#nowplaying-RS: A New Benchmark Dataset for Building Context-Aware Music Recommender Systems

Asmita Poddar
National Institute of Technology,
Rourkela, India
asmita.poddar@gmail.com

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

Yi-Hsuan Yang
Research Center for IT Innovation,
Academia Sinica, Taiwan
yang@citi.sinica.edu.tw

ABSTRACT

Music recommender systems can offer users personalized and contextualized recommendation and are therefore important for music information retrieval. An increasing number of datasets have been compiled to facilitate research on different topics, such as content-based, context-based or next-song recommendation. However, these topics are usually addressed separately using different datasets, due to the lack of a unified dataset that contains a large variety of feature types such as item features, user contexts, and timestamps. To address this issue, we propose a large-scale benchmark dataset called #nowplaying-RS, which contains 11.6 million music listening events (LEs) of 139K users and 346K tracks collected from Twitter. The dataset comes with a rich set of item content features and user context features, and the timestamps of the LEs. Moreover, some of the user context features imply the cultural origin of the users, and some others—like hashtags—give clues to the emotional state of a user underlying an LE. In this paper, we provide some statistics to give insight into the dataset, and some directions in which the dataset can be used for making music recommendation. We also provide standardized training and test sets for experimentation, and some baseline results obtained by using factorization machines.

1. INTRODUCTION

Social media applications have been gaining popularity over the years. For instance, Twitter serves 330 million monthly active users as of January, 2018. [1] Similarly, Spotify is a highly popular music streaming service that allows users to listen to music anywhere, anytime. Spotify users can tweet about the songs they are listening to using the so-called #nowplaying tweets (e.g., "#nowplaying Yellow Submarine - The Beatles #happy"). From such tweets, rich metadata about the listening events (LEs) of users can be extracted [1]. For example, the hashtag "#happy" in the above example might be a self-expression of the listener's underlying emotional state. Likewise, Spotify provides rich information, metadata and audio content features of tracks [2], that can be obtained via the Spotify API. Hence, these sites are ideal for information retrieval and analysis, especially for building music recommender systems (RS).

The performance of an RS in general, highly relies on the content of the dataset used for model training. For example, *context-based recommendation* aims at modeling how contextual factors of a user, such as time (e.g., time-of-day, day-of-week and month-of-year), location (e.g., in-door, out-door), weather, user activity (e.g., reading, exercising) and user emotion/mood (e.g., happy, sad), affect the user's preference [8–11]. As some of such contextual factors are hard to collect from users, a recent trend is to mine contextual information from the sequence of user behavior in the recent past (e.g., list of previously played songs) to infer the current user preference [12–16]. This is known specifically as *sequence-based recommendation*, or *next-song recommendation* for music. When timestamps of the user behavior are available and are exploited to divide the user history into multiple time sessions (e.g., with a long time gap between two sessions), the recommendation setting can also be referred to as *session-based recommendation* [17–19]. Datasets with millions of data entries/points were used in such recent studies (not necessarily focusing on music), especially those based on deep learning techniques (e.g., [14, 20]).

Despite exciting progress that has been made lately, we observe that context-based recommendation and sequence-based recommendation were usually addressed separately using different datasets. Moreover, most existing work did not make use of item content features (such as audio features extracted from musical audio), which can mitigate the so-called cold-start problem and improve the diversity/interpretability of the recommendation result [19, 21]. This is mostly due to the lack of a consolidated dataset containing different data types like item content features, user contexts, as well as timestamps of the user-item association. Accordingly, it is hard to investigate the dependency of different data types and jointly model them in a single framework.

Building on top of online resources from Twitter and Spotify, in this paper we propose a new dataset to address this demand. The new dataset, referred to as #nowplaying-RS hereafter, contains 11.6 million LEs of 140K users, including 350K tracks. The dataset features 6 user contextual features including hashtags and emotion information extracted from the hashtags contained in the underlying tweets, timestamps of LEs and 11 item content fea-

[1] https://www.omnicoreagency.com/twitter-statistics/ (01/18)

Dataset	Num. ratings	Num. users	Num. tracks	Context feat.	Content feat.	Time-stamps
LFM-1b [3]	1,000,000,000	120,175	585,095	√ (1)		√
MMTD [4]	1,000,000	215,375	133,968	√ (11)		√
MusicMicro [5]	594,306	136,866	71,400	√ (6)		
MLHD [6]	∼27 billion	∼583,000	∼7,000,000			√
Yahoo! Music [7]	262,810,175	1,000,990	624,961	√ (2)		√
#nowplaying [1]	46,054,607	4,150,615	1,206,499	√ (1)		√
#nowplaying-RS	11,639,541	138,781	346,273	√ (7)	√ (11)	√

Table 1. Comparison of existing datasets for music recommendation, with the number in parentheses indicating the number of features contained.

tures. We believe that the dataset may contribute to research on music RS in a number of ways: i) it is a large-scale, context-aware dataset; ii) the dataset contains timestamps for LEs, making it a valuable dataset for sequence-based recommendation; iii) the hashtags contained are a unique feature since they give an idea of the listening context and also, the emotional state of the user at the time of listening to the track, allowing for sentiment-aware approaches to recommendation and retrieval; iv) it contains item content features of the songs; and finally v) it facilitates the development of an integrated system that jointly models different data types relevant to recommendation.

The dataset, our code, and train/test splits used in our experiments are available at `http://dbis-nowplaying.uibk.ac.at/#nowplayingrs`. Please note that our methods and the resulting dataset naturally adhere to Twitter's policies and that all user data is anonymized.

Information filtering and recommendation is an integral part of the way users perceive music. Context-aware music recommendation, in particular, can find its applications in personalized music streaming, smart cars, smart speakers, etc. As #nowplaying-RS contains user-provided hashtags self-expressing their activities, thoughts, and emotions, it holds the promise to deepen our understanding of the way people interact with music in the daily lives, such as how people use music to communicate ideas, express themselves and to modulate moods.

In what follows, we firstly highlight related context-aware recommendation datasets. Then, we present our dataset: the methods of data acquisition, the availability and content, and general statistics as well as the hashtag content. Finally, we present some pre-defined train/test splits of the dataset for benchmarking and further perform some proof-of-concept experiments with #nowplaying-RS using factorization machines [22] in two different settings.

2. RELATED DATASETS

Table 1 features a comparison of the most comprehensive and popular music recommendation datasets available. We note that all of these datasets feature implicit, positive-only feedback (ratings) [23] on the tracks—i.e., information about which tracks were listened by a user. Schedl's LFM-1b dataset [3] contains one billion LEs crawled from the last.fm platform and includes artists, tracks, albums

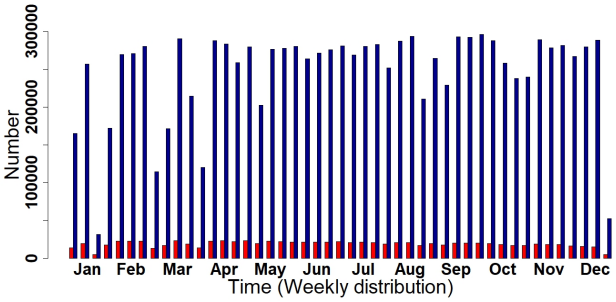


Figure 1. Barplot depicting the weekly distribution of the number of tweets (LEs) and number of users tweeting over time. Each bar represents the span of a week. Red bars represent the number of unique users and blue bars represent the number of tweets.

and extensive user information (e.g., demographic aspects or scores such as novelty or mainstreaminess to describe the user's taste). However, besides the country of the user and timestamps of LEs, the dataset does not provide any contextual data. In contrast, the million musical tweets dataset (MMTD) [4] and the MusicMicro dataset [5] come with contextual information related to time and location. The musical listening histories dataset (MLHD) [6], the Yahoo! Music ratings dataset [7] and the #nowplaying dataset [1] contain a substantial number of users, items also including timestamps of LEs; however, no contextual information is given. Only information about the source of the underlying tweet (how it was sent) is provided in the #nowplaying dataset. In comparison to the existing datasets, our #nowplaying-RS dataset provides the following unique features: First, we provide a publicly available and extensive dataset of LEs, particularly suited for (sequential) context-aware recommendations. Second, we provide a great variety of context and content information about the users and tracks, as well as clues of the underlying emotions, activities, and thoughts of the users through hashtags.

3. THE #NOWPLAYING-RS DATASET

3.1 Dataset Creation Procedure

The basis for the #nowplaying-RS dataset is the #nowplaying dataset compiled by Zangerle *et al.* [1], which contains

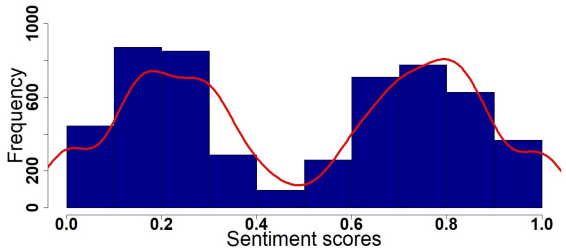
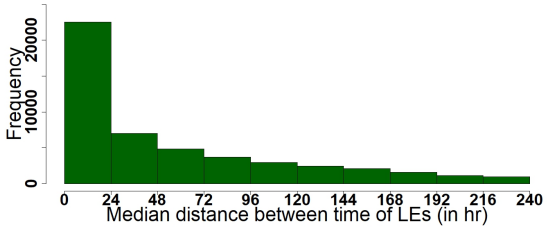
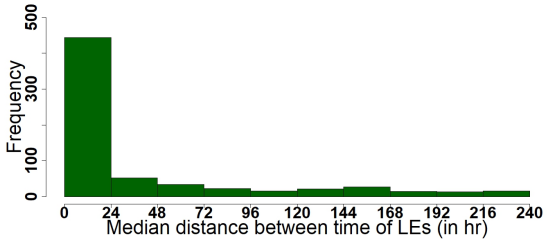


Figure 2. Histogram showing the variation of average sentiment scores of the hashtags in the dataset. The sentiment scores have been scaled in the range of [0,1].



(a) All LEs



(b) Affect-enriched LEs

Figure 3. Histogram of median time distance (in hours) per user for (a) all LEs and (b) affect-enriched LEs. The distribution of median time distances up to 10 days (240 hours) is shown here.

LEs crawled from Twitter, where LEs were extracted from individual tweets. For the creation of the proposed dataset, we extracted all LEs for the year 2014, with a total of 17 million LEs. The LEs are uniformly distributed in this time frame as shown in Figure 1. These basic LEs featuring a timestamp, artist and track were enriched with further information gathered via the Twitter API [2] about the language and time zone of the user. Furthermore, we also retrieved and added the hashtags that were used in each tweet as these may serve as contextual indicators. Particularly, since hashtags may indicate the context of the user or the song played and also, the emotional state of a user underlying an LE, we extract affective contextual information from hashtags contained in these tweets by applying an unsupervised sentiment dictionary approach. We relied on well-established dictionaries which have been widely used and evaluated [24, 25], where we chose the dictionaries that provide both the best coverage and performance in terms of accuracy: AFINN [26], Opinion Lexicon [27], SentiStrength [28], Vader [29] and the Sentiment Hashtag lexicon [30]. We only stored an entry if we are able to detect the sentiment of the tweet hashtag using at least one of the dictionaries. In total 5,290 hashtags in the dataset have been assigned with a sentiment score. Figure 2 shows the distribution the sentiment scores of these hashtags.

Previous research has shown that content features can mitigate the cold-start problem and also improve the interpretability of results in recommendation tasks [31]. Hence, we added track content features to the dataset to be able to describe tracks by means of acoustic content features. We used the Spotify API [3] to obtain the acoustic features for the contained LEs. This was done in two steps: The track API was first used to search for the track and artist to get the Spotify-ID of the given track. Then, this track-ID was used to gather the acoustic features of the track.

[2] https://developer.twitter.com/en/docs/tweets/search/api-reference

Item	Number
Listening Events	11,639,541
Users	138,781
Tracks	346,273
Artists	44,214
Unique user-track pairs	3,043,487
Unique track-artist pairs	346,273

Table 2. Statistics of the proposed dataset

Category	Hashtag
Mood	#lazy, #sleepy, #lonely, #depressed, #sad, #angry, #jealous
Activity	#driving, #swimming, #walking, #running, #studying, #focus, #sleeping, #cooking
Place of listening	#gym, #home, #car, #office, #school, #college, #hospital, #disco
Emotion	#hate, #love, #wonderful, #encore, #fun, #addictedtothis
Location	#berlin, #toronto, #india, #australia
Genre	#metal, #edm, #classical, #softrock, #pop, #jazz, #blues
Source	#radio, #itunes, #fm

Table 3. Categories of the hashtags and some examples

3.2 Content and Statistics

Table 2 lists some statistics of the dataset. As can be seen, the dataset contains 11,639,541 listening events of 138,781 users who listened to 346,273 distinct tracks performed by 44,214 distinct artists. The final dataset contains 17,560,113 hashtags associated with LEs (44,913 unique lower-cased hashtags). The most widely used hashtags are #nowplaying, #listenlive and #music comprising 63.15%, 7.37% and 0.95% of the total number of hashtag usages in the dataset. [4] Hashtags allow to infer the listening context of a song and are used in a highly diverse manner: they might be used

[3] https://developer.spotify.com/web-api/

[4] Please note that the underlying Twitter data was crawled using the search terms #nowplaying, #listento and #listeningto.

IDs	Description
User ID	Unique user ID.
Track ID	Unique track ID.
Artish ID	Unique artist ID.
Context features	**Description**
Timestamp	Exact time of creation of the tweet underyling the LE in the format YYYY/MM/DD HH/MM/SS.
Tweet language	Language in which the tweet underlying the LE was posted.
User Language	Language of the user (as stated in the user's interface settings).
Time zone	Time zone from where the tweet was posted.
Hashtags	Hashtags contained in LE; categorizes and contextualizes a tweet by a keyword.
Sentiment	Sentiment score extracted from hashtags contained in LE (ranges from 0 (negative) to 1 (positive)).
Content features	**Description**
Instrumentalness	Signifies whether a track contains vocals.
Liveness	Presence of an audience in the track recording (range is $[0, 1]$, where 1 indicates high probability of liveness).
Speechiness	Presence of spoken words in a track - whether a track contains more music or words (range is $[0, 1]$, where 0 is a track with no speech).
Danceability	Suitability of a track for dancing based on a combination of musical elements like tempo, rhythm stability, beat strength, and overall regularity (range is $[0, 1]$, where 1 is a most danceable song).
Valence	Musical positiveness conveyed by a track (range is $[0, 1]$, where 1 is a highly positive and cheerful song).
Loudness	The overall loudness of a track in decibel (dB).
Tempo	The overall estimated tempo of a track in beats per minute (BPM).
Acousticness	Probability whether a track is acoustic (range is $[0, 1]$).
Energy	Perceptual measure of intensity and activity (range is $[0, 1]$, where 1 indicates a high-energy track).
Mode	Modality (major or minor) of a track, i.e., the type of scale from which its melodic content is derived. Major is 1 and minor is 0.
Key	The key that the track is in. Integers map to pitches using standard Pitch Class notation.

Table 4. Data contained in the #nowplaying-RS dataset, including IDs, user context features and item content features

to describe the genre or context of the played song (e.g., #metal, #70s), the source of the song (e.g., a radio station), a description of the artist (e.g., #rapgod) or some notion of the perceived emotion of the user (e.g., #fun). Please see Table 3 for more examples. As for the sentiment values of hashtags associated with tweets underlying the LEs, we observe that the sentiment information allows to contextualize LEs regarding the mood of the user at the time of listening to a track. Particularly, we can derive the change of the mood context of a user over time.

Among the user context features, the timestamps can be used for not only indicating information regarding time-of-day, day-of-week and month-of-year, but also for modeling user preference using sequence-based or session-based models. We show in Figure 3 how closely spaced in time the LEs are to each other. Figure 3(a) shows the median distance between the time of successive LEs per user while Figure 3(b) shows the median distance between the time of successive LEs enriched with affect-related hashtags, per user. We see that there are larger number of users with LEs having a short time span (50% of users have a time span of 0–37 hours) between successive LEs in general. About 50% of users who have used affect-related hashtags with LEs have a time span of 0–52 hours between successive LEs. This allows us to model user preferences or mood on a nearly daily basis, making this dataset useful for sequence-based recommendation.

Table 4 gives an overview of the entire dataset including the 6 context features and 11 audio content features.

4. DATASET USE CASES AND PRE-DEFINED DATA SPLITS

To demonstrate some of the various possible uses of the dataset, we provide pre-defined data splits for two possible use cases: context-aware recommendation and context-aware next-song recommendation.

4.1 Context-aware recommendation

Context has been shown to be highly influential when it comes to the perceived utility of recommendations by users.

Hence, a dataset that provides context features, is an important step towards context-aware RS. Particularly the hashtag context and thereby, also the extracted mood information, is a novel approach towards context-aware RS that can be exploited and experimented with this dataset.

4.1.1 Creation of Training and Test Sets

Due to the implicit nature [23] of the data, the #nowplaying-RS dataset only contains positive examples. However, for both model training and evaluation, negative examples are needed. Below, we first describe how we split our dataset (positive only) into training and test sets, and then describe how we create the negative samples.

In a real-world setting [21], we are given historical (past) ratings of users and aim to predict the tracks that a user would like to listen to in the future, in a given context. Therefore, we used the timestamps to split #nowplaying-RS into the training (from Jan. 1 to Sep. 30) and test sets (from Nov. 1 to Dec. 23). The LEs during the month of Oct. may be used to create a validation set for the experiments. For data cleansing, we removed users who have listened to less than 10 tracks and tracks which have been listened to by less than 10 users, since such records cannot contribute to modeling the user preferences sufficiently. We also removed LEs that do not contain hashtags or do not exhibit any sentiment information from the dataset for the experiments. Table 5(a) depicts the characteristics of the training and test sets employed for the experiments, counting only positive examples.

For each LE in either the training or test set, we further added nine tracks as the negative samples. Based on this list, in our track ranking experiment, we aimed to rank this list of tracks, such that the positive example is ranked as the first (i.e., the most relevant) track. We used two different population methods to find the negative samples per LE: *random* population (POP_RND), where we added nine randomly chosen tracks that the user has not listened to previously and *user-based* population (POP_USER), where we randomly picked nine tracks the user has previously listened to, but in a different context, and added these to the set. The resulting set of 10 tracks can be subsequently used as input to an RS. It can be understood that track recommendation under the POP_USER population method is more difficult, because all the 10 candidate tracks are known to the user and likely the recommender has to rely on contextual features to pick the right one.

4.2 Context-aware Next-song Recommendation

Sequence-based recommendation has recently become an important research topic. Given historical user data, we aim to predict next interactions with the recommender system. For example, this year, the RecSys Challenge [5] (as part of the ACM Recommender Systems Conference) also aims to perform a playlist continuation task and hence, sequence-based recommendations based on data provided by Spotify. Similarly, the ACM WSDM Cup [6] was based on historic listening data of users. In the following, we provide a training and test set split to perform next-song recommendation.

[5] http://www.recsyschallenge.com/2018/

[6] https://wsdm-cup-2018.kkbox.events/

(a)	Number of LEs	Number of users	Number of items
Training set	257,012	3,982	22,092
Test set	104,334	1,467	13,978
(b)	Number of LEs	Number of users	Number of items
Training set	253,030	1,830	20,631
Test set	102,867	686	13,321

Table 5. Splitting of the dataset into training and test sets for (a) context-aware RS and (b) context-aware next-song RS, respectively. In our experiments, we use 0.01 of the training set as the validation set. Please note that we count only positive samples in this table.

4.2.1 Creation of Training and Test Sets

The sequence of a user's listening actions depicts the evolution in the users' taste [32]. The sequence of songs listened to by the user can be created from the timestamps available in the dataset. The mean length of sequences per user in our dataset is 123 (median length: 13). For sequence-based recommendation we can use the "previous-N" songs a user has listened to for predicting the next song that he/she would listen to. For our experiments, we took into account the user's most recent song preference into consideration by taking $N = 1$, i.e., the "last song" that a user had listened to is used as the context information to infer the "next song" that a user would likely listen to. Since, we are using $N = 1$, the user must have listened to a sequence of at least 2 songs to be considered in the training and test sets. Table 5(b) provides the statistics of these positive examples. Data cleansing steps as mentioned in Section 4.1.1 were also used here.

The negative examples corresponding to each LE were created as follows: we randomly chose 9 tracks, which the user has not listened to, i.e., using the POP_RND random population method. Concretely, a positive sample consists of: User_ID+Track_ID+Previous_Track_ID+User_context, whereas a negative sample consists of User_ID+Negative_Track_ID+Previous_Track_ID+User_context.

5. EXPERIMENTS AND RESULTS

To provide some baseline results, we experimented with track recommendation tasks based on the two uses cases mentioned in the last section using factorization machines (FM) [22], a state-of-the-art recommendation algorithm.

5.1 Evaluation Metric and Evaluated Methods

As evaluation measures, we computed the mean reciprocal rank (MRR) values over all the test LEs for the settings POP_RND and POP_USER, respectively, as we are only interested in how the ranking methods perform in regards

Method (FM with different features)		CB	Cx	(a) Context-aware POP_RND	(a) Context-aware POP_USER	(b) Next-song POP_RND
1	[Base]			**0.7755**	0.3906	0.4770
2	[Base]+Valence	√		**0.8010**	0.6325	**0.7922**
3	[Base]+Tempo	√		0.7241	0.6581	0.6071
4	[Base]+Created_at (time)		√	0.5257	**0.7985**	0.5718
5	[Base]+Timezone		√	0.4274	0.4743	0.4052
6	[Base]+Hashtag		√	0.7515	0.7814	0.7852
7	[Base]+Sentiment		√	0.7137	**0.8854**	**0.8005**

Table 6. Evaluation results in mean reciprocal rank (MRR) for (a) context-aware recommendation for the POP_RND and POP_USER settings and (b) context-aware next song recommendation for $N = 1$, using the POP_RND setting. 'CB' and 'Cx' indicate whether the method is content-based or context-based, respectively. [Base] indicates a part of the input to FM. For (a) [Base] = User_ID+Track_ID and (b) [Base] = User_ID+Track_ID+Previous_Track_ID. We highlight the top two results per column using bold font.

to ranking the ground truth track (i.e., the positive example) as high as possible in the ordered list of recommendation candidates, among the other nine negative examples. The values of MRR range from 0 to 1, with higher value indicating better result. For example, if the ground truth track is ranked at the third place, the MRR would be equal to $1/3$.

Our implementation of FM was based on libFM [22]. [7] We set the dimensionality of the factorized two-way interactions to five and performed ten learning iterations to train our FM for each experiment.

We consider the following methods in our evaluation.

- Method 1 is the baseline method (indicated as '[Base]' in Table 6) that uses User_ID and Track_ID only, for the context-aware experiment and User_ID+Track_ID +Previous_Track_ID for the next-song experiment.
- Methods 2 and 3 additionally take into account different content features of the tracks. We specifically selected valence and tempo. Valence describes the musical positiveness conveyed by a track. Hence, it can also be seen as a suitable proxy for content descriptors, as valence depends on all the other audio features giving the emotional content of the song, and thus, a reflection of user mood too [33]. Tempo, on the other hand, gives the speed or pace of a track and enables perceiving of music in an organized manner [34].
- Methods 4 and 5 take into account different context features of the user (time of creation of the tweet and timezone), providing the date and time of tweeting about a track and a sense of the location of the user.
- Finally, method 6 considers hashtags while method 7 employs the sentiment scores of the hashtags .

5.2 Result on Context-aware Recommendation

Table 6(a) shows the results obtained for context-aware recommendation. Method 1 alone (i.e., using only user IDs and item IDs) works quite well for POP_RND, but this is not the case for POP_USER. For methods 2 and 3, the content features of the tracks give better results in the POP_RND setting. This confirms that content features can mitigate the cold start problem [19, 21]. Though making recommendations is more challenging in the POP_USER setting, as we have to select a relevant track, given a context from among tracks that a user has already listened to, we find that the hashtags and the sentiment information contained in them (methods 6 and 7) contribute to better personalized recommendations as compared to the POP_RND setting. The content features do not contribute much for this setting. We believe that these preliminary results may serve as a baseline for future context-aware recommendation tasks.

5.3 Result on Next-song Recommendation

Table 6(b) shows the results obtained for context-aware next-song recommendation. For methods 2 and 3, the content features of the audio tracks contribute significantly to next-song recommendation. This suggests some acoustic coherence between the sequence of tracks a user listens to. While timestamps (i.e., method 4) give a sense of how far apart consecutive LEs are, timezone (method 5) does not contribute much to make next-song recommendation. Methods 6 and 7 perform well for next-song recommendation. This suggests that the hashtags and sentiment information give important information about how the mood of the user evolves with time on listening to the tracks. This temporal and affective information can be used for personalized playlist generation. These experimental results demonstrate that the proposed dataset can be used to study how to leverage listening history along with other contextual features of users for building a music RS.

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented the #nowplaying-RS dataset, which can be used to compare and evaluate large-scale recommendation approaches in a real-life setting. We believe that the dataset can serve as a standard in benchmarking context-aware recommendation, or at least supplement existing datasets—particularly given the diverse content and

[7] http://www.libfm.org/

context features provided by the dataset.

We have only showcased two possible use cases of the dataset in this paper. Further use cases could be to use the combination of mood-related hashtags and timestamps that allows us to track the emotion variation of users (i.e., how people use music to modulate their emotion); and to utilize the combination of user language and tweet language that allows us to do culture-aware recommendation. The combination of content, context and temporal features makes it possible to explain in greater detail why we recommend an item to a user (i.e., explainability). We would further like to explore the scope of using the various features of the dataset to improve music recommendation quality using neural networks.

7. REFERENCES

[1] E. Zangerle, M. Pichl, W. Gassler, and G. Specht, "#nowplaying music dataset: Extracting listening behavior from twitter," in *Proc. Int. Workshop on Internet-Scale Multimedia Management*, 2014, pp. 21–26.

[2] A. Germain and J. Chakareski, "Spotify Me: Facebook-assisted automatic playlist generation." in *Proc. IEEE Int. Works. Multimedia Signal Processing*, 2013, pp. 25–28.

[3] M. Schedl, "The LFM-1b dataset for music retrieval and recommendation," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2016, pp. 103–110.

[4] D. Hauger, M. Schedl, A. Košir, and M. Tkalcic, "The Million Musical Tweets Dataset: what can we learn from microblogs," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2013.

[5] M. Schedl, "Leveraging microblogs for spatiotemporal music information retrieval." in *ECIR*. Springer, 2013, pp. 796–799.

[6] G. Vigliensoni and I. Fujinaga, "The music listening histories dataset," in *Proc. Int. Soc. Music Information Retrieval Conf.*, 2017, pp. 96–102.

[7] G. Dror, N. Koenigstein, Y. Koren, and M. Weimer, "The yahoo! music dataset and kdd-cup11," in *Proc. KDD Cup*, 2012, pp. 3–18.

[8] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2, pp. 89–119, 2012.

[9] G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in *Recommender Systems Handbook*, 1st ed., F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. New York, NY, USA: Springer-Verlag New York, Inc., 2010, ch. 7, pp. 217–253.

[10] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1304–1315, Oct 2013.

[11] Y.-H. Yang and Y.-C. Teng, "Quantitative study of music listening behavior in a smartphone context," *ACM Trans. Interactive Intelligent Systems*, vol. 5, no. 3, 2015.

[12] F. Figueiredo, B. Ribeiro, J. M. Almeida, and C. Faloutsos, "TribeFlow: Mining & predicting user trajectories," in *Proc. Int. Conf. World Wide Web*, 2016, pp. 695–706.

[13] Y. Zhu, H. Li, Y. Liao, B. Wang, Z. Guan, H. Liu, and D. Cai, "What to do next: Modeling user behaviors by time-LSTM," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2017, pp. 3602–3608.

[14] T. Donkers, B. Loepp, and J. Ziegler, "Sequential user-based recurrent neural network recommendations," in *Proc. ACM RecSys*, 2017, pp. 152–160.

[15] S. Chang, Y. Zhang, J. Tang, D. Yin, Y. Chang, M. A. Hasegawa-Johnson, and T. S. Huang, "Streaming recommender systems," in *Proc. Int. Conf. World Wide Web*, 2017, pp. 381–389.

[16] R. He, W.-C. Kang, and J. McAuley, "Translation-based recommendation," in *Proc. ACM RecSys*, 2017, pp. 161–169.

[17] M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi, "Personalizing session-based recommendations with hierarchical recurrent neural networks," in *Proc. ACM RecSys*, 2017, pp. 130–137.

[18] P. Loyola, C. Liu, and Y. Hirate, "Modeling user session and intent with an attention-based encoder-decoder architecture," in *Proc. ACM RecSys*, 2017, pp. 147–151.

[19] T. X. Tuan and T. M. Phuong, "3D convolutional networks for session-based recommendation with content features," in *Proc. ACM RecSys*, 2017, pp. 138–146.

[20] C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing, "Recurrent recommender networks," in *Proc. ACM Int. Conf. Web Search and Data Mining*, 2017, pp. 495–503.

[21] S.-Y. Chou, Y.-H. Yang, and Y.-C. Lin, "Evaluating music recommendation in a real-world setting: On data splitting and evaluation metrics," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2015, pp. 1–6.

[22] S. Rendle, "Factorization machines with libfm," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 3, p. 57, 2012.

[23] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 263–272.

[24] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, "Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science*, vol. 5, no. 1, p. 23, Jul 2016.

[25] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods," in *Proc. ACM Conf. Online Social Networks*, 2013, pp. 27–38.

[26] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[27] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. the ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2004, pp. 168–177.

[28] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2544–2558, 2010.

[29] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. Conf. Weblogs and Social Media*, 2014.

[30] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," vol. 50, pp. 723–762, 2014.

[31] X. Wang and Y. Wang, "Improving content-based and hybrid music recommendation using deep learning," in *Proc. ACM Multimedia*, 2014, pp. 627–636.

[32] R. Devooght and H. Bersini, "Long and short-term recommendations with recurrent neural networks," in *Proc. ACM Conf. User Modeling, Adaptation and Personalization*, 2017, pp. 13–21.

[33] B. Den Brinker, R. Van Dinther, and J. Skowronek, "Expressed music mood classification compared with valence and arousal ratings," *EURASIP J. Audio, Speech, and Music Processing*, vol. 2012, no. 1, p. 24, 2012.

[34] A. Fernández-Sotos, A. Fernández-Caballero, and J. M. Latorre, "Influence of tempo and rhythmic unit in musical emotion regulation," *Frontiers in Computational Neuroscience*, vol. 10, 2016.