Enhancing Music Recommendation with Social Media Content: an Attentive Multimodal Autoencoder Approach

Tiancheng Shen[†], Jia Jia[†*], Yan Li[§], Hanjie Wang[§] and Bo Chen[§]
[†]Department of Computer Science and Technology, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology (BNRist)
The Institute for Artificial Intelligence, Tsinghua University
stc18@mails.tsinghua.edu.cn, jjia@tsinghua.edu.cn
[§]WeChat AI, Tencent Inc., China
{rockyanli, hankinwang, jennychen}@tencent.com

***Abstract*—Music recommendation methods predict users' music preference primarily based on historical ratings. Meanwhile, manifold personal factors of users are also important for the problem, and research efforts have been made to improve the recommendation performance with auxiliary user information. As an important indicator of users' personal traits and states, the numerous social media content (e.g., texts, images and short videos), however, is still hardly exploited. In this work, we systematically study the utilization of multimodal social media content for music recommendation. We define groups of both targeted handcrafted features and generic deep features for each modality, and further propose an Attentive Multimodal Autoencoder approach (AMAE) to learn cross-modal latent representations from the extracted features. Attention mechanism is also employed to integrate users' global and contextual music preference with alterable weights. Experiments demonstrate remarkable improvement of recommendation performance (+2.40% in Hit Ratio and +3.30% in NDCG), manifesting the effectiveness of our AMAE approach, as well as the significance of incorporating social media content data in music recommendation.**

I. Introduction

In the era of information explosion, huge amounts of digital music are accessible online, and it can be increasingly laborious for users to pick music tracks of interest from the vast music library. Hence, music recommendation has gained lots of attention, which may model users' preference, and efficiently provide the music that satisfies users' tastes. For recommendation systems, collaborative filtering (CF) is widely used, which utilizes users' historical interactions [1]. And as the most popular CF approach, matrix factorization (MF) [2], which learns a latent space to represent users and items, has become a standard model for recommendation. However, such methods may suffer from the data sparsity problem, leading to unsatisfactory performance.

Music has always been closely related to people's daily lives. Social and psychological studies show that music listening is related to manifold long and short-term factors of people, including interpersonal relationship, social identity, mood, personality, etc. [3], [4]. Research efforts have been made to enhance music recommendation with incorporation of users' auxiliary information, involving demographics, geolocations, daily activities, social relationships, etc. [5]–[8]. While improvement has been achieved, these works only consider partial information, lacking comprehensive modeling of users' states.

Nowadays, social media is extremely prevalent. According to Global Web Index, digital consumers spend an average of 2.4 hours on social networks everyday [9]. Users often share their daily activities and thoughts through social media platforms, covering almost every topic and domain. The multimodal user-generated content (UGC, e.g., texts, images, and short videos) may imply their personal traits and states, based on which analysis of social media users' personality, emotion and mental health states has achieved success [10]–[12]. Furthermore, the UGC can hopefully reflect music preference, even if it is not directly related to music. However, so far, the utilization of social media content for music recommendation is still in its infancy, which is limited to only one-sided features extracted from the textual modality [13]–[15].

This work focuses on the utilization of multimodal social media content for music recommendation. It is nontrivial owing to the following challenges: 1) Since the social media content may convey intricate connotations regarding music preference, what features should be extracted for each modality? 2) For the extracted multimodal low-level features, how to associate them with music tracks while capturing the cross-modal correlations? 3) Since music preference can be impacted by both global and contextual factors, how to combine them while distinguishing the dominating ones?

In this paper, we systematically study the problem, and deal with the challenges respectively: 1) For the content of each modality, we refer to relevant psychological researches, and define groups of handcrafted features targeted on music recommendation. Generic deep features are also analyzed to capture the implicit patterns and off-topic information. 2) We propose an Attentive Multimodal Autoencoder approach

*To whom correspondence should be addressed.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

(AMAE), which processes the extracted content features with autoencoders, and employs cross-modal loss to guarantee both the consistence and complementarity among representations of multi-modalities. 3) We devise an attention module to adaptively estimate the weights of global and contextual factors with consideration of the user and the music embedding. We combine our AMAE approach with MF model for the final prediction.

To verify our scheme, we construct a WeChat[1] dataset of 163,329 users, containing their 17,826,932 music interactions, and 45,276,160 tweets of multi-modalities. The dataset is anonymized and desensitized by Tencent, and specific users cannot be located. Extensive experiments are conducted, where our approach significantly improves the MF models (+2.40% in HR and +3.31% in NDCG), and outperforms the existing methods that utilize the social media content (+2.52% in HR and +3.30% in NDCG). We further investigate the contributions of different feature groups and the impacts of the model components in AMAE, which further validates our AMAE approach, and manifests the effectiveness of enhancing music recommendation with multimodal social media content.

We summarize the main contributions as follows:

- We systematically study the utilization of multimodal social media content for music recommendation, which is unique to the best of our knowledge. Specifically, we define groups of both handcrafted and deep content features, and analyze the problem from both global and contextual perspectives.
- We propose an AMAE approach which employs autoencoder structure and attention mechanism to learn cross-modal latent representations from content features, and to model users' global and contextual music preference.
- We conduct extensive experiments on a large real-world dataset, where encouraging results verify the effectiveness of both the content data and our AMAE approach.

The remainder of paper is organized as follows. Section 2 introduces related work. Section 3 elucidates data and features. Section 4 expounds the proposed model. Section 5 presents the experiments. Section 6 is the conclusion.

II. Related Work

A. Recommendation System

For recommendation systems, matrix factorization (MF) is a standard method [2]. Given the user-item rating matrix, it projects users and items into a shared latent space, and the user-item interaction can be modelled by the inner product of their latent vectors. Lots of variations of MF have been proposed [16], [17], while in recent years, deep learning has been employed in MF methods. For example, generalized matrix factorization (GMF) is proposed under the neural collaborative filtering framework [18], and deep matrix factorization (DMF) employs neural network architecture to learn latent embeddings of users and items [19]. Besides, MF models have also been extended to utilize extra information, such as review texts, item metadata and user neighborhood [20], [21].

On the other hand, attention mechanism has also been explored in recommendation models. Researches show that, when faced with multiple feature interactions, historical behaviors, and item components, etc., attentive modules can effectively estimate the contributions of different components and integrate them into a single representation with variable weights [15], [22], [23].

Inspired by these works, our AMAE approach is combined with a MF framework for the final prediction, and we devise an attentive module in AMAE to adaptively integrate the global and contextual factors.

B. User-Centric Music Recommendation

As revealed by psychological researches, people's music preference is related to diverse global and contextual factors, such as interpersonal relationship, social identity, mood and personality [3], [4]. For user-centric music recommendation, different types of auxiliary user information has been exploited: [5] utilized users' social relationships, [6] explored the data collected from sensors of users' mobile devices, [7] presented a venue-aware music recommender system, and [8] tried to capture the influence of user demographics on music preference. While advance has been achieved, the information utilized in these works is still not enough to comprehensively model the users.

On the other hand, with the prevalence of social media, the multimodal social media content becomes an important reflection of users' personal traits and states, which may hopefully benefit music recommendation. Accordingly, the million musical tweets dataset (MMTD) [24] was constructed, whereas only music-related content is included in MMTD, which is too sparse to perform in-depth user analysis. More recently, [13] analyzed user emotion in microblogs for music recommendation, [14] mined the embeddings of microblog texts, and [15] tried to model users' personality and emotion via their tweets and social behavior. However, so far, only one-sided features extracted from the textual modality have ever been considered for music recommendation, and the utilization of multimodal social media content remains to be further investigated.

III. Data and Features

In this section, we expatiate the multimodal music recommendation dataset and the extracted groups of features in our work. Specifically, for the content of each modality, we not only define handcrafted features targeted on music recommendation, but also consider generic deep features for comprehensive descriptions of the content data. All the extracted features are further aggregated in two different ways to describe the users from both global and contextual perspectives.

A. Data collection

In this work, we explore the utilization of multimodal social media content with WeChat as a specific platform. WeChat

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

is one of the most popular acquaintance social network platforms in China [25], where numerous content data, including texts, images, music and short videos, are posted, shared and browsed everyday. We construct a multimodal WeChat dataset based on the one in [15], where 171,254 active users were included, together with their music interactions, tweet texts, etc., during the period of 2017.10 to 2018.4. Specifically, users' music behavior, e.g., liking and sharing, was recorded in the form of $(\mathrm{uid}, \mathrm{mid}, \mathrm{timestamp})$. We further implement the dataset by collecting the images and short videos posted in the tweets. The anonymization and desensitization of data is performed by Tencent for privacy concerns, and specific users cannot be located. We further filter the data so that: 1) each user posts at least 10 tweets with images or short videos; 2) whether a user, or a music track, is involved in at least 10 interaction records. This leads to our multimodal music recommendation dataset of 163,329 users, 34,140 music tracks, and 17,826,932 user-music interactions, whose statistics are summarized in Table I.

TABLE I
STATISTICS OF THE DATASET.

Object	Count
Users	163,329
Music Tracks	34,140
User-Music Interactions	17,826,932
Tweets	45,276,160
Tweet Texts	41,136,288
Images	40,112,483
Short Videos	7,489,903

B. Textual Feature Extraction

Text is the most direct and explicit way of expression. Targeted on music recommendation, we focus on the emotion features, which may greatly impacts users' music preference [26]. We adopt a hierarchical emotion classification system of three granularities, where 2, 7 and 21 categories are included respectively in these levels [13]. Specifically, emotions are categorized into positive and negative ones, while positive emotions further include happiness, like and surprise, and negative emotions further include sadness, anger, fear and hate, and so on. A Chinese emotion lexicon from DUTIR[2] is adopted, and targeted on our dataset, the lexicon is extended with 259 emojis and 1831 common words on WeChat. Besides, we also analyze the topics and language styles of the textual content, which is closely related to users' personal traits and states [12], [27]. The Simplified Chinese Microblog Word Count (SCMBWC) [28] dictionary is employed to extract the commonly used linguistic features in sentiment analysis research, including part of speech statistics, punctuation count, topic-related word measure, cognitive process estimation, etc. After processing with Jieba Chinese text segmentation[3], we calculate the word frequency of all the above-mentioned categories to comprehensively estimate users' textual content. Macro statistics, e.g., text length and sentence count are also considered.

Despite the elaborate work in defining handcrafted features, high-level representations are also significant, which may capture the latent factors and off-topic information in an implicit way. To extract the deep features of tweet texts, we first train 200-dimensional Word2Vec embeddings [29] on the crawled corpus. As for the tweet-level representations, instead of simply averaging the vectors of words, we employ the smooth inverse frequency (SIF) weighting scheme, and subsequently remove the common components in the weighted average representations, which was shown to achieve remarkable performance in sentence embedding, especially for the social network corpus (i.e., the Twitter dataset) [30].

C. Visual Feature Extraction

Visual content, e.g., image and short video, is a more flexible and intricate way of expression, delivering rich connotations. Visual features have been revealed to effectively convey personal emotions and states [11]. In this work, for simplicity, we convert short videos into three screenshots, i.e., the frames at the beginning, in the middle, and at the end of the video. Thus, the visual content can be processed uniformly.

Inspired by the previous works on affective color psychology theories and image classification [31], we consider the following groups of handcrafted features: 1) Color: we extract the mean and contrast of hue, saturation and brightness, together with area statistics, i.e., area of warm/cold color, area of 5 brightness levels, and area of 3 saturation levels, as defined in [32]. 2) Texture and composition: we evaluate the image clarity via Tenengrad gradient, and calculate the dynamic-static description based on line slopes [33]. 3) Content: we focus on the human faces in images, which may strongly draw attention of observers. Number of faces and relative size of the biggest face are extracted via libfacedetection[4].

Besides the handcrafted visual features, deep features extracted by convolutional neural network are also considered. We scale the images to 224*224 size, and adopt ResNet50 [34] to extract 2048-dimensional representations. While the weights of ResNet50 were pretrained on ImageNet, a dataset for visual recognition task, we believe the features can embody valuable information, and be helpful for music recommendation.

D. Feature Aggregation: Global and Contextual Descriptions

Music preferences are related to both global and contextual factors. Therefore, based on the extracted attributes of single tweets, we further integrate them into features of two granularities: 1) *Global* (user-level) features: for a certain user, the extracted features of all his/her tweets in the sampling period are averaged as a global description; 2) *Contextual* (interaction record-level) features: we focus on each user-music interaction record, collect the users' tweets posted within 24 hours before the interaction behavior, and average the features of these

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

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

[4]http://github.com/ShiqiYu/libfacedetection.

filtered tweets to estimate users' contextual state regarding the interaction record. Global features are padded when no tweets were posted in the window of filtering.

In this way, we obtain both global and contextual multi-modal content features for each user-music interaction record, and can therefore analyze their correlations in a systematic manner.

IV. Methodology

In this section, we present our AMAE approach, which learns cross-modal latent representations from the extracted content features, and employs attention mechanism to integrate global and contextual factors.

A. Preliminaries

Suppose $\mathcal{U}, \mathcal{M}$ are the set of users and modalities in the dataset, respectively. Given a certain user-music record $(\text{uid}, \text{mid}, \text{timestamp})$, or (u, m, t) for short, let $\mathbf{x}_{ut}^{Ag}, \mathbf{x}_{ut}^{Ac}$ denote the global and contextual features for modality $A \in \mathcal{M}$. Let $\mathbf{x}_{ut} = \{\mathbf{x}_{ut}^{A*} \mid A \in \mathcal{M}, * \in \{g, c\}\}$ be the set of all the content features. We aim to learn a function $\hat{y}_{umt} = f(u, m, \mathbf{x}_{ut})$ to predict the user's preference y_{umt} with a score output. In this work, we follow the implicit feedback setting and the value of y_{umt} is binary.

The problem can be intuitively solved with matrix factorization (MF) methods, which make predictions by

$$\hat{y}_{umt} = \mathbf{p}_u^T \mathbf{q}_m, \tag{1}$$

where $\mathbf{p}_u, \mathbf{q}_m \in \mathbb{R}^d$ denote the latent representations of user u and music track m. Still, we intend to improve the performance with the content features.

B. Autoencoder for Multi-Modalities

Autoencoder is a powerful unsupervised deep model for feature learning. For each modality $A \in \mathcal{M}$, we train an autoencoder $\mathcal{H}^A$ to encode the extracted content features into d-dimensional represention. We employ a deep structure of K encoding layers and K decoding layers to increase the non-linearity of the model, as shown in Fig. 1. Let $h_0(\mathbf{x}_{ut}^{A*}) = \mathbf{x}_{ut}^{A*} \in \{\mathbf{x}_{ut}^{Ag}, \mathbf{x}_{ut}^{Ac}\}$ denote the input, and the encoding process can be formulated as

$$h_{i+1}(\mathbf{x}_{ut}^{A*}) = \sigma(\mathbf{W}_{e_i}^A h_i(\mathbf{x}_{ut}^{A*}) + \mathbf{b}_{e_i}^A), \tag{2}$$

where $i = 0, 1, \ldots, K-1$, and $\mathbf{W}_{e_i}^A, \mathbf{b}_{e_i}^A$ are the parameters for encoding. Similarly, let $r_K(\mathbf{x}_{ut}^{A*}) = h_K(\mathbf{x}_{ut}^{A*})$, and the decoding process is

$$r_{j-1}(\mathbf{x}_{ut}^{A*}) = \sigma(\mathbf{W}_{d_j}^A r_j(\mathbf{x}_{ut}^{A*}) + \mathbf{b}_{d_j}^A), \tag{3}$$

where $j = 1, 2, \ldots, K$, and $\mathbf{W}_{d_j}^A, \mathbf{b}_{d_j}^A$ are the parameters for decoding. Thus, we get the desired encoded representation $\mathbf{z}_{ut}^{A*} = h_K(\mathbf{x}_{ut}^{A*})$, and the reconstructed input $\hat{\mathbf{x}}_{ut}^{A*} = r_0(\mathbf{x}_{ut}^{A*})$. Typically, autoencoder $\mathcal{H}^A$ is optimized by minimizing the reconstruction loss between the raw input and decoded output

$$L_r^A = \frac{1}{2} \sum_{*\in\{g,c\}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}_u} \|\hat{\mathbf{x}}_{ut}^{A*} - \mathbf{x}_{ut}^{A*}\|^2, \tag{4}$$

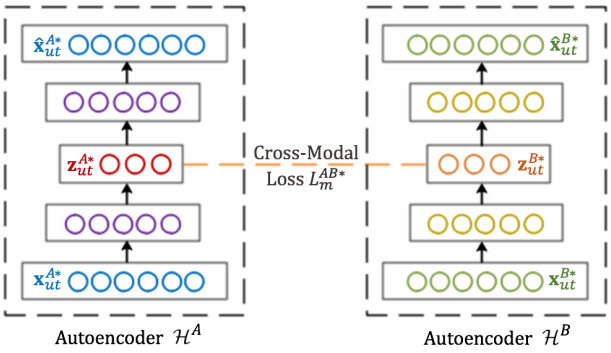


Fig. 1. Autoencoder for multi-modalities.

where $\mathcal{T}_u$ is the set of timestamps at which user u had music interactions. Thus, the model may smoothly capture the data manifold and preserve the similarity among samples [35].

Since the problem involves multi-modalities, different features are expected to be complementary, delivering information of different aspects. However, when features of different modalities are related to the same user, describing either global traits, or contextual states at the same timestamp, they should be respectively consistent. To guarantee both the consistence and the complementarity of the encoded multimodal representations, we do not simply concatenate the encoded results of the autoencoders. For modality A and B, given user u and v, we consider the joint distribution of the two modalities as

$$p_{u_t v_s}^{AB*} = \frac{1}{1 + \exp(-(\mathbf{z}_{ut}^{A*})^T \mathbf{z}_{vs}^{B*})}. \tag{5}$$

Inspired by [36], we employ the most negative sampling strategy, select the most distinctive sample against user u at timestamp t:

$$\tilde{v}, \tilde{s} = \arg\min_{v,s} p_{u_t v_s}^{AB*}, \tag{6}$$

and we aim to maximize the cross-modal likelihood

$$L_m^{AB*} = \prod_{u \in \mathcal{U}} \prod_{t \in \mathcal{T}_u} p_{u_t u_t}^{AB*} (1 - p_{u_t \tilde{v}_{\tilde{s}}}^{AB*}), \tag{7}$$

or equivalently, to minimize the cross-modal loss

$$L_m^{AB*} = \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}_u} (-\log p_{u_t u_t}^{AB*} - \log(1 - p_{u_t \tilde{v}_{\tilde{s}}}^{AB*})), \tag{8}$$

where the first term induces consistent, but non-identical encoding of multimodal features for the same user at the same timestamp, and the second term pushes them away when the features are from distinctive samples. Thus, consistent and complementary multimodal features can be learned.

Thereby, we get the final loss function for autoencoders of all modalities:

$$\begin{aligned} L = &\frac{1}{2} \sum_{A \in \mathcal{M}} \sum_{*\in\{g,c\}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}_u} \|\hat{\mathbf{x}}_{ut}^{A*} - \mathbf{x}_{ut}^{A*}\|^2 + \lambda \sum_{A \in \mathcal{M}} \Omega(\Theta_{\mathcal{H}_A}) \\ &- \sum_{\substack{A,B \in \mathcal{M} \\ A \neq B}} \sum_{*\in\{g,c\}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}_u} (\log p_{u_t u_t}^{AB*} + \log(1 - p_{u_t \tilde{v}_{\tilde{s}}}^{AB*})), \end{aligned} \tag{9}$$

where $\Omega(\Theta_{\mathcal{H}_A})$ is the regularizer.

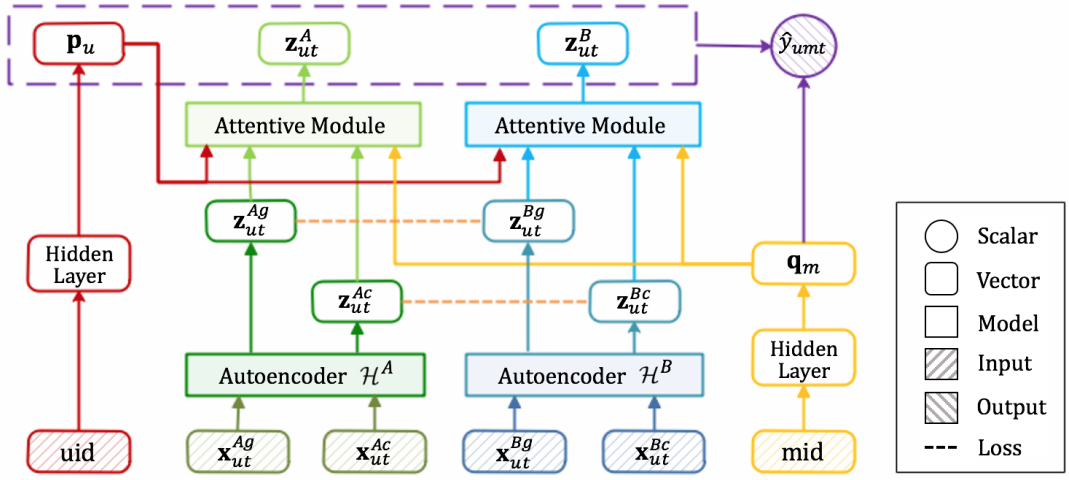


Fig. 2. Attentive Multimodal Autoencoder approach (AMAE) with MF framework. Annotations are illustrated in Section IV.

C. Attentive Module under MF Framework

For certain modality A, global and contextual features $\mathbf{x}^{Ag}_{ut}, \mathbf{x}^{Ac}_{ut}$ are processed by the same autoencoder $\mathcal{H}_A$, and there are no more constraints between $\mathbf{z}^{Ag}_{ut}$ and $\mathbf{z}^{Ac}_{ut}$. However, users' music preference can be volatile, and the contribution proportions of global and contextual features can also vary greatly [15]. In AMAE, we propose to employ an attentive module to integrate these two levels of features with alterable weights. While it is common for attentive networks to process only the content features (the features to be integrated) [15], [22], for music preference, the attention weights of global and contextual factors might be greatly impacted by the user and the music track. Therefore, we utilize the latent presentation of the user and the music track $\mathbf{p}_u$ and $\mathbf{q}_m$, and devise a two-layer network. Formally,

$$\beta^{A*}_{ut} = (\mathbf{w}^A_{\text{att}})^T \sigma(\mathbf{W}^A_{\text{att}}[\mathbf{z}^{A*}_{ut}, \mathbf{p}_u, \mathbf{q}_m] + \mathbf{b}^A_{\text{att}}) + b^A_{\text{att}}, \quad (10)$$

where $\{\mathbf{W}^A_{\text{att}}, \mathbf{b}^A_{\text{att}}\}$ and $\{\mathbf{w}^A_{\text{att}}, b^A_{\text{att}}\}$, are the parameters for the first and the second layer, respectively. We calculate the attention weights via softmax function and obtain the weighted sum for modality A by

$$\alpha^{Ag}_{ut} = \frac{\exp(\beta^{Ag}_{ut})}{\sum_{*\in\{g,c\}} \exp(\beta^{A*}_{ut})}, \quad \alpha^{Ac}_{ut} = \frac{\exp(\beta^{Ac}_{ut})}{\sum_{*\in\{g,c\}} \exp(\beta^{A*}_{ut})}, \quad (11)$$

$$\mathbf{z}^A_{ut} = \alpha^{Ag}_{ut} \mathbf{z}^{Ag}_{ut} + \alpha^{Ac}_{ut} \mathbf{z}^{Ac}_{ut}. \quad (12)$$

With integrated representations $\mathbf{z}^A_{ut}$ for each modality A, we concatenate them with the user latent vector $\mathbf{p}_u$, and make predictions via dot product with the music latent factor $\mathbf{q}_m$:

$$\hat{y}_{umt} = [\mathbf{p}_u, \mathbf{z}^A_{ut}, \mathbf{z}^B_{ut}, \ldots]^T \mathbf{q}_m. \quad (13)$$

Here, we assume $\mathbf{p}_u, \mathbf{z}^A_{ut}, \mathbf{z}^B_{ut}, \ldots \in \mathbb{R}^d$ and $\mathbf{q}_m \in \mathbb{R}^{d\times(|\mathcal{M}|+1)}$, i.e., the dimensionality of $\mathbf{q}_m$ is increased to pair with the multimodal content factors, which can be easily implemented by modifying the embedding layer size in the MF model.

Fig. 2 presents our framework, where $|\mathcal{M}|=2$, and for MF, a DMF model [19] with two hidden layers is illustrated.

We devise a two-stage process for model training. In the first stage, we train the autoencoders via Eqn 9 to deal with the raw content features. In the second stage, the weights of autoencoders are fine-tuned and the entire model is optimized based on the point-wise binary cross-entropy loss for prediction results:

$$L = -\frac{1}{N} \sum_{(u,m,t)} y_{umt} \log \hat{y}_{umt} + (1 - y_{umt}) \log(1 - \hat{y}_{umt}) + \lambda \Omega(\Theta), \quad (14)$$

where N is the number of samples in the training set, and $\Omega(\Theta)$ is the regularizer.

V. Experiments

In this section, we estimate our scheme of enhancing music recommendation with social media content and evaluate our AMAE approach with extensive experiments. Specifically, we aim to answer:

RQ1 Is it helpful to incorporate multimodal social media content in music recommendation, and how do different groups of features contribute to the problem?

RQ2 Can the hybrid structure of AMAE effectively utilize the content features to enhance recommendation performance?

RQ3 How do users' social media content correlate to their music preference?

A. Experimental Settings

1) Dataset: Experiments are conducted on the constructed multimodal WeChat dataset (Section III), including 163,329 users, 34,140 music tracks and 17,826,932 user-music interaction records. Two modalities (textual and visual modality) are involved in this dataset.

2) Evaluation Metrics: We take 80% of the user-music interaction records for model training, and the remaining for testing. As the timestamp is considered for the user-music interactions, each test case has only one positive instance. Following [15], [37], we pair each positive test instance with 99 randomly sampled negative instances. Each method predicts

the preference score for the 100 instances, and a recommendation list is therefore generated according to the rank of the score. To assess the ranked list, we employ Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [38] at the position 10. Specifically, HR indicates whether the positive instance is ranked in the top 10, and NDCG evaluates the predicted position of the positive instance.

3) Baselines: We compare the following recommendation methods:

- **ItemPop.** A naive approach which conducts recommendation simply based on the item popularity measured by the number of interactions in the training set.
- **MF.** A standard recommendation model. We implement it via degraded GMF, as illustrated in [18].
- **DMF** [19]. An advanced MF model which adopts a deep structure to learn latent factors of users and items. The cross-entropy loss is used due to our implicit feedback setting.
- **MF+AMAE.** It combines our AMAE approach with MF.
- **DMF+AMAE.** It combines our AMAE approach with DMF.

We also consider the following music recommendation methods with utilization of social media content:

- **UCFE** [13]. The optimal method in [13], named user-based collaborative filtering with emotion. User emotion is estimated via contextual tweet texts.
- **MF_S**, **PMF_S** [14]. The two optimal methods in [14]. Global textual features are extracted by stacked denoising autoencoder (SDAE) over bag of words (BoW), and thus combined with MF or PMF.

4) Parameter Settings: We set the latent factor size d as 16. The autoencoders in AMAE consist of two layers, and the intermediate layer of the encoding and decoding part has 32 neurons. ReLU activation is used for the autoencoders and the attention module, while the prediction layer uses sigmoid function. For DMF, two hidden layers are adopted. The model parameters are randomly initialized with Gaussian distribution and optimized by Adam [39] with a mini-batch size of 512. According to [15], [19], for each positive instance (u, m^+, t), three negative instances (u, m^-, t) are randomly sampled in each training epoch. To reduce the time complexity, $\tilde{v}$ and $\tilde{s}$ (Eqn. 6) are selected within the mini-batch.

B. Performance Comparison

We report the performance of the compared methods in Table II, from which we have the following observations: 1) DMF+AMAE achieves the best performance, with 0.7726 in HR and 0.5167 in NDCG. 2) While there can be a large disparity between different MF models, with the utilization of social media content via AMAE, the performance of both MF and DMF improves significantly, by at least 2.40% in HR and 3.31% in NDCG. 3) Although social media content is utilized in UCFE, MF_S and PMF_S, they do not achieve remarkable performance as compared to other methods. This is because they only exploit one-sided textual features for user-modeling.

TABLE II
PERFORMANCE OF COMPARED METHODS.

Method	HR	NDCG
ItemPop	0.6304	0.3978
MF	0.7428	0.4889
DMF	0.7545	0.5001
UCFE	0.7439	0.4890
MF_S	0.7479	0.4936
PMF_S	0.7536	0.5002
MF+AMAE	0.7621	0.5085
DMF+AMAE	**0.7726**	**0.5167**

Beyond the remarkable performance of AMAE, we further conduct in-depth analysis in the following subsections to give insights into RQ1-RQ3.

C. Feature Contribution Analysis (RQ1)

We comprehensively evaluate the contributions of the defined handcrafted and deep features. Based on the original MF framework, different groups of features are utilized with our AMAE approach, and Table III shows the results, where H, D denote the handcrafted and the deep features, and G, C denote the global and the contextual features, respectively.

TABLE III
FEATURE CONTRIBUTION ANALYSIS.

Utilized Features	MF+AMAE		DMF+AMAE	
	HR	NDCG	HR	NDCG
None	0.7428	0.4889	0.7545	0.5001
Textual-H	0.7503	0.4966	0.7613	0.5073
Textual-D	0.7519	0.4986	0.7628	0.5077
Textual-All	0.7562	0.5033	0.7663	0.5112
Visual-H	0.7455	0.4919	0.7570	0.5018
Visual-D	0.7498	0.4960	0.7616	0.5066
Visual-All	0.7513	0.4972	0.7619	0.5078
Textual+Visual-G	0.7538	0.5006	0.7662	0.5105
Textual+Visual-C	0.7565	0.5024	0.7669	0.5114
Textual+Visual	0.7621	0.5085	0.7726	0.5167

As can be seen from Table III: 1) All the extracted features positively contribute to music recommendation, which verifies our work on feature extraction. 2) For both textual and visual modality, models with deep features outperform those with handcrafted features. While a lot of work has been devoted to defining useful handcrafted features, the experimental result shows that, the social media content may convey much implicit information regarding music behavior, which remains to be further explored. 3) Still, for both modalities, the incorporation of handcrafted features and deep features result in even better performance, indicating that these two categories of features are complementary, capturing valuable information from different aspects. 4) Both textual and visual content can significantly improve the performance, which proves the deficiency of previous work where only tweet texts were exploited. 5) Both global and contextual features are beneficial for recommendation, and the contextual features seem to be slightly more effective, perhaps partially due to the padding

strategy when the contextual content is missing. Moreover, the integration of them leads to the optimal performance, manifesting that users' music preference can be impacted by both global and contextual factors, and more insights regarding the issue will be given later.

D. Model Component Analysis (RQ2)

To further investigate the effectiveness of our AMAE approach, we conduct experiments with the following components removed, respectively: 1) Reconstruction loss L_r (Eqn. 4) for the autoencoders, i.e., the decoders are removed and the features are encoded with two dense layers. 2) Cross-modal loss L_m for the multimodal latent factors, as illustrated in Eqn. 8. 3) Weights fine-tuning for the autoencoders, i.e., the autoencoders and the MF framework are trained separately, and the input of MF is replaced by fixed encoded representations $\{\mathbf{z}_{ut}^{Ag}, \mathbf{z}_{ut}^{Ac} | A \in \mathcal{M}\}$. 4) The attentive module, i.e., the global and contextual factors are directly concatenated with the user embedding for final prediction, and the dimensionality of $\mathbf{q}_m$ is changed to $d \times (2|\mathcal{M}|+1)$ accordingly.

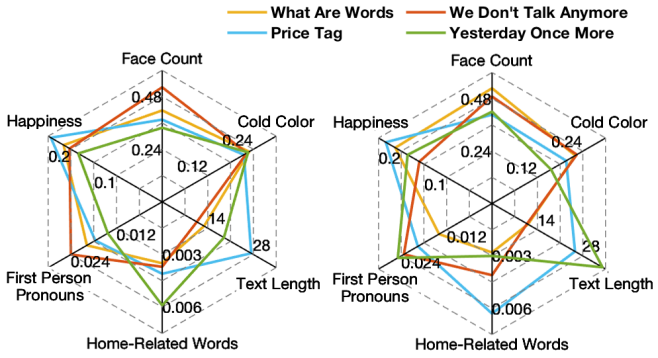
TABLE IV
MODEL COMPONENT ANALYSIS.

Removed Component	MF+AMAE		DMF+AMAE	
	HR	NDCG	HR	NDCG
None	0.7621	0.5085	0.7726	0.5167
Reconstruction Loss L_r	0.7574	0.5040	0.7683	0.5129
Cross-Modal Loss L_m	0.7601	0.5068	0.7708	0.5156
Weights Fine-tuning	0.7547	0.5015	0.7662	0.5112
Attentive Module	0.7603	0.5063	0.7714	0.5161

Table IV shows the performance of the altered models, which is unsurprisingly worse than the original AMAE: 1) Removal of reconstruction loss L_r and weights fine-tuning both hurt the performance severely, which shows that, it is effective to learn low-dimensional embeddings via autoencoder, while the autoencoder weights need to be further fine-tuned according to the prediction task, in order that the encoded representations can better adapt to music recommendation. 2) The omission of cross-modal loss L_m results in declining performance, which justifies the significance of learning both consistent and complementary latent factors for multimodal data. 3) It is effective to introduce the attention mechanism into AMAE. Since users' music preference can be influenced by miscellaneous factors, simply concatenating the global and the contextual representations with equal weights is not a perfect choice, while the attention module can solve the problem by adaptively discriminating the important factors.

E. Case Study (RQ3)

In this subsection, we intend to explore the correlations between social media content and music preference with specific examples. We choose four popular songs, and for each song, we calculate the average of the content features regarding corresponding user-music interactions. Fig. 3(a) and 3(b) illustrate several representative handcrafted features, including happiness, first person pronoun count, home-related word count, and text length from the textual modality, as well as cold color ratio and face count from the visual modality. Here, we do not include the deep features, since they do not convey explicit comprehensible meanings.



(a) Global content features (b) Contextual content features

Fig. 3. Average of social media content features regarding four popular songs.

From Fig. 3(a) and 3(b), we can discover that: 1) *What Are Words* is an affectionate song. While most corresponding content features are of medium values, larger face count is observed contextually. 2) *Price Tag* is a rousing and lively song, and the listeners are more positive and talkative, with consistent higher values in happiness and text length. 3) *We Don't Talk Anymore* is about breaking up, whose listeners tend to be especially unhappy in the contextual point of view. 4) *Yesterday Once More* is a classic song about memory, whose listeners might be senior and introverted. Globally, they tweet more about home, but less about themselves, and post fewer images with faces. However, the condition dramatically changes in the contextual perspective, where the text length, the first person pronoun count, and the warm color ratio increase sharply, as they may be tweeting heaps of words about their past experiences.

Based on these samples, it can be further summarized that: 1) Users' music preference is closely related to their social media content, and can be reflected by a wide range of features, including emotion, language style, topic of post, image color, image content, etc. 2) Global and contextual content factors impact users' music preference differently, and can be rather divergent for the same user-music interaction.

VI. CONCLUSION

In this paper, we aimed to enhance music recommendation with utilization of multimodal social media content. We constructed a large-scale multimodal dataset, defined handcrafted and deep features for each modality, and analyzed users' music preference from both global and contextual perspectives. We further proposed an AMAE approach to learn cross-modal latent representations from raw features, and to integrate global and contextual factors via attentive mechanism. Experimental results validated our approach and demonstrated the efficacy of exploiting multimodal social media content 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

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *International Conference on World Wide Web*, 2001, pp. 285–295.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[3] A. C. North, D. J. Hargreaves, and S. A. O'Neill, "The importance of music to adolescents." *British Journal of Educational Psychology*, vol. 70, no. 2, pp. 255–272, 2011.

[4] A. J. Lonsdale and A. C. North, "Why do we listen to music? a uses and gratifications analysis." *British Journal of Psychology*, vol. 102, no. 1, pp. 108–34, 2011.

[5] J. Bu, S. Tan, C. Chen, C. Wang, H. Wu, L. Zhang, and X. He, "Music recommendation by unified hypergraph:combining social media information and music content," in *ACM International Conference on Multimedia*, 2010, pp. 391–400.

[6] X. Wang, D. Rosenblum, and Y. Wang, "Context-aware mobile music recommendation for daily activities," in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 99–108.

[7] Z. Cheng and J. Shen, "On effective location-aware music recommendation," *Acm Transactions on Information Systems*, vol. 34, no. 2, pp. 1–32, 2016.

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

[9] GWI, "Global web index's flagship report on the latest trends in social media," 2019.

[10] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI'11 extended abstracts on human factors in computing systems*, 2011, pp. 253–262.

[11] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, "Modeling emotion influence in image social networks," *IEEE Transactions on Affective Computing*, vol. 6, no. 3, pp. 286–297, 2015.

[12] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T.-S. Chua, and W. Hall, "Cross-domain depression detection via harvesting social media," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2018, pp. 1611–1617.

[13] S. Deng, D. Wang, X. Li, and G. Xu, "Exploring user emotion in microblogs for music recommendation," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9284–9293, 2015.

[14] W. Ma, M. Zhang, C. Wang, C. Luo, Y. Liu, and S. Ma, "Your tweets reveal what you like: Introducing cross-media content information into multi-domain recommendation." in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2018, pp. 3484–3490.

[15] T. Shen, J. Jia, Y. Li, Y. Ma, Y. Bu, H. Wang, B. Chen, T.-S. Chua, and W. Hall, "Peia: Personality and emotion integrated attentive model for music recommendation on social media platforms," in *Proceedings of the 34th AAAI conference on artificial intelligence*, 2020. in press.

[16] A. Mnih and R. R. Salakhutdinov, "Probabilistic matrix factorization," in *Advances in neural information processing systems*, 2008, pp. 1257–1264.

[17] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2016, pp. 549–558.

[18] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[19] H.-J. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, "Deep matrix factorization models for recommender systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3203–3209.

[20] Y. Bao, H. Fang, and J. Zhang, "Topicmf: Simultaneously exploiting ratings and reviews for recommendation," in *Twenty-Eighth AAAI conference on artificial intelligence*, 2014, pp. 2–8.

[21] L. Hu, A. Sun, and Y. Liu, "Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 2014, pp. 345–354.

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

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

[24] D. Hauger, M. Schedl, A. Košir, and M. Tkalčič, "The million musical tweet dataset: what we can learn from microblogs," in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR*, 2013, pp. 189–194.

[25] B. Zhang, Q. Wu, X. Chen, and L. Chen, "Information cascades over diffusion-restricted social network: A data-driven analysis," in *IEEE Conference on Computer Communications Workshops*, 2019, pp. 151–156.

[26] B. Ferwerda, M. Schedl, and M. Tkalcic, "Personality & emotional states: Understanding users' music listening needs," in *UMAP 2015 Extended Proceedings*, 2015.

[27] A. D. Kramer, "The spread of emotion via facebook," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 767–770.

[28] R. Gao, B. Hao, H. Li, Y. Gao, and T. Zhu, "Developing simplified chinese psychological linguistic analysis dictionary for microblog," in *International Conference on Brain and Health Informatics*, 2013, pp. 359–368.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[30] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proceedings of the 5th International Conference on Learning Representations*, 2017, pp. 1–16.

[31] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[32] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, 2006, pp. 3534–3539.

[33] N. Bianchi-Berthouze, "K-dime: an affective image filtering system," *IEEE MultiMedia*, vol. 10, no. 3, pp. 103–106, 2003.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] D. Wang, C. Peng, and W. Zhu, "Structural deep network embedding," in *the 22nd ACM SIGKDD International Conference*, 2016, pp. 1225–1234.

[36] H. Gao and H. Huang, "Deep attributed network embedding," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2018, pp. 3364–3370.

[37] X. He, Z. He, J. Song, Z. Liu, Y.-G. Jiang, and T.-S. Chua, "Nais: Neural attentive item similarity model for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2354–2366, 2018.

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

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, 2015, pp. 1–15.