Personality Bias of Music Recommendation Algorithms

Alessandro B. Melchiorre
Johannes Kepler University Linz and Linz Institute of Technology, Austria
alessandro.melchiorre@jku.at

Eva Zangerle
University of Innsbruck, Austria
eva.zangerle@uibk.ac.at

Markus Schedl
Johannes Kepler University Linz (JKU) and Linz Institute of Technology (LIT), Austria
markus.schedl@jku.at

ABSTRACT

Recommender systems, like other tools that make use of machine learning, are known to create or increase certain biases. Earlier work has already unveiled different performance of recommender systems for different user groups, depending on gender, age, country, and consumption behavior. In this work, we study user bias in terms of another aspect, i.e., users' personality. We investigate to which extent state-of-the-art recommendation algorithms yield different accuracy scores depending on the users' personality traits. We focus on the music domain and create a dataset of Twitter users' music consumption behavior and personality traits, measuring the latter in terms of the OCEAN model. Investigating recall@K and NDCG@K of the recommendation algorithms SLIM, embarrassingly shallow autoencoders for sparse data (EASE), and variational autoencoders for collaborative filtering (Mult-VAE) on this dataset, we find several significant differences in performance between user groups scoring high vs. groups scoring low on several personality traits.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; *Music retrieval*; • **Applied computing** → Psychology.

KEYWORDS

music recommender systems, personality, bias, neural networks, dataset

ACM Reference Format:
Alessandro B. Melchiorre, Eva Zangerle, and Markus Schedl. 2020. Personality Bias of Music Recommendation Algorithms. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20), September 22–26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3383313.3412223

1 INTRODUCTION

Recommender systems in the multimedia domain—in particular in the music domain—have been shown to exhibit various kinds of biases, most notably on the item level (e.g., long-tail items are less frequently recommended [1, 5, 15]) and on the user level (e.g., users of a certain gender, belonging to a certain age group, or living in a certain country receive recommendations of different quality [24]). However, one important user characteristic that has not been studied yet under the perspective of recommender systems bias is personality. Personality traits are stable over a longer period of time and can, therefore, be considered in a way similar to gender when it comes to investigating bias [7]. Against this background, we address the following research questions: *Do state-of-the-art recommender algorithms yield different performance scores for different user groups in terms of personality traits? If so, how can these differences be characterized?*

In the study presented here, we focus on the music domain since some personality traits have already been shown to correlate with music preferences [22] and usage of music [6]. We, therefore, speculate that music listeners with different personality profiles might be treated differently by music recommender systems.

In this paper, we present related literature (Section 2), detail our methodology and data (Section 3), describe experimental setup and results (Section 4), present conclusions, limitations, and future research avenues (Section 5).

2 RELATED WORK

Related literature can be categorized into recommender systems research that considers personality in the recommendation process and research on bias and fairness of recommender systems.

Personality traits are a psychological construct that remains stable over the years [7]. They are known to influence our preferences and consumption behaviors, e.g., towards music [12]. Research that integrates users' personality into the recommendation process has emerged only recently, though [27]. The most common personality model adopted in recommender systems research is the OCEAN model [18], which describes personality traits along five dimensions: *openness to experience* (conventional vs. creative thinking), *conscientiousness* (disorganized vs. organized behavior), *extraversion* (engagement with the external world), *agreeableness* (need for social harmony), and *neuroticism* (emotional instability).

While *personality-aware recommender systems* have been proposed in domains other than music (e.g., movies [19], food/recipes [2], and computer games [28]), we focus our discussion on music recommendation due to the scope of this paper. Lu and Tintarev propose a system that adapts according to users' personality traits and their diversity needs [17]. To this end, results of a collaborative filtering recommender are re-ranked with respect to the level of diversity each item, i.e., song, contributes to the recommendation list. Intra-list diversity is computed on item features such as music key, genre, and number of artists. Based on previously identified correlates between personality traits and diversity needs, the authors map each personality trait to a desired level of diversity and

RecSys '20, September 22–26, 2020, Virtual Event, Brazil

ACM ISBN 978-1-4503-7583-2/20/09…$15.00
https://doi.org/10.1145/3383313.3412223

integrate this information as weighting term into the objective function used for re-ranking. Fernández-Tobías et al. present different personality-aware recommender systems to alleviate the cold-start problem in book, movie, and music recommendation [11]. In particular, they propose a matrix factorization approach for model-based collaborative filtering that integrates a user latent factor describing personality traits in terms of the five dimensions of the OCEAN model.

The concept of *fairness* requires systems not to discriminate against either a group [21] or individuals [9] in terms of recommendation quality. Establishing fairness typically involves identifying discriminated individuals or groups and, subsequently, developing algorithms that eliminate this discrimination [4]. Burke extended the concept of fairness to multisided fairness, noting that recommender systems have to consider the interests of all stakeholders of the system [3].

Recent research revealed a popularity bias in current recommendation algorithms. In particular, it was shown that users are recommended items that do not match their preference towards a certain popularity level (niche songs/artists are undervalued) [1, 15]. Ekstrand et al. investigated demographic biases in collaborative filtering scenarios with regards to age and gender and found that biases do not necessarily correlate with user group size [10]. Schedl et al. showed that users of different gender, age, and country receive (music) recommendations of different quality [24]. Our work, in contrast, is the first to investigate biases that may result from different personality traits.

3 MATERIALS AND METHODS

In the following, we describe the creation of the used dataset (Section 3.1) and its composition (Section 3.2), the investigated recommendation algorithms (Section 3.3), and the evaluation metrics we adopt (Section 3.4). We publicly release the dataset and code needed to reproduce the experiments at https://github.com/CPJKU/pers_bias.

3.1 Data Acquisition

To obtain *behavioral data on music consumption* as well as information on users' personality, we exploit microblogs shared on Twitter, and particularly leverage so-called #nowplaying tweets in which users tweet about the music they are currently listening to. Along the lines of [13, 29], we utilize #nowplaying tweets stemming from 2018 and 2019 (256,705,566 tweets in total, gathered via the Twitter Streaming API,[1] searching for the keywords #nowplaying, #listento, or #listeningto). To extract track and artist information from those tweets, we use the MusicBrainz database[2] [26], an openly available database of music metadata. It provides metadata on artists, recordings, releases, etc., which is obtained through crowd-sourcing. For extracting artist names and track titles from tweets, we firstly strip URLs, mentions, and hashtags from the tweet text. Subsequently, we tokenize the text and identify the longest subsequence of tokens that corresponds to an artist entry in the MusicBrainz database. If we detect a matching artist, we remove the tokens constituting the artist name from the tweet and try to match the remaining text to a track of the detected artist, again using MusicBrainz metadata. If we cannot match a tweet against both, a track name and an artist name, we discard it.

We further refine the dataset by heuristically removing alleged radio stations through a careful check of the occurrence of certain words in the tweets, the number of shared links, and the number of listening events (user–item interactions). We identify a set of words hinting at radios (e.g., #listenlive and radio) and drop a "user" if at least half of their tweets contain any of these words. Since radio stations tend to share many tweets with links in it, we also drop a user if the majority of the user's tweets contain at least one link. Lastly, we remove all users above the 99.99% percentile of the number of listening events as radios commonly create an exorbitant number of listening events.

To obtain *personality* information of the users, we query the Twitter API[3] to get their most recent 1,000 tweets, excluding retweets.[4] Users with private or deleted profiles are discarded. These tweets are then fed to the IBM Personality Insight API,[5] which returns the personality estimates for each user according to the OCEAN model [18] (cf. Section 2), scaled to [0,1] in terms of percentile ranges. To achieve the maximum accuracy for trait prediction with the service,[6] we only keep users that tweet in English and use more than 3,000 words across their tweets. Lastly, we drop users with fewer than 5 listening events, as commonly done in related research [16, 23], and to enable the evaluation protocol (80:20 training/test split) detailed in Section 4.1.

3.2 Dataset Description

The processing steps described above eventually lead to a final dataset comprising 395,056 total listening events, 18,310 users with personality values, and 15,753 unique tracks. Basic statistics on the behavioral data in our final dataset, i.e., related to user–item interactions, can be found in Table 1. On the right side, a statistical summary of the number of tracks per user (user playcounts) and the number of users per track (track playcounts) is provided. The distribution of users' personality traits among the [0,1]-normalized scores is depicted in Figure 1, where the vertical lines denote the median values. To assess whether users with different personality profiles are treated differently by the mentioned approaches, we perform a median split over each personality trait, thus, effectively, creating two groups of users for each trait: the *high* group, scoring above the median, and the *low* group, scoring lower. Table 2 shows a set of statistics for each personality trait (in the columns) and user group (high vs. low on that trait; in the upper and lower part of the table, respectively). We observe that the total number of unique tracks is quite similar, regardless of personality trait and user group (within range [15,600, 15,700]), except for highly neurotic people who cover fewer items in their listening habits. In terms of the number of listening events, except for neuroticism, the high groups consistently show higher numbers, with a particularly pronounced

[1] https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
[2] https://musicbrainz.org
[3] https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline
[4] Personality is assumed to be stable through time, so it is reasonable to use recent behavioral data to predict personality traits.
[5] https://www.ibm.com/watson/services/personality-insights
[6] https://cloud.ibm.com/docs/personality-insights?topic=personality-thatut#sufficient

difference between high and low groups for the traits extraversion and openness. This does not seem overly surprising since we expect that people who are extraverted and open to experience will listen to (and share) more music than introverts and less open users.

3.3 Recommendation Approaches

We investigate to what extent the following three state-of-the-art recommendation approaches for implicit data yield different accuracy measures, depending on users' personality traits. They have been shown, in extensive experiments, to perform well [8]; and we adapt them, where necessary, to cope with the non-binary nature of our interaction data. This selection of algorithms allows us to investigate both deep learning (non-linear) and traditional (linear) approaches:

- *Sparse Linear Methods (SLIM)* [20]: SLIM is a linear model that aims to compute top-n recommendations by factorizing the item–item co-occurrence matrix under non-negativity and L_1 and L_2 normalization constraints. The learned item coefficients are then used to sparsely aggregate past user interactions and predict the items the user will interact with in the future.
- *Embarrassingly Shallow Autoencoders (EASE)* [25]: EASE is a shallow linear model that could be considered as an extension of SLIM. Since EASE keeps only the L_2 norm constraint, a closed-form solution exists, making it computationally more efficient to train the model.
- *Variational Autoencoders (Mult-VAE)* [16]: Mult-VAE is a variational autoencoder architecture, i.e., a non-linear, probabilistic model, that uses multinomial conditional likelihood for collaborative filtering. Annealing is used to apply regularization for the learning objective.

3.4 Evaluation Metrics

We assess performance using *recall@K* and *normalized discounted cumulative gain@K* (NDCG@K) and report values averaged over all user groups in the test set.[7] Recall@K for user u is defined as

$$Recall@K(u) = \frac{1}{\min(K, N_u)} \sum_{i=1}^{K} rel(i) \tag{1}$$

where N_u is the number of items in the test set which are relevant to u, K is the length of the recommendation list, and $rel(i)$ is an indicator function signaling whether the recommended track at rank i is relevant to u (i.e., $rel(u) = 1$) or not relevant to u (i.e., $rel(u) = 0$). NDCG@K is defined as

$$NDCG@K(u) = \frac{DCG@K(u)}{IDCG@K(u)} \tag{2}$$

where $IDCG@K(u)$ is the ideal $DCG@K$ for user u, obtained when all items in u's test set are ranked in decreasing order of their play count, and $DCG@K(u)$ is the discounted cumulative gain at position k for user u, given by

$$DCG@K(u) = \sum_{i=1}^{K} \frac{rel(i)}{\log_2(i+1)} \tag{3}$$

where $rel(i)$ is the same indicator function as above.

In our experiments, we compute recall@K and NDCG@K for $K = \{5, 10, 50\}$, to model different user needs, ranging from a user interested in only a few top recommendations to a user who inspects a large part of the recommendation list.

4 EXPERIMENTS AND RESULTS

4.1 Experimental Setup

For our experiments, we apply a similar data splitting procedure as used in [16], i.e., we split the users in training/validation/test sets (80%/10%/10%) and for the held-out users we use 80% of their items for training and the remaining 20% as test set to compute the metrics.

We select the hyperparameters of the algorithms under investigation by performing a grid search over different parameters and optimizing for NDCG@50 across all validation users. For SLIM, we explore different α values (sum of the L_1 and L_2 coefficients) and L_1 *ratios* (ratio of L_1 coefficient in α). In detail, we search α in {.5, .1, .01, .001} and L_1 *ratio* in {1, .1, .01}. For EASE, we explore different weights for the L_2 norm in {1, 10, 10^2, $5 \cdot 10^2$, 10^3, 10^4, 10^5, 10^6, 10^7}. For Mult-VAE, we re-use most of the hyperparameters proposed in the original paper [16], except for the architecture and the annealing procedure. We set the total number of epochs to 100. We explore different (symmetric) architectures,[8] comprising 0 or 1 hidden layer(s) with fewer than 500 units for each layer.[9] As for the annealing procedure, we either anneal the regularization parameter through the end of training or stop it half-way by changing the annealing steps in {10,000, 20,000}. In addition, we explore caps for the annealing procedure in {0.5, 1}.

After validation, the best model is selected and evaluated for each user group (defined by trait and high vs. low characteristic) in the test set.

We conduct all experiments across 10 (random) splits of users among the sets, using 10 different seeds for splitting. Results are then averaged across the seeds.[10]

4.2 Results and Discussion

Tables 3 and 4 show the results for all algorithms, personality traits, and user groups, in terms of recall@K and NDCG@K, respectively. The values represent the performance scores averaged across the 10 runs/seeds. Note that the standard deviation of the results across these 10 runs is very low,[11] indicating that results are robust and stable across runs.

To assess the statistical significance of the differences between the high and low user group for each trait, we apply the two-tailed

[7] Note that we will investigate beyond-accuracy metrics [14], such as coverage and diversity, as part of future work.

[8] I-100-I, I-500-I, I-200-50-200-I, I-200-100-200-I, I-500-200-500-I, where I is the total number of tracks.

[9] Increasing the layers and/or the units did not improve results.

[10] Note that the random split stated previously could in theory create unbalanced training/validation/test sets where some user groups may be underrepresented. We also carried out additional experiments where we enforced an equal split in each set for each group (one trait at the time). Results were consistent with the findings reported in this paper.

[11] Standard deviations are 0.0044, 0.0044, and 0.0042 for NDCG@5, 10, and 50, respectively; 0.0049, 0.0052, and 0.0064 for recall@5, 10, and 50, respectively.

No. LEs	No. tracks	No. users		Mean	Std.	Min.	25%	50%	75%	Max.
395,056	15,753	18,310	User playcounts	21.6	34.3	5.0	8.0	12.0	21.0	950.0
			Track playcounts	25.1	33.1	8.0	11.0	16.0	26.0	986.0

Table 1: Statistical summary of the behavioral data (users sharing listening events) in our dataset.

Group		Agr.	Con.	Ext.	Neu.	Ope.
High	No. unique tracks/user (mean and std.)	19.1 ± 24.4	19.2 ± 25.5	20.0 ± 26.3	16.2 ± 18.4	19.5 ± 24.9
	No. unique tracks	15,694	15,674	15,655	15,429	15,652
	No. listening events	208,054	206,179	217,895	177,892	209,741
Low	No. unique tracks/user (mean and std.)	17.3 ± 21.7	17.2 ± 20.4	16.4 ± 19.2	20.3 ± 26.9	16.9 ± 21.1
	No. unique tracks	15,664	15,695	15,672	15,607	15,619
	No. listening events	187,002	188,877	177,161	217,164	185,315

Table 2: Mean and standard deviation of the number of unique tracks per user, for each personality trait and group; as well as total numbers of unique tracks and listening events created by all users in the low and in the high group.

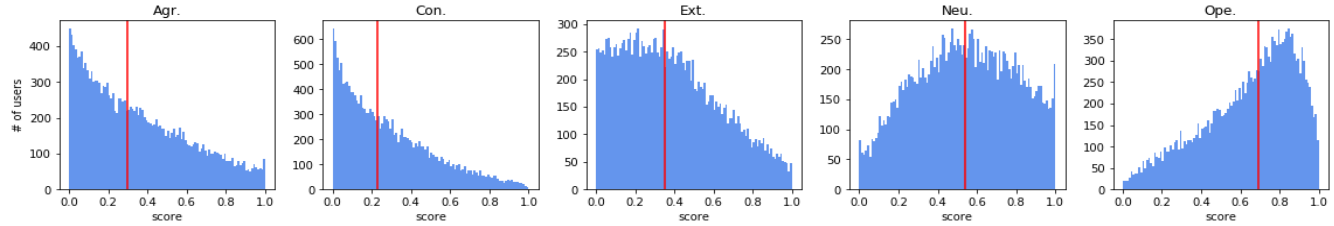

Figure 1: Distribution of personality traits. The x-axis represents the (scaled) score for each trait while the y-axis represents the number of users. The red line represents the median for each trait.

		@5			@10			@50		
Trait	Algorithm	All	High	Low	All	High	Low	All	High	Low
Agr.	EASE	0.0366	0.0348	0.0385	0.0537	0.0534	0.0540	0.1122	0.1129	0.1113
	SLIM	0.0334	0.0320	0.0348	0.0486	0.0478	0.0494	0.1014	**0.1025***	**0.1002***
	Mult-VAE	0.0433	**0.0443***	**0.0423***	0.0634	**0.0655*****	**0.0611*****	0.1456	**0.1504*****	**0.1407*****
Con.	EASE	0.0366	**0.0328***	**0.0406***	0.0537	**0.0495***	**0.0580***	0.1122	0.1096	0.1148
	SLIM	0.0334	**0.0292*****	**0.0377*****	0.0486	**0.0447***	**0.0527***	0.1014	0.0989	0.1040
	Mult-VAE	0.0433	0.0405	0.0462	0.0634	0.0602	0.0665	0.1456	0.1424	0.1488
Ext.	EASE	0.0366	**0.0312****	**0.0420****	0.0537	**0.0467***	**0.0605***	0.1122	0.1032	0.1211
	SLIM	0.0334	**0.0284****	**0.0384****	0.0486	**0.0425***	**0.0547***	0.1014	0.0926	0.1101
	Mult-VAE	0.0433	**0.0378****	**0.0488****	0.0634	0.0568	0.0698	0.1456	0.1348	0.1560
Neu.	EASE	0.0366	**0.0422*****	**0.0311*****	0.0537	**0.0608****	**0.0466****	0.1122	0.1216	0.1028
	SLIM	0.0334	**0.0396*****	**0.0272*****	0.0486	**0.0562*****	**0.0411*****	0.1014	**0.1128***	**0.0900***
	Mult-VAE	0.0433	**0.0500*****	**0.0367*****	0.0634	**0.0721*****	**0.0547*****	0.1456	**0.1588****	**0.1324****
Ope.	EASE	0.0366	**0.0265*****	**0.0468*****	0.0537	**0.0410*****	**0.0663*****	0.1122	**0.0935*****	**0.1307*****
	SLIM	0.0334	**0.0232*****	**0.0436*****	0.0486	**0.0366*****	**0.0605*****	0.1014	**0.0841*****	**0.1186*****
	Mult-VAE	0.0433	**0.0316*****	**0.0550*****	0.0634	**0.0479*****	**0.0787*****	0.1456	**0.1232*****	**0.1678*****

Table 3: Recall@5, 10, and 50 for each algorithm, personality trait, and group (high vs. low; and for all users). Significant differences between high and low groups are marked in bold and with an asterisk (Mann-Whitney-U test, * $p < .05$, ** $p < .01$, *** $p < .001$).

Mann-Whitney-U test on the high and low user scores [12] (NDCG@K and recall@K) and highlight their respective means in the tables in bold, with asterisks denoting the different alpha levels. The most notable observation is that for the traits neuroticism and openness most of all differences between the high and low groups are highly significant ($p < .001$), both in terms of recall@{5, 10, 50}, and NDCG@{5, 10, 50}. As a second observation, we find that the direction of difference in performance is nearly always consistent between all investigated algorithms, i.e., all algorithms treat the high vs. low groups unfairly in the same manner or direction; though the absolute value of the difference varies between algorithms, of course.

While performance for highly neurotic users is consistently better than the performance of low-neurotic-users, the opposite is true for all the other traits. These results seem to be correlated with the data consumption statistics shown previously (cf. Table 2), namely,

[12] The results follow the same trend when using the Fisher's method to aggregate the p-values across the seeds, although with decreased significance level except for openness.

Trait	Algorithm	@5 All	@5 High	@5 Low	@10 All	@10 High	@10 Low	@50 All	@50 High	@50 Low
Agr.	EASE	0.0311	0.0295	0.0327	0.0392	0.0386	0.0399	0.0576	**0.0575***	**0.0577***
	SLIM	0.0279	0.0263	0.0295	0.0351	0.0340	0.0363	0.0517	**0.0514****	**0.0520****
	Mult-VAE	0.0380	**0.0385***	**0.0374***	0.0474	**0.0485*****	**0.0462*****	0.0724	**0.0747*****	**0.0701*****
Con.	EASE	0.0311	**0.0274***	**0.0349***	0.0392	**0.0352***	**0.0433***	0.0576	0.0542	0.0611
	SLIM	0.0279	**0.0241*****	**0.0319*****	0.0351	**0.0312***	**0.0391***	0.0517	0.0484	0.0551
	Mult-VAE	0.0380	0.0353	0.0407	0.0474	0.0445	0.0503	0.0724	0.0697	0.0752
Ext.	EASE	0.0311	**0.0266****	**0.0355****	0.0392	**0.0342***	**0.0441***	0.0576	0.0525	0.0626
	SLIM	0.0279	**0.0242****	**0.0317****	0.0351	**0.0310***	**0.0392***	0.0517	0.0474	0.0560
	Mult-VAE	0.0380	**0.0340****	**0.0417****	0.0474	0.0433	0.0513	0.0724	0.0678	0.0769
Neu.	EASE	0.0311	**0.0366*****	**0.0257*****	0.0392	**0.0454****	**0.0331****	0.0576	0.0639	0.0513
	SLIM	0.0279	**0.0335*****	**0.0224*****	0.0351	**0.0413*****	**0.0290*****	0.0517	0.0585	0.0449
	Mult-VAE	0.0380	**0.0436*****	**0.0324*****	0.0474	**0.0539*****	**0.0409*****	0.0724	**0.0798***	**0.0652***
Ope.	EASE	0.0311	**0.0221*****	**0.0400*****	0.0392	**0.0293*****	**0.0491*****	0.0576	**0.0463*****	**0.0688*****
	SLIM	0.0279	**0.0196*****	**0.0363*****	0.0351	**0.0261*****	**0.0441*****	0.0517	**0.0413*****	**0.0620*****
	Mult-VAE	0.0380	**0.0285*****	**0.0473*****	0.0474	**0.0366*****	**0.0581*****	0.0724	**0.0600*****	**0.0848*****

Table 4: NDCG@5, 10, and 50 for each algorithm, personality trait, and group (high vs. low; and for all users). Significant differences between high and low groups are marked in bold and with an asterisk (Mann-Whitney-U test, * $p < .05$, ** $p < .01$, *** $p < .001$).

a higher number of listening events and higher number of average tracks per user suggest a negative impact on the performance metrics. Furthermore, for conscientiousness and extraversion, the unfair treatment of user groups mostly appears for EASE and SLIM but not for Mult-VAE, while the opposite is true for agreeableness. This suggests that different models trained on the same data will lead to different kind of biases.

To finally answer our research questions: *Do state-of-the-art recommender algorithms yield different performance scores for different user groups in terms of personality traits?* They do indeed for some personality traits, in terms of recall@K and NDCG@K; most notably for the traits openness and neuroticism, and to a smaller extent for the other traits. *If so, how can these differences be characterized?* Scoring low on the personality trait results in higher performance for openness, extraversion, and conscientiousness, but in lower performance for neuroticism and agreeableness.

5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this work, we presented a first study to investigate the extent to which state-of-the-art recommendation algorithms (EASE, SLIM, and Mult-VAE) treat users with different personality traits in different ways, in terms of accuracy metrics (recall@K and NDCG@K). We found highly significant differences ($p < .001$) in both performance scores in particular for the traits neuroticism and openness as well as significant differences at $p < .01$ and $p < .05$ for the other traits.

While results are noteworthy, we also identify several *limitations* of the study at hand. First, like every research that leverages user data shared in online social networks, results obtained for Twitter users may not generalize to the population at large, or even to other platforms. Also, since Twitter's Streaming API only provides access to a small percentage of all shared tweets, the data is incomplete, though still substantial in size. Third, since we rely on self-disclosed information of Twitter users, the listening data we extract from their tweets may not accurately reflect the actual behavior of users, rather how the users want to be perceived (e.g., by avoiding to share guilty pleasure songs).

There are several directions we contemplate for *future research*. In the initial study presented here, we identified certain biases in terms of unequal treatment of different personality groups. However, the exact origin of these biases still needs to be investigated further. In particular, to which extent differences in accuracy can be explained by different consumption patterns of users with different personality (data bias), and to which extent these differences are introduced by the recommender system itself (algorithmic bias) remains an open question that will be addressed in the future. In addition, we plan to include beyond-accuracy metrics [14], e.g., diversity, serendipity, and coverage in our investigation. Finally, we would like to investigate the extent to which results generalize to platforms other than Twitter and additional recommendation algorithms.

ACKNOWLEDGMENTS

This research is supported by Know-Center Graz, through project "Theory-inspired Recommender Systems".

REFERENCES

[1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. In *Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark, September 20, 2019 (CEUR Workshop Proceedings)*, Robin Burke, Himan Abdollahpouri, Edward C. Malthouse, K. P. Thai, and Yongfeng Zhang (Eds.), Vol. 2440. CEUR-WS.org. http://ceur-ws.org/Vol-2440/paper4.pdf

[2] Ifeoma Adaji, Czarina Sharmaine, Simone Debrowney, Kiemute Oyibo, and Julita Vassileva. 2018. Personality Based Recipe Recommendation Using Recipe Network Graphs. In *Social Computing and Social Media. Technologies and Analytics*, Gabriele Meiselwitz (Ed.). Springer International Publishing, Cham, 161–170.

[3] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[4] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*. 202–214.

[5] Òscar Celma. 2010. *Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space.* Springer. https://doi.org/10.1007/978-3-642-13287-2

[6] Tomas Chamorro-Premuzic and Adrian Furnham. 2007. Personality and music: can traits explain how people use music in everyday life? *British Journal of Psychology* 98 (May 2007), 175–185.

[7] Deborah A. Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters* 115, 1 (2012), 11–15. https://doi.org/10.1016/j.econlet.2011.11.015

[8] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. 2019. A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research. *CoRR* abs/1911.07698 (2019). arXiv:1911.07698 http://arxiv.org/abs/1911.07698

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[10] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. 172–186.

[11] Ignacio Fernández-Tobías, Matthias Braunhofer, Mehdi Elahi, Francesco Ricci, and Iván Cantador. 2016. Alleviating the New User Problem in Collaborative Filtering by Exploiting Personality Information. *User Modeling and User-Adapted Interaction* 26, 2-3 (June 2016), 221–255. https://doi.org/10.1007/s11257-016-9172-z

[12] Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. 2017. Personality Traits and Music Genres: What Do People Prefer to Listen To?. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. ACM, New York, NY, USA, 285–288. https://doi.org/10.1145/3079628.3079693

[13] David Hauger, Markus Schedl, Andrej Košir, and Marko Tkalčič. 2013. The million musical tweet dataset: what we can learn from microblogs. International Society for Music Information Retrieval.

[14] Marius Kaminskas and Derek Bridge. 2017. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2017), 2:1–2:42. https://doi.org/10.1145/2926720

[15] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II (Lecture Notes in Computer Science)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.), Vol. 12036. Springer, 35–42. https://doi.org/10.1007/978-3-030-45442-5_5

[16] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 689–698. https://doi.org/10.1145/3178876.3186150

[17] Feng Lu and Nava Tintarev. 2018. A Diversity Adjusting Strategy with Personality for Music Recommendation. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, co-located with ACM Conference on Recommender Systems (RecSys 2018)*.

[18] Robert R. McCrae and Oliver P. John. 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality* 60 (June 1992), 175–215. Issue 2. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[19] Orestis Nalmpantis and Christos Tjortjis. 2017. The 50/50 Recommender: A Method Incorporating Personality into Movie Recommender Systems. In *Engineering Applications of Neural Networks*, Giacomo Boracchi, Lazaros Iliadis, Chrisina Jayne, and Aristidis Likas (Eds.). Springer International Publishing, Cham, 498–507.

[20] Xia Ning and George Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu (Eds.). IEEE Computer Society, 497–506. https://doi.org/10.1109/ICDM.2011.134

[21] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

[22] Peter J. Rentfrow and Samuel D. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256.

[23] Markus Schedl. 2017. Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *Int. J. Multim. Inf. Retr.* 6, 1 (2017), 71–84. https://doi.org/10.1007/s13735-017-0118-y

[24] Markus Schedl, David Hauger, Katayoun Farrahi, and Marko Tkalcic. 2015. On the Influence of User Characteristics on Music Recommendation Algorithms. In *Advances in Information Retrieval - 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings (Lecture Notes in Computer Science)*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.), Vol. 9022. 339–345. https://doi.org/10.1007/978-3-319-16354-3_37

[25] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3251–3257. https://doi.org/10.1145/3308558.3313710

[26] Aaron Swartz. 2002. Musicbrainz: A semantic web service. *IEEE Intelligent Systems* 17, 1 (2002), 76–77.

[27] Marko Tkalcic and Li Chen. 2015. Personality and recommender systems. In *Recommender systems handbook*. Springer, 715–739.

[28] Hsin-Chang Yang and Zi-Rui Huang. 2019. Mining personality traits from social messages for game recommender systems. *Knowledge-Based Systems* 165 (2019), 157–168. https://doi.org/10.1016/j.knosys.2018.11.025

[29] Eva Zangerle, Martin Pichl, Wolfgang Gassler, and Günther Specht. 2014. # nowplaying music dataset: Extracting listening behavior from twitter. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*. 21–26.