PEIA: Personality and Emotion Integrated Attentive Model for Music Recommendation on Social Media Platforms

Tiancheng Shen,[1] Jia Jia,[1*] Yan Li,[2] Yihui Ma,[1] Yaohua Bu,[1]
Hanjie Wang,[2] Bo Chen,[2] Tat-Seng Chua,[3] Wendy Hall[4]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Beijing National Research Center for Information Science and Technology (BNRist)
The Institute for Artificial Intelligence, Tsinghua University
[2]WeChat AI, Tencent Inc, China, [3]School of Computing, National University of Singapore
[4]Electronics and Computer Science, University of Southampton

Abstract

With the rapid expansion of digital music formats, it's indispensable to recommend users with their favorite music. For music recommendation, users' personality and emotion greatly affect their music preference, respectively in a long-term and short-term manner, while rich social media data provides effective feedback on these information. In this paper, aiming at music recommendation on social media platforms, we propose a Personality and Emotion Integrated Attentive model (PEIA), which fully utilizes social media data to comprehensively model users' long-term taste (personality) and short-term preference (emotion). Specifically, it takes full advantage of personality-oriented user features, emotion-oriented user features and music features of multi-faceted attributes. Hierarchical attention is employed to distinguish the important factors when incorporating the latent representations of users' personality and emotion. Extensive experiments on a large real-world dataset of 171,254 users demonstrate the effectiveness of our PEIA model which achieves an NDCG of 0.5369, outperforming the state-of-the-art methods. We also perform detailed parameter analysis and feature contribution analysis, which further verify our scheme and demonstrate the significance of co-modeling of user personality and emotion in music recommendation.

1 Introduction

With the rapid development of informatization, Internet has become the major source of retrieving multimedia information. Music, as a significant approach of communication and expression, has been consumed by people as a common daily-life activity (Song, Dixon, and Pearce 2012). Although massive amounts of digital music have been accessible to people, to pick the favorite music out of the complete database, however, is difficult and time-consuming for users. Hence, targeted on music applications, stores and communities, research efforts on music recommendation have been made to generate potentially desired music playlists for users, of which the most common approaches are collaborative filtering (CF) and content-based methods (CBM). Specifically, CF recommends items via the choice of similar users, while CBM utilizes the acoustic signals and the track metadata (Sarwar et al. 2001; Dieleman and Schrauwen 2013; Lee et al. 2018).

Besides the tracks, user traits also play an important role in music recommendation. Researches show that people listen to music for multiple purposes, including interpersonal relationships promotion, moods optimization, identity development and surveillance (Lonsdale and North 2011). Music preference, on the one hand, is related to long-term traits including a wide array of personality dimensions, self-views, and cognitive abilities (Rentfrow and Gosling 2003). On the other hand, since music is an emotion-loaded type of content, the emotion context also impacts music preferences, but in a short-term manner (Ferwerda, Schedl, and Tkalcic 2015). Some researches also work towards user-centric music recommendation, incorporating either personality-oriented or emotion-oriented features (Deng et al. 2015; Cheng and Tang 2016; Cheng et al. 2017; Dhahri, Matsumoto, and Hoashi 2018). While advance has been achieved, these works only adopt limited information for one-sided modeling of users, lacking systematic analysis of user traits targeted on music recommendation. This inspires us that: can we model the users systematically from perspectives of both personality and emotion, with music content incorporated as well, to improve music recommendation?

Nowadays, people are increasingly relying on social media platforms like WeChat [1] and Twitter [2] to share their daily lives, which may reflect their personal traits and states. This makes it possible to incorporate users' personality and emotion together for music recommendation. Still, it is a non-trivial task owing to the following challenges: (1) Since social media data is quite complicated, how to effectively capture useful information for in-depth user-modeling? (2) Since the problem involves multi-faceted features, how to model the user-music correlations with all of them, as well as adaptively discriminating the most important factors?

In this work, we focus on music recommendation on social media platforms. We first construct a WeChat dataset containing 171,254 active users, 35,993 popular music tracks and 18,508,966 user-music-interaction records. The

[*] Corresponding author: J. Jia (jjia@mail.tsinghua.edu.cn)

[1] https://weixin.qq.com/.

[2] https://twitter.com/.

dataset is anonymized and desensitized by Tencent, and specific users cannot be located. Then, we deeply investigate the feature contributions, and explore the correlations between user traits and music preferences. We further propose a Personality and Emotion Integrated Attentive model (PEIA) to deal with the challenges respectively. (1) Beyond simply IDs, we comprehensively analyze each user by extracting personality-oriented and emotion-oriented features, involving demographic, textual and social behavioral attributes. For each music track, we also consider its acoustic features, metadata, lyric and emotion. (2) We adopt a deep framework to incorporate all the features, and employ hierarchical attention to estimate the importance of different feature interactions, as well as the weights of users' long-term taste (personality) and short-term preference (emotion).

We conduct extensive experiments and our model significantly outperforms other generic approaches (+1.80% in NDCG) and several music-oriented recommendation methods (+4.90% in NDCG). Case study is also conducted to reveal the inherent correlations between user traits and music preferences. All the experimental results verify our scheme and demonstrate the effectiveness of co-modeling of user personality and emotion in music recommendation.

The main contributions of this paper are as follows:

- We construct a large-scale music recommendation dataset on WeChat, which, compared to existing datasets like MSD (Bertin-Mahieux et al. 2011), contains elaborate personality-oriented user features, emotion-oriented user features and music features for in-depth modeling.
- We reveal several key factors of users' music preference, e.g. time and age, which may provide reference for music recommendation on social media platforms.
- We propose a PEIA model which employs hierarchical attention under deep framework to learn the correlations among user personality, user emotion and music, and achieves remarkable performance on a large real-world dataset.

2 Related Work

2.1 Recommendation System

Recommendation system is a well-researched topic and a wide variety of methods have been developed. While traditional models like factorization machines (FM) (Rendle 2010) perform well in learning low-order interactions, deep neural network has been widely applied in recommender systems in recent years (Qu et al. 2016; Zhang, Du, and Wang 2016). It is found that models incorporating linear kernel and neural network can combine the benefits of memorization and generalization for better recommendation (Cheng et al. 2016). Accordingly, based on user and item embeddings, DeepFM (Cheng et al. 2016) combines factorization machines and neural network, and DCN (Wang et al. 2017) introduces a cross network and a deep network in parallel. xDeepFM (Lian et al. 2018) employs a Compressed Interaction Network (CIN) to learns high-order feature interactions explicitly. Besides, attention mechanism, which allows variable contributions of different parts when compressing them into a single representation, has been introduced in recommendation models like AFM (Xiao et al. 2017), ACF (Chen et al. 2017) and DIN (Zhou et al. 2018) to discriminate the importance of different feature interactions, historical behaviors, and item components.

Inspired by these generic methods, we devise our PEIA model under a deep framework. Specifically, aiming at music recommendation, hierarchical attention is employed to estimate the weights of latent representations regarding personality and emotion.

2.2 Music Recommendation

Music Recommendation is an essential branch of recommendation system, where the utilization of audio and metadata is quite natural, and such methods are called content-based. For example, Dieleman and Schrauwen (2013) used a convolutional neural network to predict the latent factors from music audio. Lee et al. (2018) proposed deep content-user embedding model which combines the user-item interaction and the music audio content.

Moreover, music preferences are shown to be related to long-term traits including a wide array of personality dimensions (Rentfrow and Gosling 2003). For user-centric music recommendation, Cheng and Tang (2016) combined acoustic features with user personalities, and Cheng et al. (2017) tried to capture the influence of user-specific information on music preferences. Since music is an emotion-loaded type of content, the emotion context also impacts music preferences, but in a short-term manner (Ferwerda, Schedl, and Tkalcic 2015). Correspondingly, Deng et al. (2015) explored user emotion in microblogs for music recommendation and Dhahri, Matsumoto, and Hoashi (2018) built a personalized mood-aware song map according to implicit user feedback. To combine the long-term and short-term factors, some methods model users' global and contextual music preferences from their listening records with music embeddings (Wang et al. 2018; Sachdeva, Gupta, and Pudi 2018).

While advance has been achieved in the aforesaid works, they only adopt limited information, and the systematical modeling of user is especially deficient. In our PEIA model, rich social media data is fully utilized for elaborate user modeling from perspectives of both personality and emotion, while music content is also incorporated, so as to enhance music recommendation.

3 Data and Features

Users' music preferences are shown to be related to personality and emotion. In this section, with WeChat platform as a specific example, we elucidate the dataset and features in our work. Specifically, we define personality-oriented user features, emotion-oriented user features and music features of multi-faceted attributes.

3.1 Data Collection

WeChat is a social media platform extremely prevalent all over China, with over one billion monthly active users (Tencent 2019). Besides instant messaging, via the WeChat Moments, users can also post, browse and interact about content of multiple media, including text, image, music, video, etc. Following the typical implicit feedback settings, we uniformly record a **"user-music interaction"** $(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{t})$ when the certain user u likes, shares, collects or listens to for over one minute the certain music track m at timestamp t.

We focus on the active users and the popular music tracks on WeChat. During 2017.10 to 2018.4, we crawl 18,508,966 user-music-interactions, involving 171,254 different users and 35,993 different music tracks, where: 1) user's repeated interactions with the same music track within a day are recorded only once; 2) each user has at least 10 interaction records; 3) each music track is interacted by at least 10 users. Focusing on these users, corresponding 46,575,922 tweets, 120,322,771 social interactions, 96,451,009 article reading records, etc., are also gathered for in-depth user modeling. Regarding the privacy issue, the data is anonymized and desensitized by Tencent, and specific users cannot be located.

3.2 Personality-Oriented User Feature Extraction

Personality is an enduring and stable factor that embodies people's behavioral patterns in the long term. We analyze users' demographics and multiple behaviors over the whole sampling period as personality indicators. Specifically, instead of directly predicting users' personality, we employ the extracted indicators to get multi-dimensional personality-oriented features for deep model learning.

Demographics. Inspired by (Lonsdale and North 2011), we extract each user's gender, age and geographic location. Since real-name authentication is compulsory in WeChat, such demographic features are relatively reliable.

Textual Features. WeChat users often post tweets in Moments, where they directly express themselves with attached texts. We process the textual content with Jieba Chinese text segmentation[3], and as a long-term descriptor, extract 100-dimensional Doc2Vec features with Gensim (Le and Mikolov 2014).

Social Behavioral Features. Social interaction is a core issue in WeChat, where varieties of social behaviors are supported. We extract number of contacts, number of tweets and social interaction (e.g., comments and likes) frequency to evaluate users' social engagement. Detailed social relationships and interaction content are not gathered due to privacy concerns. Temporal distribution of posting is also studied as a reflection of daily schedule.

Article Reading Features. Article reading behavior is an important indicator of users' personal interests. We define 23 topics (e.g., sports, Internet, beauty & fashion), train a text classifier with FastText (Joulin et al. 2016) for article topic prediction, and thus calculate each user's reading frequency in the 23 topic dimensions.

3.3 Emotion-Oriented User Feature Extraction

Emotion is a changeable and transient factor which is closely related to the context. We set a time window before each user-music interaction record and conduct emotion-oriented user feature extraction over the targeted behaviors. To balance the timeliness of emotion and the sufficiency of data, the time window is set as 24 hours.

Temporal Features. We categorize the timestamp of each user-music interaction according to time of the day (morning, afternoon, evening and midnight) and week (weekday and weekend).

Emotion Vectors. Similar to (Deng et al. 2015), based on users' textual content of Moment tweets, we adopt a Chinese emotion lexicon from DUTIR[4], which includes 27,466 words with emotion tags of three different granularities, i.e., 2d-emotion, 7d-emotion and 21d-emotion, as shown in Table 1. We implement the lexicon with 259 emojis and 1831 words which are common on WeChat, and count the average frequency of emotion words per tweet for each emotion category. The results of three granularities are concatenated to form a 30-dimensional emotion vector. For the user-music interaction records with no tweets posted in the time window, which account for 17.8% of the dataset, the emotion vector is calculated over the user's all tweets in the dataset.

Table 1: Emotion classification system.

2d-emotion	7d-emotion	21d-emotion
(1) Positive	(1) Happy	(1) Joyful; (2) Relieved
	(2) Like	(3) Respect; (4) Praise; (5) Believe; (6) Like; (7) Hopeful
	(3) Surprised	(8) Surprised
(2) Negative	(4) Angry	(9) Angry
	(5) Sad	(10) Sad; (11) Disappointed; (12) Remorse; (13) Miss
	(6) Fear	(14) Fear; (15) Ashamed; (16) Flustered

	(7) Hate	(17) Disgusted; (18) Annoyed;
		(19) Reproach; (20) Jealousy;
		(21) Suspect

3.4 Music Feature Extraction

While each music track can be uniquely identified by its ID, this is far too vague for its profiling. We comprehensively study each track by analyzing acoustic features, metadata, lyric and emotion.

Metadata. Metadata plays an important role in music recommendation (Wang et al. 2018). For each music track, we consider its artist, genre, language and release year. Artists that appear in only one track are merged as an "other" artist and 3721 unique artists are eventually involved.

Acoustic Features. Acoustic measures of music have been extensively studied (Berenzweig et al. 2004; Mckay and Fujinaga 2008; Eyben et al. 2015). In this work, we employ openSMILE (Eyben et al. 2013), an open-source multimedia feature extractor to extract the "emobase" set of 988-dimensional acoustic features. Specifically, we extract low-level descriptors (LLD, e.g., intensity, loudness, MFCC, Pitch), compute delta coefficients and apply several functionals (e.g. range, mean, skewness, kurtosis).

Lyric. We study the semantic information of music tracks by analyzing the lyrics. We employ Baidu Translator API[5] to translate all the lyrics into Chinese, perform text segmen-

[3] https://github.com/fxsjy/jieba.

[4] http://ir.dlut.edu.cn/.

[5] http://api.fanyi.baidu.com.

tation with Jieba and extract 100-dimensional Doc2Vec features with Gensim.

Emotion Vectors. Consistent with the emotion vectors of users, we focus on the translated lyrics, and extract 30-dimensional music emotion vectors in the same way we deal with user emotions. Specifically, for the tracks with no lyrics (i.e., pure music, which accounts for 6.4% of the music tracks), we train a linear regression model on the rest music tracks with acoustic features as input, and thus complement the missing features.

4 Methodology

After feature extraction for user-modeling and item-profiling, in this section, we present the PEIA model, which employs hierarchical attention under deep framework to learn the correlations among user personality, user emotion and music. Meanwhile, a linear kernel and a deep network are fused in PEIA to learn both low- and high-order feature interactions across users and music tracks.

For a certain user-music interaction (u,m,t), suppose $\mathbf{p}$, $\mathbf{e}, \mathbf{m}$ are the personality-oriented features, emotion-oriented features and music features, respectively. Let y_{umt} denote the user preference corresponding to the sample, where an implicit feedback setting of binary values is considered in this work. We aim to learn a prediction function $\hat{y}_{umt} = f(\mathbf{p}, \mathbf{e}, \mathbf{m})$ to estimate user's preference with a score output.

4.1 Hierarchical Attentional Feature Interaction

Suppose the three feature groups contain P, E, M fields, respectively, i.e. $\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_P]$, $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_E]$ and $\mathbf{m} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_M]$. Each field can be either sparse (e.g., user ID) or dense (e.g., emotion vectors). We first respectively leverage the feature vectors of all fields into d-dimensional dense embeddings of a common latent space:

$$\begin{aligned} \mathbf{p'}_i &= \mathbf{W}_{p_i}\mathbf{p}_i + \mathbf{b}_{p_i}, \\ \mathbf{e'}_j &= \mathbf{W}_{e_j}\mathbf{e}_j + \mathbf{b}_{e_j}, \\ \mathbf{m'}_k &= \mathbf{W}_{m_k}\mathbf{m}_k + \mathbf{b}_{m_k}, \end{aligned} \tag{1}$$

where $i = 1, 2, \ldots, P, j = 1, 2, \ldots, E, k = 1, 2, \ldots, M$ and $\mathbf{W}_{p_i}, \mathbf{W}_{e_j}, \mathbf{W}_{m_k}, \mathbf{b}_{p_i}, \mathbf{b}_{e_j}, \mathbf{b}_{m_k}$ are the embedding parameters (for one-hot sparse vectors, $\mathbf{b} = \mathbf{0}$). In this way, we reduce the input dimensionality and pave the way for further processing of vector operation.

Feature interactions can be estimated via inner product or Hadamard product of each pair of feature vectors (Rendle 2010; Guo et al. 2017; Xiao et al. 2017), and we adopt Hadamard product which allows more flexible expressions. Specifically, we focus on feature interactions between user and music. Let $\mathbf{l}_{ik}$ denote the long-term taste factor, i.e., interactions between personality-oriented features and music features, and $\mathbf{s}_{jk}$ denote the short-term preference factor, i.e., interactions between emotion-oriented features and music features. Formally,

$$\begin{aligned} \mathbf{l}_{ik} &= \mathbf{p'}_i \odot \mathbf{m'}_k = [p'_{i_1}m'_{k_1}, p'_{i_2}m'_{k_2}, \ldots, p'_{i_d}m'_{k_d}], \\ \mathbf{s}_{jk} &= \mathbf{e'}_j \odot \mathbf{m'}_k = [e'_{j_1}m'_{k_1}, e'_{j_2}m'_{k_2}, \ldots, e'_{j_d}m'_{k_d}], \end{aligned} \tag{2}$$

where $\mathbf{l}_{ik}, \mathbf{s}_{jk} \in \mathbb{R}^d$ and $\odot$ denotes the Hadamard product.

The attention mechanism discriminates the importance of different components when compressing them into a single representation. PEIA employs hierarchical attention over the factors of long-term taste and short-term preference. Firstly, we perform a weighted sum on the interacted pairs of features:

$$\mathbf{l}_{\text{att}} = \sum_{i=1}^{P}\sum_{k=1}^{M} \alpha_{ik}\mathbf{l}_{ik}, \quad \mathbf{s}_{\text{att}} = \sum_{j=1}^{E}\sum_{k=1}^{M} \beta_{jk}\mathbf{s}_{jk}, \tag{3}$$

where α_{ik} and β_{jk} are the attention scores for feature interaction $\mathbf{l}_{ik}$ and $\mathbf{s}_{jk}$ respectively. Here, the attention scores are calculated via a two-layer attention network:

$$\begin{aligned} \alpha'_{ik} &= \mathbf{w}_l^T \sigma(\mathbf{W}_l \mathbf{l}_{ik} + \mathbf{b}_l) + b_l, \\ \beta'_{jk} &= \mathbf{w}_s^T \sigma(\mathbf{W}_s \mathbf{s}_{jk} + \mathbf{b}_s) + b_s, \end{aligned} \tag{4}$$

where $\mathbf{W}_l, \mathbf{W}_s \in \mathbb{R}^{t\times d}$, $\mathbf{b}_l, \mathbf{b}_s \in \mathbb{R}^t$ are the first layer parameters, $\mathbf{w}_l, \mathbf{w}_s \in \mathbb{R}^t$, $b_l, b_s \in \mathbb{R}$ are the second layer parameters and t is the size of hidden layer. Normalization of softmax is further adopted for the attention scores in Eqn 3:

$$\alpha_{ik} = \frac{\exp(\alpha'_{\text{ik}})}{\sum_{i=1}^{P}\sum_{k=1}^{M}\exp(\alpha'_{\text{ik}})}, \ \beta_{jk} = \frac{\exp(\beta'_{\text{jk}})}{\sum_{j=1}^{E}\sum_{k=1}^{M}\exp(\beta'_{\text{jk}})}. \tag{5}$$

In this way, we get the final feature interaction factors of user's long-term taste and short-term preference, i.e., $\mathbf{l}_{\text{att}}$ and $\mathbf{s}_{\text{att}}$. It is notable that two attention networks are separately applied for the two factors, since users' personality and emotion may follow different patterns of correlation with music tracks. We further combine the two interaction factors while distinguishing the overall contribution of user's personality and emotion:

$$\mathbf{z}_{\text{att}} = \gamma_l \mathbf{l}_{\text{att}} + \gamma_s \mathbf{s}_{\text{att}}, \tag{6}$$

where the attention scores are obtained similarly:

$$\begin{aligned} &\gamma'_l = \mathbf{w}_z^T \sigma(\mathbf{W}_z \mathbf{l}_{\text{att}} + \mathbf{b}_z) + b_z, \ \gamma'_s = \mathbf{w}_z^T \sigma(\mathbf{W}_z \mathbf{s}_{\text{att}} + \mathbf{b}_z) + b_z, \\ &\gamma_l = \frac{\exp(\gamma'_l)}{\exp(\gamma'_l) + \exp(\gamma'_s)}, \ \gamma_s = \frac{\exp(\gamma'_s)}{\exp(\gamma'_l) + \exp(\gamma'_s)}. \end{aligned} \tag{7}$$

4.2 Fusion with Deep Module

It is found that the fusion of linear kernel and deep model can combine the benefits of memorization and generalization for better recommendation (Cheng et al. 2016). While the result of hierarchical attentional feature interaction $\mathbf{z}_{\text{att}}$ can be directly used for prediction, a deep module is incorporated in PEIA. Similar to DeepFM and xDeepFM, we employ a deep neural network (DNN) with H hidden layers to model the high-order feature interactions in an implicit and non-linear manner. The input of the DNN is the concatenation of all the embedding vectors, denoted as $\mathbf{z}'_0$:

$$\mathbf{z}'_0 = [\mathbf{p}'_1, \ldots, \mathbf{p}'_P, \mathbf{e}'_1, \ldots, \mathbf{e}'_E, \mathbf{m}'_1, \ldots, \mathbf{m}'_M]. \tag{8}$$

The processing of DNN can be formally defined as

$$\mathbf{z}'_{h+1} = \sigma(\mathbf{W}_{d_h}\mathbf{z'}_h + \mathbf{b}_{d_h}), \tag{9}$$

where $h = 0, 1, \ldots, H-1$ and $\mathbf{W}_{d_h}, \mathbf{b}_{d_h}$ are the parameters. It is worth mentioning that the hierarchical attentional

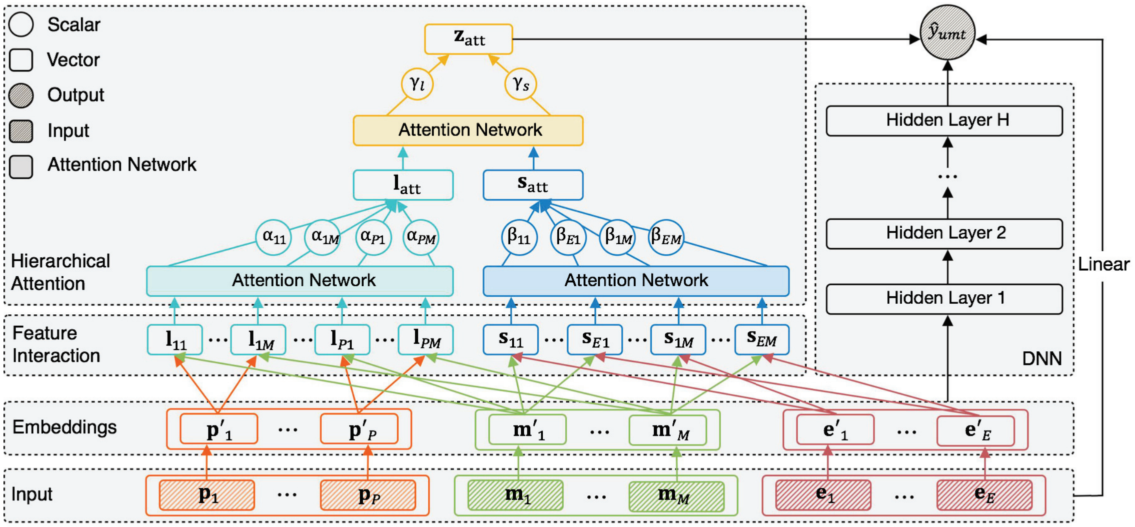

Scalar
Vector
Output
Input
Attention Network
Hierarchical Attention
Feature Interaction
Embeddings
Input
$\mathbf{z}_{\text{att}}$
γ_l
γ_s
Attention Network
$\mathbf{l}_{\text{att}}$
$\mathbf{s}_{\text{att}}$
α_{11} α_{1M} α_{P1} α_{PM}
β_{11} β_{E1} β_{1M} β_{EM}
Attention Network
Attention Network
$\mathbf{l}_{11}$ ⋯ $\mathbf{l}_{1M}$ ⋯ $\mathbf{l}_{P1}$ ⋯ $\mathbf{l}_{PM}$
$\mathbf{s}_{11}$ ⋯ $\mathbf{s}_{E1}$ ⋯ $\mathbf{s}_{1M}$ ⋯ $\mathbf{s}_{EM}$
$\mathbf{p'}_1$ ⋯ $\mathbf{p'}_P$
$\mathbf{m'}_1$ ⋯ $\mathbf{m'}_M$
$\mathbf{e'}_1$ ⋯ $\mathbf{e'}_E$
$\mathbf{p}_1$ ⋯ $\mathbf{p}_P$
$\mathbf{m}_1$ ⋯ $\mathbf{m}_M$
$\mathbf{e}_1$ ⋯ $\mathbf{e}_E$
$\hat{y}_{umt}$
Hidden Layer H
Hidden Layer 2
Hidden Layer 1
DNN
Linear

Figure 1: Personality and Emotion Integrated Attentive (PEIA) model. Annotations are in illustrated in Section 4.

feature interaction part and the DNN part share the same feature embedding, which may enable the embedding layer to learn both low- and high-order feature interactions from raw features (Guo et al. 2017).

Let $\mathbf{z}_0 = [\mathbf{p}_1, \dots, \mathbf{p}_P, \mathbf{e}_1, \dots, \mathbf{e}_E, \mathbf{m}_1, \dots, \mathbf{m}_M]$ denote the concatenation of raw features, we finally incorporate $\mathbf{z}_{\text{att}}$, $\mathbf{z}'_H$ and $\mathbf{z}_0$ for the resulting output of PEIA:

$$\hat{y}_{umt} = \sigma(\mathbf{w}_0^T \mathbf{z}_0 + \mathbf{w}_{\text{att}}^T \mathbf{z}_{\text{att}} + \mathbf{w}_H^T \mathbf{z}'_H + b). \tag{10}$$

For the training of PEIA, we take the point-wise binary cross-entropy loss

$$L = -\frac{1}{N} \sum_{(u,m,t)} y_{umt} \log \hat{y}_{umt} + (1 - y_{umt}) \log(1 - \hat{y}_{umt}) + \lambda ||\Theta||^2, \tag{11}$$

where N is the number of user-music interaction records in the training set, Θ is the set of model parameters, and λ is the regularization parameter. With regard to the natural scarcity of negative feedbacks in our problem, for each positive sample (u, m^+, t), we randomly select music track m^- which is never observed to be interacted by u, to form negative sample (u, m^-, t). The negative samples are selected in each training epoch and may vary in different epochs.

Figure 1 illustrates our PEIA model.

5 Experiments

In this section, we estimate our scheme of music recommendation on social media platforms and evaluate our PEIA model with extensive experiments. Specifically, we aim to answer:

RQ1 Does the hybrid model of PEIA effectively capture the correlations across users and music for recommendation?

RQ2 Is it effective to extract multi-faceted features from social media to improve music recommendation, and can PEIA capture the important factors via hierarchical attention?

RQ3 How do personality and emotion affect users' music preference?

5.1 Experimental Settings

Dataset. We conduct experiments on the desensitized WeChat dataset (Section 3), including 171,254 users, 35,993 music tracks and 18,508,966 user-music interaction records. Considering user ID and music ID, the number of fields of personality-oriented user features, emotion-oriented user features and music features are $P = 5$, $E = 2$, $M = 5$.

Evaluation Metrics. We take 80% of the user-music interaction records for model training, and the remaining for testing. While a certain user-music pair may involve several interaction records, we guarantee that all of them belong to either training set or test set. With timestamp considered, there is only one positive instance for each test case. Following (He et al. 2018), each positive test instance is paired with 99 randomly sampled negative instances, and each method predicts preference scores for the 100 instances. The performance is estimated by Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) (He et al. 2015) at the position 10.

Baselines. We compare the proposed PEIA model with the following generic recommendation methods:

- **ItemPop.** A non-personalized method which generates recommendation list according to the popularity (i.e., the number of interactions) of each music track.
- **FM.** It models first- and second-order feature interactions.
- **DNN.** A deep neural network with the concatenation of feature embeddings as input.
- **AFM (Xiao et al. 2017).** It improves FM by discriminating the importance of different feature interactions via a neural attention network.
- **DeepFM (Guo et al. 2017).** It combines factorization machines and neural network.
- **xDeepFM (Lian et al. 2018).** It combines the explicit high-order interaction module with implicit interaction module and traditional FM module.

- **PEIA-I.** It removes the second-level attention form PEIA and has $\gamma_l = \gamma_s = 0.5$ in Eqn 6.

We also consider the following music-oriented recommendation methods:

- **UCFE (Deng et al. 2015).** An emotion-aware method based on collaborative filtering.
- **MEM (Wang et al. 2018).** It models users' global and contextual music preferences from their listening records with music embeddings.
- **DCUE (Lee et al. 2018).** Short for Deep Content-User Embedding model. It is a hybrid method that utilizes user-item interaction and music audio content.

Parameter Settings. The deep module consists of two hidden layers, each with 200 neurons. ReLU activation is adopted for the deep module and the attention network, while the prediction layer uses sigmoid function. By default, we have $d = t = 32$ as the embedding size and the attention size. For each positive sample, three negative instances are randomly sampled regarding the same user and the same timestamp. The model parameters are randomly initialized with Gaussian distribution and optimized by Adam (Kingma and Ba 2014) with a mini-batch size of 512. For xDeepFM, the number of cross layers is 3.

5.2 Performance Evalutaion (RQ1)

Table 2: Performance of compared methods.

Method	HR	NDCG	Method	HR	NDCG
ItemPop	0.6335	0.4022			
FM	0.7653	0.5092	UCFE	0.7442	0.4901
DNN	0.7765	0.5223	MEM	0.7683	0.5118
AFM	0.7836	0.5274	DCUE	0.7530	0.4978
DeepFM	0.7776	0.5226	PEIA-I	0.7924	0.5357
xDeepFM	0.7792	0.5248	PEIA	**0.7941**	**0.5369**

Table 2 shows the performance of compared methods and PEIA performs the best, achieving 0.7941 and 0.5369 in HR and NDCG. We have the following observations: (1) Despite different detailed structures, hybrid methods (DeepFM, xDeepFM, PEIA-I and PEIA) generally outperform the single methods (FM and DNN), which justifies that the combination of linear model and deep model can enhance recommendation performance. (2) Methods with attention mechanism (AFM and PEIA-I) achieve remarkable performance as compared with other models, manifesting the complexity of feature interactions, and the importance of distinguishing them with different weights. (3) PEIA further improves PEIA-I moderately, demonstrating the effectiveness of the hierarchical attention in our PEIA model. (4) PEIA significantly outperforms other music-oriented recommendation methods, which verifies our scheme of modeling from perspectives of user personality and emotion, and proves that PEIA can take full advantage of the multi-faceted data.

We further investigate the models with extensive parameter analysis and Figure 2 presents the result of two parameters: (1) **Embedding size *d*.** We test the models with different values of d. As shown in Figure 2(a) and 2(b), it is not enough to represent the latent vector within only 4 dimensions, while too large embeddings may lead to over-fitting. (2) **Number of negative samples.** We change the ratio between negative and positive samples. Figure 2(c) and 2(d) illustrate the result, where the natural setting with equal portions of positive and negative samples does not lead to the best performance, and the optimal ratio is around 2 to 4.

Moreover, it is notable that in all cases, PEIA achieves the best performance, indicating a positive reply to RQ1.

5.3 Feature Contribution & Attention Analysis (RQ2)

The extraction of multi-faceted features is an important part in our PEIA approach. In this section, we aim to estimate the effectiveness of these features, as well as their interactions.

We test the PEIA model with one feature field removed each time, and report the performance in Table 4. We also investigate the importance of feature interactions and report the average attention scores $\gamma_l \alpha_{ik}, \gamma_s \beta_{jk}$ in Table 5, where we disregard the ID of user and music to focus on the extracted features, and abbreviations are used, e.g., P-D for Personality-Demographic. It can be summarized that: (1) All the extracted features positively contribute to music recommendation, which validates our work on user-modeling and item-profiling. (2) Removal of user demographics, user emotion vectors and music metadata hurts the performance severely, demonstrating their significance in music recommendation. (3) The impact of user text and music audio is relatively minor. While such fields are helpful, more effective feature extraction targeted on music recommendation may be in need, which is left as a future work. (4) For each feature field, the sum of attention scores in Table 5 is generally consistent with its impact on performance in Table 4, proving that PEIA can correctly estimate the importance of different factors. (5) The standard deviations of attention scores are quite large, indicating the effectiveness of attention mechanism, which may assign weights for feature interactions dynamically in different situations.

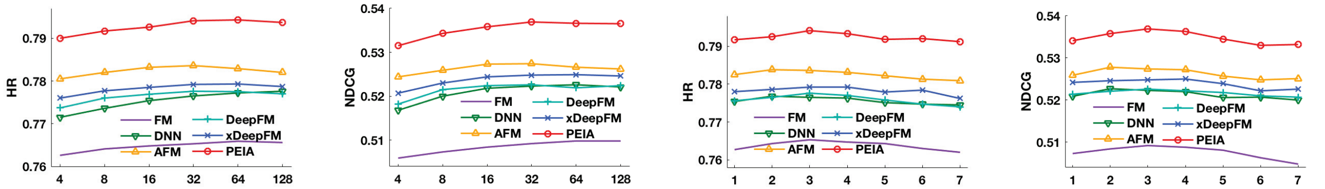
Combining with the result in Table 3, it is obvious that the answer to RQ2 is also affirmative.

5.4 Case Study (RQ3)

In this section, we aim to explore the correlations between user traits and music preferences, from both a statistical and an individual point of view.

Based on personality-oriented user features, we focus on four groups of users, i.e., the male, the female, the young ($\leq$25 years old) and the elderly ($\geq$50 years old). We calculate the average of several representative acoustic features of their interacted tracks, and Figure 3(a) shows the result. Remarkable difference in the 1st Mel-Frequency Cepstral Coefficient (MFCC[1]) can be observed, indicating different music preferences of different user groups. It can also be concluded that, compared with the young, the elderly prefer high-pitched and rhythmic tracks, which have higher values in F0 and zero cross rate. Gender differences are unexpectedly less significant in other acoustic features.

Targeted on emotion-oriented user features, we analyze the temporal change of music preferences during a day.



(a) d – HR (b) d – NDCG (c) Negative samples – HR (d) Negative samples – NDCG

Figure 2: Parameter analysis.

Table 3: Examples of user's tweets and music behavior.

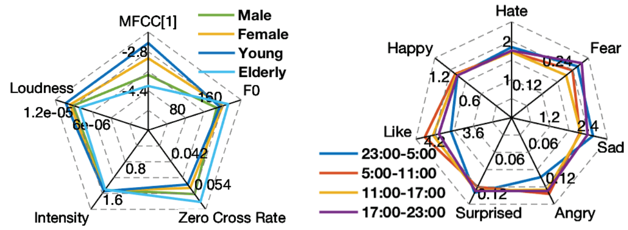
Timestamp	Recent Tweets	Interacted Track	Attention Score
2017-12-17 19:25	大吉大利，今晚吃鸡😝 (*Winner winner, chicken dinner*😝)	*Lucky Strike*	$\gamma_l = 0.79, \gamma_s = 0.21$
2017-12-19 15:10	雪越下越大 (*The snow is getting heavier*)	*Sugar*	$\gamma_l = 0.88, \gamma_s = 0.12$
2017-12-25 23:46	深夜矫情[皱眉] (*Sentimental at midnight* [Moue])	*Five Hundred Miles*	$\gamma_l = 0.32, \gamma_s = 0.68$

Table 4: Feature contribution analysis.

Type	**Removed Field**	**HR**	**NDCG**
User-Personality	Demographic	0.7815	0.5250
	Textual	0.7925	0.5356
	Social Behavioral	0.7880	0.5311
	Article Reading	0.7893	0.5320
User-Emotion	Temporal	0.7884	0.5309
	Emotion Vector	0.7842	0.5281
Music	Metadata	0.7723	0.5158
	Acoustic	0.7921	0.5353
	Lyric	0.7874	0.5312
	Emotion Vector	0.7907	0.5334

Table 5: Attention scores of feature interactions (%).

	M-M	M-A	M-L	M-E	Sum
P-D	11.8±4.9	2.3±1.3	5.2±2.2	3.7±1.4	23.1±4.2
P-T	2.4±2.3	0.1±0.1	1.2±1.1	0.1±0.2	3.8±2.0
P-S	8.6±3.8	1.1±0.6	3.3±1.7	2.0±1.0	15.0±5.2
P-A	9.1±3.6	1.8±0.8	4.0±1.8	1.9±0.8	16.9±3.5
E-T	6.1±4.6	1.0±0.9	2.5±2.1	0.9±0.7	10.6±5.7
E-E	13.0±3.9	4.2±1.9	6.7±2.7	6.7±1.4	30.5±3.6
Sum	51.0±16.7	10.5±4.6	23.1±9.4	15.4±4.6	



(a) Demographics – Acoustics (b) Time – Music Emotion

Figure 3: User traits vs. music preferences.

Specifically, we focus on the emotion vectors of interacted tracks, and calculate the average of 7d-emotion (Table 1) for four time buckets. As shown in Figure 3(b), people tend to choose arousing tracks with more expression of "like" and "anger" in the daytime, while depressive tracks with fear and sadness are more popular at night, which is consistent with people's daily emotional variance (Eaton and Funder 2001).

In Table 3, we show specific examples of a certain user's music behavior, together with corresponding recent tweets. We can imply from the first and the second record that, cheerful songs by *Maroon 5* are regularly enjoyed by the user. However, for the third record, the interacted track changes greatly to a lyrical one as the user gets sentimental at midnight. In this case, user's choice of music is impacted by both long-term taste (personality) and short-term preference (emotion), while the PEIA model may adaptively estimate the situation with different attention scores γ_l, γ_s.

6 Conclusion

In this paper, we aimed at music recommendation on social media platforms with incorporation of users' long term taste (personality) and short-term preference (emotion). We constructed a large-scale dataset, systematically extracted multi-faceted features, and explored several key factors of users' music preference. We further proposed a PEIA model which employs hierarchical attention under deep framework to learn the user-music correlations. Experimental results verified our scheme and demonstrated the significance of co-modeling of user personality and emotion in music recommendation.

Acknowledgments

This work is supported by National Key Research and Development Plan (2016YFB1001200), the state key program of the National Natural Science Foundation of China (61831022), National Natural and Science Foundation of China (61521002), and Beijing Academy of Artificial Intelligence (BAAI).

References

Berenzweig, A.; Logan, B.; Ellis, D. P. W.; and Whitman, B. P. W. 2004. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal* 28(2):63–76.

Bertin-Mahieux, T.; Ellis, D. P.; Whitman, B.; and Lamere, P. 2011. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*.

Chen, J.; Zhang, H.; He, X.; Nie, L.; Liu, W.; and Chua, T.-S. 2017. A entive collaborative filtering: Multimedia recommendation with item-and component-level a ention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*.

Cheng, R., and Tang, B. 2016. A music recommendation system based on acoustic features and user personalities. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 203–213.

Cheng, H. T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; and Ispir, M. 2016. Wide & deep learning for recommender systems. 7–10.

Cheng, Z.; Shen, J.; Nie, L.; Chua, T.-S.; and Kankanhalli, M. 2017. Exploring user-specific information in music retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 655–664.

Deng, S.; Wang, D.; Li, X.; and Xu, G. 2015. Exploring user emotion in microblogs for music recommendation. *Expert Systems with Applications* 42(23):9284–9293.

Dhahri, C.; Matsumoto, K.; and Hoashi, K. 2018. Mood-aware music recommendation via adaptive song embedding. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 135–138.

Dieleman, S., and Schrauwen, B. 2013. Deep content-based music recommendation. *Advances in Neural Information Processing Systems* 26:2643–2651.

Eaton, L. G., and Funder, D. C. 2001. Emotional experience in daily life: valence, variability, and rate of change. *Emotion* 1(4):413–421.

Eyben, F.; Weninger, F.; Gross, F.; and Schuller, B. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, 835–838.

Eyben, F.; Salomao, G. L.; Sundberg, J.; Scherer, K. R.; and Schuller, B. W. 2015. Emotion in the singing voice: a deeper look at acoustic features in the light of automatic classification. *Eurasip Journal on Audio Speech & Music Processing* 2015(1):19.

Ferwerda, B.; Schedl, M.; and Tkalcic, M. 2015. Personality & emotional states: Understanding users' music listening needs.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*.

He, X.; Chen, T.; Kan, M.-Y.; and Chen, X. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1661–1670.

He, X.; He, Z.; Song, J.; Liu, Z.; Jiang, Y.-G.; and Chua, T.-S. 2018. Nais: Neural attentive item similarity model for recommendation. *IEEE Transactions on Knowledge and Data Engineering* 30(12):2354–2366.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016. Fasttext.zip: Compressing text classification models. *CoRR* abs/1612.03651.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *Computer Science*.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, 1188–1196.

Lee, J.; Lee, K.; Park, J.; Park, J.; and Nam, J. 2018. Deep content-user embedding model for music recommendation. *arXiv preprint arXiv:1807.06786*.

Lian, J.; Zhou, X.; Zhang, F.; Chen, Z.; Xie, X.; and Sun, G. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1754–1763.

Lonsdale, A. J., and North, A. C. 2011. Why do we listen to music? a uses and gratifications analysis. *British Journal of Psychology* 102(1):108–34.

Mckay, C., and Fujinaga, I. 2008. Combining features extracted from audio, symbolic and cultural sources. In *International Conference on Music Information Retrieval*, 597–602.

Qu, Y.; Cai, H.; Ren, K.; Zhang, W.; Yu, Y.; Wen, Y.; and Wang, J. 2016. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 1149–1154.

Rendle, S. 2010. Factorization machines. In *2010 IEEE International Conference on Data Mining*, 995–1000.

Rentfrow, P. J., and Gosling, S. D. 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):1236.

Sachdeva, N.; Gupta, K.; and Pudi, V. 2018. Attentive neural architecture incorporating song features for music recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 417–421.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *International Conference on World Wide Web*, 285–295.

Song, Y.; Dixon, S.; and Pearce, M. 2012. A survey of music recommendation systems and future perspectives. In *The International Symposium on Computer Music Modeling and Retrieval*.

Tencent. 2019. 2018 annual report.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, 12.

Wang, D.; Deng, S.; Zhang, X.; and Xu, G. 2018. Learning to embed music and metadata for context-aware music recommendation. *World Wide Web* 21(5):1399–1423.

Xiao, J.; Ye, H.; He, X.; Zhang, H.; Wu, F.; and Chua, T.-S. 2017. Attentional factorization machines: learning the weight of feature interactions via attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3119–3125.

Zhang, W.; Du, T.; and Wang, J. 2016. Deep learning over multi-field categorical data. In *European conference on information retrieval*, 45–57. Springer.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1059–1068.