other users. Again, we considered a sentiment score equal to 1.0 as a high sentiment. The formula for computing precision is as follows:

$$Precision = 100 \times \frac{number\ \ of\ \ recommended\ \ personal\ \ top\ \ items}{number\ \ of\ \ recommended\ \ items} \tag{6}$$

It is worth noting that the personal top items considered for the precision metric were items with high sentiment levels that were extracted from the reviews of the target user. For the hit rate metric, however, the global top items were items with high (average) sentiment levels that were from all user reviews.

It is worth noting that the personal top items considered for the precision metric were items with high sentiment levels that were extracted from the reviews of the target user. For the hit rate metric, however, the global top items were items with high (average) sentiment levels that were from all user reviews.

For the experiments, we executed the implementation on a Ubuntu 16.04 machine with eight Intel Xeon E5620 CPUs clocked at 2.4 GHz on each of the cores and 16 GB available RAM memory. For training models, a Tesla T4 instance was used. The results obtained from the experiments are discussed in the next section.

## 4. Results

We conducted a set of experiments to address the formulated research questions. In this section, the observed results for each of the experiments are described.

### *4.1. Experiment A: Exploratory analysis*

As a preliminary study, we analyzed the considered datasets to obtain a better understanding of the distribution of data, including the extracted sentiments and the ratings. The results are plotted in Fig. 2 and 3. It can be seen that, there is a clear difference between the two datasets. While the distribution of the video items in the Video Games dataset resembles a typical normal distribution (see Fig. 3), the distribution of music items in Digital Music dataset reflects a half normal distribution (see Fig. 2). In the Digital Music dataset, as the values of user ratings increase (Fig. 2–up) or the values of sentiment increase (Fig. 2–bottom), the number of music items grows almost exponentially. For the sentiment, the peak of the distribution appears to be at 1 (out of 1), and for the rating, the peak is observed around the value of 5 (out of 5). This basically indicates a higher probability to observe items with more positive ratings and more positive sentiment.

In the Video Games dataset, as the value of rating increase (Fig. 3–up) or the value of sentiment increases (Fig. 3–bottom), the number of items reaches a peak and then goes down to a lower value. For the sentiment, the peak appears at around 0.8, and for the rating, the peak is observed around the value of 4.5. This observation may indicate that users express different behaviours when consuming music items than when consuming video game items. According to the results, music listeners commonly provide feedback on a narrow range of the *best* music that perfectly matches their musical taste. In contrast, video game players provide feedback on a wider range of game items, including not only the *best* games but also the fairly *good* games.

In the follow-up experiment, we analyzed the reviews of the top items in each dataset, which were evaluated as *best* from the users' points of view. These items can be found in two different categories (a): items with the highest average ratings, or (b): items with the highest sentiment levels in their reviews. Through this analysis, we checked the reviews of the top items in each of the categories and identified a set of unique words (terms) that appeared with the greatest frequency in the

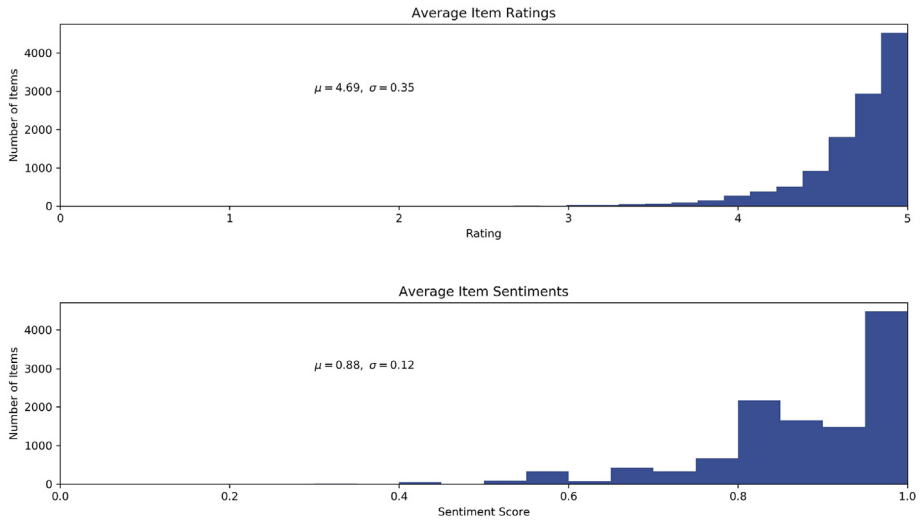



**Fig. 2.** Distribution of the average sentiment of music items in Amazon **Digital Music** dataset.

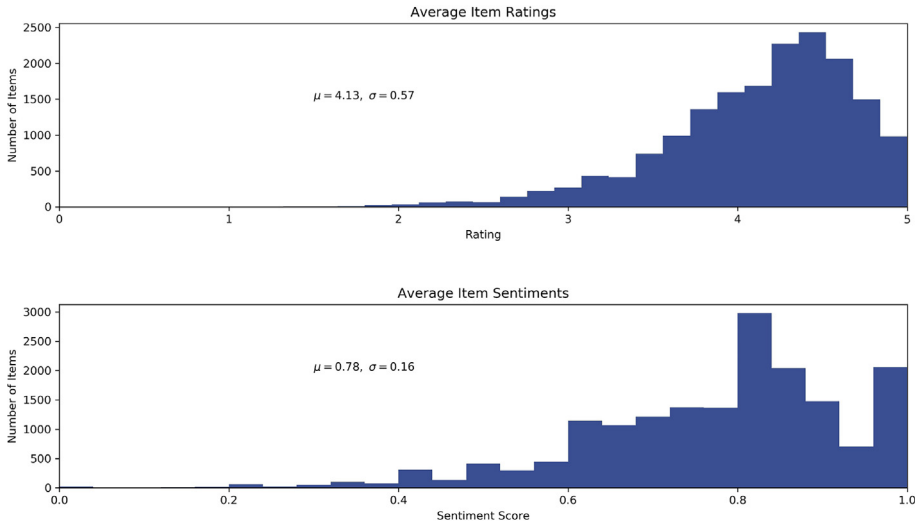
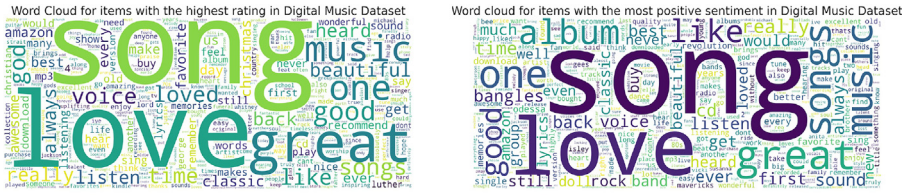



**Fig. 3.** Distribution of the average sentiment of game items in Amazon **Video Games** dataset.

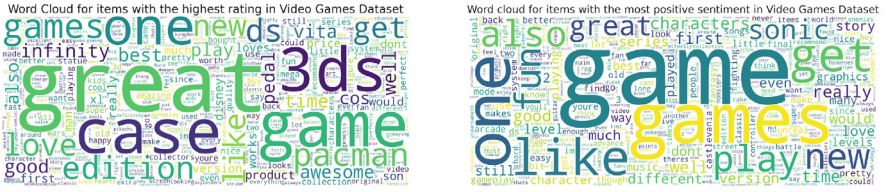
reviews. We were interested to determine whether the top items in the categories differed with regard to the words used in the reviews. We considered the *tf-idf* metric and computed it for all of the words and then generated a WordCloud figure for the top items in both datasets. The results are shown in Fig. 4 and 5.

As can be seen, in the Digital Music dataset, the WordCloud of the top items with the highest ratings (Fig. 4–left) does not substantially differ from the WordCloud of the top items with the highest levels of sentiments (Fig. 4–right). However, in the Video Games dataset, a clear difference can be observed between the WordCloud of the top items with the highest ratings (Fig. 5–left) and the top items with the highest sentiment levels (Fig. 5-right). This may indicate that, in the music domain, listeners were prone using a more similar set of words when providing written feedback (reviews) of the items, whether or not the items are highly rated or received positive sentiments. In contrast, in the video game domain, players were likely to use a diverse set of words in their written feedback for highly rated items in comparison to the items that receive positive sentiment. Both of these observations are interesting and reflect the differences in feedback signals provided by users in different application domains.




(a) Reviews for top items (w.r.t. rating) (b) Reviews for top items (w.r.t. sentiment)

**Fig. 4.** WordCloud of reviews for **digital music** items.




(a) Reviews for top items (w.r.t. rating) (b) Reviews for top items (w.r.t. sentiment)

**Fig. 5.** WordCloud of reviews for **video game** items.

### *4.2. Experiment B: Correlation analysis*

In this experiment, we investigated potential correlations among items with regard to their sentiment levels and the ratings received from users, considering both datasets. For that, we computed the *Pearson correlation* between the average of the ratings an item received from users and the average of the sentiments extracted from the reviews of that item. According to our analysis, a high overall correlation was observed in the Video Game dataset than in the Digital Music dataset. While the average Pearson correlation for music items was 0.284 (p-value $< 0.01$), this value was 0.648 (p-value $< 0.01$) for game items. This may mean that users are more likely to write reviews that are aligned with their ratings for video games than they are for music items.

In order to better illustrate this observation, we plotted the average values for the ratings and sentiments for the items in each dataset. Fig. 6 and 7 show the obtained results for both the Amazon Digital Music and Amazon Video Games datasets. First, as can be seen, there is a concentration of data around the higher ratings in both datasets. This was an expected observation because the users typically tend to provide feedback (either ratings or reviews) on items that they have consumed and found interesting. Second, there is a clearer linear correlation in the Video Games dataset than in the Digital Music dataset. Again, this indicates the difference between the digital music application domain and the video game domain when it comes to the *rating* and *reviewing* behaviors of the users. Accordingly, while music listeners are less consistent in providing reviews and ratings for music items, game players represent a higher level of consistency in reviewing and rating game items.

For the sake of comparison, we also explored whether the correlation results are dependent on a particular SA approach. For this purpose, we considered a set of commonly employed SA techniques, including *SentiStrength*, *TextBlob*, and *Vader*. The results (shown in Table 3) confirm our reported observation (i.e, relatively low correlation scores in the case of the Amazon Music Dataset, and moderate correlation scores in the case of the Amazon Games Dataset). Moreover, the observed results revealed no substantial differences among the considered techniques in terms of the correlation scores. The results also showed that BERT-sentiment yielded the highest correlation score in Amazon Games dataset while TextBlob achieved the highest correlation score in Amazon Music dataset.

Furthermore, we investigated whether or not the top items with the highest ratings were different from the items with the highest sentiment levels in their reviews. For that, we selected the topmost items in both of the datasets in terms of average ratings as well as average sentiments in their reviews. For the sake of comparison, we also considered the items with the largest number of reviews. The three lists of top items are presented in Table 4 and 5 for the Digital Music and Video Games datasets, respectively.

As can be seen, these lists are distinctively different, indicating that users may provide higher ratings for items that are different from the items with the most positive reviews. As a conclusion of the results of this experiment, the ratings and sentiments of the reviews were not necessarily correlated, and, can therefore be indicative of different sides of user profiles and potentially represent different perspectives of user preferences.

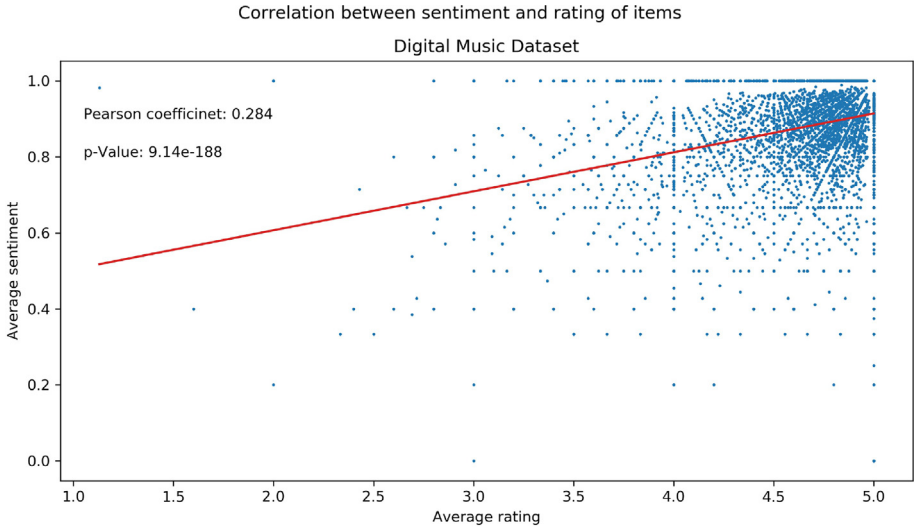



**Fig. 6.** Correlation between average sentiment of music items and their average ratings provided by users based on Amazon Digital Music dataset.

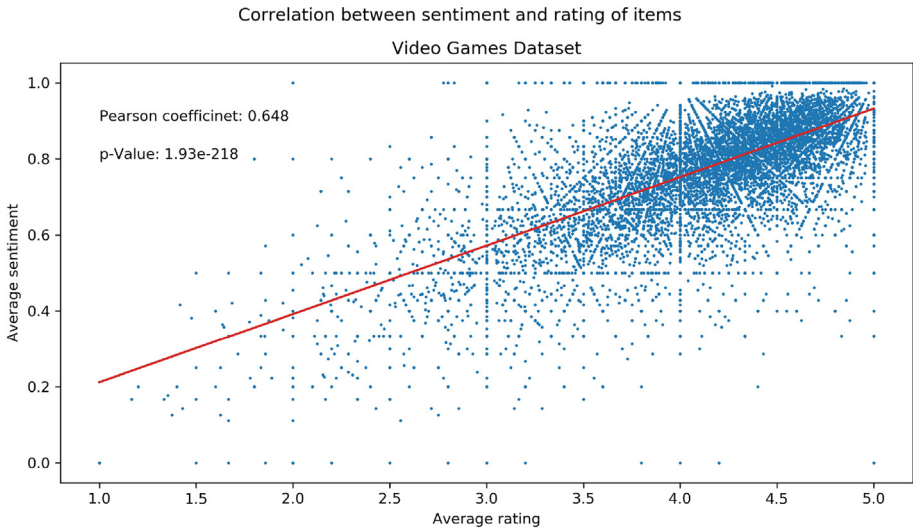



**Fig. 7.** Correlation of the average sentiment of games with average ratings provided by users based on Amazon Video Games dataset.

**Table 3**
Correlation between the sentiment of reviews and their respective ratings using different SA methods.

| Method | Amazon Music Dataset | Amazon Games Dataset |
|---|---|---|
| **BERT Sentiment (our method)** | 0.284 | 0.648 |
| **SentiStrength** | 0.299 | 0.551 |
| **TextBlob** | 0.377 | 0.530 |
| **Vader** | 0.117 | 0.476 |

### *4.3. Experiment C: Sentiment analysis*

In the third experiment, we analyzed the performance of different recommender systems in order to understand which one encompassed more items with positive sentiments in their reviews. Accordingly, we checked the output of different recommender systems, including the hybrid ones, and computed the percentage of recommended items with positive sentiments. To keep the results more reliable, we excluded items that had received less than 10 reviews from users. Again,

**Table 4**
Comparison of Top Items in **Amazon Digital Music** Dataset.

| | Top Ratings | Top #Reviews | Top Sentiment |
|---|---|---|---|
| 1 | Stranglehold | Happy | Ah! Leah! |
| 2 | It's the Most Wonderful Time of the Year | Blurred Lines [feat. T.I.] | No One |
| 3 | Tush | Roar | Love Stuff |
| 4 | Where Were You (When the World Stopped Turning) | I Can Only Imagine | Hello It's Me |
| 5 | Try a Little Tenderness | Hello It's Me | Brand New Key |

**Table 5**
Comparison of Top Items in **Amazon Video Games** Dataset.

| | Top Ratings | Top #Reviews | Top Sentiment |
|---|---|---|---|
| 1 | New Super Mario Bros 2 | Diablo III & Zack | Wiki Quest for Barbaros' Treasure |
| 2 | Mamas & Papas Soft Toy | God of War III | Trauma Center: Under the Knife 2 |
| 3 | Advance Wars: Dual Strike | Mario Kart | Rock Band |
| 4 | Disney Infinity 3.0 Edition | The Last of Us Remastered | Rune Factory 2: A Fantasy Harvest Moon |
| 5 | Wario amiibo | StarCraft II: Wings of Liberty | Advance Wars: Days of Ruin |

when measuring the *hit rate*, we considered a recommended item a *hit* if the extracted sentiment was 1, and considered it a *miss* if the sentiment was less. We note that although the minimum value of the hit rate metric is 0 and the maximum value is 1, we mapped the values to the range of [0–100].

Fig. 8 presents the results of the experiment, showing that the scores computed in the Digital Music dataset were larger than those for the Video Games dataset.

This may reflect that recommender systems in the digital music domain may recommend items that received more positive reviews in comparison to the recommender systems in the video games domain. In addition, as expected, the ideal recommender system (recommending items with ratings 4 or 5), achieved the highest hit rate in both datasets, thus recommending the most items with positive sentiments. In the Video Games dataset, the ideal recommender obtained the hit rate value of 11.33, and in the Digital Music dataset, the recommender obtained the value of 15.22. This means that, on average, fewer than 16% of the top music items (in terms of ratings) have received reviews with positive sentiment. This value was even lower for video games at less than 12%. This was a rather unexpected result and potentially indicates an inconsistency between the ratings and reviews. Comparing the non-ideal recommender systems, the best results were achieved by the hybrid recommendation techniques, i.e., YoutubeRanker and DeepFM. In the Video Games dataset, YoutubeRanker technique had obtained value of 4.01, while DeepFM obtained value of 3.75. In the Digital Music dataset, both of the recommendation techniques performed well, obtaining hit rate values of 14.32 and 9.68, respectively. The worst results were seen with recommendations based on ALS, obtaining the hit rate values of 2.65 in the Digital Music dataset and 0.77 in the Video Games dataset.

### 4.4. Experiment D: Hybrid recommendation with sentiment

In the previous experiment, we measured the impact of incorporating sentiment data on the quality of recommendations, measured in terms of Precision@10. For that, we compared the quality of recommendations with different hybrid recommender systems using the sentiment of reviews.

The results are depicted in Fig. 9. It can be seen that the quality of recommendations significantly improved with both the Video Games and Digital Music datasets. In the Video Games dataset, the precision of YoutubeRanker model was 0.015 when the sentiment score was not incorporated in the recommendation. When sentiment was incorporated, this value doubled and reached 0.039 when sentiment has been used (≈160% improvement). For the DeepFM, the precision score without sentiment was 0.016, and the precision was boosted to the value of 0.033 by utilizing the sentiment data (≈106% improvement).

With the Digital Music dataset, the precision values of the YoutubeRanker model without and with sentiment were 0.022 and 0.048, respectively (≈118% improvement). With the DeepFM, the precision values without and with sentiment data were 0.003 and 0.010, respectively (≈233% improvement).

### 4.5. Experiment E: Extended comparison

The previous experiment clearly showed the effectiveness of incorporating sentiment data into the recommendation process. Accordingly, the quality of two recommender techniques (i.e., DeepFM and YoutubeRanker) performing the best in a *hybrid* setting, was evaluated with and without using sentiment data. We further extended this experiment and compared these two techniques with three state-of-the-art sentiment-aware algorithms capable of utilizing different forms of input data and incorporating them into the recommendation process (i.e., DeepCoNN [45], D-attn [47], and, ESCOFILTS [26]).

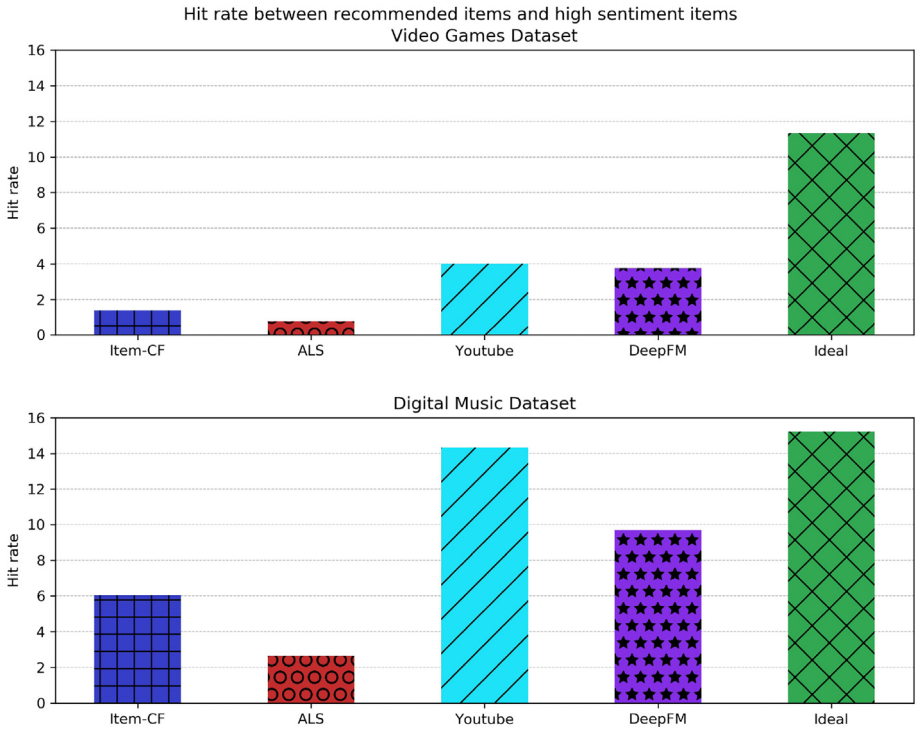



**Fig. 8.** Comparison of different recommender systems in terms of the *hit rate* metric, computed with respect to the sentiments scores.

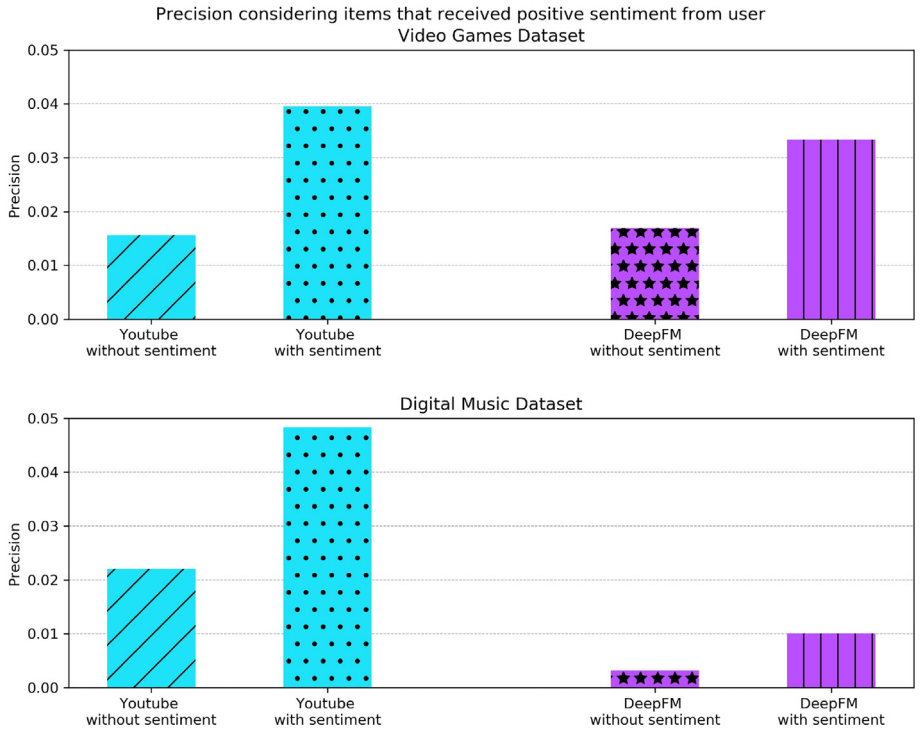



**Fig. 9.** Comparison of different hybrid recommender systems with and without usage of the sentiment information withing item reviews.

For the sake of a comprehensive evaluation, we considered the PR-AUC metric in this experiment. To compute this metric, Precision and Recall scores were first measured for different thresholds, and then the area under this curve was computed.

The higher the PR-AUC score, the better the quality of the recommendation. PR-AUC scores were computed in two scenarios: (i) The first considered the ratings from users as representative of true feedback from users, and (ii) the second considered the sentiment of reviews from users as representative of the true feedback of users.

The results are presented in Table 6. Please note that DeepFM-S refers to a version of the DeepFM model that incorporates sentiment scores as input data. Similarly, YoutubeRanker-S refers to the YoutubeRanker model that incorporates sentiment scores as input data.

Comparing the results, the overall best performance was observed for the YoutubeRanker-S and ESCOFILT recommender algorithms. More particularly, in scenario (i), where user ratings were considered for evaluation, YoutubeRanker-S algorithm outperformed all the other algorithms in terms of PR-AUC. scenario (ii), where sentiments of the user reviews were considered for evaluation, the ESCOFILT algorithm outperformed all the other algorithms. All of the differences are statistically significant with $p\text{-value} < 0.01$. The only exception was the marginal difference between the results of DeepCoNN and YoutubeRanker-S (in Video Games dataset) with $p\text{-value} < 0.1$. As the table shows, we obtained larger PR-AUC scores in the scenario where ratings were used as feedback, considering both of the datasets. The difference was slightly higher in the Digital Music dataset than in the Video Games dataset.

In summary, our results showed that using different feedback data (e.g., sentiments of user reviews or user ratings) in the evaluation process can result in differences in the measured performance of recommender algorithms. Such differences may depend on the dataset collected within different application domains. While this is an interesting observation, further analysis would be beneficial for better investigating the level, cause, and effect of such differences.

Finally, we made a comparison of different models in terms of computation time (in milliseconds) for each model to predict on a 64 mini-batch sample. According to our measurement, D-Attn, ESCOFILT, and ALS achieves the best computation times with scores of 0.12 ms, 1.31 ms, and 4.21 ms, respectively. The observed computation times for the other models were 24.43 ms, 48.49 ms, 56.48 ms, and 62.99 ms for Item-CF, DeepCoNN, DeepFM-S, and YoutubeRanker-S, respectively.

### 4.6. Discussion

This section provides a discussion of the observed results in addressing the formulated research questions:

1. *The sentiments of the reviews provided by users are not necessarily correlated with the ratings from users.* While there might be a moderate correlation observed in the video games domain, our observation has shown almost no correlation in the digital music domain. This result may reflect differences in the types of data collected and used by recommender systems operating in different application domains. This indicates the importance of considering different sources of user feedback in different forms in order to obtain a better picture of user profiles in recommender systems. These observations address RQ1.
2. *Sentiment of the reviews can be considered a substitute for other types of user feedback (e.g., ratings) for evaluating recommendation quality.* We considered different metrics to evaluate the proposed recommendation technique in two different scenarios: (i) When the ratings of users are indicative of their true preferences, and (ii) When the sentiments obtained from reviews are indicative of users' true preferences. We would argue that our adopted methodology can better reflect reality than traditional forms of evaluation that are solely based on user ratings as ground truth. These observations address RQ2.
3. *Hybrid recommender systems that incorporate sentiment in the recommendation process yield superior performance than recommender systems without the sentiment component.* We considered different recommender algorithms and compared them in two different scenarios: one with the usage of sentiment and one without the usage of sentiment in the recommendation process. Our results showed that the quality of recommendations was substantially improved when the sentiment of the review is taken into account with the ratings of users. The observed results were consistent in both of the considered application domains, digital music, and video games. These observations (and the following ones) are secondary results that we report in addition to the previously described primary results.
4. *Different recommender algorithms are different in their recommendation output in terms of the sentiment of reviews.* We assessed the performances of a wide range of recommender algorithms, including YoutubeRanker, DeepFM, DeepCoNN, D-Attn, ESCOFILT, Item-CF, and ALS, in terms of various metrics, i.e., hit Rate, Precision, and PR-AUC, computed according to the user ratings or sentiment scores. The results showed that using the sentiment of users in a hybrid recommendation process based on YoutubeRanker algorithm (in the Digital Music dataset), and, ESCOFILT algorithm (in the Video Games dataset) achieved superior performances in comparison to the other algorithms. This indicates that hybrid recommender systems using these algorithms result in recommendations of items that receive not only higher ratings from users, but also positive sentiment extracted from their reviews.
5. *There is inconsistency in the observed behaviors of users when providing feedback (e.g., ratings and reviews) in different application domains.* Music listeners mostly provide feedback to a narrow range of the best music while game players provide feedback to a wider range of games, including ones that are good enough but not necessarily the best. A potential explanation for this phenomenon may be related to the different costs of item consumption (e.g., time of consumption) in these two application domains. While assessing a music item can be quick and relatively cheap, assessing the quality of a video game can be time-consuming and requires more effort and money. This may motivate the video game players more than music listeners to provide their feedback on games, particularly when consumed games did not present the expected quality or did not match the player's personal taste.

**Table 6**
Comparing the performances of recommendation algorithms in terms of area under the precision-recall curve (PR-AUC). The PR-AUC were computed under different scenarios, such as when ratings are considered the true preferences of users and when the sentiments of reviews are considered the true preferences of users. The differences in the presented values were significant with $p\text{-value} < 0.01$. The only exception was the marginal difference between DeepCoNN and YoutubeRanker-S (Video Games dataset) with $p\text{-value} < 0.1$.

| Recommender | PR-AUC$_{rating}$ | | PR-AUC$_{sentiment}$ | |
|---|---|---|---|---|
| | **Music** | **Video Games** | **Music** | **Video Games** |
| **DeepFM-S** | 0.984 | 0.910 | 0.918 | 0.862 |
| **YoutubeRanker-S** | **0.986** | **0.920** | 0.924 | 0.873 |
| **DeepCoNN** | 0.967 | 0.846 | 0.901 | 0.816 |
| **D-attn** | 0.933 | 0.773 | 0.916 | 0.794 |
| **ESCOFILT** | 0.974 | 0.912 | **0.942** | **0.885** |

Overall, the reported observations are promising as they represent the potential power of the sentiment data extracted from the user reviews and incorporating them into the hybrid recommender systems.

## 5. Conclusion

Traditional recommender systems have primarily relied on using *ratings* as the main feedback received from users when learning preferences and building profiles. While ratings can be a valuable source of data, it may not necessarily draw a full picture of the actual preferences of users and their specific tastes.

In this paper, we proposed a novel form of *hybrid* recommender system capable of going beyond the rating-based feedback and learning from the sentiment scores computed based on user reviews for the items. Sentiment of the reviews have been shown to be important signals representing the users' actual feedback on the assessed items. In order to evaluate the effectiveness of the proposed recommender system, we conducted a set of experiments utilizing two well-known datasets, the Amazon Video Games, and Amazon Digital Music dataset.

Our experiments showed that the user ratings for items do not necessarily correlate with the sentiment of reviews written for the same items. Moreover, recommender systems may generate different recommendations in terms of sentiment scores and can thus perform differently when it comes to sentiment in reviews. In addition to that, incorporation of sentiment data within the hybrid recommendation process can substantially increase the quality of the recommendation and hence offer benefits to users.

In the future, we plan to extend this work by utilizing different datasets with a larger number of users and items. We also plan to assess the accuracy of different methods for SA and measure the impact of each on the quality of the recommendation. Finally, we plan to design a live demo of the proposed hybrid recommender system utilizing the sentiment data and evaluate its quality in an online setup with real users. This can help to better analyze the quality of the proposed recommendation technique in a different scenario and application domain.

## CRediT authorship contribution statement

**Mehdi Elahi:** Supervision, Formal analysis, Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Danial Khosh Kholgh:** Writing - review & editing, Software, Investigation, Validation, Visualization. **Mohammad Sina Kiarostami:** Conceptualization, Resources, Investigation, Writing - review & editing. **Mourad Oussalah:** Supervision, Formal analysis, Conceptualization, Methodology, Writing - original draft, Writing - review & editing. **Sorush Saghari:** Software, Investigation, Validation.

## Data availability

Data will be made available on request.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

and Reference heading should be left justified, bold, with the first letter capitalized but have no numbers. Text below continues as normal.

## References


[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE transactions on knowledge and data engineering 17 (6) (2005) 734–749.
[2] M. Elahi, D.K. Kholgh, M.S. Kiarostami, S. Saghari, S.P. Rad, M. Tkalčič, Investigating the impact of recommender systems on user-based and item-based popularity bias, Information Processing & Management 58 (5) (2021) 102655.
[3] J. Ben Schafer, J. Konstan, J. Ried, E-commerce recommendation applications, Data Mining and Knowledge Discovery (2001) 115–153.
[4] J. Chen, N. Song, Y. Su, S. Zhao, Y. Zhang, Learning user sentiment orientation in social networks for sentiment analysis, Information Sciences 616 (2022) 526–538.
[5] F. Tang, L. Fu, B. Yao, W. Xu, Aspect based fine-grained sentiment analysis for online reviews, Information Sciences 488 (2019) 190–204.
[6] Y. Pan, Y. Huo, J. Tang, Y. Zeng, B. Chen, Exploiting relational tag expansion for dynamic user profile in a tag-aware ranking recommender system, Information Sciences 545 (2021) 448–464.
[7] A. Beheshti, S. Ghodratnama, M. Elahi, H. Farhood, Social Data Analytics, CRC Press, 2022.
[8] R. Kuo, C.-K. Chen, S.-H. Keng, Application of hybrid metaheuristic with perturbation-based k-nearest neighbors algorithm and densest imputation to collaborative filtering in recommender systems, Information Sciences 575 (2021) 90–115.
[9] B. Bhavitha, A.P. Rodrigues, Chiplunkar, Comparative study of machine learning techniques in sentimental analysis, in: International conference on inventive communication and computational technologies, 2017, pp. 216–221.
[10] Y. Rao, Q. Li, X. Mao, L. Wenyin, Sentiment topic models for social emotion mining, Information Sciences 266 (2014) 90–100.
[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
[12] J. Kamps, M. Marx, R.J. Mokken, M. De Rijke, Using wordnet to measure semantic orientation of adjectives, in: Proceedings of 4th International Conference on Language Resources and Evaluation, 2004, pp. 1115–1118.
[13] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in: Proceedings of 12th International World Wide Web Conference, 2003, pp. 519–528.
[14] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, I. Androutsopoulos, Semeval-2015 task 12: Aspect based sentiment analysis, in: Proceedings of the 9th international workshop on semantic evaluation (SemEval, 2015, pp. 486–495.
[15] J. Herce-Zelaya, C. Porcel, J. Bernabé-Moreno, A. Tejeda-Lorente, E. Herrera-Viedma, New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests, Information Sciences 536 (2020) 156–170.
[16] F.B. Moghaddam, M. Elahi, Cold start solutions for recommendation systems (2019).
[17] K. Zhang, H. Qian, Q. Liu, Z. Zhang, J. Zhou, J. Ma, E. Chen, Sifn, A sentiment-aware interactive fusion network for review-based item recommendation, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3627–3631.
[18] C.N. Dang, M.N. Moreno-García, F.D.l. Prieta, An approach to integrating sentiment analysis into recommender systems, Sensors 21 (16) (2021) 5666.
[19] G. Preethi, P.V. Krishna, M.S. Obaidat, V. Saritha, S. Yenduri, Application of deep learning to sentiment analysis for recommender system on cloud, in: 2017 International conference on computer, information and telecommunication systems (CITS), IEEE, 2017, pp. 93–97.
[20] T.P. Sahu, S. Ahuja, Sentiment analysis of movie reviews: A study on feature selection & classification algorithms, in: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, 2016, pp. 1–6.
[21] X. Ma, X. Lei, G. Zhao, X. Qian, Rating prediction by exploring user's preference and sentiment, Multimedia Tools and Applications 77 (6) (2018) 6425–6444.
[22] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, S. Ma, 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, pp. 83–92.
[23] Q. Diao, M. Qiu, C.-Y. Wu, A.J. Smola, J. Jiang, C. Wang, 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, 2014, pp. 193–202.
[24] K. Haruna, M. Akmar Ismail, S. Suhendroyono, D. Damiasih, A. Pierewan, e. a. Chiroma H., Context aware recommender system: A review of recent developmental process and future research direction, Applied Sciences 7 (2) (2017) 1211–1221.
[25] Z. Cheng, Y. Ding, L. Zhu, M. Kankanhalli, Aspect-aware latent factor model: Rating prediction with ratings and reviews, in: Proceedings of the 2018 world wide web conference, 2018, pp. 639–648.
[26] R.A. Pugoy, H.-Y. Kao, Unsupervised extractive summarization-based representations for accurate and explainable collaborative filtering, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 Long Papers), Association for Computational Linguistics, 2021, pp. 2981–2990, https://doi.org/10.18653/v1/2021.acl-long.232.
[27] X. Lei, X. Qian, G. Zhao, Rating prediction based on social sentiment from textual reviews, IEEE Transactions on Multimedia 18 (9) (2016) 1910–1921.
[28] F. Peleja, P. Dias, J. Martins F.and Magalhães, A recommender system for the tv on the web: integrating unrated reviews and movie ratings, Multimedia Systems 19(6) (2013) 543–558.
[29] Y. Wang, M. Wang, W. Xu, A sentiment-enhanced hybrid recommender system for movie recommendation: a big data analytics framework, Wireless Communications and Mobile Computing (2018).
[30] P. Cremonesi, C. Francalanci, A. Poli, R. Pagano, L. Mazzoni, A. Maggioni, M. Elahi, Social network based short-term stock trading system, arXiv preprint arXiv:1801.05295.
[31] J. Treynor, F. Black, How to use security analysis to improve portfolio selection, Journal of Business 46 (1) (1973) 66–86.
[32] M. Braunhofer, M. Elahi, F. Ricci, Techniques for cold-starting context-aware mobile recommender systems for tourism, Intelligenza Artificiale 8 (2) (2014) 129–143.
[33] L. Console, I. Torre, I. Lombardi, S. Gioria, V. Surano, Personalized and adaptive services on board a car: an application for tourist information, Journal of Intelligent Information Systems 21 (3) (2003) 249–284.
[34] B. Agarwal, N. Mittal, P. Bansal, S. Garg, Sentiment analysis using common-sense and context information, Computational intelligence and neuroscience (2015).
[35] J. Ni, J. Li, J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 188–197.
[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45.
[37] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 142–150. http://www.aclweb.org/anthology/P11-1015.
[38] A. Klimashevskaia, M. Elahi, D. Jannach, C. Trattner, L. Skjærven, Mitigating popularity bias in recommendation: Potential and limits of calibration approaches, in: International Workshop on Algorithmic Bias in Search and Recommendation, Springer, 2022, pp. 82–90.
[39] G. Guo, J. Zhang, Z. Sun, N. Yorke-Smith, Librec: A java library for recommender systems., in: UMAP Workshops, Vol. 4, Citeseer, 2015, pp. 38–45.
[40] G. Linden, B. Smith, J. York, Amazon. com recommendations: Item-to-item collaborative filtering, IEEE Internet computing 7 (1) (2003) 76–80.
[41] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, arXiv preprint arXiv:1301.7363.

[42] G. Takács, D. Tikk, Alternating least squares for personalized ranking, in: Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 83–90. doi:10.1145/2365952.2365972.
[43] P. Covington, J. Adams, E. Sargin, Deep neural networks for youtube recommendations, in: Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 191–198.
[44] H. Guo, R. Tang, Y. Ye, Z. Li, X. He, Deepfm: A factorization-machine based neural network for ctr prediction, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press, 2017, pp. 1725–1731.
[45] L. Zheng, V. Noroozi, P.S. Yu, 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, pp. 425–434.
[46] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.
[47] S. Seo, J. Huang, H. Yang, Y. Liu, Interpretable convolutional neural networks with dual local and global attention for review rating prediction, in: Proceedings of the eleventh ACM conference on recommender systems, 2017, pp. 297–305.
[48] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.
[49] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162. https://aclanthology.org/D14-1162.
[50] D. Miller, Leveraging bert for extractive text summarization on lectures, arXiv preprint arXiv:1906.04165.