ALF-200k: Towards Extensive Multimodal Analyses of Music Tracks and Playlists

Eva Zangerle, Michael Tschuggnall, Stefan Wurzinger, Günther Specht

Department of Computer Science
Universität Innsbruck
firstname.lastname@uibk.ac.at

Abstract. In recent years, approaches in music information retrieval have been based on multimodal analyses of music incorporating audio as well as lyrics features. Because most of those approaches are lacking reusable, high-quality datasets, in this work we propose ALF-200k, a publicly available, novel dataset including 176 audio and lyrics features of more than 200,000 tracks and their attribution to more than 11,000 user-created playlists. While the dataset is of general purpose and thus, may be used in experiments for diverse music information retrieval problems, we present a first multimodal study on playlist features and particularly analyze, which type of features are shared within specific playlists and thus, characterize it. We show that while acoustic features act as the major glue between tracks contained in a playlists, also lyrics features are a powerful means to attribute tracks to playlists.

Keywords: music information retrieval, multimodal dataset, lyrics features, audio features, playlist analyses, classification

1 Introduction

With the advent of music streaming platforms, the way users consume music has changed fundamentally. Users stream music from large online music collections and listen to it using a variety of devices [1]. Platforms like Spotify[1] naturally also provide means for creating personal playlists. The analysis of such playlists has mostly been performed from either an (automatic) playlist generation perspective (e.g., [2]) or an organizational perspective (e.g., [3]). Also, features extracted from tracks (either from the track's audio signal or from metadata) have been utilized for music classification tasks (e.g., [4, 5]). Similarly, multimodal approaches that combine audio with lyrics features have been proposed for tasks like genre classification (e.g., [5]) or emotion recognition (e.g., [6]). Nevertheless, especially when incorporating song lyrics, most of the datasets used either lack quantity or quality, due to the variety and quality of online lyrics sources. In this work, we at first present the ALF-200k dataset (ALF stands for Acoustic and Lyrics Features), a novel dataset tackling this problem by providing an extensive set of more than 200,000 music tracks together with their occurrences in

[1] https://www.spotify.com, accessed October 2017

user's playlists and 176 pre-computed audio and lyrics features. As a first case study incorporating this dataset, we set out to analyze user-generated playlists regarding features that are shared among the tracks within a given playlist, i.e., features that characterize the playlist, are utilized implicitly and in an presumably unconscious manner. Particularly, we perform a multimodal classification task on the characteristics of playlists gathered from Spotify and analyze these in regards to their predictive power. By modeling the analyses as a classification task on a per-playlist basis, we show that acoustic features act as the major glue between tracks contained in the same playlist. We foresee that the dataset and the proposed collection approach may contribute to improving the collection of correct and comprehensive lyrics and audio features in future research.

2 Related Work

Multimodal approaches incorporating both the audio signal and lyrics have been shown to perform well [5–7] for genre classification. Mayer et al. [5, 7] use rhyme, part-of-speech, bag-of-words and text statistics for genre classification. They showed that lyrics features can be used orthogonally to audio features and that they can be beneficial in determining different genres. Other approaches solely rely on features extracted from lyrics: Fell and Sporleder [8] propose an n-gram model incorporating vocabulary, style, semantics and song structure for genre classification. Hu et al. [6] propose to use basic text features, lyrics content (bag-of-words), linguistic features, psychological categories, contained sentiment and text-stylistic features for music emotion recognition. However, for none of these tasks datasets of sufficient size are publicly available. The dataset most similar to ALF-200k, the Million Song Dataset (MSD) [9], features one million songs, according artists, last.fm tags and similarities. The musixmatch extension to the MSD dataset provides a mapping between the MSD dataset and lyrics on the musixmatch platform. However, while the MSD contains audio features, no lyrics features are provided. On the other side, solely lyrics features have been utilized for mood detection, e.g., by the MoodyLyrics dataset [10]. Nevertheless, to our knowledge, the proposed large-scale ALF-200k dataset is novel in that it combines rich lyrics and audio features at scale.

3 Dataset

In the following, we present the methods utilized for creating the ALF-200k dataset. To foster reproducibility and repeatability, we make our code and data publicly available on GitHub[2].

Generally, we aim to curate a dataset containing tracks, respective lyrics and audio features and playlists of users containing these tracks. We therefore rely on the dataset collected by Pichl et al. [11], which contains 18,000 playlists created by 1,016 users, resulting in a total of 670,000 distinct tracks.

[2] https://github.com/dbis-uibk/ALF200k

As for the corresponding lyrics features, we propose the following crawling method to ensure reliable, correct and complete lyrics for the analyses: At first, we utilize the provided Spofiy IDs of Pichl's dataset to gather artist names and titles of the according tracks. Along the lines of previous research [4, 5], we subsequently search for corresponding lyrics on the following user-contributed lyrics databases. Concretely, we utilize ChartLyrics, LYRICSnMUSIC, LyricWikia, eLyrics.net, LYRICSMODE, METROLYRICS, Mp3lyrics, SING365, SONGLYRICS and Songtexte.com. While the former three platforms provide an API that allows for gathering lyrics based on artist name and track title, the latter seven do not provide any interface and hence, have to be scraped by gathering the HTML code of the underlying websites. After having gathered the lyrics from the proposed platforms, all tracks with non-English lyrics are removed as a number of features are not available for other languages (e.g., uncommon or slang words). In a next step, we clean the obtained lyrics by removing non-UTF8 characters, superfluous white-spaces and also by removing typical characteristics of online lyrics like track structure annotations (e.g., verse/chorus/interlude/...), references and abbreviations of repetitions (e.g., "–Chorus (x2)–"), annotations of background voices (e.g., "yeah yeah yeah") or track remarks (e.g., "written by" or "Duration: 3:01"). Subsequently, we incorporate only lyrics into the ALF-200k dataset that are confirmed by at least three of the crawled lyrics platforms. Therefore, we compute the similarity of all found lyrics versions for a given track and rely on word bigrams as a representation of each crawled lyrics. Next, we apply the Jaccard similarity coefficient on the set of word bigrams representing the lyrics for all pairs of lyrics. Finally, we choose the version for which at least three sources share a high similarity according to an empirically estimated threshold. If less than three sources confirm a specific lyrics variant, the respective track is removed from the dataset as it would not be possible to reliably extract lyrics features from it. This presents us with a total of 226,747 lyrics.

As we also aim to include extracted features in the dataset, we rely on *audio and lyrics features* to represent tracks, as these have been shown to be orthogonal and beneficial in multimodal approaches [6, 5, 12]. As for *audio content descriptors* of tracks, we rely on standard acoustic features retrieved via the Spotify Track API[3]. These content features are extracted and aggregated from the audio signal of a track and comprise: *danceability* (how suitable a track is for dancing), *energy* (perceived intensity and activity), *speechiness* (presence of spoken words in a track), *acousticness* (confidence whether track is acoustic), *instrumentalness* (prediction whether track contains no vocals), *tempo* (in beats per minute), *valence* (musical positiveness conveyed), *liveness* (prediction whether track was recorded live or in studio), *duration* (total time of track) and *loudness* (sound intensity in decibels). Besides acoustic features, we also incorporate more than hundred different *lyrics features* which have been shown to be beneficial for track classifications [8, 4]. We thereby included four different types of lyrics features: lexical [6], linguistic [5, 13, 8], syntactic [8] and semantic [6, 14, 8, 15]

[3] https://developer.spotify.com/web-api/, accessed October 2017

features. Due to space constraints, we provide a detailed overview of all features in Table 1 and refer the interested reader to the according papers.

Type	#	**Features**
Acoustic (AU)	10	danceability, energy, speechiness, liveness, acousticness, valence, tempo, duration, loudness, instrumentalness
Lexical (LX)	34	bag-of-words* (4), token count, unique token ratios (3), avg. token length, repeated token ratio, hapax dis-/tris-/legomenon, unique tokens/line, avg. tokens/line, line counts (5), words/lines/chars per min., punctuation and digit ratios (9), stop words ratio, stop words/line
Linguistic (LI)	39	uncommon words ratios (2), slang words ratio, lemma ratio, Rhyme Analyzer features (24), echoisms (3), repetitive structures (8)
Semantic (SE)	55	Regressive imagery (RI) conceptual thought features (7), RI emotion features (7), RI primordial thought features (29), SentiStrength scores (3), AFINN scores (4), Opinion Lexicon scores, VADER scores (4)
Syntactic (SY)	38	POS bag-of-words*, pronouns frequencies (7), POS frequencies (6), text chunks (23), past tense ratio

Table 1. Extracted Lyrics Features (# refers to the number of features contained; bag-of-word features (marked with *) are counted as one feature each, despite that they amount to hundreds of features depending on the lyrics).

4 Case Study: User Playlist Characteristics

As a first case study based on the ALF-200k dataset, we are interested in finding features that are shared among tracks within playlists. Therefore, we apply the following method: for each playlist of size s (i.e., playlists containing s tracks), we add s random tracks that are not contained in the original playlist. This allows us to evaluate the binary classification performance by measuring the accuracy at which any given track in the test set was predicted to be part of the playlist or not. By utilizing 5-fold cross-evaluation, the performance of classifiers is measured by computing the average classification accuracy, averaged across all folds. We rely on a set of standard classification approaches provided by the Weka framework [16]: BayesNet, Naïve Bayes, KNN, SVM with different kernels (linear, C-SVM, nu-SVM), J48 decision trees and PART, utilizing the respective standard parameter configurations.

In a preliminary experiment, we determined the minimum required length of a playlist to contain enough reasonable data (tracks) and removed all playlists that do not fulfill the minimum required playlist size (8 tracks), which results in a dataset comprising 7,903 playlists. For each playlist in the dataset, we apply a 5-fold cross-validation and measure the prediction accuracy.

Table 2 lists the average accuracies of all individual feature sets and combinations thereof. Being in line with previous findings (e.g., [5]), the best result is achieved by the SVMs with linear and nu-kernel and reaches 70% by utilizing only acoustic features (AU), slightly outperforming the set of all available

Feature Set	**BayesNet**	**J48**	**kNN**	**LibLinear**	**LibSVM (C)**	**LibSVM (nu)**	**Naïve Bayes**	**PART**	**Max.**
AU	0.65	0.66	0.66	0.70	0.62	0.70	0.65	0.66	**0.70**
all features	0.69	0.62	0.58	0.68	0.55	0.68	0.69	0.62	**0.69**
AU+LX+LI+SE	0.69	0.62	0.57	0.67	0.55	0.67	0.69	0.62	**0.69**
AU+LX+LI	0.69	0.63	0.56	0.67	0.55	0.67	0.69	0.63	**0.69**
AU+LX	0.69	0.63	0.55	0.66	0.54	0.67	0.69	0.63	0.69
LX+LI+SE+SY	0.66	0.60	0.57	0.67	0.55	0.67	0.66	0.60	**0.67**
LX+LI+SE	0.66	0.60	0.56	0.66	0.55	0.66	0.66	0.60	**0.66**
LX+LI	0.65	0.60	0.55	0.65	0.54	0.66	0.65	0.60	**0.66**
LX	0.65	0.60	0.54	0.65	0.54	0.65	0.64	0.60	**0.65**
LI	0.57	0.58	0.57	0.61	0.53	0.61	0.57	0.58	**0.61**
SY	0.59	0.57	0.57	0.60	0.52	0.60	0.59	0.57	**0.60**
SE	0.55	0.56	0.55	0.57	0.51	0.58	0.55	0.56	**0.58**
Baseline	***0.50***	***0.50***	***0.50***	***0.50***	***0.50***	***0.50***	***0.50***	***0.50***	***0.50***

Table 2. Average Accuracies For Each Classifier, Sorted By Maximum Accuracy (Max.)

features. At a first glance, this indicates that acoustic features represent the main characteristic that holds playlists together. Except for Naïve Bayes and BayesNet, AU reached the best accuracy values for all classifiers. As can be seen in Table 2, the feature sets achieving the worst accuracy results are SE, SY and LI. These findings suggest that—when inspected individually—semantic (e.g., contained sentiment or psychological categories), syntactic or linguistic features (uncommon or slang words) are not able to fully capture what actually makes the tracks of a playlist cohesive. Nevertheless, the best combination without relying on acoustic metrics, i.e., combining only textual features extracted from song lyrics, gains an accuracy of 67%, which is only slightly inferior to the best result. Thus, our preliminary experiments demonstrate that lyrics within playlists are homogeneous to a substantial extent and that they can be used to attribute tracks to playlists. Finally, we also note that the analysis conducted is not able to capture user-specific contextual motivations to put certain tracks in a playlist (besides the mere characteristics of tracks). Users may also create playlists to remember certain events and the music they associate with this occasion as, e.g., their wedding or holidays, where the cohesive features of the playlist do not necessarily lie in the track's characteristics, but rather in the perceived emotion and evoked memories. Nevertheless, we believe that this study can provide interesting and relevant insights into the composition of playlists on streaming platforms from a multimodal perspective.

5 Conclusion and Future Work

In this paper, we presented ALF-200k, a novel, publicly available dataset for multimodal music classification problems, containing over 200,000 tracks including precomputed audio features as well as hundreds of metrics extracted from high-quality lyrics. In an exemplary case study we analyzed multimodal features, particularly focusing on detecting features that are shared within playlists and hence, characterize those playlists. As for future work, we are highly interested in utilizing and learning from the dataset to be able to automatically group playlists per genre, user and also per context [17]. Furthermore, we aim to evaluate the characteristics of Spotify playlists to the quality criterion applied for playlist recommendation tasks [2].

References

1. Zhang, B., Kreitz, G., Isaksson, M., Ubillos, J., Urdaneta, G., Pouwelse, J.A., Epema, D.: Understanding user behavior in spotify. In: 2013 Proc. IEEE INFOCOM. (April 2013) 220–224
2. Bonnin, G., Jannach, D.: Automated generation of music playlists: Survey and experiments. ACM Computing Surveys (CSUR) **47**(2) (2015) 26
3. Kamalzadeh, M., Baur, D., Möller, T.: A survey on music listening and management behaviours. In: Proc. ISMIR 2012. (2012)
4. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. American music **183**(5,049) (2009) 2–209
5. Mayer, R., Neumayer, R., Rauber, A.: Combination of audio and lyrics features for genre classification in digital audio collections. In: Proc. ACM MM. (2008) 159–168
6. Hu, X., Downie, J.S.: Improving mood classification in music digital libraries by combining lyrics and audio. In: Proc. JCDL 2010, ACM (2010) 159–168
7. Mayer, R., Neumayer, R., Rauber, A.: Rhyme and style features for musical genre classification by song lyrics. In: Proc. ISMIR 2008. (2008) 337–342
8. Fell, M., Sporleder, C.: Lyrics-based analysis and classification of music. In: COLING. Volume 2014. (2014) 620–631
9. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: Ismir. Volume 2. (2011) 10
10. Çano, E., Morisio, M.: Moodylyrics: A sentiment annotated lyrics dataset. In: Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence. ISMSI '17, New York, NY, USA, ACM (2017) 118–124
11. Pichl, M., Zangerle, E., Specht, G.: Understanding Playlist Creation on Music Streaming Platforms. In: Proc. IEEE Symposium on Multimedia, IEEE (2016)
12. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proc. ICMLA 2008, IEEE (2008) 688–693
13. Hirjee, H., Brown, D.G.: Rhyme analyzer: An analysis tool for rap lyrics. In: Proc. ISMIR 2010. (2010)
14. Martindale, C.: Romantic progression: The psychology of literary history. (1976)
15. Ribeiro, F.N., Araújo, M., Gonçalves, P., André Gonçalves, M., Benevenuto, F.: Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science **5**(1) (Jul 2016) 23
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter **11**(1) (2009) 10–18
17. Pichl, M., Zangerle, E., Specht, G.: Towards a context-aware music recommendation approach: What is hidden in the playlist name? In: Proc. ICDM Workshops. (2015) 1360–1365