Improving Recommendation Systems with User Personality Inferred from Product Reviews

Xinyuan Lu[1,2]   Min-Yen Kan[2]

[1]Integrative Sciences and Engineering Programme (ISEP), NUS Graduate School
[2]School of Computing, National University of Singapore, Singapore

luxinyuan@u.nus.edu   kanmy@comp.nus.edu.sg

ABSTRACT

Personality is a psychological factor that reflects people's preferences, which in turn influences their decision-making. We hypothesize that accurate modeling of users' personalities improves recommendation systems' performance. However, acquiring such personality profiles is both sensitive and expensive. We address this problem by introducing a novel method to automatically extract personality profiles from public product review text. We then design and assess three context-aware recommendation architectures that leverage the profiles to test our hypothesis.

Experiments on our two newly contributed personality datasets — *Amazon-beauty* and *Amazon-music* — validate our hypothesis, showing performance boosts of 3–28%. Our analysis uncovers that varying personality types contribute differently to recommendation performance: *open* and *extroverted* personalities are most helpful in music recommendation, while a *conscientious* personality is most helpful in beauty product recommendation.

CCS CONCEPTS

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

KEYWORDS

Recommendation Systems, Psychology, Personality, Review Texts

1 INTRODUCTION

Online recommendation systems are algorithms that help users to find their favorite items. In recommendation systems, the user's

profile is important as people with different ages, educational backgrounds exhibit different preferences. Besides static attributes such as gender, the user's psychological factors, especially personality, can be viewed as a user's dynamic profile are also vital in recommendations.

People with similar personalities are more likely to have similar interests and preferences [23]. Therefore, accurate modeling of the user's personality plays a vital role in recommendation systems. For example, in movie recommendation, an outgoing person may favour watching comedic movies over romantic ones [23]. Other studies [11] have shown that in music recommendation, a user's degree of openness strongly determines their preference for energetic music genres. These examples show that personality traits can influence users' preferences.

While we can see that personality traits motivate users' preferences, there are challenges that need to be solved before one can utilize the traits in the recommendation. First, collecting personality data is time-consuming. The current best practice for collecting personality data requires conducting a user study via an ethically-cleared questionnaire with informed consent. Subsequent training of assessors is also needed. The entire collection process can take months [19] and also be an expensive process in terms of effort.

Second, processing personality data raises sensitivity and privacy concerns. If handled incorrectly, such data can be misused by users intentionally, resulting in a violation of privacy protection policies and biased performance of the recommendation systems. For example, a scandal emerged when a Facebook app illegally collected 87 million users' personality information to manipulate their voting choices in the U.S. presidential election in March 2018 [9]. Such risks make the balance between collecting and utilizing users' personality information challenging. This issue has stalled progress in this emerging field of research.

Due to the problem above, the third challenge is a lack of personality-grounded datasets in the existing work. One notable exception is the website *myPersonality*[1], which contained personality data and the likes of Facebook users. However, in 2018, *myPersonality*'s founders decided to discontinue the project as "complying with various regulations [had] become too burdensome". To the best of our knowledge, there are thus few datasets suitable for testing the effect of personality factors in recommendation systems research.

In this study, we explore methods to overcome these challenges discussed above and contribute to personality-based recommendation research. We identify a new source for inferring a user's personality traits: user-generated content, specifically e-commerce review texts. Studies show that review texts can reflect a user's personality since individuals manifest their personality through

---

[1]https://sites.google.com/michalkosinski.com/mypersonality

their choice of words [17]. For example, when showing their dislikes on the same shampoo, an agreeable person may comment *"I bought this shampoo for my husband. The smell is not good."*, while a neurotic and aggressive person might comment *"Arrived opened and leaking all over the box. Tried shampoo but didn't help at all. Still so itchy!!!"*. In addition, review text is easy to obtain and made publicly available by users in full disclosure on online commercial websites, which helps to solve both time and privacy issues.

In our experiments, we explore the possibility of automatically inferring users' personality traits from their review texts and then use this information to help recommendations. We do this by leveraging an Application Programming Interface (API) to automatically analyze the user's personality. There already exist deployed, production-level APIs for automatic personality detection, such as IBM Personality Insights[2], Humantic AI[3], and Receptiviti[4] that purport to yield personality profiles. In our work here, we use the Receptiviti API, because it is a widely-validated and widely-used psychology-based language analysis platform for understanding human emotion, personality, motivation, and psychology from language. Receptiviti's API outputs scores for the commonly-used OCEAN personality model: five values, one for each of the five personality aspects of Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism (each corresponding to one letter of "OCEAN"). Finally, this inferred personality is fed as input to a recommendation system, to test whether it can improve recommendation performance.

To conduct our study, we first construct two new datasets extending from an existing Amazon review dataset, in the beauty and music domains. We first extract the user reviews that are between 30 to 80 words. Afterward, we concatenate all the valid review texts of each user and input their concatenation to the Receptiviti API to output each user's inferred personality scores. As a quality check, we evaluate the accuracy of personality detection, by plotting the personality distribution for each dataset. We observe the users with extremely high/low personality scores and find that these are reliable indicators of personality, and use the such confidently labeled output as ground truth (silver data).

We incorporate these personality scores into the recommendation process and investigate their effect on current neural-based recommendation systems. We observe consistent improvements on the performance of such recommendation systems. When we consider different personality groups, we find that *extroversion* and *agreeableness* benefit the recommendation performances across all the domains. However, we lack an in-depth understanding of how these personalities affect recommendations and users' behavior. This points to future directions in utilizing other auxiliary information to infer users' personality traits, *e.g.*, users' browsing histories.

In summary, our contributions are:

- We construct two new datasets in the music and beauty domains that combine users' public product reviews alongside automatically inferred personality scores. This directly

addresses the lack of personality-based datasets for recommendation, while avoiding privacy issues by using public data.
- We conduct empirical experiments over these datasets, finding that leveraging personality information indeed improves the recommendation performance, from 3% to 28%.
- We analyze the influence of personality traits in these domains and find the personality traits of extroversion and agreeableness improve the recommendation performance across all domains.

2 RELATED WORK

The current study investigates how to extract personality traits from texts and how personality traits can be utilized in recommendation systems. Therefore, the review below focuses on the literature that discusses personality detection and personality-based recommendation systems. We first give an introduction of the OCEAN personality models (Section 2.1) before reviewing two topics related to our work: personality detection (Section 2.2) and personality-based recommendation systems (Section 2.3).

2.1 The OCEAN Model

Personality involves a pattern of behavior that is not likely to change over a short period of time [1]. It can be detected either explicitly by a questionnaire or implicitly by observing user behaviors [12]. The most commonly-used model describing personality traits is the OCEAN model [12], which we use to model a user's personality traits. The five fundamental personality dimensions defined by OCEAN are:

(1) **Openness to Experience (O)**, which describes the breadth and depth of people's life, including the originality and complexity of their experiences. Individuals with high openness tend to be knowledgeable, analytical, and more investigative.
(2) **Conscientiousness (C)** This trait involves how individuals control, regulate and direct their impulses. For example, highly conscientious people are usually cautious.
(3) **Extroversion (E)** Extroversion indicates how much people are in touch with the outside world. Extroverts are more willing to talk to others about their thoughts.
(4) **Agreeableness (A)** This trait reflects individual differences and social harmony in cooperation. Highly agreeable people are more willing to share tasks than to complete tasks independently.
(5) **Neuroticism (N)** This refers to the tendency of experiencing negative emotions. People with high neuroticism are often in a bad mood, therefore they prefer to respond emotionally.

2.2 Personality Detection

There are two common ways to measure a person's personality traits using a personality model: personality assessment questionnaires and automatic personality detection.

*2.2.1 **Personality Assessment Questionnaires**.* Self-reporting personality questionnaires are commonly used to reveal personality differences among individuals. Responses to questions usually take the form of a five-point Likert scale (*strongly agree, agree, disagree,*

and *strongly disagree*). Such personality inventories differ with respect to the number and content of their questions. Common long questionnaires include the NEO Five-Factor Inventory (60 items) [2], NEO-Personality-Inventory Revised (240 items) [6], and the Big-Five Inventory (BFI, 44 items) [18]. Practitioners prefer using shorter instruments, such as the BFI-10 and Ten-Item Personality Inventory (TIPI) [7, 20], as they are time-saving and easier to fill out.

However, using questionnaires for self-report has two major drawbacks. First, questions that assess personality are often quite subjective such as "Do you easily get nervous?". Answers for such questions are easily affected by a self-bias [16] or reference-group effects [24]. For example, an introverted engineer might think he is an extrovert if he/she is working with a group of individuals that may be more introverted. Consequently, the results of questionnaires are often hard to reproduce. Second, assessing a personality by questionnaires can be inconvenient, as the subjects are necessary to participate in the studies.

2.2.2 **Automatic Personality Detection**. To make personality detection more convenient and reproducible, practitioners prefer automated personality detection, which infers a personality type based on user data. Such methods are less accurate than personality questionnaires — as it relies on the input user data manifesting personality traits — but has the advantage of not requiring inputs to be answers to questionnaires. For example, social media posts that exhibit opinions and viewpoints are a prime source of text data useful for personality detection. Individuals have different language use behaviors that reveal personality traits [10]. Automatic, text-based personality detection infer users' personalities by analyzing users' word choice (lexical selection) and sentence structure (grammatical selection). Such technology has been sufficiently proven, making them commonplace and deployed at scale in production, and available as a service through cloud-based application APIs.

We study whether knowing users' personality information can lead to better recommendations, and also, how users' personality information be best modeled in recommendation systems to improve performance. While large-scale recommendation datasets exist, they universally lack users' personality information. It is infeasible to ask to obtain this information via questionnaires since the identity of users is usually confidential. Therefore, we utilize automatic personality detection to infer personality from product reviews written by users. Product reviews are ideal: they are widely available on online commercial websites, they often demonstrate personality traits, and they are public (the texts are meant to be read by others). Hence, they can serve as a good source for automatically detecting personality in an economic but accurate way.

2.3 Personality-based Recommendation Systems

Since personality traits are characteristics that do not change sharply over time and do not depend on a certain context or stimulus, they are more easily used to create personalized recommendation systems. Earlier work by Winoto and Tang [21] [4] focused on extending Matrix Factorization by adding a personality latent factor. Their model used implicit feedback data, such as user–item interactions, beyond just ratings. They only considered the OCEAN scores as one attribute, so the effects that are attributable just to personality are

not clear. Besides, personality traits have been used to determine a neighborhood of similar users by calculating personality similarity. Thus, for example, Asabere *et al.* [3] proposed a recommendation system for attendees of conferences that integrates personality traits and social bonds of attendees.

In their work, user similarity was equated as personality similarity, calculated by Pearson's correlation between two users' OCEAN scores. They demonstrated that the system accuracy improves with a larger user base, due to the higher likelihood of finding other users with similar personalities (high correlation).

Researchers have also associated user personality scores with items. Yang and Huang [23] attributed items (here, computer games) with a personality that is an average of its users. This latent representation can then be used to recommend items to users with a similar personality as that of the other users of that item. This may make sense when certain items are used primarily by certain personality types (as in computer games) but are less compelling for items that may be used by many personality types. Lastly, in social media recommendations, Wu *et al.* [22] proposed an approach for recommending interest groups by integrating personality. The personality-based similarity was defined as the Euclidean distance between two users' personality scores. However, it combines the personality signal linearly in the recommendation process, which we feel may be limiting.

In summary, compared with other context attributes (e.g., purchase history), personality information helps to capture the users' potential interests rather than recommending a similar purchased item. However, the previous works used the OCEAN personality as a linear similar score which lacks the capability of capturing more nuanced information latent in personality scores. Different from the methods above, we propose two novel ways of adding personality features into the recommendation system: 1) taking the most salient personality trait as a learnable vector and 2) calculating a user's personality embedding as a weighted sum of a user's OCEAN personality features, which is a learnable embedding within the recommendation system.

3 DATASET CONSTRUCTION

We construct two new datasets, as extensions of the existing, well-known Amazon review dataset. We first automatically infer users' personality traits from users' review texts as review texts can reflect the personality through word usage. They are also publicly-available text on online commercial websites, allowing researchers to have legal access to textual data where experimentation can be replicated. Based upon the parent Amazon review dataset, we construct two new domain-specific datasets: an *Amazon-beauty* and an *Amazon-music* dataset. These contain Amazon reviews of products in the beauty and music domains, alongside their posting users' inferred personality scores.

3.1 Data Source

The Amazon dataset[5] [15] is widely used for training and evaluating recommendation systems. It contains a large number of item descriptions, ratings, and product reviews collected from the Amazon online commercial website. The Amazon dataset is divided

---

[5]https://nijianmo.github.io/amazon/index.html

according to the domain. In our study, we choose two domains: *beauty* and *music*. We construct datasets separately for these two domains since we want to study whether personality has different influences on users' behaviours for different domains. Studies have shown that people with different personalities prefer different kinds of music [11]. For example, people with a high degree of openness like to listen to rock music, while neurotic people like jazz. Therefore, we choose *music* as one of the domains to be studied. In order to study the role of personality in different domains, we randomly select *beauty* for comparison. Table 1 shows a sample of the original Amazon dataset, which contains the user (*reviewerID*, *reviewerName*), the product's Amazon Standard Identification Number (*asin*), the review text for the product (*reviewText*), and the overall rating given to the product (*overall*).

reviewerID	A2SUAM1J3GNN38
asin	0000013714
reviewerName	J.McDonald
vote	5
style	Format:Hardcover
reviewText	I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read because we think the book was published for singing from more than playing form. Greate purchase though!
overall	5.0

Table 1: An example of Receptiviti score for a specific, anonymized user.

3.2 Dataset Construction

Since we do not know the personality for each user in the Amazon dataset, we need to infer them. We first retrieve each user's review texts and then use the Receptiviti API[6], a computational language psychology platform for understanding human behavior, to infer a personality. The API can take a long piece of human-written text (more than 300 words), and output a faceted personality score with 35 factors, including OCEAN scores.

For each user that wrote reviews in either of the two domains, we collect all his/her review texts and concatenate them together into a single document. Afterward, we send the concatenated text to Receptiviti to infer a personality. We select the personality scores corresponding to the five-dimensional personality traits defined in the OCEAN model [12] (Table 2). Each personality score is normalized to a range from 1 to 100. The higher the score, the more overt the personality trait. Note that each of the five OCEAN scores is independent of the other.

User ID	AGR	CON	NEU	EXT	OPEN
A2GBIFL43U1LKJ	54.05	34.87	25.96	54.39	42.71

Table 2: An example of Receptiviti score for a specific, anonymized user.

---
[6]https://www.receptiviti.com/

To improve the personality prediction process, we only analyze the personality traits for *active* users who bought many products and wrote a sufficient number of product reviews. To be specific, we select users that 1) wrote product reviews for at least 10 different items they purchased, and where 2) each product review contains between 30 to 80 words. Table 3 shows the statistics after the filtration. For example, using these criteria, 1,791 *active* users are selected for the *Amazon-music* dataset. Each user in the *Amazon-music* dataset has an average of 990.48 review words over all of his/her reviews, averaging 51.01 words for each review.

3.3 Dataset Statistics

Aside from our constructed *Amazon-beauty* and *Amazon-music dataset*, we also include an existing dataset *Personality 2018* in our study. Personality 2018[7] [14] is a version of the MovieLens dataset that includes each user's personality information obtained through questionnaires. It contains 21,776 movies, 339,000 ratings, and 678 users with the OCEAN personality questionnaire scores from 1 to 7. This dataset is included to study the difference between questionnaire-based personality trait scores with our review-based automatic personality trait detection scores.

Table 3 shows the final statistics of the datasets used in our study. We can observe that the *Amazon-beauty* / *Amazon-music* dataset has the largest / smallest percentage of interactions. The *Personality2018* dataset contains the largest number of items and the smallest number of users. We can see that these datasets differ in domains, number of users, items, and interactions, which facilitates the study of personality-based recommendation across a wide spectrum of settings.

Dataset	Amazon-beauty	Amazon-music	Personality'18
# of items	85	8,895	21,776
# of users	991	1,791	678
# of ratings	5,269	28,399	339,000
# of interactions	6.26%	0.18%	2.30%
Avg. words/user	990.48	466.43	-
Avg. words/review	51.01	51.18	-

Table 3: Statistics of the three datasets used in our study.

4 METHODS

Based on our constructed dataset, we conduct experiments to study whether the recommendation system can benefit from incorporating personality traits. We choose the Neural Collaborative Filtering (NCF) [8] as the foundation model of our study because it is the fundamental neural-based model for the recommendation. Specifically, we design a personality-enhanced version of NCF [8] to compare with the vanilla NCF, alongside several other baselines.

4.1 Neural Collaborative Filtering (NCF)

NCF [8] is the first deep-learning-based recommendation algorithm. Different from traditional collaborative filtering algorithms, the model encodes the user and item into latent vectors and then

---
[7]https://grouplens.org/datasets/personality-2018/

projects them through a Multi-layer Perceptron (MLP) to predict a probability score, representing the probability that a user would buy a target item. In our implementation, we use a 4-layer MLP and a 16-dimensional user and item embedding.

4.2 Personality-enhanced NCF

We then propose three different ways to incorporate the personality information into the NCF model, as shown in Fig. 1. We first design *NCF+Most salient Personality* model by adding the most salient personality trait as input into NCF. We also design *NCF + Soft-labeled Personality* and *NCF + Hard-coded Personality* to incorporate all the five personality traits of OCEAN. The difference between the two latter versions is that the personality vector in *NCF + Soft-labeled Personality* is learnable, while in *NCF + Hard-coded Personality*, the vector is predetermined and fixed.
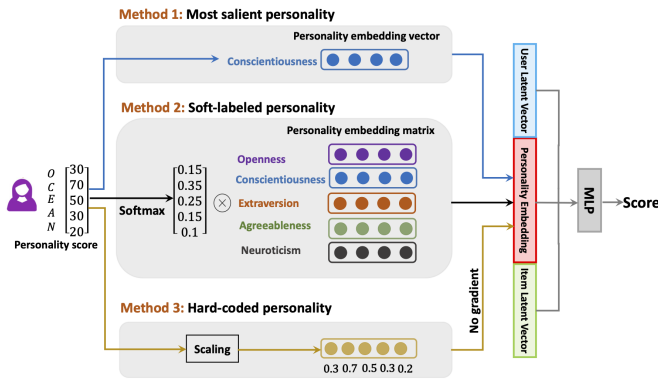

Figure 1: The overall structure of our model. In this example, the user's OCEAN score is {30,70,50,30,20}. TheNCF + Most salient personality selects the personality with the highest score, i.e., *conscientiousness* as the personality embedding vector. NCF + Soft-labeled personality takes all five OCEAN scores as a personality embedding matrix. NCF + Hard-coded personality predetermines and fixes the personality vector as {0.3,0.7,0.5,0.3,0.2}

1. **NCF + Most Salient Personality.** In this model, we introduce a 4-dimensional *personality vector* for each of the five types of personalities, which are learned during training. We treat the most salient personality as the user's personality label and concatenate the corresponding personality vector with the user's latent vector.

2. **NCF + Soft-labeled Personality.** In this model, we make full use of all five personality trait scores. We first apply a *Softmax* function to map the personality scores into a probability distribution of personality. Afterward, the probability distribution is used as the weight to calculate the weighted sum of the five personality vectors. The output vector is then concatenated with the user's latent vector as the input of the MLP.

3. **NCF + Hard-coded Personality.** This model also considers all the user's five personality traits information. However, instead of introducing learnable personality vectors, we directly scale each

personality score to sum to a unit value (here, 100) to get a hard-coded 5-dimensional vector to represent the user's personality information. This vector is concatenated with the user's latent vector, but is fixed during training.

5 EXPERIMENTS

We evaluate our proposed method on our three datasets by answering the following four research questions. We first evaluate whether we can accurately detect personality from texts (RQ1, Section 5.1). Afterward, we analyze the distribution of personality in review texts (RQ2, Section 5.2). Then, we explore whether adding personality information can improve recommendation performance (RQ3, Section 5.3). Finally, we analyze the influence of personality information on different domains (RQ4, Section 5.4).

5.1 Can we accurately detect personality from texts? (RQ1)

To evaluate whether we can accurately detect personality traits from texts, we analyze the personality scores inferred by the Receptiviti API for each user. Since there are over 2,500 users in total in our two constructed datasets, it is time-consuming to manually evaluate them all. As a compromise, we choose to manually examine the users that receive extremely high scores for certain personality traits. We believe those examples are more easily evaluated by humans. Specifically, for each personality trait, we select the users that receive the top 10 highest scores on this type. We analyze both the *Amazon-beauty* and the *Amazon-music* datasets, resulting in a total of 100 samples. These samples are evaluated by two graduate students. Both were trained with a detailed explanation of the OCEAN personality model. We ask them to choose whether the sampled review texts accurately match their inferred personality, choosing between three options of *yes*, *no*, or *not sure*. We then calculate the accuracy of the samples and the inter-annotator agreement between the two annotators using Cohen's Kappa [5]. We find that the inferred personality matches with the review text in 81% of the *Amazon-beauty* samples, and 79% of the samples from *Amazon-music*. The average Cohen's Kappa is 0.70. We take this to indicate that the Receptiviti API can indeed infer users' personality traits from review texts with generally high accuracy.

Table 4 shows examples of review texts with their inferred personality scores. We observe that people with different personalities have different language habits. For example, extroverts tend to use the words "love" and exclamation marks because they are characterized by a strong tendency to express their affection. People who are agreeable are usually bought items for other people, *e.g.,* "my kids" and "my wife", perhaps due to their inclusiveness. Conscientious people usually talk about their own experience and feelings before recommending the items to others, *e.g.,* "I have had this shower gel once before" or "Don't just take my word for it". This is perhaps because they are usually cautious.

5.2 What is the distribution of users' personalities? (RQ2)

We further analyze the personality distribution for all users by plotting the score histograms for each personality trait in the *Amazon-beauty* dataset and the *Amazon-music* dataset in Fig. 2.

Personality label	Personality Score	Review Texts
Openness	63.07	Near perfect exfoliating gloves my only complaint is a matter of preference rather than product defect. I prefer the harder surface area to use on round areas of the body or potentially harder like the feet, elbows, etc.
Openness	62.62	Azur is always my favorite in the Thymes collection because of its clean, fresh scent. I like that my skin feels moisturized when using this product in the shower.
Conscientiousness	75.38	I have had this shower gel once before, and it's amazing. Hard to find, too. One of The Body Shop's best scents, and it's usually only available seasonally! wish they sold it in bigger bottles, but I was happy to find it.
Conscientiousness	71.02	Don't just take my word for it, you must try it. A dear friend got me this from Italy 12 years ago and has been using it since, very hard to find it in the US. This shower cream will transform your shower experience.
Extroversion	75.06	Love this shampoo! Recommended by a friend! The color really lasts!!!
Extroversion	72.90	Looked all over to find where to purchase this product and we are very happy to be able to finally find it. The PRELL Conditioner is by far the best you can buy. We love it!!
Agreeableness	80.06	Great product - my wife loves it
Agreeableness	78.18	Great deal and leaves my kids smelling awesome! I bought a box of them years ago and we still have some left!!! Great deal and leaves my kids smelling awesome!
Neuroticism	67.81	Too expensive for such poor quality. There was no improvement and I am starting to think my scalp is worse off than it was before I started using this product. I do agree with other reviews that it feels watered.
Neuroticism	62.28	Nope. It smells like artificial bananas, and this smell does linger. It's pure liquid, there is no thickness to it at all, it's like pouring banana water on your head that lathers. It does not help with an itchy scalp either.

Table 4: The data sample of extreme personality cases to the annotators. Each data sample contains the user's personality labels, personality scores, and review texts.

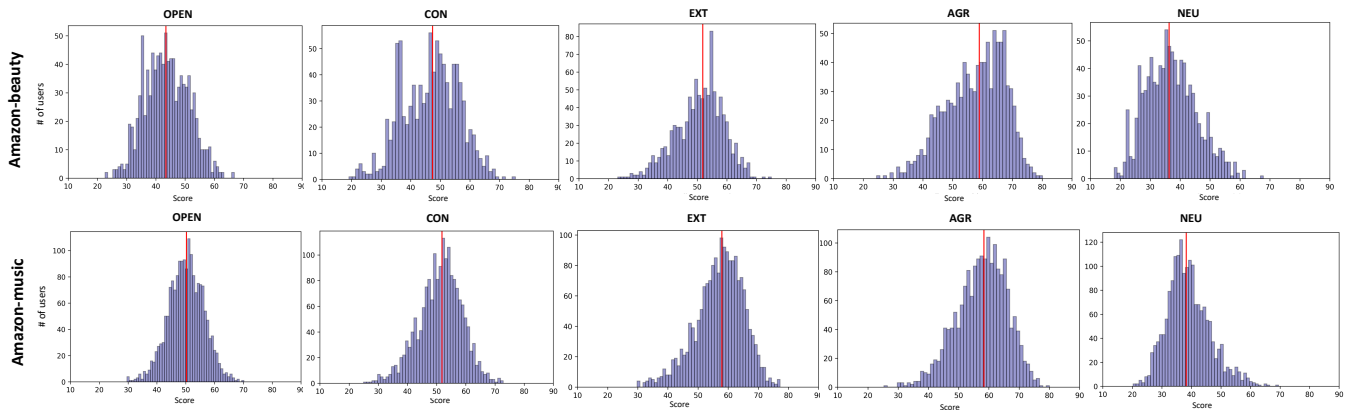**Figure 2: Distribution of personality traits in *Amazon-beauty* and *Amazon-music* datasets. The x-axis represents the score for each trait; the y-axis represents the number of users. The red line represents the median for each trait.**

We observe a similar trend in both domains: agreeable people have the highest median score, and neurotic people have the lowest median score. A possible reason is that neurotic people are more

introverted and are less likely to publish their opinions publicly, while agreeable people are more willing to share their thoughts. Another observation is that the personalities of each dataset are generally bell-curved. This indicates that each personality trait is normally distributed.

We also examine the difference in personality distributions between the two domains. In the *Amazon-music* dataset, the average scores of extroversion and openness are higher than those in the *Amazon-beauty* dataset. This indicates that the personality characteristics of extroverts and open people are more obvious in the music domain than in the beauty domain.

From the above figures, we draw the following conclusions. First, the personality traits of users are not evenly distributed. There are more instances of people with certain personality traits (*e.g.,* agreeableness) than others (*e.g.,* neuroticism). A possible reason is that people with certain personalities are more willing to write product reviews. 2) The distributions for the two domains are generally the same, with higher agreeable scores and lower neurotic scores. However, there is a slight difference. For example, the scores of extroverts in music are generally higher than that in the beauty domain. This could be explained by the possibility that people who are passionate about music may be more emotional.

5.3 Does incorporating personality improve recommendation performance? (RQ3)

Next, we want to explore whether adding the induced user personality benefits the recommendation quality. To this end, we compare the personality-enhanced NCF with the following two baseline models that do not utilize personality information.

1. **NCF with random personality (NCF + Random).** We randomly assign each user with a random personality label, regardless of his/her original, inferred personality scores.

2. **NCF with same personality (NCF + Same).** We assign each user to a single personality trait. To be specific, we assume every user is "open" and assign the corresponding personality vector to NCF. Although the personality vector does not provide any additional signal to the model in this case, it can serve as a placeholder to keep the network structure identical to the personality-enhanced model, resulting in a fair comparison.

Evaluation Metrics. We use two metrics to measure the performance of our proposed recommendation models: *Hit Rate (HR) @ K* and *Normalized Discounted Cumulative Gain (NDCG) @ K (K = 3, 5, 10)*. Larger HR and NDCG demonstrate better accuracy.

Experiment Results. Table 5 shows the experimental results in the *Amazon-beauty* and the *Amazon-music*, and *Personality2018* datasets, respectively. In *Amazon-beauty* and *Amazon-music*, we find that the three personality-enhanced NCF models outperform the two baseline models, in terms of both NDCG and HR. Especially, the first three rows show that the NCF with the most salient personality label outperforms NCF with the same or random personality label. This indicates that adding personality information into NCF improves recommendation performance. From the last three rows, we further find that NCF + Soft-labeled/Hard-coded outperforms NCF + Most salient personality in terms of NDCG. This shows that utilizing all five personality traits is better than using the most salient personality trait in NCF.

In the *Personality 2018* dataset, the trend in the *Amazon-beauty* and *Amazon-music* also holds for it. For example, the NCF + Soft-labeled model outperforms the other models, showing that adding personality information improves performance. However, the improvement in the *Personality 2018* is less obvious than that in the *Amazon-beauty* dataset. We hypothesize the reason might be due to the difference in the sizes of the datasets. Since *Amazon-beauty* is a small dataset, adding personality information may better help to address the data sparsity problem, therefore exhibiting a better performance gain.

5.4 How does personality information improve the performance of recommendation system? (RQ4)

To gain a better understanding of the improvement brought by incorporating personality, we separately evaluate the HR and NDCG for the five personality traits, as shown in Table 6 . "+" represents the NCF+Soft-labeled model (with personality information), and "-" represents the NCF+Same model (without personality information). We make two major observations.

First, the improvement brought by adding personality is prominent for the *Amazon-beauty* dataset, over all five personality traits. In particular, the trait of conscientiousness (CON) has the highest gain in terms of both HR (+21%) and NDCG (+57%). However, in the *Amazon-music* dataset, *openness* (+27%), *agreeableness* (+10%), *extroversion* (+5%) improve while *neuroticism* (−18%) and *conscientiousness* (−12%) decreases.

Second, for the *Personality2018* dataset, the improvement brought by adding personality is not obvious: only *conscientiousness*, *extroversion*, and *agreeableness* have shown minor performance gain. From the above breakdown analysis, we find that adding personality information can benefit certain personality traits better than others. However, the personality trait that improves the most differs greatly across the three datasets. This indicates that although improvements are observed in terms of empirical results, the mechanism of how personality influences the recommendation still deserves more in-depth investigation.

6 DISCUSSION

In this work, we make a preliminary attempt to explore how to automatically infer users' personality traits from product reviews and how the inferred traits can benefit the state-of-the-art automated recommendation processes. Although we observe that recommendation performance is indeed boosted by incorporating personality information, we believe there are several limitations. In the following, we discuss these limitations with potential future directions.

First, we believe capturing personality from the review texts may lead to selective bias. Introverts are less likely to share their thoughts online while extroverts are more likely to share experiences. This results in an imbalanced personality distribution in our collected data. As shown in the analysis in RQ2 (Section 5.2), extroversion is the most common personality trait of users in our datasets. To address this, in future works, we could utilize other context information to infer users' personalities such as a user's purchase history. Such user behaviours can also reflect personality;

Algorithms	Amazon-beauty						Amazon-music						Personality2018					
Rating	H@3	H@5	H@10	N@3	N@5	N@10	H@3	H@5	H@10	N@3	N@5	N@10	H@3	H@5	H@10	N@3	N@5	N@10
NCF+Random	0.923	0.965	0.975	0.675	0.605	0.660	0.159	0.224	0.339	0.117	0.143	0.171	0.510	0.628	0.777	0.406	0.454	0.504
NCF+Same	0.918	0.967	0.975	0.683	0.630	0.662	0.160	0.224	0.340	0.122	0.149	0.167	0.511	0.622	0.777	0.403	0.454	0.502
NCF+Most-Salient	0.939	**0.969**	**0.977**	0.714	0.676	0.707	0.156	0.226	0.343	**0.164**	0.145	0.174	0.516	0.631	0.795	0.415	0.463	0.511
NCF+Soft-labeled	0.936	0.965	0.973	0.810	**0.867**	0.831	0.156	0.225	**0.348**	0.113	0.141	0.175	**0.528**	**0.656**	**0.805**	**0.421**	**0.471**	**0.511**
NCF+Hard-Coded	**0.948**	0.961	0.977	**0.849**	0.826	**0.848**	0.175	0.232	0.345	0.147	**0.160**	**0.189**	0.503	0.622	0.758	0.398	0.447	0.498

Table 5: Hit Rate(H) and NDCG(N) @K in the *Amazon-beauty, Amazon-music,* and *Personality 2018* datasets. The best performance is bolded.

Trait		Amazon-beauty		Amazon-music		Personality2018	
		HR	NDCG	HR	NDCG	HR	NDCG
OPEN	+	**0.833**	**0.729**	**0.330**	**0.205**	**0.535**	0.420
	-	0.750	0.545	0.313	0.161	0.547	**0.422**
CON	+	**0.883**	**0.769**	0.228	0.132	**0.475**	0.358
	-	0.727	0.490	**0.279**	**0.150**	0.441	**0.361**
EXT	+	**0.970**	**0.882**	**0.319**	**0.181**	**0.611**	**0.412**
	-	0.872	0.600	0.317	0.169	0.556	0.411
AGR	+	**0.968**	**0.878**	**0.332**	**0.198**	**0.621**	**0.512**
	-	0.864	0.593	0.308	0.185	0.552	0.430
NEU	+	**0.933**	**0.835**	0.397	0.230	0.489	0.390
	-	0.833	0.536	0.397	**0.254**	**0.511**	**0.415**

Table 6: HR and NDCG results group by 5 personality traits in *Amazon-beauty, Amazon-music,* and *Personality2018* datasets. "+" represents the NCF+Soft-labeled model (with personality information), and "-" represents the NCF+Same model (without personality information).The best performance is in bold.

for example, open people are more likely to follow popular trends which can be reflected in their purchase history.

Second, we only conduct experiments on a single basic model, NCF, which may loss of generalization. More advanced models graph recommendation models can be used in the future. Third, we conduct empirical experiments on whether personality information benefits recommendation. However, more in-depth investigation is necessary on how personality affects recommendation and users' behavior. In the future, we could conduct a user study to find the causal relationship between personality and recommendation. To be specific, we can develop different marketing strategies for users with different personalities. By observing the effects of different strategies on users' behavior, we can gain a better understanding of how personality affects recommendation. Fourth, we find that the *openness*, *conscientiousness* and *neuroticism* features do not have a noticeable impact on the recommendation performance. A possible reason is that OCEAN only contains five types of personality, which might be insufficient to provide enough useful signals to recommendations. A possible solution is to use a more fine-grained personality model than OCEAN; *e.g.,* the MBTI personality model which has a richer, 16-facet personality profile.

Last, the five personalities are encoded independently of each other in our model. But there is a correlation between these personality traits in real life; *e.g.,* a majority of extroverts are also open. In the future, we can make use of the relationship between personalities, perhaps by defining a hierarchical structure of personality traits and employing graph-based neural networks to encode them.

7 CONCLUSION AND FUTURE WORKS

In this work, we explore a new way of automatically extracting personality information from review texts and applying it to recommendation systems. We first construct two new datasets based on the Amazon dataset in the beauty and music domains and include OCEAN personality scores automatically inferred by the Receptiviti API, a commercial service. We then analyze the accuracy of using texts to obtain personality profiles and output personality score distributions. To explore the effectiveness of using personality in current recommendation systems, we conduct a few experiments with the standard neural collaborative filtering (NCF) recommendation algorithm and our variants, finding that incorporating personality information improves recommendation performance by 3% to 28%. In terms of the relationship between personality and domain, we find that *openness*, *extroversion*, and *agreeableness* are helpful in music recommendation, while *conscientiousness* is most helpful in the beauty recommendation.

In the future, more advanced models graph recommendation models can be used in the experiments. In addition, collecting more information beyond review texts (*e.g.,* purchase history, browsing history) is a potential direction. Moreover, except for the accuracy-based performance, it is possible to improve the fairness by using the OCEAN model [13]. To explore the inner relationship between personality and recommendation systems, doing a user study is also a possible way to further validate the findings.

ACKNOWLEDGEMENT

REFERENCES

[1] G.W. Allport. 1961. *Pattern and Growth in Personality.* Holt, Rinehart and Winston. https://books.google.com.sg/books?id=GVRAAAAAIAAJ

[2] Anton Aluja, Oscar Garcia, Jerome Rossier, and Luis F. Garcia. 2005. Comparison of the NEO-FFI, the NEO-FFI-R and an alternative short version of the NEO-PI-R (NEO-60) in Swiss and Spanish samples. *Personality and Individual Differences* 38, 3 (2005), 591–604. https://doi.org/10.1016/j.paid.2004.05.014

[3] Nana Yaw Asabere, Amevi Acakpovi, and Mathias Bennet Michael. 2018. Improving Socially-Aware Recommendation Accuracy Through Personality. *IEEE Trans. Affect. Comput.* 9, 3 (2018), 351–361. https://doi.org/10.1109/TAFFC.2017.2695605

[4] Deger Ayata, Yusuf Yaslan, and Mustafa E Kamasak. 2018. Emotion based music recommendation system using wearable physiological sensors. *IEEE transactions on consumer electronics* 64, 2 (2018), 196–203.

[5] J. Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin* 70 4 (1968), 213–20.

[6] Paul T Costa Jr and Robert R McCrae. 2008. *The Revised Neo Personality Inventory (neo-pi-r).* Sage Publications, Inc.

[7] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017,* Rick

Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 173–182. https://doi.org/10.1145/3038912.3052569

[9] Joanne Hinds, Emma J. Williams, and Adam N. Joinson. 2020. "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. *Int. J. Hum. Comput. Stud.* 143 (2020), 102498. https://doi.org/10.1016/j.ijhcs.2020.102498

[10] Jacob B Hirsh and Jordan B Peterson. 2009. Personality and language use in self-narratives. *Journal of research in personality* 43, 3 (2009), 524–527.

[11] Mahesh Babu Mariappan, Myunghoon Suk, and Balakrishnan Prabhakaran. 2012. FaceFetch: A User Emotion Driven Multimedia Content Recommendation System Based on Facial Expression Recognition. In *2012 IEEE International Symposium on Multimedia, ISM 2012, Irvine, CA, USA, December 10-12, 2012*. IEEE Computer Society, 84–87. https://doi.org/10.1109/ISM.2012.24

[12] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.

[13] Alessandro B Melchiorre, Eva Zangerle, and Markus Schedl. 2020. Personality bias of music recommendation algorithms. In *Fourteenth ACM conference on recommender systems*. 533–538.

[14] Tien T. Nguyen, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2018. User Personality and User Satisfaction with Recommender Systems. *Inf. Syst. Frontiers* 20, 6 (2018), 1173–1189. https://doi.org/10.1007/s10796-017-9782-y

[15] Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 188–197. https://doi.org/10.18653/v1/D19-1018

[16] Cynthia A. Pedregon, Roberta L. Farley, Allison Davis, James M. Wood, and Russell D. Clark. 2012. Social desirability, personality questionnaires, and the "better than average" effect. *Personality and Individual Differences* 52, 2 (2012), 213–217. https://doi.org/10.1016/j.paid.2011.10.022

[17] James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77, 6 (1999), 1296.

[18] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[19] Sanja Stajner and Seren Yenikent. 2020. A Survey of Automatic Personality Detection from Texts. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 6284–6295. https://doi.org/10.18653/v1/2020.coling-main.553

[20] Ewa Topolewska, Ewa Skimina, WŁODZIMIERZ Strus, Jan Cieciuch, and Tomasz Rowiński. 2014. The short IPIP-BFM-20 questionnaire for measuring the Big Five. *Roczniki Psychologiczne* 17, 2 (2014), 385–402.

[21] Pinata Winoto and Tiffany Ya Tang. 2010. The role of user mood in movie recommendations. *Expert Syst. Appl.* 37, 8 (2010), 6086–6092. https://doi.org/10.1016/j.eswa.2010.02.117

[22] Wen Wu, Li Chen, and Yu Zhao. 2018. Personalizing recommendation diversity based on user personality. *User Model. User Adapt. Interact.* 28, 3 (2018), 237–276. https://doi.org/10.1007/s11257-018-9205-x

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

[24] Wu Youyou, David Stillwell, H. Andrew Schwartz, and Michal Kosinski. 2017. Birds of a Feather Do Flock Together: Behavior-Based Personality-Assessment Method Reveals Personality Similarity Among Couples and Friends. *Psychological Science* 28, 3 (2017), 276–284. https://doi.org/10.1177/0956797616678187 arXiv:https://doi.org/10.1177/0956797616678187 PMID: 28059682.