

Leave No User Behind

Towards Improving the Utility of Recommender Systems for Non-mainstream Users

Li, Roger Zhe; Urbano, Julián; Hanjalic, A.

DOI

[10.1145/3437963.3441769](https://doi.org/10.1145/3437963.3441769)

Publication date

2021

Document Version

Final published version

Published in

WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining

Citation (APA)

Li, R. Z., Urbano, J., & Hanjalic, A. (2021). Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users. In *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 103-111). ACM.
<https://doi.org/10.1145/3437963.3441769>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users

Roger Zhe Li
Delft University of Technology
Delft, The Netherlands
Z.Li-9@tudelft.nl

Julián Urbano
Delft University of Technology
Delft, The Netherlands
J.Urbano@tudelft.nl

Alan Hanjalic
Delft University of Technology
Delft, The Netherlands
A.Hanjalic@tudelft.nl

ABSTRACT

In a collaborative-filtering recommendation scenario, biases in the data will likely propagate in the learned recommendations. In this paper we focus on the so-called mainstream bias: the tendency of a recommender system to provide better recommendations to users who have a mainstream taste, as opposed to non-mainstream users. We propose NAE CF, a conceptually simple but effective idea to address this bias. The idea consists of adding an autoencoder (AE) layer when learning user and item representations with text-based Convolutional Neural Networks. The AEs, one for the users and one for the items, serve as adversaries to the process of minimizing the rating prediction error when learning how to recommend. They enforce that the specific unique properties of all users and items are sufficiently well incorporated and preserved in the learned representations. These representations, extracted as the bottlenecks of the corresponding AEs, are expected to be less biased towards mainstream users, and to provide more balanced recommendation utility across all users. Our experimental results confirm these expectations, significantly improving the recommendations for non-mainstream users while maintaining the recommendation quality for mainstream users. Our results emphasize the importance of deploying extensive content-based features, such as online reviews, in order to better represent users and items to maximize the de-biasing effect.

KEYWORDS

Recommender Systems; Mainstream Bias; User Fairness

ACM Reference Format:

Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-mainstream Users. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21)*, March 8–12, 2021, Virtual Event, Israel. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3437963.3441769>

1 INTRODUCTION

Collaborative Filtering (CF) models are the most investigated and deployed models in the domain of recommender systems [4]. These models assume that users who had similar preferences on items

of a specific kind (e.g. books, movies) in the past may continue having similar preferences on other items of the same kind. The preferences of the users are expressed through their explicit (e.g. ratings) or implicit (e.g. clicks, downloads) interactions with items.

Among the CF models, *Matrix Factorization* (MF) [15], which tries to find a representation of both users and items in the same latent factor space, has long been the most successful and most widely deployed CF model. More recently, generalized factorization models, such as factorization machines [30], have been proposed, exploiting input beyond user-item interactions to learn the latent space. Exploiting more input, such as contextual features and other types of useful side information about users and items, was shown to further improve the recommendation quality. The potential for further improvement, for instance by relying on more abundant input data including audio, visual and textual item descriptions or social network dynamics, came only when deep neural networks (DNN) [10] entered the recommendation domain and enabled more sophisticated user/item representation space learning. In particular, textual data acquired from websites have been extensively exploited for this purpose, allowing users to leave review comments for items along with ratings. For this type of data, DNN-based user/item modeling utilizing NLP techniques has been shown to achieve significantly higher recommendation performance [8, 37, 44, 45] as well as provide convincing explanations [6, 7, 26, 40].

While these developments have greatly contributed to the improvement of the overall recommendation accuracy, one problem has remained largely unsolved, namely the presence of various types of biases in the learned recommendation models. In this paper we focus on the bias towards the so-called *mainstream users*. A mainstream user often prefers items liked by many people and also reacts negatively to items widely disliked by others [33]. Contrary to this, non-mainstream users typically show interest on rarely-visited items or have an opposite attitude towards widely accepted or rejected items. Such a “grey sheep” property [46] makes these users different from others, making it difficult for a CF algorithm to identify similar peers. This leads to recommendations of a generally lower quality, because recommendations for these users are based on neighbors with insufficiently similar preferences. Furthermore, non-mainstream users are typically a minority and the numerous consistent user-item interactions within the cluster of mainstream users are likely to be dominant in steering the process of learning the user/item representation space. Because of this, the non-mainstream users and their preferred (“outlier”) items become underrepresented in such a space, leading to inequality of the recommendation performance across the user population. This is the *mainstream bias*, the tendency to provide better recommendations to the mainstream users. Such bias could make non-mainstream



This work is licensed under a Creative Commons Attribution International 4.0 License.

WSDM '21, March 8–12, 2021, Virtual Event, Israel.
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8297-7/21/03.
<https://doi.org/10.1145/3437963.3441769>

users draw insufficient utility from a recommender system and could discourage them from using it anymore. This could lead to online businesses starting to lose customers. For the information and news recommender systems, however, we foresee even more serious consequences. Recommender systems may namely become less inclusive with respect to non-mainstream opinions and (e.g. political) views and in this way contribute to undesired long-term effects, like intellectual segregation and societal polarization.

In this paper we build on the success of DNNs for developing recommendation models and propose a simple but effective solution towards neutralizing the mainstream bias. With our new recommendation model, referred to as **Neural AutoEncoder Collaborative Filtering (NAECF)**, we introduce adversarial conditions to the process of learning the recommendation algorithm, which in this specific case is realized as the minimization of the rating prediction error. The adversarial conditions are imposed by *autoencoders* [5], a deep learning architecture widely used for recommendation [18, 25, 35], added to a state-of-the-art DNN-based recommendation framework. They enforce that the user and item representations are learned in a way such that they preserve their specific and unique properties before being fed to the rating predictor. Since this preservation is achieved for all users, mainstream or not, the autoencoders prevent that the learned representations are biased towards the users with a mainstream taste. The results of experiments conducted on different domains and scales of the real-world datasets from Amazon [21] show that the representations learned in this way indeed help to de-bias the produced recommendations (predicted ratings in this case). Compared to the case without deploying the adversarial conditions, our proposed method produces significantly better recommendations for non-mainstream users while largely maintaining the recommendation quality for mainstream users. We clearly show that this performance improvement largely stems from adding adversarial conditions to the process of user and item representation learning. In addition, our experiments demonstrate the benefit the non-mainstream users draw from the application of content-based features such as online reviews, further highlighting their value for achieving high recommendation quality across the user community.

The proposed NAECF approach is, to the best of our knowledge, the first to enforce preserving the unique user and item properties as the adversary to the process of learning how to recommend. This allows us to effectively address the mainstream bias in recommender systems, which has not been extensively studied this far.

A recommender system can be designed to either predict ratings or rank items. Despite the latter is picking up momentum in the field, we still choose in this paper to follow the rating prediction paradigm. The main reason for this lies in the core of our contribution, which is to investigate how a state-of-the-art recommendation framework may be extended in order to de-bias the process of generating recommendations. Since the framework we build upon was evaluated in terms of rating prediction, we follow this same paradigm in this paper. Nonetheless, the user-item representation space generated by the autoencoders can serve to predict both ranking and ratings, so we do not consider our choice to limit the broad application of our proposal to the recommendation practice.

The results of the paper can be fully reproduced with data and code available online¹.

2 RELATED WORK

Our work relates mainly to two topics: biases in recommender systems and review-based user/item modeling.

2.1 Biases in Recommender Systems

Potential biases in the training data have already been recognized in early work on matrix factorization for recommendation. Koren et al. [15] introduced a correction in the dot-product rating prediction formula to incorporate rating biases across users, that is, how the rating scale is interpreted by different users. Another bias related to ratings is the anchoring bias; it emerges from the influence of previous recommendations to a user on that user's future ratings. Adomavicius et al. [3] explored two approaches to neutralize this bias. The first approach involves computational post-hoc adjustments of the ratings that are known to be biased. The second approach involves a user interface by which the system tries to prevent this bias *during* rating. A different sort of bias is the popularity bias, due to which popular items may be recommended more frequently than other, less popular items (e.g., long-tail). Abdollahpouri et al. [1] proposed an add-on to a general collaborative filtering algorithm by which a trade-off between accuracy and long-tail coverage can be tuned. More recently, the discussion about biases has increasingly been conducted in the context of resolving ethical and societal issues when deploying recommender systems in practice, such as polarization [29], fairness [22] and discrimination [19], giving a further boost to the research on this topic.

In this paper we focus on the aforementioned mainstream bias. While being conceptually close to popularity bias [2, 17], there is an important difference between the two. Popularity bias could lead to a separation between more and less popular items, similar to the separation of items into those being interesting to mainstream and non-mainstream users. However, popularity bias is not informative regarding the way a recommender system serves different groups of users. According to Steck [36], users may tend to provide feedback on popular items simply by following (being influenced by) other users. In this way, their preferences are likely to be unconsciously driven away from their real interest. By focusing on mainstreamness, we explicitly look at the bias in the user population.

Kowald et al. [17] demonstrated that non-mainstream music listeners are likely to receive the worst recommendations. Schedl and Bauer [34] investigated music preferences across age groups. They observed that, although only taking a small proportion of users, kids and adolescents have significantly different preferences from other age groups in terms of music genres, and the recommendation performance on these two groups is also distinctive among all users. To repair the unfairness caused by the mainstream bias, several recent works aim at identifying non-mainstream music listeners and using the power of cultural aspects [24, 33] and human memories [16] to better profile these underrepresented users in recommender systems. Despite the reported progress, existing methods to alleviate the mainstream bias usually rely on their specific definitions of mainstreamness, which may limit the findings. Furthermore,

¹<https://github.com/roger-zhe-li/wsdm21-mainstream>

these methods tend to split users into different mainstreamness groups for individual training. This setting may cause the loss of recommendation accuracy due to not exploiting the cross-group collaborative information. The approach proposed in this paper aims at neutralizing the mainstream bias in a more generic fashion and without divisions within the user population.

2.2 Review-based User/Item Modeling

Supported by the rapid development of natural language processing (NLP) techniques, online reviews have increasingly been identified as an important source of useful information for addressing data sparsity issues in recommendation. Exploiting these reviews has led to several advanced recommendation concepts, pioneered by Collaborative Deep Learning (CDL) [39]. This concept introduces a hierarchical Bayesian model using Stacked Denoising Autoencoders (SDAE) to reconstruct the rating matrix from encoded textual reviews. Another method, DeepCoNN [45], unifies the processes of learning the user/item representation and rating prediction in an end-to-end model. The unification is achieved through a combination of Convolutional Neural Networks (CNN) and factorization machines. Due to the sequential nature of reviews, Recurrent Neural Networks (RNN) and attention models are also widely used for user and item feature learning. Wu et al. [42] trained the review representations and ratings jointly within a Long Short-Term Memory (LSTM) framework for movie recommendation. Chen et al. [6] extended the DeepCoNN concept by incorporating attention factors into NARRE, a DeepCoNN-based framework, to provide convincing explanations. MPCN [37] is another attention-based model, which uses two hierarchical attention layers to infer the review importance. Although the use of text reviews has partially resolved the data sparsity issues, a more direct way to achieve this is to increase the scale of the training data. As an example, AugCF [41] was proposed on top of DeepCoNN to augment review and rating data using Generative Adversarial Networks [11]. All models mentioned above represent users and items following the same principle, and the representations are derived from the same data source. Contrary to this, NPA [43] and NeuHash-CF [12] represent users and items in different ways. While they model the items using content-based information, the users are represented by one-hot coded userID.

Despite the remarks expressed in literature state that reviews serve the recommendation better as regularizers than features [32], the models mentioned above have been reported to achieve remarkable overall recommendation accuracy, showing the benefit of using textual review data as input. In this paper, we look at online reviews from a different angle and further than accuracy alone. We analyze their value in achieving better user representations that allow us to balance the recommendation quality across users. We show that, with our proposed recommendation model, reviews can be instrumental in neutralizing the mainstream bias.

3 PROPOSED MODEL: NAE CF

The architecture of the proposed NAE CF model is illustrated in Fig. 1. The scheme shows that with NAE CF we pursue two learning goals simultaneously: maximizing the recommendation accuracy and reconstructing the users' and items' original feature vectors

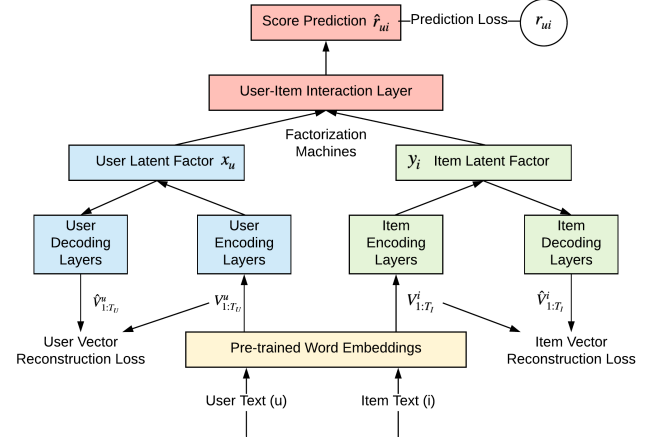


Figure 1: Overall architecture of NAE CF.

in the autoencoders. These feature vectors consist of the texts of user reviews, so we refer to the process taking place in the two AEs as “text reconstruction”. Recommendation accuracy may be achieved by optimizing for rating prediction or ranking prediction. Since DeepCoNN [45], the strongest baseline for comparison, is designed for rating prediction, we also take rating prediction as the criterion for recommendation optimization. This allows us to assess specifically the effect of enforcing user and item reconstruction as an adversarial condition to recommendation optimization on the mainstream bias. If the effect is there, it can also be expected if a ranking prediction scheme is expanded in the same way.

3.1 Model Formulation

The data we use consist of tuples (u, i, r_{ui}, c_{ui}) , representing a user u providing a rating r_{ui} to item i and leaving a review text c_{ui} for said item. Based on Fig. 1, we see the realization of the overall goal of NAE CF by minimizing the following loss:

$$L = L_R + w (L_U + L_I) , \quad (1)$$

where L_R , L_U and L_I are, respectively, the mean rating prediction loss, and the mean text reconstruction losses for users and items. The constant w is a weight determining the relative influence of user and item AEs compared to the rating prediction module. The three losses are defined by the following expressions:

$$L_R = \frac{1}{N_R} \sum_{u,i} \text{loss}_R(u, i) \quad (2)$$

$$L_U = \frac{1}{N_U} \sum_u \text{loss}_U(u) \quad (3)$$

$$L_I = \frac{1}{N_I} \sum_i \text{loss}_I(i) , \quad (4)$$

where N_R , N_U and N_I represent the number of interactions, users and items in the training set, respectively. Normalizing by these terms makes the effect of the weight w invariant to the statistics of the dataset.

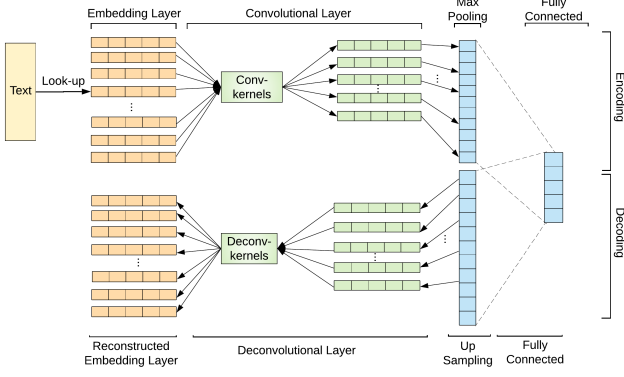


Figure 2: Architecture of the convolutional autoencoder for text feature transformation and extraction.

3.2 Learning for Rating Prediction

In NAE CF, the rating prediction loss for an individual user-item interaction is computed as a traditional squared loss

$$\text{loss}_R(u, i) = \left(\frac{r_{ui} - \hat{r}_{ui}}{r_{\max} - r_{\min}} \right)^2, \quad (5)$$

where \hat{r}_{ui} is the predicted rating given by user u to item i . The loss is normalized by the limits of the rating scale used in the dataset, so that L_R is bounded between 0 and 1. The prediction is computed for the interaction $\hat{z} = (\mathbf{x}_u, \mathbf{y}_i)$ between vectors \mathbf{x}_u and \mathbf{y}_i , encoding the user and item text feature representations, respectively. \mathbf{x}_u and \mathbf{y}_i are the latent factors, low-rank representations of users and items, extracted as the bottlenecks of the corresponding AEs, as indicated by the green and blue blocks in Fig. 1. We follow the settings of DeepCoNN [45] with a Factorization Machine layer [31], and compute rating prediction as

$$\hat{r}_{ui} = \hat{a}_0 + \sum_{m=1}^{|\hat{z}|} \hat{a}_m \hat{z}_m + \sum_{m=1}^{|\hat{z}|} \sum_{n=m+1}^{|\hat{z}|} \langle \hat{v}_m, \hat{v}_n \rangle \hat{z}_m \hat{z}_n \quad (6)$$

where \hat{a}_0 denotes the global bias and \hat{a}_m denotes the strength of first order interactions in \hat{z} . Second order interactions are modeled by $\langle \hat{v}_m, \hat{v}_n \rangle = \sum_{f=1}^{|\hat{z}|} \hat{v}_{m,f} \hat{v}_{n,f}$.

3.3 Learning for Textual Feature Transformation and Reconstruction

The latent factors \mathbf{x}_u and \mathbf{y}_i are used not only for rating prediction, as indicated in the previous section, but also to reconstruct the original user and item representations in the computation of text reconstruction losses. We use an encoder to generate latent factors, which takes an initial user representation V_u or item representation V_i . We deploy the strategy proposed by DeepCoNN [45], that applies TextCNNs [13] for feature transformation. For an arbitrary user u , we extract all review texts they authored and concatenate them into a single long document. Similar to the top NLP models like BERT [9] and GPT-2 [28], here we adopt a cutoff length T_U to truncate words exceeding the limit. For users with fewer than T_U words, we pad empty words, denoted by <UNK> and initialized by zeros. This way all users are represented by the same number of review tokens.

Table 1: Statistics of the datasets.

Dataset	#users	#items	#ratings	Sparsity	#words
Instant videos	5,130	1,685	37,126	99.57	19M
Digital music	5,541	3,568	64,706	99.67	73M
BeerAdvocate	3,703	37,580	393,035	99.72	198M

Then we introduce a look-up layer to get the initial individual word embeddings from a pre-trained model. By concatenating them, we obtain user embedding $V_{1:T_U}^u$. Similarly, we obtain item embedding $V_{1:T_I}^i$ for item i .

Encoding these initial $V_{1:T_U}^u$ and $V_{1:T_I}^i$ embeddings results in the latent factors, which then serve as input to a decoder that we introduce to create reconstructed embeddings $\hat{V}_{1:T_U}^u$ and $\hat{V}_{1:T_I}^i$. The architecture of the decoder is symmetric to the encoder with deconvolution and unpooling layers, as shown in Fig. 2. All hyper-parameters used in the decoding stage are the same as in the encoding stage.

The success of reconstructing initial user and item embeddings is modeled by the text reconstruction losses, which are computed for each user and item. In order to have scores on a bounded scale, we rely on the cosine similarity to measure text reconstruction loss. Unlike most cases in text analysis where embedding values are positive, the original pre-trained embeddings we use in this paper do have negative values, making the cosine similarity range from -1 to 1. Therefore, we also normalize cosine similarities so that the scales of L_U and L_I are comparable to that of L_R . This leads to the following formulation of the individual text reconstruction losses:

$$\text{loss}_U(u) = \left(\frac{1 - \cos(V_{1:T_U}^u, \hat{V}_{1:T_U}^u)}{2} \right)^2, \quad (7)$$

$$\text{loss}_I(i) = \left(\frac{1 - \cos(V_{1:T_I}^i, \hat{V}_{1:T_I}^i)}{2} \right)^2, \quad (8)$$

where \cos stands for the cosine similarity between the original vectors and the reconstructed ones.

3.4 Model Learning

We use Adaptive Moment Estimation (Adam) [14] to minimize the overall loss function in Eq. (1). This way, training converges fast and the learning rate is adapted during the process.

4 EXPERIMENTAL DESIGN

Here we present a series of experiments designed to evaluate the proposed NAE CF model through the following research questions:

- RQ1: Does NAE CF improve the recommendation for non-mainstream users, creating a better balance across users?
- RQ2: What is the effect of using reviews and textual feature transformations on mainstream and non-mainstream users?
- RQ3: What is the correlation between recommendation accuracy and the difficulty of user feature reconstruction?

4.1 Data and Metrics

In this paper we focus on improving the recommendation for non-mainstream users, and investigate the power of text reviews for this purpose. Therefore, the selected datasets are all review-based

(see Table 1). We use two Amazon real-world datasets² covering different recommendation domains, namely instant videos and digital music, and another dataset from BeerAdvocate [20]³. The ratings all range from 1 to 5. However, in the Amazon datasets ratings are integers, while in the BeerAdvocate dataset they are multiples of 0.5. Users in the Amazon datasets have at least 5 interactions. To align with this setting, we filter the BeerAdvocate dataset using the same threshold. Due to unavailability of computational resources, we randomly sampled 25% of users to form a BeerAdvocate subset.

Following the original setting of DeepCoNN [45] and its latest related research [32], we use the Google News pre-trained word vectors [23] to generate pre-trained word embeddings. Each word in the review is thus represented as a 300-dimension vector.

We evaluate the rating prediction accuracy by computing the conventional Root-Mean-Square Error on the test set:

$$rRMSE = \sqrt{\frac{\sum_{u,i} (r_{ui} - \hat{r}_{ui})^2}{N}}. \quad (9)$$

where N is the number of ratings. To evaluate recommendation performance for individual users, we also report per-user RMSE ($uRMSE$, as opposed to $rRMSE$) for further investigation. We cap the predicted ratings to $[1, 5]$, so there are no out-of-bounds values.

4.2 Baselines

We compare the performance of our proposed NAEF model with two related recommendation models:

- **Matrix Factorization [15]**. We use MF as a classical, pure similarity-based CF baseline. All non-textual hyper-parameters in NAEF are reused.
- **DeepCoNN [45]**. This is the pioneering work and state-of-the-art method that introduces deep learning techniques to build a text-based recommender system. User and item features are extracted in parallel, and their interaction is realized by means of factorization machines (FM). Although there are other text-based models following a similar architecture, such as NARRE [6] and [41], that may outperform DeepCoNN, the components they added for better recommendation performance are mainly attention layers or data augmentation modules, introducing no significant change in the model architecture. Therefore, to focus on the effect of autoencoders in NAEF, we still adopt DeepCoNN as the strongest and most relevant baseline.

4.3 Experimental Protocol

We randomly split the datasets into training, validation and test sets with proportions 80%, 10% and 10%, respectively. To address the influence of the data splitting strategy, we set 10 different random seeds and thus use 10 different splits. While all users have at least 5 interactions in total, a random split may distribute these interactions unevenly across sets, such that there may be users with only one rating in the training set. To address this potentially unreliable situation, we only account for users with at least 3 interactions in the training set for evaluation.

²<http://jmcauley.ucsd.edu/data/amazon/>

³<http://snap.stanford.edu/data/web-BeerAdvocate.html>

We first do a grid search on the two Amazon datasets separately to fix the hyper-parameters on DeepCoNN. Then we reuse them for the investigation of NAEF. The hyper-parameters tuned are listed below, with the optimal values indicated in bold:

- Number of latent factors for DeepCoNN and NAEF: {5, 10, 20, 50, **100**, 200, 500}. All latent factors are initialized with a Uniform distribution between -0.01 and 0.01 .
- Learning rate: {**0.00001**, 0.0001, 0.001, 0.01, 0.03, 0.1}.
- Dropout rate to avoid overfitting: {0, 0.1, **0.2**, 0.5}.
- Batch size: {32, 64, 128, **256**, 512, 1024}.
- Number of words: {128, 256, **512**, 1024, 2048}.
- Length of CNN kernels: {2, 3, 4}.

Using the DeepCoNN architecture as reference, we investigate the impact of text reconstruction loss with different weights. Since our main concern in this paper is the effect of adding adversarial conditions via AEs to the original DeepCoNN setting, we weight the user and item AEs with the same weight w , as shown in Eq. (1). Specifically, we consider weight values in the set $\{0, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$. Note that NAEF reduces to DeepCoNN when $w = 0$. Similar to the fine-tuning of the hyper-parameters, the optimal weight is selected on the validation set.

Autoencoders act as adversaries to the rating prediction process, so their activation may lower the overall validation $rRMSE$. Therefore, we deploy a two-stage training strategy: we set $w = 0$ in the first 50 epochs as a pre-training process to get the model ready to train for NAEF, and then change w to the value we are tuning for the next 50 epochs.

Since NAEF does not chase the best overall performance, but rather a better balance across users, we follow a different validation strategy for w . First, we separate users in bins based on their $uRMSE$ score with DeepCoNN; to stress more on the performance for non-mainstream users, we use the 4 uneven bins defined by percentiles 10, 50 and 90 of the $uRMSE$ distribution. The performance gain with respect to DeepCoNN is then computed using these bins as strata, assigning smaller importance to the first and last bins, that is, users with a good recommendation and users who are extremely difficult to model. This way, the assessment of model capability is better aligned with our purposes. Gain is thus defined as follows:

$$\Delta = 0.1\Delta_1 + 0.4\Delta_2 + 0.4\Delta_3 + 0.1\Delta_4, \quad (10)$$

where Δ_b indicates the mean $uRMSE$ difference between DeepCoNN and NAEF in user bin b , and bin weights reflect the fraction of users they contain out of the total sample. A positive Δ value means NAEF improves upon DeepCoNN.

All models are implemented in PyTorch [27], with CUDA and CuDNN for acceleration on an NVIDIA GeForce GTX 1080Ti GPU.

5 RESULTS

In this section, we present and analyze the experimental results. As a summary, Table 2 presents the mean performance of all models over the 10 splits per dataset. It can be seen that DeepCoNN and NAEFs show significantly better recommendation accuracy than MF (paired t -test, $p < 0.05$ [38]), and that NAEF and DeepCoNN perform similarly overall, provided that the weight of the text reconstruction loss is not too high. Furthermore, in Section 5.1

Table 2: $rRMSE$ over 10 data splits for all recommendation models in all three datasets ($mean \pm std.dev.$). Bold for best results per dataset. * for results statistically different from the best (t -test, $p < 0.05$).

Dataset	MF	DeepCoNN	NAECF						
			$w = 0.1$	$w = 0.2$	$w = 0.5$	$w = 1.0$	$w = 2.0$	$w = 5.0$	$w = 10.0$
Instant Video	$1.1600 \pm 0.0264^*$	0.9744 ± 0.0145	0.9732 ± 0.0149	0.9749 ± 0.0122	0.9754 ± 0.0159	0.9757 ± 0.0169	0.9798 ± 0.0212	$0.9896 \pm 0.0221^*$	$0.9967 \pm 0.0221^*$
Digital Music	$1.0466 \pm 0.0097^*$	0.9078 ± 0.0138	0.9083 ± 0.0115	0.9106 ± 0.0128	0.9097 ± 0.0114	0.9104 ± 0.0128	0.9118 ± 0.0134	0.9167 ± 0.0108	$0.9219 \pm 0.0146^*$
BeerAdvocate	$1.0442 \pm 0.0048^*$	0.6722 ± 0.0090	0.6707 ± 0.0064	0.6692 ± 0.0035	$0.6746 \pm 0.0059^*$	$0.6756 \pm 0.0082^*$	$0.6785 \pm 0.0098^*$	$0.6899 \pm 0.0137^*$	$0.7068 \pm 0.0278^*$

Table 3: Weights w yielding the best performance gain per split on the validation set.

Dataset	Split									
	1	2	3	4	5	6	7	8	9	10
Instant Video	2	5	0.1	10	1	5	0.1	0.1	0	0.5
Digital Music	0	0	5	5	2	0.1	2	0.1	0.1	0.5
BeerAdvocate	0.1	0.2	0	0.5	0	0.2	0.5	0.1	0.1	0.2

we show that NAECF, while maintaining similar overall recommendation quality as DeepCoNN, manages to create a significantly better balance across users thanks to the introduction of the user and item reconstruction losses as adversaries to the rating prediction optimization. Finally, in Section 5.2 we dive deeper into the ability of the autocorrelates to reconstruct users from the learned representations, and how this correlates with the recommendation performance per user. This analysis sheds more light on the mechanics underlying NAECF and the reported results.

5.1 Performance Balance Across Users

In order to answer research questions RQ1 and RQ2, we investigate the effect of autoencoders and text reviews on the recommendations for non-mainstream and mainstream users.

5.1.1 Effect of autoencoders. As an adversarial learning model, NAECF has two conflicting goals: minimizing the text reconstruction losses L_U and L_I versus minimizing the rating prediction loss L_R . If the weight of the text reconstruction loss is too small, autoencoders cannot exert sufficient influence on the training process, making them ineffective regarding the mainstream bias. Conversely, if the text reconstruction loss dominates the training process, we expect to have a significant drop in terms of overall rating prediction accuracy. Following the validation process in Section 4.3, we chose the weight w with the best gain Δ on the validation set as the optimal one.

Table 3 reports the optimal validation-set weight per split. As the table shows, in 5 of the 30 splits a weight $w = 0$ achieved the best gain; note that such cases correspond to a simple re-training of DeepCoNN. However, in the vast majority of cases a weight different from zero yielded a better performance gain, although there does not appear to be a single optimal weight for the autoencoders in NAECF. Overall, this suggests that w is a hyperparameter to tune on a case by case basis, and that autoencoders are expected to help when the characteristics of the data allow for it; sometimes they do not lead to a substantial gain over DeepCoNN. Furthermore, and based on detailed $rRMSE$ results not reported in the paper, we see that the weights with the best performance gains often lead to lower overall performance (7, 6, and 6 out of 10 seeds in three

Table 4: Test-set performance gains averaged over splits (higher is better): per-bin gain Δ_b and overall gain Δ . * for gains statistically different from 0 (t -test, $p < 0.05$).

Dataset	Δ_1	Δ_2	Δ_3	Δ_4	Δ
Instant Video	-0.0035	0.0256*	0.0267*	-0.0308*	0.0175*
Digital Music	0.0036	0.0184*	0.0106*	-0.0167*	0.0103*
BeerAdvocate	0.0119	0.0117*	0.0063*	-0.0115*	0.0073*

datasets). This contrast shows that a high score on an overall accuracy metric like $rRMSE$ does not necessarily reflect a good balance across individual users.

After the optimal weights are chosen on the validation set, we turn our attention to the corresponding test-set results. In Table 4 we report the average performance gains of NAECF over DeepCoNN, both per bin and overall. The table shows that users in the central bins (ie. central 80% of users) receive a statistically significant performance gain on all datasets, which is exactly the users that we specifically target in NAECF. For the two Amazon datasets, these are also the bins receiving the largest gains; for the BeerAdvocate dataset it is the first bin that has the highest gain, though the difference is not significant from the second bin. In fact, the gains and losses observed for the 10% of users in the first bin are not statistically different from zero, which means that users that already receive good performance are neither helped nor punished by NAECF. Therefore, the application of autoencoders as adversaries to the rating prediction problem does not sacrifice performance for the mainstream users. Finally, we observe that the 10% of users in the last bin do receive a statistically significant performance loss. While unfortunate, such loss is a collateral damage on a minority of users who are hard to satisfy anyway, in benefit of the bulk of users who now receive better recommendations. Averaging gains across bins, as indicated in Eq. (10), we see that NAECF yields statistically better results than DeepCoNN on all datasets. This indicates the overall success for NAECF to create a better balance across users. In general, we help most of the non-mainstream users without hurting mainstream users.

Figure 3 shows the test-set performance gain for the different data splits. We can first notice that the optimal weight, selected based on the gain on the validation set, turned into a slight loss in the test set for only one split in the Instant Video dataset ($\Delta = -0.0054$), and one split in the Digital Music dataset ($\Delta = -0.001$). Detailed results not reported in the paper show that this is mainly due to a drop in Δ_4 , representing the users that are hard to optimize for in any case. In 5 cases the optimal weight was $w = 0$, which yields a gain $\Delta = 0$.⁴ For the majority of cases though (23 out of 30

⁴Due to the stochastic nature of the training process, one retraining of DeepCoNN may yield a slight gain with respect to another, but it should be zero on expectation. Therefore, we set $\Delta = 0$ when $w = 0$.

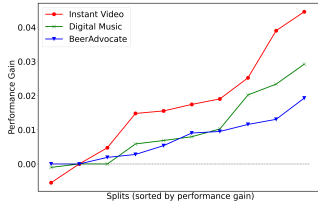


Figure 3: Test-set performance gain Δ on each of the 10 data splits, sorted within dataset.

splits), the optimal weight selection achieves a higher gain Δ and therefore helps achieve a better overall balance in recommendation performance across mainstream and non-mainstream users. If we consider only top 90% users to select the best weight, there are in total 28 out of 30 splits (except 2 in the Digital Music dataset) where NAECF shows superiority over DeepCoNN. However, this does not mean that a higher weight is always better for reaching a balance across users. For the BeerAdvocate dataset, 83% of the top-3 weights selected via the validation process are not larger than 0.5. For the two Amazon datasets, and although the optimal weights distribute over all weight candidates, unreported results still show that weights no larger than 2.0 take 85% of the top-3 best results. This observation matches our expectation that a mild weight value is more likely to bring a better trade-off between the overall recommendation accuracy and the balance across users. Based on these observations, we provide a positive answer to RQ1.

5.1.2 Effect of text reviews. We hypothesize that exploiting elaborate user- and item-related information, in our case in the form of online reviews, not only contributes to overall recommendation performance [6, 45], but also to neutralizing the mainstream bias. While non-mainstream users are relatively underrepresented in the user space, it should at least help if their individual representations are as elaborate as possible to model their preferences better.

In order to verify this hypothesis, we investigate the effect of this additional information compared to the case where it is not used, such as in a classical collaborative-filtering models like MF. We deliberately do not compare MF with NAECF because we would confound the use of text reviews for boosting recommendation accuracy and balancing across users. Instead, we choose to compare with DeepCoNN, which may anyway be regarded as a special case of NAECF and, architecture-wise it is the closest to collaborative filtering in the NAECF family. As such, a superiority of DeepCoNN over MF will indirectly mean a superiority of NAECF as well.

Fig. 4 shows the $uRMSE$ improvements on the test set made by DeepCoNN, compared to MF on all 3 datasets. As the figure shows, our expectations are met on all three datasets. The improvement on $uRMSE$ scores has a clearly positive correlation with the baseline $uRMSE$ achieved by MF, meaning that it is the users who received worse recommendations in MF, the ones who benefit the most from the inclusion of textual features in DeepCoNN. We note that, close to the origin of the plots, we see that DeepCoNN leads to slight performance loss for the users for which MF achieved the best performance. This is however an artifact of the evaluation process. Note that users with a $uRMSE$ close to zero in MF have almost

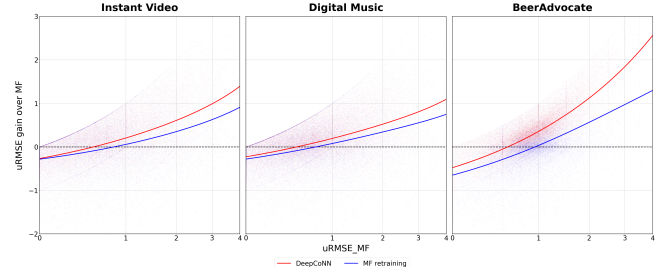


Figure 4: $uRMSE$ gain over MF (positive is better) of DeepCoNN and a retrained MF model, on all 10 data splits. Curves represent a spline-smoothed fit.

no room for improvement, so any other model we compare with will probably perform worse. Similarly, other models will likely perform better for the users with very high $uRMSE$ in MF, because it is just not possible to perform worse. To illustrate and account for this effect, Fig. 4 also compares with a retrained MF model, displaying both the overall correlation and the loss close to the origin. These serve as a sort of baseline to assess the improvement of DeepCoNN (ie. rather than comparing the red curve with the $y = 0$ axis, compare it with the blue curve). We can thus confirm the superiority of DeepCoNN, as $uRMSE$ scores are always higher than on a retrained MF model.

Finally, and similar to the comparison between DeepCoNN and NAECF stated in Eq. (10), here we also compare DeepCoNN and MF in terms of gain Δ . The values on three datasets are 0.1175, 0.1017 and 0.3282, respectively. Such a significant improvement shows the effectiveness of review-based features in creating balance across different users, by which we provide an answer to RQ2.

In summary, we confirmed that NAECF creates a better balance across users by significantly improving the recommendation accuracy for non-mainstream users, subject to a good selection of the weight hyper-parameter. We also compared the review-based DeepCoNN and the CF-based MF, and found that the improvement stems mainly from a better optimization for non-mainstream users who are harder to handle in bare collaborative filtering. This way, NAECF’s superiority lies in the use of review text, not only to boost rating prediction, but also as an adversary to ensure better user representation. Ultimately, these findings direct an open question to the correlation between the text reconstruction loss and the recommendation accuracy, which we study next.

5.2 User Feature Reconstruction

Mainstream users are generally active and display common behavioral patterns. This makes it easier for them to be matched with proper neighbors in collaborative filtering, ultimately giving them more accurate recommendations. At the same time, good performance on similarity-based user modeling will make them easier to reconstruct in the NAECF autoencoders, and should therefore have a lower text reconstruction loss after training. This should be reflected by a positive correlation between $uRMSE$ scores and user reconstruction losses $loss_U$. Because DeepCoNN does not contain any text reconstruction module, the loss should be randomly

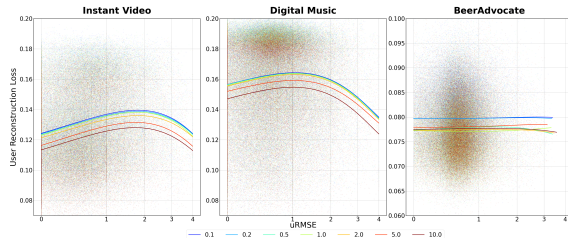


Figure 5: NAECF $loss_U$ by test-set $uRMSE$, for each weight w . Lines represent a spline-smoothed fit.

distributed and uncorrelated with $uRMSE$; we verified this in the data but do not report it here. However, intuition tells us that mainstream users should have low reconstruction losses. The failure of DeepCoNN to reflect this expectation means there is room for improvement to create a better balance across users, which is confirmed by our findings in the previous section.

Therefore, we now look into the correlation between user reconstruction loss $loss_U$ and $uRMSE$ recommendation accuracy. Fig. 5 shows the relationship for each of the evaluated weights w . We can see clear differences across datasets, but there are several qualitative commonalities. First, the majority of users are not mainstream even if they have a rather low $uRMSE$ score. Second, higher weights generally lead to lower reconstruction losses and therefore to better user representations. This is expected because a high weight makes the text reconstruction losses dominate the overall loss in Eq. (1), but the figure further shows that the relative relationship between $loss_U$ and $uRMSE$ is pretty consistent across weights. Interestingly, the BeerAdvocate dataset shows some fluctuations with high weights. This evidences that the optimal weight needs proper tuning, because excessively high weights lead to substantial performance loss and $uRMSE$ scores become less stable as a consequence. As reported in Table 3, the optimal weights for this dataset are rather small indeed in comparison with the other two datasets.

We also followed the earlier approach of dividing users in four bins according to the $uRMSE$ distribution. Fig. 6 similarly shows the relationship between $loss_U$ and $uRMSE$ with all the evaluated weights, but differentiating among user bins. We can clearly observe that, as expected, the relationship is monotonically positive except for the last bin in the Digital Music dataset. This confirms again that users who are better represented receive more accurate recommendations. As reported in Table 4, mainstream users in bin 1 do not always benefit from NAECF because they already receive good recommendations and there is little room for improvement, regardless of how well they are reconstructed. Fig. 6 confirms this specially in the two Amazon datasets, where bin 1 users receive nearly perfect recommendations. But NAECF improves performance specially for the 80% of non-mainstream users in bins 2 and 3, because those are harder to represent to begin with. Fig. 6 confirms that these users generally have the highest reconstruction losses indeed. Together with the correlations in Fig. 5, we see the relationship between the mainstreamness of users and the difficulty to represent them. Notwithstanding, the bottom 10% of users in bin 4 are too extreme to find proper representations, so the autoencoders hardly work for them. We even observed in Fig. 6 a negative correlation on the

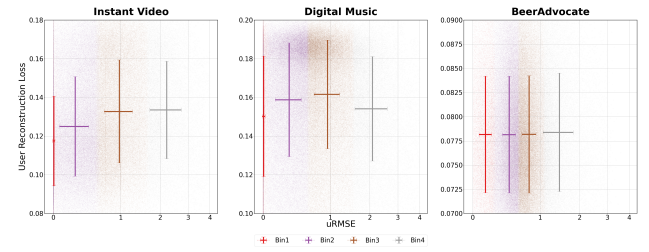


Figure 6: NAECF $loss_U$ by test-set $uRMSE$, with all evaluated weights w . Error bars show standard deviations per user bin.

Instant Video and Digital Music datasets when $uRMSE$ is large. This confirms that NAECF sacrifices performance for these extreme users in favor of the others. Although unfortunate, we find this behavior acceptable because these users often display such particular tastes and patterns that it is hard for them to benefit from virtually any CF-based recommendation model.

6 CONCLUSION AND FUTURE WORK

Rating accuracy has long been an important criterion to evaluate recommender systems, if not the most important. Previous research has therefore focused mainly on maximizing the overall performance averaged over users. However, traditional collaborative filtering methods focus more strongly on recommending items that have positive interactions by similar users. In this situation, it is hard for CF models to work well with non-mainstream users that have special tastes or habits. Because non-mainstream users are rather a minority, this problem may not have a strong effect on the overall accuracy, yet it may create an unfair imbalance across users. To address this problem, we proposed a conceptually simple but effective model named NAECF, which minimizes the rating prediction loss while keeping the user and item properties preserved in the learned user and item representations. Preservation of user and item properties is imposed as an adversarial condition by minimizing reconstruction losses in addition to rating prediction error. This prevents these representations from being biased towards mainstream users.

We conducted experiments on three real-world datasets, and found that NAECF achieves an overall rating accuracy that is on par with the state-of-the-art. However, its strength is in the better balance it achieves across users thanks to a significant improvement of the recommendation accuracy for non-mainstream users, without significantly harming the mainstream ones. This improvement is achieved through an optimal trade-off between rating prediction and text reconstruction. Our results confirm a clear correlation between how well users are represented and the quality of their recommendations, evidencing that side information may be instrumental not only for boosting overall accuracy, but also to minimize possible biases in the learned models.

Future work will be conducted in several directions. First, we will investigate whether the conclusions drawn here for rating prediction generalize to the ranking paradigm, which is gaining popularity in the recommendation field. Second, in this paper we treated users and items as equally important through a single text

reconstruction weight. One may argue that improving the representation of users alone is not enough, because the model also needs a good item representation to know what to recommend. However, users and items may have different impacts, and we would like to explore this question by implementing two weights in the NAEFCF loss. Third, we introduced side information from text reviews in order to achieve a better balance across users. However, text reviews are just an example of additional content-based resources such as images and demographic information that can be used to achieve a similar function. We would like to further investigate the effect of other side information in the future and, perhaps more importantly, how to effectively incorporate such information in NAEFCF to eliminate the mainstream-bias. Finally, we are also interested in combining NAEFCF with explainable recommendation, so that we can provide convincing explanations to non-mainstream users.

REFERENCES

- [1] Himan Abdollahpour, Robin Burke, and Bamshad Mobasher. 2018. Popularity-Aware Item Weighting for Long-Tail Recommendation. *arXiv preprint arXiv:1802.05382* (2018).
- [2] Himan Abdollahpour, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *RMSE@RecSys (CEUR Workshop Proceedings)*, Vol. 2440. CEUR-WS.org.
- [3] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2014. De-biasing user preference ratings in recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 2.
- [4] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering* 6 (2005), 734–749.
- [5] Dana H Ballard. 1987. Modular Learning in Neural Networks. In *AAAI* 279–284.
- [6] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *World Wide Web Conference. International World Wide Web Conferences Steering Committee*, 1583–1592.
- [7] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems* 37, 2 (2019).
- [8] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xueming Song, and Mohan S Kankanhalli. 2018. A 3NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *IJCAI* 3748–3754.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.
- [10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.
- [11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NIPS*. 2672–2680.
- [12] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2020. Content-aware Neural Hashing for Cold-start Recommendation. In *SIGIR*. ACM, 971–980.
- [13] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP. ACL*, 1746–1751.
- [14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- [15] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.
- [16] Dominik Kowald, Elisabeth Lex, and Markus Schedl. 2020. Utilizing Human Memory Processes to Model Genre Preferences for Personalized Music Recommendations. *CoRR abs/2003.10699* (2020).
- [17] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *ECIR (2) (Lecture Notes in Computer Science)*, Vol. 12036. Springer, 35–42.
- [18] Chen Ma, Peng Kang, Bin Wu, Qinglong Wang, and Xue Liu. 2019. Gated Attentive-Autoencoder for Content-Aware Recommendation. In *ACM International Conference on Web Search and Data Mining*. 519–527.
- [19] Masoud Mansoury, Himan Abdollahpour, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. 2020. Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems. In *FLAIRS Conference*. AAAI Press, 193–196.
- [20] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *IEEE International Conference on Data Mining*. IEEE, 1020–1025.
- [21] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.
- [22] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *CIKM*. ACM, 2243–2251.
- [23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [24] Peter Müllner. 2019. *Studying Non-Mainstream Music Listening Behavior For Fair Music Recommendations*. Ph.D. Dissertation. Graz University of Technology.
- [25] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1933–1942.
- [26] Deng Pan, Xiangrui Li, Xin Li, and Dongxiao Zhu. 2020. Explainable Recommendation via Interpretable Feature Mapping and Evaluation of Explainability. In *IJCAI*. ijcai.org, 2690–2696.
- [27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.
- [28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
- [29] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In *WSDM*. ACM, 231–239.
- [30] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 995–1000.
- [31] Steffen Rendle. 2012. Factorization Machines with libFM. *ACM Trans. Intell. Syst. Technol.* 3, 3 (2012), 57:1–57:22.
- [32] Naveen Sachdeva and Julian McAuley. 2020. How Useful are Reviews for Recommendation? A Critical Review and Potential Improvements. In *SIGIR*. ACM, 1845–1848.
- [33] Markus Schedl and Christine Bauer. 2018. An analysis of global and regional mainstreamness for personalized music recommender systems. *Journal of Mobile Multimedia* 14, 1 (2018), 95–112.
- [34] Markus Schedl and Christine Bauer. 2019. Online Music Listening Culture of Kids and Adolescents: Listening Analysis and Music Recommendation Tailored to the Young. *CoRR abs/1912.11564* (2019).
- [35] Suvasish Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. 2015. Autorec: Autoencoders meet collaborative filtering. In *International Conference on World Wide Web*. ACM, 111–112.
- [36] Harald Steck. 2011. Item popularity and recommendation accuracy. In *ACM conference on Recommender systems*. ACM, 125–132.
- [37] Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-Pointer Co-Attention Networks for Recommendation. In *KDD*. ACM, 2309–2318.
- [38] Julián Urbano, Harllely Lima, and Alan Hanjalic. 2019. Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors. In *ACM SIGIR*. 505–514.
- [39] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1235–1244.
- [40] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 165–174.
- [41] Qinyong Wang, Hongzhi Yin, Hao Wang, Quoc Viet Hung Nguyen, Zi Huang, and Lizhen Cui. 2019. Enhancing Collaborative Filtering with Generative Augmentation. In *KDD*. ACM, 548–556.
- [42] Chao-Yuan Wu, Amr Ahmed, Alex Beutel, and Alexander J. Smola. 2017. Joint Training of Ratings and Reviews with Recurrent Recommender Networks. In *ICLR (Workshop)*. OpenReview.net.
- [43] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *KDD*. ACM, 2576–2584.
- [44] Susen Yang, Yong Liu, Yinan Zhang, Chunyan Miao, Zaiqing Nie, and Juyong Zhang. 2020. Learning Hierarchical Review Graph Representation for Recommendation. *arXiv preprint arXiv:2004.11588* (2020).
- [45] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *ACM International Conference on Web Search and Data Mining*. ACM, 425–434.
- [46] Yong Zheng, Mayur Agnani, and Mili Singh. 2017. Identification of Grey Sheep Users by Histogram Intersection in Recommender Systems. In *ADMA (Lecture Notes in Computer Science)*, Vol. 10604. Springer, 148–161.