PERSONALITY BIAS OF MUSIC RECOMMENDATION ALGORITHMS

Alessandro B. Melchiorre[1,2], Eva Zangerle[3], Markus Schedl[1,2]

[1]Institute of Computational Perception, Johannes Kepler University [2]Linz Institute of Technology (LIT)
[3]University of Innsbruck

JOHANNES KEPLER UNIVERSITY LINZ
Institute of Computational Perception

Introduction

Previous studies have shown that Recommender Systems may underserve specific user groups based on sensitive aspects, e.g. provide better or worse recommendations depending on the user's *age* or *gender*. In our paper, we study this bias from the perspective of the user's **personality**. Focusing on the music domain we ask the following question:

Do state-of-the-art recommender system algorithms treat users differently depending on their personalities?

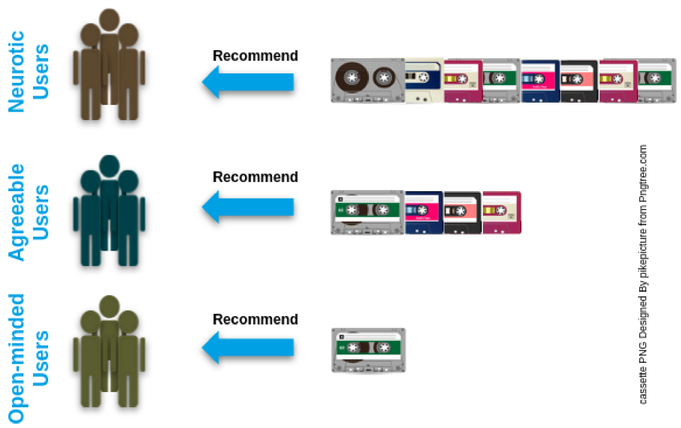

Fig. 1: Depending on their personalities, users may receive better or worse recommendations.

Personality User Groups

We employ the Big Five model to assess user's personalities which measures personality over the following five traits: **Openness**, **Conscientiousness**, **Extraversion**, **Agreeableness**, **Neuroticism**. For each trait, a user is assigned to either the **High** or **Low** group whether their score is higher or lower than the median for that trait. We then compute the recommendation quality measures for the two groups over each trait and inquiry whether they are different or not.

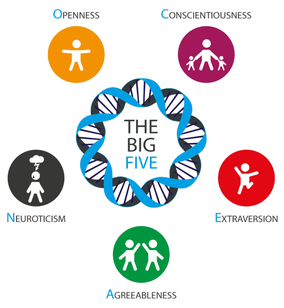

Fig. 2: The Big Five Personality Traits

Data

To obtain behavioural data on music consumption and the users' personality, we exploit **microblogs shared on Twitter**. We fetch the tweets shared by users with the #nowplaying hashtag in the years 2018-2019 and extract the corresponding music track. For each user, we use their most recent tweets to extract their personality scores from text using the IBM Personality Insight tool.[1] Below we show the statistics of the dataset and the user groups.

General Stats	# listening events		# tracks	# users
	395,056		15,753	18,310
Trait	# unique tracks/user (mean and std.)		# listening events	
	High	Low	High	Low
Agr.	19.1 ± 24.4	17.3 ± 21.7	208,054	187,002
Con.	19.2 ± 25.5	17.2 ± 20.4	206,179	188,877
Ext.	20.0 ± 26.3	16.4 ± 19.2	217,895	177,161
Neu.	16.2 ± 18.4	20.3 ± 26.9	177,892	217,164
Ope.	19.5 ± 24.9	16.9 ± 21.1	209,741	185,315

Table 1: Statistics of the dataset and of the user groups.

Methodology

We investigate the following algorithms:

- Embarrassingly Shallow Autoencoders (**EASE**)
- Sparse Linear Models (**SLIM**)
- Variational Autoencoders for Collaborative Filtering (**Mult-VAE**)

Recommendation quality is evaluated through **Recall@K** and **NDCG@K** for K = {5, 10, 50}. All experiments are repeated across 10 different seeds. To assess unequal treatment of the High vs. Low groups for each trait, we use the **two-tailed Mann-Whithney-U** test and test for significance at difference alpha levels.

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

Results

The results for NDCG@K are shown below. Significant differences are highlighted and shown with the respective alpha values. Results for Recall@K show similar tendencies. Noteworthy observations:

- **Highly neurotic and highly agreeable** users receive better recommendation compared to low-neurotic and low-agreeable users, while the opposite is true for the other traits.
- The **Openness and Neuroticism** traits show the most significant differences between the High and Low groups.
- For **Agreeableness and Conscientiousness**, not all algorithms show unequal treatments of the users.
- The algorithms tend to agree on the direction of the bias (e.g. worse recommendation for the High group).

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

Table 2: Significant differences between high and low groups are marked in bold and with an asterisk (Mann-Whitney-U test, * $p < .05$, ** $p < .01$, *** $p < .001$).

Conclusion and Future Work

In our work, we tested if state-of-the-art recommender system algorithms (EASE, SLIM, Mult-VAE) treat users differently depending on their personality.

We found highly significant differences ($p < .001$) in both NDCG@K and Recall@K for the traits neuroticism and openness as well as significant differences at $p < .01$ and $p < .05$ for the other traits.

As for directions for future research, we contemplate to investigate the origin of these biases, generalize our results to other social platforms and additional algorithms, and include beyond-accuracy metrics such as diversity, serendipity, and familiarity.

[1] https://www.ibm.com/watson/services/personality-insights/

alessandro.melchiorre@jku.at

https://github.com/CPJKU/pers_bias