Music4All-Onion — A Large-Scale Multi-Faceted Content-Centric Music Recommendation Dataset

Marta Moscati
marta.moscati@jku.at
Johannes Kepler University Linz
Linz, Austria

Emilia Parada-Cabaleiro
emilia.parada-cabaleiro@jku.at
Johannes Kepler University Linz and
Linz Institute of Technology
Linz, Austria

Yashar Deldjoo
yashar.deldjoo@poliba.it
Polytechnic University of Bari
Bari, Italy

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

Markus Schedl
markus.schedl@jku.at
Johannes Kepler University Linz and
Linz Institute of Technology
Linz, Austria

ABSTRACT

When we appreciate a piece of music, it is most naturally because of its content, including rhythmic, tonal, and timbral elements as well as its lyrics and semantics. This suggests that the human affinity for music is inherently *content-driven*. This kind of information is, however, still frequently neglected by mainstream recommendation models based on collaborative filtering that rely solely on user-item interactions to recommend items to users. A major reason for this neglect is the lack of standardized datasets that provide *both* collaborative and content information.

The work at hand addresses this shortcoming by introducing *Music4All-Onion*, a large-scale, multi-modal music dataset. The dataset expands the Music4All dataset by including 26 additional audio, video, and metadata characteristics for 109,269 music pieces. In addition, it provides a set of 252,984,396 listening records of 119,140 users, extracted from the online music platform Last.fm, which allows leveraging user-item interactions as well. We organize distinct item content features in an onion model according to their semantics, and perform a comprehensive examination of the impact of different layers of this model (e.g., audio features, user-generated content, and derivative content) on content-driven music recommendation, demonstrating how various content features influence accuracy, novelty, and fairness of music recommendation systems. In summary, with Music4All-Onion, we seek to bridge the gap between collaborative filtering music recommender systems and content-centric music recommendation requirements.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Multimedia and multimodal retrieval**; *Collaborative filtering*.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557656

KEYWORDS

Multimedia Content Analysis, Music Information Retrieval, Music Recommender Systems, Audio Features, Onion Model, Audio Signal, Lyrics, Natural Language Processing, Video, Image

ACM Reference Format:
Marta Moscati, Emilia Parada-Cabaleiro, Yashar Deldjoo, Eva Zangerle, and Markus Schedl. 2022. Music4All-Onion — A Large-Scale Multi-Faceted Content-Centric Music Recommendation Dataset. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3511808.3557656

1 INTRODUCTION AND MOTIVATION

With the spiraling increase of digital content available to users, and likewise interaction data between users and content items, Recommender Systems (RSs) have become ubiquitous. Compared to other domains, Music Recommender Systems (MRS) are characterized by a large item-set size and a high sparsity of user-item interactions, making them prone to issues such as the cold-start problem and popularity biases. Those issues are often mitigated with Content-Based Recommenders (CBRs) that leverage item features, as opposed to Collaborative Filtering (CF), which relies entirely on user-item interaction data. Music consumption is also characterized by the fact that human music perception happens at different levels of semantics, and often involves not only the listened audio signal but also textual or visual input. Additionally, owing to the developments in Music Information Retrieval (MIR), many techniques [4, 12] allow the audio-based extraction of features characterizing music items at different semantic levels. Because of these aspects, MRSs are particularly apt for research in CBRs [7–10].

One big obstacle to the development of advanced content-based MRSs is the lack of comprehensive, standardized, and large-scale datasets, providing features characterizing the items at different semantic levels. Another one is understanding how feature semantics affect recommendation, which requires a categorization of features depending on their semantic charge. We address both points in this paper. First, we present Music4All-Onion, a dataset that enhances the established Music4All [29] and LFM-2b [30] datasets, by including several additional item features. Second, we propose an onion model to organize item features according to their semantics, thus helping the interpretation of the impact of item features on the

Table 1: Publicly available content-centric multi-faceted datasets: AF=audio features (LL=low-level, HL=high-level)

Resource	Modalities	Songs
AB Genre [3]	AF (HL+LL), genre	1,935,991
ALF-200k [42]	AF (HL), lyrics features	226,747
MuMu [23]	AF (HL+LL), genre, image, text, ratings, product similarities	147,295
Music4All [29]	AF (HL), genre, tags	109,269
MusiClef [31]	AF (HL+LL), genre, mood, tags, artist descriptions	1,355
MIREX Mood [25]	AF (HL+LL), score, lyrics	193
URMP [17]	Audio, video	44
Music4All-Onion	AF (HL+LL), lyrics embeddings, genre, tags, video embeddings	109,269

recommendation task. We benchmark, in terms of accuracy and beyond-accuracy metrics, these newly categorized features by comparing the performance of CBRs fueled by these features among each other and with pure CF models. Our analysis shows that content features improve recommendation accuracy with respect to pure CF, and that multi-modal CBRs, leveraging several features simultaneously, achieve the best performance. We also show that optimal selection of content features depends on the objective of the MRS, e.g., maximizing accuracy, diversity, or fairness.

Our contribution is, therefore, three-fold: First, we introduce Music4All-Onion, a large-scale multi-faceted dataset for music recommendation. Second, we propose an onion model for categorizing features according to their semantics. Based on these two contributions, we show how multi-modality improves recommendation, and how different features can be leveraged to optimize for accuracy and beyond-accuracy metrics. We provide the Music4All-Onion dataset and accompanying source codes for the conducted experiments at http://www.cp.jku.at/datasets/Music4All-Onion.

2 RELATED RESOURCES

While there are many datasets in the fields of RSs and MIR, only a few combine multiple modalities: those publicly available vastly differ in terms of size and covered modalities (see Table 1).

The AcousticBrainz (AB) Genre dataset [3] provides audio features (via AcousticBrainz) and genre information for up to 1,935,991 songs. Genre labels are collected from four different sources, where genres are organized hierarchically into main genres and subgenres. The ALF-200k dataset [42] combines acoustic and lyrics features, resulting in 176 high-level (HL) features for each of the 226,747 included tracks. The Multimodal Music dataset (MuMu) [23, 24] is based on the Million Song Dataset (MSD) [1] and the Amazon Reviews dataset [20] and encompasses 147,295 songs. The MuMu dataset combines information on purchases of albums (recovered from Amazon) with information on individual tracks, hence providing audio features (extracted from AcousticBrainz), multi-label genre annotations, album reviews, average rating per album, selling rank, similar products, and URL of the cover image of the album. The proposed Music4All-Onion dataset extends the Music4All dataset, which provides 109,269 songs and high-level acoustic features (extracted from Spotify), genres, and Last.fm tags. The MusiClef dataset [31] contains 1,355 popular music songs and provides

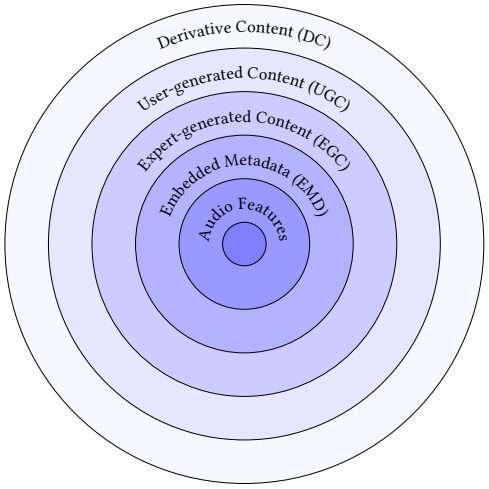

Figure 1: Onion model of music features.

high- and low-level (LL) audio features (e.g., MFCCs and block-level features), manually annotated genre and mood labels by domain experts, Last.fm tags, and textual artist descriptions crawled from various websites. The MIREX mood dataset [25] is based on the mood tags used in the MIREX mood classification task [14]. Songs annotated with these tags are retrieved from AllMusic and extended with mood tags, lyrics, and MIDI data for each of the 193 songs. The University of Rochester Multimodal Music Performance (URMP) dataset [17] provides 44 multi-instrument classical music pieces, where for each track, the audio for the individual tracks, the musical score in MIDI format, the audio and video recording of the assembled mixture, and frame and note-level pitches are contained.

Music4All-Onion both interlinks and substantially extends the established Music4All [29] and LFM-2b [30] datasets. In contrast to existing datasets, Music4All-Onion is large-scale *and* provides features extracted from audio, video, and metadata for 109,269 music tracks. In addition, it includes 252,984,396 listening records of 119,140 users of Last.fm. This combination of rich content features across multiple modalities and extensive collaborative information (listening records) makes it a unique resource for RSs research.

3 ONION MODEL OF MUSIC FEATURES

We present an onion model (as depicted in Figure 1) proposed by Deldjoo et al. [10] to categorize music content features in layers that reflect a transition from highly objective features with a low semantic charge (the inner layers) to more subjective and semantically meaningful features (the outer layers). The innermost layer corresponds to features extracted from the raw audio signal, commonly adopting traditional MIR signal processing techniques. Features in the Embedded Metadata (EMD) layer contain descriptive and technical metadata such as artist, track, and album name, or lyrics. Expert-Generated Content (EGC) refers to attributes assigned by or filtered with information from users with training or experience in the music domain, while User-Generated Content (UGC) encompasses information attached to items by general users. The Derivative Content (DC) layer refers to works created in relation to the original.

Table 2: Features introduced in Music4All-Onion, categorized according to the layers of the onion model.

Layer	Features
Audio	Acoustic (Short-term, Block-level)
EMD	Lyrics (preprocessed, tf-idf, word2vec, emotions)
EGC	Genres (tf-idf)
UGC	Tags (dictionary with Last-FM API weights, tf-idf)
DC	Video (VGG19, Incp3, Resnet)

4 THE MUSIC4ALL-ONION DATASET

In this section, we describe the features provided by Music4All-Onion, categorized according to the onion model introduced in Section 3. These are summarized in Table 2. We also describe the additional set of listening events obtained by matching Music4All-Onion with the popular LFM-2b dataset [30].

4.1 Audio

Low-level (LL) and high-level (HL) features are extracted from the audio signal. We divide these features into short-term and block-level features, depending on the length of the sequence considered.

4.1.1 Short-term features. Short-term features are extracted at frame-level and subsequently aggregated to obtain a feature vector per instance. The 13 Mel Frequency Cepstral Coefficients (MFCCs), encoding timbre information, are extracted with `kaldi` [26], and aggregated in three different ways: as statistical summarization by concatenating the mean and flattened covariance matrix as described in [16], as Bag of Audio Words (BoAW) computed with `openXBOW` [33], and as i-vectors [11] of three different dimensionalities. Using `openSMILE` [12], we also extract pitch- and emotion-related [36] features. Pitch features are aggregated as BoAWs, while emotion-related features are aggregated using the statistical aggregators of `openSMILE`, as well as with BoAWs. We also extract two large-scale hand-crafted feature sets: ComParE [34] extracted with `openSMILE`, which contain 6,373 features computed from MFCCs, spectral, prosodic, and voice quality descriptors through statistical aggregation; and spectral, time-domain, rhythm, and tonal frame descriptors aggregated via mean and standard deviation, extracted with Essentia [4].

4.1.2 Block-level features. Compared to short-term features, block-level features (BLFs) [35] are computed on *longer* sequences (several seconds) of spectrograms, and then aggregated using percentiles. We compute the six features defined in [35], capturing spectral, harmonic, rhythmic, and tonal music characteristics.

4.2 Embedded Metadata (EMD)

In addition to existing artist, album, and track names, we extract different representations of song lyrics.

4.2.1 Preprocessed lyrics. We provide a preprocessed version of the lyrics of Music4All, obtained after lowercasing, removing superfluous white spaces, consecutive newlines, and annotation (e.g., [guitar] or [chorus]), duplicating segments (e.g., [3x] or [x2]), translating lyrics from other languages to English, replacing numbers with English words, substituting English contractions with spelled-out forms, removing special characters and stopwords, applying lemmatization and stemming.

4.2.2 Lyrics embeddings. We provide two vector representations of the preprocessed lyrics: word2vec and tf-idf. The word2vec representation is obtained by first mapping each word to its 300-dimensional pre-trained word2vec embedding [21], and then averaging each component over the set of words. For tf-idf, tf is defined in terms of absolute word counts while idf is defined as

$$\mathrm{idf}(t) = \log\left[\frac{1+n}{1+\mathrm{df}(t)}\right] + 1, \tag{1}$$

n being the total number of tracks, and df being the fraction of lyrics documents (one document for each track) in which the term appears. The resulting tf-idf vectors are L_2-normalized.

4.2.3 Lyrics emotions. Emotional content is represented by mapping words from the lyrics onto valence, arousal, and dominance values according to the extended *Affective Norms for English Words* [41] lexicon. Words not present in this lexicon are mapped with the National Research Council Canada (NRC) lexicon [22]. In addition, we compute the polarity compound measure according to the Valence Aware Dictionary for sEntiment Reasoning (VADER) [15] using NLTK [2]. The values are aggregated as BoWs.

4.3 Expert-Generated Content (EGC)

The maintainers of the original Music4All dataset infer track genre by filtering out the tags appearing on the track Last.fm page, by the genres defined on Every Noise at Once,[1] and provide the resulting genre list per track. We convert those lists into tf-idf representations. We exclude genres that are associated with only one track. The tf of a specific genre of a track is defined as one divided by the number of genres attached to the track, idf is defined in Equation 1.

4.4 User-Generated Content (UGC)

The Music4All dataset comes with lists of track tags crawled from the Last.fm website. Those give no information on how often each tag was associated with the track. To fill this gap, we provide the tags retrieved with the Last.fm API,[2] which attaches a weight ($\in \{0, \ldots, 100\}$) to each tag depending on the frequency of its occurrence for the track under consideration, i.e., how many users assigned the tag to the track. We further convert the tags into a tf-idf representation by first filtering out tags with more than 50 characters (to remove tags consisting of sentences or extracts of the lyrics) and removing tags appearing in less than 5 tracks (to remove tags that are only meaningful to a very restricted subset of users). We then transform the tags for each track into a tf-idf representation, where tf is defined as the Last.fm tag weight divided by the sum of all weights of the track and idf as in Equation 1.

4.5 Derivative Content (DC)

Since official music videos frequently feature additional artistic contributions, such as those of directors or occasionally actors, and since many YouTube videos do not correspond to the official ones, but rather are covers or videos created by users of YouTube, we consider videos of songs uploaded to YouTube to be DC. For 98,877 out of 109,269 tracks, YouTube videos are available; we download them and extract image frames at the rate of 1 Hz. Each frame is then converted to three different vector representations using pretrained versions of VGG19 [38], Inception v3 [39], and Resnet [13], and aggregated to track level using maximum and mean.

[1] https://everynoise.com/

[2] https://www.last.fm/api/show/track.getTopTags

Table 3: Performance of the recommenders in terms of accuracy and beyond-accuracy metrics sorted descendingly on NDCG. For each metric, the best value is marked in bold, the second-best value in italic, and the worse is underlined.

Model	Feature/Aggregation	Layer	NDCG	Recall	User Entropy	Item Entropy	Coverage	Novelty
BiVAE	TWB	Mixed	**0.0641**	**0.0646**	*0.3800*	0.0760	0.0041	-1.5788
BiVAE	Video (VGG19)	DC	*0.0575*	*0.0604*	0.3573	0.1304	0.0517	-1.5723
BiVAE	Lyrics (tf-idf)	EMD	0.0558	0.0582	**0.3836**	0.1251	0.0438	-1.5734
BiVAE	Audio (i-vec256)	Audio	0.0545	0.0555	0.3783	0.0884	0.0434	-1.5769
MostPop	—	—	0.0537	0.0524	0.3622	0.0000	0.0016	-1.5850
BPR	—	—	0.0534	0.0515	0.3617	0.0000	0.0017	-1.5850
BiVAE	Genres (tf-idf)	EGC	0.0501	0.0549	0.3738	**0.1865**	**0.0601**	**-1.5656**
BiVAE	tags (tf-idf)	UGC	0.0499	0.0536	0.3586	*0.1776*	*0.0526*	*-1.5671*
BiVAE	—	—	0.0495	0.0557	0.3747	0.1574	0.0468	-1.5694

4.6 LFM-2b Listening Events

While Music4All-Onion provides a plethora of item features, the LFM-2b dataset [30] provides comprehensive information on the users. To enable leveraging both user and item features, and to provide an extra set of listening events, we match the tracks of the two datasets on their Spotify Uniform Resource Indicator (URI). Of the 109,269 tracks of Music4All, 56,512 (about 51%) also appear in LFM-2b. Restricting the listening events of LFM-2b to those tracks results in 252,984,396 listening events and 50,016,042 unique pairs of (user, item), corresponding to at least one listening event.

5 BENCHMARKING

We showcase the impact of Music4All-Onion by comparing the performance of Bilateral Variational Autoencoders (BiVAEs) [40] leveraging features from different layers of the onion model (i-vectors of 256 components, tf-idf of lyrics, genres and tags, and VGG19 representation of videos) to learn the priors of the item VAE. This CBR is built on VAEs, which have been proven to be successful for recommendation tasks [18, 19, 37, 40]. Furthermore, we consider two CF models: BiVAE with Gaussian priors (i.e., not leveraging any item feature) and matrix factorization with Bayesian Personalized Ranking (BPR) [27], as well as a non-personalized algorithm recommending the overall most popular items to all users (MostPop). In addition to accuracy metrics (NDCG@10 and Recall@10), we include several beyond-accuracy metrics, defined in [32], that we briefly describe here. Item- and user-entropy measure how well relevant recommendations are spread on the set of items and users, respectively, with higher values of entropy indicating *fairer* recommenders [6]. Coverage is the fraction of items in the catalog appearing at least once in the top-10 recommendations. Novelty is a measure of how likely the recommender is to make unpopular recommendations. The BiVAE models and BPR are trained using the library Cornac [28], while for the evaluation we rely on our implementation of the metrics since Cornac does not provide beyond-accuracy metrics, and since it does not allow evaluating the performance of models that are not included in the library, a feature required for optimization of and comparison with the aggregated model introduced below. The dimensionality of all latent representations is set to 10. For BiVAE, the encoders consist of a hidden layer with 20 nodes and tanh activation, and are trained for 100 epochs with a batch size of 128 and a learning rate of 0.001. The regularization hyperparameter in BPR is set to 0.01 and the model is trained for 200 iterations. To evaluate the impact of multi-modality on recommendation, we also consider a late fusion of all the BiVAE-based models with a generalization of Borda count rank-aggregation that we name *Truncated Weighted Borda* (TWB) [5]: ranking points are assigned to the top-50 items of every individual model and weighted with L_1-normalized weights. The combination of weights is optimized on NDCG, performing a grid-search with weights $\omega \in \{0, 0.2, \ldots, 1\}$.

The set of songs consists of the 79,072 items for which all features are available. The corresponding listening events of Music4All (4.2M) are binarized by assigning 1 to the (user, item) pairs with listening counts greater than or equal to 2, resulting in 707,284 positive user-item interactions. These are split into a train (60%), a validation (20%), and a test (20%) set. The weights for TWB are optimized on the validation set. All reported results refer to the test set. By inspecting the results in Table 3, the following conclusions can be drawn. On **accuracy**, the best values of NDCG and recall are achieved by the aggregated model, with a combination of weights $(\omega_{\text{Aud}}, \omega_{\text{Lyr}}, \omega_{\text{Gen}}, \omega_{\text{Tag}}, \omega_{\text{Vid}}, \omega_{\text{CF}}) = (0.2, 0.2, 0, 0.2, 0.4, 0)$, indicating apositive impact of multi-modality on music recommendation. The results show a trade-off between **accuracy** and **beyond-accuracy/fairness** metrics, indicating that different modalities can be leveraged depending on the evaluation dimension to be optimized, which can depend on the interests of the various stakeholders of the MRS. Optimizing consumer-side measures such as accuracy and user-fairness may result in the selection of TWB or video features, which negatively impacts provider-centric metrics (coverage, item-fairness). For a two-sided equitable ecosystem, the system designer may choose the weights accordingly.

6 CONCLUSIONS

We introduce Music4All-Onion, a large-scale multi-modal dataset providing 26 content features for 109,269 songs. We also propose an onion model to organize these features according to their differing semantic charge. A set of experiments shows that content-based MRSs should leverage different features, depending on which objective is to be optimized, and that content-based music recommenders tend to outperform pure CF algorithms in terms of accuracy, with multi-modal variants achieving the best performance.

ACKNOWLEDGMENTS

This research was funded in whole, or in part, by the Austrian Science Funds (FWF): P33526 and DFH-23. The authors are grateful to Giulio Davide Carparelli for providing support for the experiments.

REFERENCES

[1] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. In *Proc. of ISMIR* (Miami Beach, FL, USA). 591–596.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc., Sebastopol, CA, USA.

[3] Dmitry Bogdanov, Alastair Porter, Hendrik Schreiber, Julián Urbano, and Sergio Oramas. 2019. The Acousticbrainz Genre Dataset: Multi-Source, Multi-Level, Multi-Label, and Large-Scale. In *Proc. of ISMIR* (Delft, Netherlands). 360–367.

[4] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José Ricardo Zapata, and Xavier Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval. In *Proc. of ISMIR* (Curitiba, PR, Brazil). 493–498.

[5] Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. Audio-Visual Encoding of Multimedia Content for Enhancing Movie Recommendations. In *Proc. of ACM RecSys* (Vancouver, British Columbia, Canada). 455–459.

[6] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogin, Alessandro Difonzo, and Dario Zanzonelli. 2022. A Survey of Research on Fair Recommender Systems. *arXiv preprint arXiv:2205.11127* (2022).

[7] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2018. Content-Based Multimedia Recommendation Systems: Definition and Application Domains. In *Proc. of Italian Information Retrieval Workshop* (Rome, Italy).

[8] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–38.

[9] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, Yinwei Wei, and Xiangnan He. 2022. Multimedia Recommender Systems: Algorithms and Challenges. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, New York, NY, USA, 973–1014.

[10] Yashar Deldjoo, Markus Schedl, and Peter Knees. 2021. Content-driven Music Recommendation: Evolution, State of the Art, and Challenges. *arXiv:2107.11803 preprint* (2021).

[11] Hamid Eghbal-Zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. 2015. I-Vectors for Timbre-Based Music Similarity and Music Artist Classification. In *Proc. of ISMIR* (Málaga, Spain). 554–560.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. OpenSMILE: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proc. of ACM Multimedia* (Florence, Italy). 1459–1462.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. of IEEE CVPR* (Las Vegas, NV, USA). 770–778.

[14] Xiao Hu, J. S. Downie, Cyril Laurier, Mert Bay, and Andreas F. Ehmann. 2008. The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In *Proc. of ISMIR* (Philadelphia, PA, USA). 462–467.

[15] Clayton Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proc. of AAAI ICWSM* (Michigan, MI, USA). 216–225.

[16] Peter Knees and Markus Schedl. 2016. *Music Similarity and Retrieval: An Introduction to Audio- and Web-Based Strategies.* Springer, Berlin, Germany.

[17] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. 2018. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Transactions on Multimedia* 21, 2 (2018), 522–535.

[18] Xiaopeng Li and James She. 2017. Collaborative Variational Autoencoder for Recommender Systems. In *Proc. of ACM SIGKDD* (Halifax, Nova Scotia, Canada). 305–314.

[19] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *Proc. of ACM WWW* (Lyon, France). 689–698.

[20] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proc. of ACM SIGIR* (Santiago, Chile). 43–52.

[21] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of ICLR* (Scottsdale, AZ, USA).

[22] Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proc. of ACL* (Melbourne, Australia). 174–184.

[23] Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra. 2018. Multimodal Deep Learning or Music Genre Classification. *Transactions of the International Society for Music Information Retrieval* 1, 1 (2018), 4–21.

[24] Sergio Oramas, Oriol Nieto, Francesco Barbieri, and Xavier Serra. 2017. Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features. In *Proc. of ISMIR* (Suzhou, China). 23–30.

[25] Renato Eduardo Silva Panda, Ricardo Malheiro, Bruno Rocha, António Pedro Oliveira, and Rui Pedro Paiva. 2013. Multi-Modal Music Emotion Recognition: A New dataset, Methodology and Comparative Analysis. In *Proc. of CMMR* (Marseille, France). 570–582.

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi Speech Recognition Toolkit. In *Proc. of IEEE ASRU* (Waikoloa, HI, USA).

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *Proc. of UAI* (Montreal, Quebec, Canada). 452–461.

[28] Aghiles Salah, Quoc-Tuan Truong, and Hady W. Lauw. 2020. Cornac: A Comparative Framework for Multimodal Recommender Systems. *Journal of Machine Learning Research* 21, 95 (2020), 1–5.

[29] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. 2020. Music4all: A New Music Database and its Applications. In *Proc. of IEEE IWSSIP* (Niteroi, Rio de Janeiro, Brazil). 399–404.

[30] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. 2022. LFM-2b: A Dataset of Enriched Music Listening Events for Recommender Systems Research and Fairness Analysis. In *Proc. of ACM SIGIR CHIIR* (Regensburg, Germany). 337–341.

[31] Markus Schedl, Nicola Orio, Cynthia CS Liem, and Geoffroy Peeters. 2013. A Professionally Annotated and Enriched Multimodal Dataset on Popular Music. In *Proc. of ACM MMSys* (Oslo, Norway). 78–83.

[32] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current Challenges and Visions in Music Recommender Systems Research. *International Journal of Multimedia Information Retrieval* 7, 2 (2018), 95–116.

[33] Maximilian Schmitt and Björn Schuller. 2017. OpenXBOW: Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. *Journal of Machine Learning Research* 18, 1 (2017), 3370–3374.

[34] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. 2013. The Interspeech Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. of Interspeech* (Lyon, France). 148–152.

[35] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. 2010. Automatic Music Tag Classification Based on Block-Level Features. In *Proc. of SMC* (Barcelona, Spain).

[36] Tiancheng Shen, Jia Jia, Yan Li, Yihui Ma, Yaohua Bu, Hanjie Wang, Bo Chen, Tat-Seng Chua, and Wendy Hall. 2020. PEIA: Personality and Emotion Integrated Attentive Model for Music Recommendation on Social Media Platforms. In *Proc. of AAAI Conference on Artificial Intelligence* (New York, NY, USA). 206–213.

[37] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. 2020. RecVAE: A New Variational Autoencoder for Top-N Recommendations with Implicit Feedback. In *Proc. of ACM WSDM* (Houston, TX, USA). 528–536.

[38] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. of ICLR* (San Diego, CA, USA).

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper with Convolutions. In *Proc. of IEEE CVPR* (Boston, MA, USA). 1–9.

[40] Quoc-Tuan Truong, Aghiles Salah, and Hady W. Lauw. 2021. Bilateral Variational Autoencoder for Collaborative Filtering. In *Proc. of ACM WSDM* (Jerusalem, Israel). 292–300.

[41] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas. *Behavior Research Methods* 45, 4 (2013), 1191–1207.

[42] Eva Zangerle, Michael Tschuggnall, Stefan Wurzinger, and Günther Specht. 2018. Alf-200k: Towards Extensive Multimodal Analyses of Music Tracks and Playlists. In *Proc. of ECIR* (Grenoble, France). 584–590.