CROWDSOURCING EMOTIONS IN MUSIC DOMAIN

Erion Çano and Maurizio Morisio

Department of Control and Computer Engineering, Polytechnic University of Turin,
Duca degli Abruzzi, 24, 10129 Torino, Italy

ABSTRACT

An important source of intelligence for music emotion recognition today comes from user-provided community tags about songs or artists. Recent crowdsourcing approaches such as harvesting social tags, design of collaborative games and web services or the use of Mechanical Turk, are becoming popular in the literature. They provide a cheap, quick and efficient method, contrary to professional labeling of songs which is expensive and does not scale for creating large datasets. In this paper we discuss the viability of various crowdsourcing instruments providing examples from research works. We also share our own experience, illustrating the steps we followed using tags collected from Last.fm for the creation of two music mood datasets which are rendered public. While processing affect tags of Last.fm, we observed that they tend to be biased towards positive emotions; the resulting dataset thus contain more positive songs than negative ones.

KEYWORDS

Social User Tags, Crowdsourcing Emotions, Music Emotion Recognition, Affective Computing

1. INTRODUCTION

Music Information Retrieval (MIR) and Music Emotion Recognition (MER) are two important research directions that are changing the way people find and listen to music. They put together data mining or machine learning techniques with several types of music features and metadata. With the exponential growth of user feedback found in the web, social tags are becoming very important for powering music searches, generating high performance music recommendations, building classifiers that identify genre, instrumentation of even emotions in songs etc. In music listening context, a social tag is just a free text label that a user applies in music-related objects like songs, artists or albums. These tags capture contextual and descriptive information about the resource they are associated to. In most of the cases, there is no limit in number of tags that can be assigned to an object, and no vocabulary restriction in forming tags. As a consequence tags usually do contain irrelevant information or noise as well. Noisy and imperfect as they might be, social tags are a source of human-generated contextual information that is become an essential part of the solution to many MIR and MER problems. For this reason, many researchers and developers are experimenting with different ways of obtaining tags.

In fact a typical problem is finding subjects who can generate descriptive tags or labels about songs. A small set of several hundred songs can require thousands of song labels generated by hundreds of people which is hardly feasible. The high cognitive load makes this process time consuming [1] and cross agreement is also difficult to achieve due to the subjective and ambiguous nature of music perception [2]. Outsourcing hand-labeling of songs to professionals results in high quality labeling but is very expensive. Organizations like Pandora have employed many musical experts as part of Music Genome Project[1]. They continuously tag songs with a large vocabulary of musically relevant words. However they are unwilling to make their datasets

[1] https://www.pandora.com/about/mgp

DOI : 10.5121/ijaia.2017.8403

public. A recent tendency which seems very promising for alleviating this problem is the ever growing phenomenon of crowdsourcing which reveals itself in various forms. There are already various crowdsourcing marketplaces like Amazon Mechanical Turk[2] which represent a quick and relatively cheap alternative for subjective user feedback about musical items. Studies like [3] and [4] suggest that this method is viable if properly applied. Other crowdsourcing approaches that are already being explored are harvesting free tags from social forums like Last.fm[3], collaborative games, web services etc. Last.fm is a popular music community that has rendered public most of its music collections and corresponding user-generated tags [5]. Is is highly popular among researchers who have been experimenting with its tags in several studies [6, 7, 8, 9]. In many cases researchers utilize music tags to build labeled datasets which are essential for training and testing MIR and MER algorithms. In other cases collected tags are used to compare efficiency of different crowdsorucing approaches or reliability of user community tags.

In this paper we discuss the different and popular crowdsourcing methods that are being used to collect human judgement about music emotions. We show examples from literature that use MTurk campaigns or design collaborative games for making the experience of music tagging more attractive. There are also many examples that discuss ways of collecting and processing social tags. We also present different models of music emotions, such as that of Russell [10] and share our experience of using Last.fm tags for creating 2 music emotion datasets. These datasets are published for research use and can be freely downloaded from http://softeng.polito.it/erion/. The first dataset (MoodyLyrics4Q) includes 5075 songs divided in 4 emotion categories. The second dataset (MoodyLyricsPN) is a bigger collection of 5940 positive and 2589 negative songs. There was a high bias towards positive emotions and songs which is reflected in emotion category sizes. The rest of this paper is structured as follows: Section 2 discusses crowdsorucing as a new research and work paradigm. Section 3 presents the most popular music emotion models that are popular in the literature. In section 4 we describe the various crowdsorucing approaches that are being successfully applied to collect human judgment in music domain. In section 5 we share our experience of creating the 2 music mood datasets and also present their characteristics. Finally, section 6 concludes.

2. CROWDSOURCING AS A NEW PARADIGM

Crowdsourcing is a phrase that was first coined by Jeff Howe in [11]. In that article he presents this cheap (or even free) and network powered labor paradigm by describing several successful examples where distributed efforts from enthusiasts have proven very successful in solving certain problems. These examples show how intelligence of the crowds is helping various companies to shorten development cycles of products, lower production costs and find R&D solutions that are brilliant and very cheap at the same time. The basic principles behind crowdsourcing paradigm as presented by Jurowiecki [12] in his book “The Wisdom of Crowds”, are diversity and independence of opinions, decentralization of works and aggregation of opinions. As illustrated in [13], the interested or requesting parties could be working groups, institutions, communities, industries, governments, or global societies. They publish problems or job requests in web platforms which serve as distributed labor networks or marketplaces. Some of these networks such as InnoCentive, NineSigma, iStockphoto or Your Encore target specialized or talented subjects, especially for R&D creative problems in specific areas [14]. There are also other examples of crowdsourcing campaigns conceived as challenges with very generous rewards. Netflix $1M Prize[4] was a very popular challenge in computer science realm. They offered the prize for the team that would propose a movie recommendation algorithm with prediction error 10% lower than the state of the art. This challenge had positive public impact,

[2] http://mturk.com
[3] https://www.last.fm
[4] http://www.netflixprize.com/

boosting work in the field of recommender systems which are now part of almost every commercial or advertising web platforms. Our focus here is on crowdsourcing mechanisms that can provide subjective human feedback about songs. To this end, other platforms like MTurk have been proved highly effective. MTurk doesn't address complex or innovative projects but is mostly a marketplace for short and simple microtasks that any person who is online can find and solve. In fact any "worker" on MTurk can browse and select different Human Intelligence Tasks (HITs) based on their nature, complexity or payment incentive. Each participant is paid by the requester of the task upon successful accomplishment of that task and it is also possible to assign bonus compensations to workers who ensure results of a high quality. This arrangement is appropriate for tasks that require massive social participation or activities which aim to collect subjective feedback to be used as experimental data, which is what we discuss in this paper.

The high popularity of MTurk has attracted interest among academics. For example, to check the linguistic diversity of MTurk workers, authors in [15] conducted a survey organized as a set of paid translation tasks, targeting bilingual workers and including words from 100 languages. Based on their results, they recommend 13 languages (e.g., Gujarati, French, German, Italian etc.) that have vast populations of workers who provide fast and qualitative results. They also report that India is the country with the highest number of workers. Authors in [16] discuss several issues regarding the use of MTurk participants in social, psychological or linguistic experiments. They especially address two concerns: Participation of reemerging (or so called nonnaïveté) workers in same or related experiments and the reactive tendency of researchers to excessively exclude prolific workers. They recommend researchers to assess whether participants have been involved in similar surveys before, especially for experiments in which the naïveté assumption of participants is highly important.

Despite the benefits of crowdsrurcing and the popularity of many initiatives, there are certain barriers that have hindered several other initiatives from taking off. In [17] the author lists some unsuccessful crowdsourcing initiatives such as Gambrian House or CrowdSpirit and also tries to spot out generic obstructive factors of crowdsourcing paradigm. According to him, the very first problem of a crowdsourcing initiative could be the generation of interest among Internet users. Furthermore, initiatives that pass the first obstacle are not guaranteed to succeed in convincing people to contribute or stay engaged in the several tasks or projects being requested. Other problems of crowdsourcing could face are "not invented here" syndrome of organizations, labour exploitation discussions, legislative issues related to copyright, employment, data security and privacy etc. As more and more companies and individuals embark in this work and production model, new ethical and legal regulations or standards of best practice will hopefully emerge as well. In section 4 we illustrate with real examples how different crowdsourcing approaches have been used to collect human judgement about various and especially emotional characteristics of musical pieces.

3. ORGANIZING MUSIC EMOTIONS

Prior to collecting descriptions about emotional aspects of music, it is essential to select and use a generic model or taxonomy of moods expressed or induced by songs. This is particularly true when the goal is to create ground-truth categorizations of songs for training and testing Music MER systems. The resulting collections should follow popular emotion models to be used by many researchers in various tasks. In this context, psychological models of emotion induced by music are a useful instrument to simplify the emotion space and have a intuitive and manageable set of classes. We therefore observed the literature for music emotion models that are both practical and widely accepted among psychologists. Apparently there are two types of music emotion models: categorical and dimensional. Categorical models represent music emotions using labels or short textual descriptors. Those labels that are semantically close are clustered together to form a category. Dimensional models on the other hand, represent music emotions using

numerical values of few parameters like Valence, Arousal etc. Having a look in the literature, we find many works trying to depict categorical or dimensional models of emotions. An early study was conducted by Hevner [18] in 1936. She describes a categorical model of 66 affect adjectives clustered in 8 groups as presented in Figure 1. Hevner's model hasn't been used much in its basic form. Nevertheless it has been used as a reference point for many studies that have also used categorical models of affect. Among dimensional models, the most popular is probably the model of Russell which is based on valence and arousal [10]. High and low (or positive and negative) values of these 2 parameters create a space of 4 categories as shown in Figure 2.

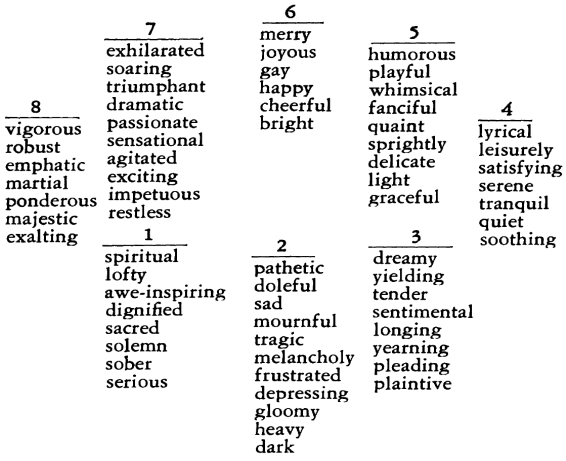

Figure 1. Model of Hevner

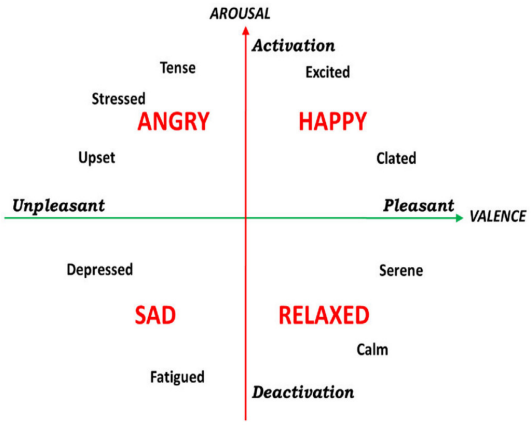

Figure 2. Mood classes in model of Russell

Models of Henver and Russell represent theoretical works of psychologists and do not certainly reflect the reality of everyday music listening. Various works try to verify to what scale such expert models match the semantic models obtained from crowdsourcing community user tags. Authors harvest mood tags from music listening communities and examine mood term cooccurence in songs. In [19] for example, authors construct an affect taxonomy of 5 clusters by analyzing AllMusic[5] user tags. That taxonomy illustrated in Figure 3 has been used in MIREX AMC task[6] since 2007. It however reveals problems like the overlaps between clusters 2 and 4. Those overlaps are related to the semantic similarity between *fun* and *humorous* terms [20]. Also,

[5] http://www.allmusic.com
[6] http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification

cluster 1 and cluster 5 share acoustic similarities and are often confused with each other. Same authors utilize Last.fm tags to construct a simplified representation of 3 clusters presented in [21]. They put together 19 basic mood tags of Last.fm and 2554 tracks of USPOP[7] song collection. To reach to the model they perform K-means clustering with 3 to 12 clusters of tags among all songs. The representation with only 3 clusters of terms results the optimal choice, even though it seems oversimplified. Authors however recommend this approach as a practical guide for similar relevant studies. A similar study was conducted in [22] where authors merge audio features with Last.fm tags. They perform clustering of all 178 mood terms that are found in AllMusic portal, reducing the mood space in 4 categories very similar to those of Russell's models. They conclude that user tag semantic features are high-level and valuable to complement the low-level audio features for higher accuracy.

Clusters	Mood Adjectives
Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Figure 3. Model used in MIREX AMC task

Another relevant work was conducted in [23] utilizing tracks and tags found in Last.fm. After selecting the most appropriate mood terms and tracks, they apply unsupervised clustering and Expected Maximization algorithm to the document-term matrix. According to their results, the optimal number of term clusters is 4. Their 4 clusters of emotion terms are very similar to the 4 clusters of the planar valence-Arousal model proposed by Russell (happy, angry, sad, relaxed). These results confirm that emotion models derived from user community affect tags are in agreement with the basic emotion models of psychologists and can be practically useful for sentiment analysis or music mood recognition tasks.

4. CROWDSOURCING MUSIC TAGS

In this section we present examples from literature that employ popular crowdsourcing approaches to harvest subjective user tags of musical pieces. We see that there are various motives that push researchers to work with music tags. In many cases they collect tags to build ground-truth labeled datasets which are essential for training supervised MER systems. In other cases they explore the reliability of user tags for building effective MIR systems or helping music searches. There are even studies where researchers try to compare quality and effectiveness of the different crowdsorucing approaches they try. In [24] for example, they explore and examine 5 approaches concurrently: user surveys, harvesting social tags, annotation games, mining websites and content-based autotagging.

4.1 MUSIC TAGS FROM MTURK WORKERS

As we mentioned above, MTurk has gained popularity as a very useful marketplace of micro tasks. In [25] for example, authors employ MTurk workers to gather multiple tags about

[7] https://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html

experimental musical tracks. They explore the similarity of human generated tags about different parts of the same musical track. Tags of different categories (genre, mood, instrument, feel, and other) are collected and a conditional restricted Boltzmann machine is built to model their relationship. Based on their results, authors report that different parts of same song tend to be described differently, particularly as they are more and more displaced one from another. Other works such as [26] have created music datasets by fusing textual and musical features together. They extract and use mixed features of 100 popular songs annotated from Amazon MTurk workers. The resulting dataset is available for research upon request to the authors. Authors in [4] perform a comparative analysis between mood annotations collected from MoodSwings, a collaborative game they developed (see next section), and annotations crowdsourced from paid MTurk workers. They follow the 2-dimensional arousal-valence mood representation model of 4 categories. Based on their statistical analysis, they report consistencies between MoodSwings and MTurk data and conclude that crowdsourcing mood tags is a viable method for ground truth dataset generation. Their dataset consisting of 240 song clips was released for public use[8]. In [3] we find another study that examines quality of music tags crowdsourced using MTurk. Authors here contrast MTurk data with those of MIREX AMC 2007 task and report similar distribution on the MIREX clusters. Authors conclude that generally, MTurk crowdsourcing can serve as a cheap and applicable option for music mood ground truth data creation. However particular attention should be paid to possible problems such as spamming that can diminish annotation quality. The viability of Amazon MTurk as a means for collecting human judgment about music is also explored in [27]. The authors submitted 1047 music excerpt similarity HITs and after integrity checking, collected 6732 unique judgements. They compare agreement rate between the 6732 similarity judgments of MTurk with judgements for the same query-candidate pairs obtained from Evalutron6000, a web-based system that aids collection and analysis of music similarity data [28]. Authors conclude that MTurk produces analogous results to using Evalutron6000 and that MTurk may be used as a reliable means for music similarity judgements.

4.2 MUSIC TAGS FROM COLLABORATIVE GAMES

The so-called “games with a purpose” represent an interesting crowdsorucing form that is also becoming popular in recent literature. Playing such games is obviously more attractive for users rather than completing repetitive and monotonous HITs in MTurk. One of the first implementations of a game-based crowdsourcing systems was an online game called ESP [29] which was used to label images. Two players collaborated with each other to label Internet images based on their content. In music domain, one of the first online games built to collect tags is MajorMiner[9] described in [30]. Users participate in this entertaining experience by listening to 10 second clips and providing audio-related and objective tags from a list. The collected data can be later used as ground truth for training music classifiers and recommenders. Based on the data they collected, authors report that players agree on many musical characteristics. They also make a comparison of their tags with those obtained from Last.fm. According to their results, MajorMiner tags are of a higher quality and Last.fm user tags tend to be more noisy, containing a lot of non-musical descriptors. They thus suggest that combining high quality tags which are scarce with social tags from sites like Last.fm which are more numerous could lead to more robust and effective music description systems. Another contribution in the category “game with a purpose” is presented in [31]. The authors describe ListenGame, a collaborative multi-player game designed to collect semantic word descriptors of audio content. Each user is asked to listen to a 15 seconds clip selected from 250 popular western songs and then chose the best and the worst word to describe the clip out of 6 words per semantic category (e.g., instrumentation, mood etc.). Authors use the collected tags to train a supervised multiclass labeling model which provides annotations for new unknown songs. Another collaborative online game designed to

[8] http://music.ece.drexel.edu/research/emotion/moodswingsturk
[9] http://majorminer.org/info/intro

crowdsource tags about muscal clips is Tag A Tune presented in [32]. Authors pretend that Tag A Tune is better than both Maj or Miner and Listen Game in 2 aspects: Players are involved in a richer audio environment that is not limited to songs only and instead of playing against a database or other players, they are coupled with a partner with whom they have to tag tunes agreeably. According to them, the effectiveness of Tag A Tune depends on the entertaining capabilities of the game to attract a massive and continuous number of players and on the agreement rate of paired players in tagging audio clips. Based on a user survey they conducted, users usually consider the game to be attractive and the collected labels are descriptive and meaningful.

MoodSwings may be the first collaborative game that specifically addressed crowdsourcing of emotion tags from online players [33]. This game differs from the above 3 in that it is designed to assess the time-varying emotional characteristics of music by gathering labels in a per-second basis. Those labels are obtained from user ratings in the two-dimensinal valence-arousal plane. Based on collected data, authors report a bias towards high arousal and valence and that users mostly provide emotional points located near the center of valence-arousal quadrants (moderate emotions), avoiding extreme points. The last game we discuss here is Emotify which was presented in [34]. In this very recent work authors crowdsource emotion tags for the creation of a music mood dataset. For categorizing emotions they utilize a model of 9 terms called GEMS (Geneva Emotional Music Scale) that was specifically created to represent emotions induced by music. It means that user emotional measures of clips in Emotify are discrete, contrary to MoodSwings where they were continuous (points in a 2-dimensional plane). A valuable outcome of this work is the dataset of emotion labels for 400 musical excerpts which is rendered public and can be used to train and test MER algorithms.

4.3 MUSIC TAGS FROM COMMUNITY USERS

The growth of web social media in the last 15 years has obviously changed the way most people listen to music. Music listening and appraisal has become social and collective with platforms like Spotify, Pandora, Last.fm and other serving songs to millions of users every day. These platforms offer a wide range of features like creating playlists, sharing favorite songs with other users or tagging songs. Last.fm is obviously the most popular among academics, mainly because of its open API which has granted researchers access to its titles and tags. In fact millions of music listeners in Last.fm continue to provide different types of tags about their favorite songs and artists. There are several motives that push them to provide tags about songs. In [35] and [36] authors discuss those reasons that appear to be fundamental and not restricted to music domain. Creating context and organization of tasks is one such reason. Many users tag songs to augment contextual relevance and enhance organization of information. This way they assist their own future search and retrieval tasks. For example, users may group their favorite songs based on tags to facilitate their everyday music listening. This way they also make a social contribution helping search tasks of other users as well. Opinion expression is also an important reason for tagging songs. A music listener may apply tags to songs to share his/her musical opinion and tastes. Social exposure, self-presentation or even attraction of attention are somehow related to opinion expression and emphasize the social and collective nature of music listening. Despite the usefulness of tags for MIR community, they also reveal some problems that need to be addressed. Polysemy of provided tags is one such problem, especially in those platforms that do not enforce tagging limitations. There is no common vocabulary of tags and misspellings or junk words represent a source of noise. The subjective nature of music listening and tagging leads to popularity bias which is another problem. Newer and unknown artists or songs tend to receive fewer tags whereas those that are highly popular attract most of attention. Careful and intelligent data preprocessing is thus essential to overcome the flaws of music tags, especially when they are used to build intelligent MIR or MER systems.

Distribution of Last.fm tag types was first examined in [37]. Here we find a quantitative exploration of the different types of tags the users provide. Most of them (68%) are related to *Genre* of songs. *Locale* counts for 12% of the total followed by *Mood* with 5%, *Instrumentation* with 4% and *Opinion* also with 4%. There are numerous examples from academic works where social community tags about songs are successfully used to create emotion taxonomies, datasets or even real intelligent applications. One of the first works that crowdsourced user generated mood tags of popular songs is [19]. Here authors report the uneven distribution of mood term vocabulary and conclude that many of the terms are highly interrelated or express different aspects of a common and more general mood class. Also in [38], authors use Last.fm tags to create a large dataset of 5296 tracks and 18 emotion categories. To tag the tracks they employ a binary approach for all the mood categories, with songs having or not tags of a certain category. They make use of this dataset in [39] to evaluate the audio-text classifier they construct. A similar work is found in [40] where they use AMG tags to create a dataset of lyrics based on valence-arousal model of Russell. Tags are first cleared and categorized in one of the 4 quadrants of the model using valence and arousal norms of ANEW [20]. Then songs are classified based on the category of tags they have mostly received. Annotation quality was further validated by 3 persons. This is one of the few public lyrics datasets of a reasonable size (771 lyrics) that are available. In [41] authors describe Musiclef, a professionally created multimodal dataset. It contains metadata, audio features, Last.fm tags, web pages and expert labels for 1355 popular songs. Those songs have been annotated using an initial set of 188 terms which was finally reduced to 94.

In [42] authors try to compare semantic quality captured by social community tags such as those of Last.fm with that of music editorial annotations such as I Like Music (ILM). Based on their results they conclude that semantic emotion models are effective to predict emotions in both Last.fm and ILM datasets. They also infer that models of mood prediction can be build based on one corpus and effectively applied to another. In [43] they present a music generator which can be parameterized along valence and arousal axes. What matter from our perspective is their system validation procedure based on crowdsourced tags. They invite users to listen to 30 seconds music clips produced by the generator and then provide mood tags. In total they collected 2020 tags and report a slight bias towards positive valence and high arousal. There are also examples of using crowdsourced Last.fm tags for improving music search and recommendation systems. In [44] they explore the role of different types of tags in boosting web search operations. Based on their experiments, authors conclude that in music domain, more than 50% of tags bring new information to the resource they annotate. They also report that most of the tags help search operations and tagging behavior reveals same characteristics as searching. On the other hand, in [45] authors utilize the role of music tags in combination with listening habits to create musical profiles of users and improve recommendations. They report that adding tags to their music recommender helps solving problems such as cold-start and data sparsity.

4.4 Other Examples Of Crowdsourcing In Music

There are also examples of collecting music characteristics based on other strategies like traditional question-based surveys, online web services etc. In [46] for example, authors describe Songle, a web service that enriches music listening experience and improves itself by means of the error-correction user contributions. When using Songle, any user plays musical pieces visualizing 4 types of descriptions at the same time: structure of the track (chorus and repeated sections), beat structure (bar lines and beats), melody line (frequency of vocal melody) and chords (root note and chord type). These music-understanding visualizations are however erroneous because of the highly diverse and complex sound mixtures. Here is where crowdsourcing of user error corrections shows up. Each user who finds an error in visualizations can correct it by selecting from a candidate list or by providing an alternative description on Songle's interface. The resulting corrections are shared and utilized to improve the experience of

future users. In [47] they collect, process and publish audio content features of 500 popular western songs. For the annotation process they utilized a question based survey and paid participants who were asked to provide feedback about each song they listened to. The questions included 135 concepts about 6 music aspects such as genre, emotion, instrument etc. Emotion category comprised 18 possible labels such as happy, calming, bizarre etc. The numerous literature examples of using music community tags that we saw in this section, emphasize the high and rich semantic value they contain and the many ways they can be exploited to improve MIR or MER systems. In the following section we share our experience in creating two music mood datasets based on emotion tags crowdsourced from Last.fm.

5. DATASET CREATION FROM AFFECT TERMS

As we previously mentioned, one of the reasons that makes tags useful, is the possibility to use them for creating ground-truth labeled datasets of songs to train and test MER systems. Recently there is high attention on corpus-based methods that involve machine or deep learning techniques [48]. There are studies that successfully predict music emotions based on text features only [49, 50, 51] utilizing complex models. Large datasets of songs labeled with emotion categories are an essential prerequisite to train and exploit those classification models. Such music datasets should be:

1. Highly polarized to serve as ground truth
2. Labeled following a popular mood taxonomy
3. As large as possible (say more than 1000 titles)
4. Publicly available for cross-interpretation of results

In [52] we created a textual dataset based on content words. It is a rich set of lyrics that can be used to analyze text features. It however lacks human judgment about emotionality of songs, and therefore cannot be used as a ground truth set. Here we share our experience creating a dataset that fulfils all above requirements based on tags collected from Last.fm.

5.1 FOLKSONOMY OF MUSIC EMOTIONS

As discussed in Section 3, different models of music emotions have been proposed. For our dataset we utilized a folksonomy of 4 categories that is very similar to the one described in [23]. We used ***happy***, ***angry***, ***sad*** and ***relaxed*** (or ***Q1***, ***Q2***, ***Q3*** and ***Q4*** respectively) as representative terms for each cluster, in consonance with the popular planar representation of Figure 2. Doing so we complied with the second requirement listed above. First we picked up about 150 emotion terms from studies cited in the previous sections and also the current 289 mood terms in AllMusic website. We performed a manual selection process, accepting only terms that are clearly part of one of the 4 clusters. For a precise and objective selection we also consulted ANEW. Throughout this process we dropped many terms that do not necessarily or clearly describe affect or emotions (e.g., *patriotic*, *technical* etc.). We also found ambiguity in the categorization of certain terms by other similar studies we consulted. Those terms were removed as well. For example, *intense*, *rousing* and *passionate* have been set into 'angry' cluster in [23]. On the other hand, in [22] they appear as synonyms of 'happy'. Same happens with *spooky*, *wry*, *boisterous*, *sentimental* and *confident*; they also appear into different emotion categories. Considering valence and arousal norms in ANEW, we also dropped out terms that appear in the borders of neighbor clusters. To illustrate, *energetic*, *gritty* and *upbeat* appear between Q1 and Q2, *provocative* and *paranoid* between Q2 and Q3, *sentimental* and *yearning* appear between Q3 and Q4 whereas *elegant* is in the middle of Q1 and Q4. At the end of this phase we reached at the representation of Table 1 which appeared to be the optimal one. This representation includes the 10 most appropriate emotion terms in each cluster.

Table 1. Clusters of terms.

Q1-Happy	Q2-Angry	Q3-Sad	Q4-Relaxed
happy	angry	sad	relaxed
happiness	aggressive	bittersweet	tender
joyous	outrageous	bitter	soothing
bright	fierce	tragic	peaceful
cheerful	anxious	depressing	gentle
humorous	rebellious	sadness	soft
fun	tense	gloomy	quiet
merry	fiery	miserable	calm
exciting	hostile	funeral	mellow
silly	anger	sorrow	delicate

5.2 DATA PROCESSING AND STATISTICS

We started from a large collection of songs so that we could reach to a big final set and thus fulfill the third requirement. We included MSD collection which is probably the largest set of titles for research in music domain [53]. Created to be a reference point for evaluating results, it also supports scaling MIR or MER algorithms to commercial sizes. There are 943334 tracks in the collection, making it a great source for analyzing human perception of music by means of user tags. Playlist dataset on the other hand is a smaller but more recent collection of 75,262 songs crawled from yes.com, a website that provides music playlists from many radio stations in the United States. Authors of Playlist used the dataset to validate a method for automatic playlist generation they developed [54]. Merging Playlist and MSD we obtained a set of 1018596 songs with some duplicates that were removed. First we crawled all Last.fm tags for each artist-title entry of the collection. Afterwards we started data processing by dropping out songs which had no tags at all. This way we obtained 539702 songs with at least one Last.fm tag. At this point we analyzed tag frequency and distribution finding a total of 217768 unique tags appearing 4711936 times. The distribution was highly imbalanced with top hundred tags summing up to 1930923 entries, or 40.1% of the total. Top 200 tags appeared in 2385356 entries which is more than half (50.6%). Also 88109 or 40.46% of the tags appeared only once. They were mostly typos or junk patterns like ”111111111”, ”zzzzzzzzz” etc. There was an average of 9.8 tags for each song. Such uneven distribution of tags across tracks has previously been reported in [55] and [19]. Top 30 tags are shown in Table 2 together with their appearance frequency. Top tag is obviously *rock* showing up 139295 times. From Table 2 we can see that among top tags, those describing genre are dominant. Same as in [37], we analyzed distribution of top 100 tags in different categories such as genre, mood, instrument, epoch, opinion etc. Here we got a slightly different picture presented in Table 3. We see that genre tags are still the most frequent with 36% of the total. However there is also a considerable uprise of opinion and mood that make up 16.2% and 14.4% respectively. Nevertheless it is important to note that this numbers were sampled from our collection of songs and do not necessarily reflect an overall tendency in Last.fm tag distribution. Our focus here is in emotion tags most frequent of which are presented in Table 4. From the 40 terms shown in Table 1, only 11 also appear in this list. There are however many other terms that are highly synonymous. We also see that positive tags

Table 2. Thirty most frequent tags.

Rank	Tag	Freq	Rank	Tag	Freq
1	rock	139295	16	mellow	26890
2	pop	79083	17	american	26396
3	alternative	63885	18	folk	25898
4	indie	57298	19	chill	25632
5	electronic	48413	20	electronic	25239
6	favorites	45883	21	blues	25005
7	love	42826	22	british	24350
8	jazz	39918	23	favorite	24026
9	dance	36385	24	instrumental	23951
10	beautiful	32257	25	oldies	23902
11	metal	31450	26	80s	23429
12	00s	31432	27	punk	23233
13	soul	30450	28	90s	23018
14	awesome	30251	29	cool	21565
15	chillout	29334	30	country	19498

are clearly more numerous than negative ones. There are 8 term from quadrants Q1 and Q4 (high valence) and only 3 from Q2 and Q3 (low valence). The most popular affect term is *mellow* appearing 26890 times. As we can see, users are more predisposed to give feedback when perceiving positive emotions in music. Word cloud of affect tags is illustrated in Figure 4. Moving on with data processing, we kept only tags assigned to at least 20 songs, same as in [26]. We removed tags related to genre (e.g., *rock*, *pop*, *indie*), instrumentation (guitar, electronic), epoch (00s, 90s) or other tags not related to mood. We also removed ambiguous tags like *love* or *rocking* and tags that express opinion such as *great*, *good*, *bad* or *fail*, same as they did in [38]. It is not possible to know if tag *love* means that the song is about love or that the user loves that song. Similarly it is not possible to infer any emotionality from opinion tags such as *great*. It may mean that the song is positive but it is not necessarily the case. A melancholic song may be great as well. The process was finalized by removing all entries left with no tags, reducing the set from 539702 to 288708 entries.

Table 3. Distribution of tag classes.

Category	Frequency	Examples
Genre	36 %	rock, pop, jazz
Opinion	16.2 %	beautiful, favourite, good
Mood	14.4 %	happy, sad, fun
Instrument	9.7 %	guitar, instrumental, electronic
Epoch	7.2 %	00s, 90s, 80s
Locale	5.5 %	american, usa, british
Other	11 %	soundtrack, patriotic

5.3 Annotation Scheme And Results

To utilize as many tags as possible and produce a big dataset (third requirement), we extended the basic 10 terms of each cluster with their related forms derived from lemmatization. For example, it makes sense to assume that *relaxing*, *relax* and *relaxation* tags express the same opinion as *relaxed* which is part of cluster 4. Doing so we reached to a final set of 147 words that were the most meaningful from music emotion perspective. At this point we proceeded with the

identification of the tracks that could be distinctly fitted in one of the 4 mood clusters based on the tags they had received.

Table 4. Thirty most frequent mood tags.

Rank	Tag	Freq	Rank	Tag	Freq
16	mellow	26890	103	soft	7164
40	funk	16324	107	energetic	6827
45	fun	14777	109	groovy	6771
50	happy	13633	127	uplifting	5188
52	sad	13391	138	calm	4769
59	melancholy	12025	145	emotional	4515
63	smooth	11494	153	funny	4034
66	relax	10838	157	cute	3993
68	upbeat	10641	227	quirky	2606
69	relaxing	10513	230	moody	2549
78	melancholic	9392	231	quiet	2538
90	atmospheric	8149	236	bittersweet	2458
93	sweet	8006	241	angry	2361
96	dark	7668	242	soothing	2361
99	dreamy	7296	291	sentimental	1937

We started by identifying and counting the extended set of tags in each song. For each song we got 4 counters, representing the number of tags from each mood category.

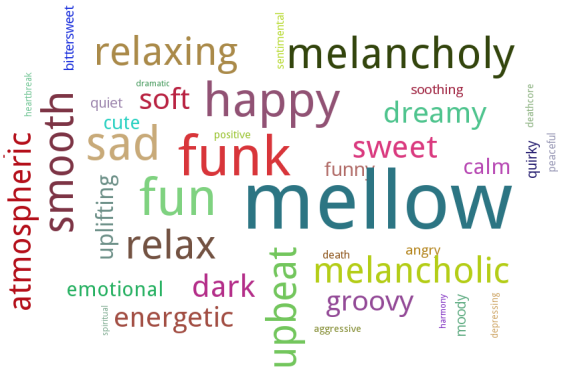

Figure 4. Cloud of most frequent affect tags

To keep in highly polarized songs only and thus fulfill the first requirement, we implemented a tight scheme denoted as ***4-0 or 6-1 or 9-2 or 14-3***. It means that a song is set to quadrant Qx if either one of the following conditions is fulfilled:

- It has 4 or more tags of Qx and no tags of any other quadrant
- It has 6 up to 8 tags of Qx and at most 1 tag of any other quadrant
- It has 9 up to 13 tags of Qx and at most 2 tags of any other quadrant
- It has 14 or more tags of Qx and at most 3 tags of any other quadrant

Songs with 3 or even fewer tags or not fulfilling one of the above conditions were discarded. The final set was a collection of 1986 happy, 574 angry, 783 sad and 1732 relaxed songs for a total of 5075. From this numbers we can see that the dataset we obtained is clearly imbalanced, with more songs being reported as positive (3718 in Q1 and Q4) and fewer as negative (only 1357 in

Q2 and Q3). This is something we expected, since as we mentioned above, tag distribution was imbalanced in the same way. A positive-negative representation of songs is clearly oversimplified and does not reveal much about song emotionality. Nevertheless such datasets are usually highly polarized. Positive and negative terms can be better distinguished from each other and the resulting datasets might be very useful for training and exercising many sentiment analysis or machine learning algorithms. For this reason we decided to create another datasets which divides song emotions as positive or negative only. We added more terms in the two categories, terms that couldn't be used with the 4 class annotation scheme. For example, tags like *passionate*, *confident* and *elegant* are positive, even though they are not distinctly happy or relaxed. Same happens with *wry*, *paranoid* and *spooky* on the negative side. We used valence norm of ANEW as an indicator of term positivity and reached to a final set of 557 terms. Given the fact that positive and negative terms were more numerous, for pos-neg classification we implemented ***5-0 or 8-1 or 12-2 or 16-3*** scheme which is even tighter. A song is considered to have positive or negative mood if it has 5 or more, 8-11, 12-15, or more than 15 tags of that category and 0, at most 1, 2, or at most 3 tags of the other category. Using this scheme we got a set of 2589 negative and 5940 positive songs for a total of 8529. Same as above, we see that positive songs are more numerous.

6. DISCUSSION

In this paper we presented the various crowdsourcing approaches that are being experimented for collecting subjective human judgment about emotionality of musical pieces. We described crowdsourcing as a new and emerging paradigm that is replacing other traditional research, work or production approaches. We also presented many literature works which apply crowdsourcing to harvest music tags from users in various forms, from MTurk campaigns to attractive collaborative games. According to several studies, these strategies if correctly applied are viable, cheap (sometimes even free) and effective. We also discussed different popular music emotion models that can be used to simplify emotion categories in MIR or MER studies and applications. Lastly, we illustrated the steps we followed for the creation of 2 music mood datasets we named MoodyLyrics4Q and MoodyLyricsPN by crowdsourcing tags from Last.fm. Analyzing Last.fm tags of songs, we observed that despite the growth of opinion and mood tags, genre tags are still the most numerous. Furthermore, those tags that express positive emotions (happy and relaxed) are dominant. For the classification of songs we used a tight scheme that labels each song based on its tag counters, guaranteeing polarized collections of songs in each emotion cluster. MoodyLyrics4Q and MoodyLyricsPN are publically available for research use. Researchers are invited to provide feedback or further extend them.

ACKNOWLEDGEMENTS

This work was supported by a fellowship from TIM[10]. Computational resources were provided by HPC@POLITO[11], a project of Academic Computing within the Department of Control and Computer Engineering at Politecnico di Torino.

REFERENCES

[1] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. IEEE Transactions on Multimedia, 13(2):303–319, April 2011.

[2] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. State of the art report: Music emotion recognition: A state of the art review. In Proceedings of the 11th International Society for Music Information Retrieval Conference, pages 255–266, Utrecht, The Netherlands, August 9-13 2010.

[10] https://www.tim.it/
[11] http://hpc.polito.it

[3] J. H. Lee and X. Hu. Generating ground truth for music mood classification using mechanical turk. In Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, pages 129–138, New York, NY, USA, 2012. ACM.

[4] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional musical mood annotation. In A. Klapuri and C. Leider, editors, ISMIR, pages 549–554. University of Miami, 2011.

[5] P. Lamere and E. Pampalk. Social tags and music information retrieval. In ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008, page 24, 2008.

[6] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In J. S. Downie and R. C. Veltkamp, editors, ISMIR, pages 619–624. International Society for Music Information Retrieval, 2010.

[7] Schedl, M., Orio, N., Liem, C., Peeters, G., A professionally annotated and enriched multimodal data set on popular music. In Proceedings of the 4th ACM Multimedia Systems Conference, pp. 78-83, ACM, February 2013

[8] Saari, P., Eerola, T., Semantic computing of moods based on tags in social media of music. IEEE Transactions on Knowledge and Data Engineering, 26(10), 2548-2560, 2014

[9] Malheiro, R., Panda, R., Gomes, P., Paiva, R., Music Emotion Recognition from Lyrics: A Comparative Study. 6th International Workshop on Machine Learning and Music (MML13).

[10] J. A. Russell. A circumplex model of affect. Journal of Personality and Social Psychology, 39:1161–1178, 1980.

[11] J. Howe. The Rise of Crowdsourcing. Wired Magazine, no. 14, pp. 1-5, 2006

[12] Surowiecki, J., The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday, 2005.

[13] Zhao, Y., Zhu, Q., Evaluation on crowdsourcing research: Current status and future direction, Information Systems Frontiers, July 2014, Volume 16, Issue 3, pp 417–434

[14] Brabham, D. C., Crowdsourcing as a Model for Problem Solving - An Introduction and Cases, Convergence 14.1 (2008): 75-90

[15] Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The Language Demographics of Amazon Mechanical Turk. Transactions Of The Association For Computational Linguistics, 2, 79-92.

[16] Chandler, J., Mueller, P. & Paolacci, G. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behav Res (2014) 46: 112. doi:10.3758/s13428-013-0365-7

[17] Simula, H., The Rise and Fall of Crowdsourcing?, Proceedings of the 2013 46th Hawaii International Conference on System Sciences, pp. 2783-279.

[18] K. Hevner. Experimental studies of the elements of expression in music. The American Journal of Psychology, 48(2):246–268, 1936.

[19] X. Hu and J. S. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In Proceedings of the 8th International Conference on Music Information Retrieval, pages 67–72, Vienna, Austria, September 23-27 2007.

[20] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

[21] X. Hu, M. Bay, and J. Downie. Creating a simplified music mood classification groundtruth set. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007), 2007.

[22] K. Bischoff, C. S. Firan, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo. Music mood and theme classification - a hybrid approach. In Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009, pages 657–662, 2009.

[23] C. Laurier, M. Sordo, J. Serr, and P. Herrera. Music mood representations from social tags. In K. Hirata, G. Tzanetakis, and K. Yoshii, editors, ISMIR, pages 381–386. International Society for Music Information Retrieval, 2009.

[24] Turnbull, D., Barrington, L., & Lanckriet, G. R., Five Approaches to Collecting Tags for Music. In ISMIR 2008, Vol. 8, pp. 225-230

[25] Mandel, M., Douglas, E., Bengio, Y., Learning tags that vary within a song, In Proceedings of the 11th International Conference on Music Information Retrieval, ISMIR 2010, Utrecht, Netherlands, pp. 399-404

[26] R. Mihalcea and C. Strapparava. Lyrics, music, and emotions. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea, pages 590–599, 2012.

[27] Lee, J.H., Crowdsourcing Music Similarity Judgments using Mechanical Turk. ISMIR 2010, pp. 183-188

[28] Gruzd, Anatoliy A., Downie, J. S., Jones, M. C., Lee, J. H., Evalutron 6000: Collecting Music Relevance Judgments, 7th ACM/IEEE Joint Conference on Digital Libraries, Vancuver, Canada, June 2007, pp. 507-507

[29] Von Ahn, L., Dabbish, L., Labeling images with a computer game. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04). ACM, New York, NY, USA, pp. 319-326.

[30] Mandel, M., Ellis, D., A Web-Based Game for Collecting Music Metadata, Journal of New Music Research, Vol. 37, No. 2, pp. 151-165

[31] Turnbull, D., Liu, R., Barrington, L. & Lanckriet, G. A Game-Based Approach for Collecting Semantic Annotations of Music. In Proceedings of the 8th International Conference on Music Information Retrieval. Vienna, Austria. September 23-27. pp. 535-538.

[32] Law, E. L., Von Ahn, L., Dannenberg, R. B., Crawford, M., TagATune: A Game for Music and Sound Annotation. In ISMIR 2007, Vol. 3, p. 2.

[33] Emelle, L., Kim, Y.E., & Schmidt, E.M. MoodSwings: A Collaborative Game for Music Mood Label Collection. ISMIR 2008.

[34] Aljanaki, A., Wiering, F. and Veltkamp, R.C., Studying emotion induced by music through a crowdsourcing game. Information Processing & Management, 2016, 52(1), pp.115-128.

[35] Marlow, C., Naaman, M., Boyd, D., Davis, M., HT06, tagging paper, taxonomy, Flickr, academic article, to read. In Proceedings of the seventeenth conference on Hypertext and hypermedia, pp. 31-40, ACM, August 2006.

[36] Ames, M., Naaman, M., Why we tag: motivations for annotation in mobile and online media. In Proceedings of the SIGCHI conference on Human factors in computing systems, ACM, April 2007, pp. 971-980

[37] Lamere, P., Social tagging and music information retrieval. Journal of new music research, 2008, No. 37(2), pp. 101-114.

[38] X. Hu, J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. In K. Hirata, G. Tzanetakis, and K. Yoshii, editors, ISMIR, pages 411–416. International Society for Music Information Retrieval, 2009.

[39] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL '10, pages 159–168, New York, NY, USA, 2010. ACM.

[40] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva. Classification and regression of music lyrics: Emotionally-significant features. In A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, J. Bernardino, and J. Filipe, editors, KDIR, pages 45–55. SciTePress, 2016.

[41] M. Schedl, C. C. Liem, G. Peeters, and N. Orio. A Professionally Annotated and Enriched Multimodal Data Set on Popular Music. In Proceedings of the 4th ACM Multimedia Systems Conference (MMSys 2013), Oslo, Norway, February–March 2013.

[42] Saari, P., Barthet, M., Fazekas, G., Eerola, T., & Sandler, M. Semantic models of musical mood: Comparison between crowd-sourced and curated editorial tags. In Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, IEEE, July 2013, pp. 1-6.

[43] Scirea, M., Nelson, M. J., Togelius, J., Moody music generator: Characterising control parameters using crowdsourcing. In International Conference on Evolutionary and Biologically Inspired Music and Art, pp. 200-211, Springer International Publishing, April 2015.

[44] Bischoff, K., Firan, C. S., Nejdl, W. & Paiu, R., Can all tags be used for search?. CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management (p./pp. 193--202), New York, NY, USA: ACM. ISBN: 978-1-59593-991-3

[45] Kim, H. H., Kim, D., & Jo, J. A unified music recommender system using listening habits and semantics of tags. International Journal of Intelligent Information and Database Systems 4, 8(1), 2014, pp. 14-30.

[46] Goto M., Yoshii K., Fujihara H., Mauch M., Nakano T., Songle: A Web Service for Active Music Listening Improved by User Contributions. In ISMIR 2011, pp. 311-316

[47] D. Turnbull, L. Barrington, D. A. Torres, and G. R. G. Lanckriet. Semantic annotation and retrieval of music and sound effects. IEEE Trans. Audio, Speech & Language Processing, 16(2):467–476, 2008.
[48] D. Tang, B. Qin, and T. Liu. Deep learning for sentiment analysis: Successful approaches and future challenges. Wiley Int. Rev. Data Min. and Knowl. Disc., 5(6):292–303, Nov. 2015.
[49] Giz H. He, J. Jin, Y. Xiong, B. Chen, W. Sun, and L. Zhao. Language Feature Mining for Music Emotion Classification via Supervised Learning from Lyrics, pages 426–435. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
[50] M. van Zaanen and P. Kanters. Automatic mood classification using tf*idf based on lyrics. In J. S. Downie and R. C. Veltkamp, editors, ISMIR, pages 75–80. International Society for Music Information Retrieval, 2010.
[51] H.-C. Kwon and M. Kim. Lyrics-based emotion classification using feature selection by partial syntactic analysis. 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI 2011), 00:960–964, 2011.
[52] E. Çano and M. Morisio, Moodylyrics: A sentiment annotated lyrics dataset, in Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, ISMSI '17, ACM, Hong Kong, March 2017, pp. 118–124. doi:10.1145/3059336.3059340.
[53] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In Proceedings of the 12th International Conference on Music Information Retrieval, ISMIR 2011.
[54] S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. Playlist prediction via metric embedding. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pages 714–722, New York, NY, USA, 2012. ACM.
[55] Y.-C. Lin, Y.-H. Yang, and H. H. Chen. Exploiting online music tags for music emotion classification. TOMCCAP, 7(Supplement):26, 2011.