MUSIC MOOD DATASET CREATION BASED ON LAST.FM TAGS

Erion Çano and Maurizio Morisio

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

ABSTRACT

Music emotion recognition today is based on techniques that require high quality and large emotionally labeled sets of songs to train algorithms. Manual and professional annotations of songs are costly and hardly accomplished. There is a high need for datasets that are public, highly polarized, large in size and following popular emotion representation models. In this paper we present the steps we followed to create two such datasets using intelligence of last.fm community tags. In the first dataset, songs are categorized based on an emotion space of four clusters we adopted from literature observations. The second dataset discriminates between positive and negative songs only. We also observed that last.fm mood tags are biased towards positive emotions. This imbalance of tags was reflected in cluster sizes of the resulting datasets we obtained; they contain more positive songs than negative ones.

KEYWORDS

Music Sentiment Analysis, Ground-truth Dataset, User Affect Tags, Semantic Mood Spaces

1. INTRODUCTION

Music sentiment analysis or music mood recognition has to do with utilizing machine learning, data mining and other techniques to classify songs in 2 (pos vs. neg) or more emotion categories with highest possible accuracy. Several types of features such as audio, lyrics or metadata can be used or combined together. Recently there is high attention on corpus-based methods that involve machine or deep learning techniques [1]. There are studies that successfully predict music emotions based on lyrics features only [2, 3, 4] utilizing complex models. Large datasets of songs labeled with emotion or mood 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 (at least 1000 lyrics)
4. Publicly available for cross-interpretation of results

It is costly and not feasible to prepare large datasets manually. Consequently, many researchers experiment with small datasets of fewer than 1000 songs, or large and professional datasets that are not rendered public. An alternative method for quick and large dataset creation is to crowdsource subjective user feedback from Amazon Mechanical Turk[1]. MTurk workers are

[1] http://mturk.com

David C. Wyld et al. (Eds) : CCSEIT, AIAP, DMDB, ICBB, CNSA - 2017
pp. 15– 26, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70603

typically asked to listen to music excerpts and provide descriptors about its emotionality. Studies like [5] and [6] suggest that this method is viable if properly applied. Another tendency is to collect intelligence from the flourishing and exponentially growing social community networks. Last.fm[2] is a community of music listeners, very rich in tags which are unstructured text labels that are assigned to songs [7]. Last.fm tags have already been used in many studies like [8, 9, 10, 11] to address various music emotion recognition issues. Nevertheless, none of their datasets has been rendered public.

Actually it is hard to believe that still today, no lyrics emotion dataset fulfills all 4 requirements listed above. An important work in the domain of movies is [12] where authors create a dataset of movie reviews and corresponding positive or negative label based on IMDB user feedback. Inspired by that work, here we utilize Playlist[3] and Million Song Dataset (MSD)[4] combined with last.fm user tags to create 2 datasets of song lyrics and corresponding emotion labels. We first categorized tags in 4 mood categories (Happy, Angry, Sad, Relaxed) that are described in section 3. Afterwards, to ensure high polarity, we classified tracks based on tag counters using a tight scheme. The first dataset (MoodyLyrics4Q) includes 5075 songs and fully complies with the 4 requisites listed above. 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 as consequence of the same bias of user tags each track had received. We also observed that even though there is a noticeable growth of opinion and mood tags, genre tags keep being the most numerous.

Currently we are working with lyrics for sentiment analysis tasks. However the mood classification of songs we provide here can be used by any researchers who have access to audio files or features as well. Both datasets can be downloaded from http://softeng.polito.it/erion/. The rest of this paper is structured as follows: Section 2 provides related studies creating and using music datasets for solving music emotion recognition tasks. Section 3 present the most popular music emotion models and the one we utilize here. In section 4 we describe data processing steps we followed whereas section 5 presents annotation schemes we used and the 2 resulting datasets in numbers. Finally, section 6 concludes.

2. BACKGROUND

Creating datasets of emotionally annotated songs is not an easy task. The principal obstacle is the subjective and ambiguous nature of music perception [13]. Appreciation of music emotionality requires human evaluators to assign each song one or more mood labels from a set of predefined categories. The high cognitive load makes this process time consuming and cross agreement is also difficult to achieve [14]. Another complication is the fact that despite the many interdisciplinary attempts of musicologist, psychologist or neuroscientists, there is still no consensus about a common representation model of music emotions.

One of the first works to examine popular songs and their user generated mood tags is [15]. Authors utilize metadata and tags of AllMusic[5] songs to create a practical categorical representation of music emotions. They observe the large and unevenly distributed mood term vocabulary size and report that many of the terms are highly interrelated or express different aspects of a common and more general mood class. They also propose a categorical music mood representation of 5 classes and a total of 29 most popular terms and recommend that reducing vocabulary of mood terms in a set of classes rather than using excessive individual mood terms is more viable and reasonable. The many works that followed mostly utilize self-created datasets to

[2] https://www.last.fm

[3] http://www.cs.cornell.edu/~shuochen/lme/data_page.html

[4] https://labrosa.ee.columbia.edu/millionsong/

[5] http://www.allmusic.com

explore different methods for music emotion recognition. In [16] authors use last.fm tags to create a large dataset of 5296 songs and 18 mood categories. Their mood categories consist of tags that are synonymous. For the annotation, they employ a binary approach for all the mood categories, with songs having or not tags of a certain category. They utilize this dataset in [17] to validate their text-audio multimodal classifier. Although big in size and systematically processed, this dataset is not distributed for public use. As noted above, another way for gathering human feedback about music mood is crowdsourcing with Amazon MTurk. In [5] authors try to answer whether that method is viable or not. They contrast MTurk data with those of MIREX AMC 2007 task[6] and report similar distribution on the MIREX clusters. Authors conclude that generally, MTurk crowdsourcing can serve as an 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. Also in [6], authors perform a comparative analysis between mood annotations collected from MoodSwings, a collaborative game they developed, 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 was released for public use but consists of 240 song clips only[7].

AMG tags have been used in [18] to create a dataset of lyrics based on Valence-Arousal model of Russell [19]. 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 received. Annotation quality was further validated by 3 persons. This is one of the few public lyrics datasets of a reasonable size (771 lyrics). In [21] they collect, process and publish audio content features of 500 popular western songs from different artists. 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. In [22] 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. A public audio dataset is created and used in [23] where they experiment on multilabel mood classification task using audio features. There is a total of 593 songs annotated by 3 music experts using the 6 categories of Tellegen-Watson-Clark model [24], an emotion framework that is not very popular in MIR literature.

Several other works such as [25] have created multimodal 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 dataset is available for research upon request to the authors. However it is very small (100 songs only) and thus cannot be used as a serious experimentation set. In [26] the 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. This categorization is highly superfluous and not very reliable. For example, are ’alarm’ or ’military’ real mood descriptors? In the next section we present a literature overview about popular music emotion representation models and the mood space we adopted here.

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

[7] http://music.ece.drexel.edu/research/emotion/moodswingsturk

3. MODELS OF MUSIC EMOTIONS

Psychological models of emotion in music are a useful instrument to reduce emotion space into a practical set of categories. Generally there are two types of music emotion models: Categorical and dimensional. The former represent music emotions by means of labels or short text descriptors. Labels that are semantically synonymous are grouped together to form a mood category. The later describe music emotions using numerical values of few dimensions like Valence, Arousal etc. A seminal study was conducted by Hevner [27] in 1936 and describes a categorical model of 66 mood adjectives organized in 8 groups as shown in figure 1. This model has not been used much in its basic form. However it has been a reference point for several studies using categorical models. The most popular dimensional model on the other hand is probably the model of Russell which is based on valence and arousal [19]. High and low (or positiv and negarive, based on normalization scale) values of these 2 dimensions create a space of 4 mood classes as depicted in figure 2. The models of Henver and Russell represent theoretical works of experts and do not necessarily reflect the reality of everyday music listening and appraisal. Several studies try to verify to what extent such expert models agree with semantic models derived from community user tags by examining mood term co-occurence in songs.

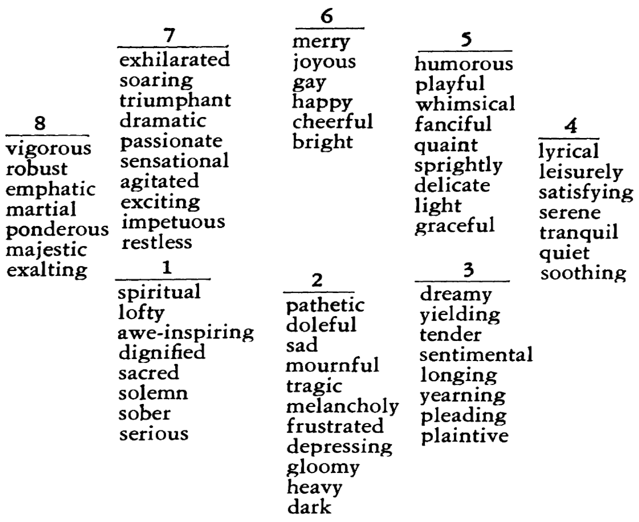

Figure 1. Model of Hevner

The model of 5 classes described in [15] was derived from analyzing AMG user tags and has been used in MIREX AMC task since 2007. It however suffers from overlaps between clusters 2 and 4. These overlaps that were first reported in [28] are a result of semantic similarity between *fun* and *humorous* terms. Furthermore, clusters 1 and cluster 5 share acoustic similarities and are often confused with each other. Same authors explore last.fm tags to derive a simplified representation of 3 categories that is described in [11]. They utilize 19 basic mood tags of last.fm and 2554 tracks of USPOP collection, and perform K-means clustering with 3 to 12 clusters. The representation with 3 clusters seems the optimal choice also verified by Principal Component Analysis method. Being aware of the fact that this representation of 3 mood clusters is over-simplified, they suggest that this approach should be used as a practical guide for similar studies. A study that has relevance for us was conducted in [10] where they merge audio features with last.fm tags. Authors perform clustering of all 178 AllMusic mood terms and reduce the mood space in 4 classes very similar to those of Russell's models. They conclude that high-level user tag features are valuable to complement low-level audio features for better accuracy. Another highly relevant work was conducted in [9] utilizing last.fm tracks and tags. After selecting the most appropriate mood terms and tracks, authors apply unsupervised clustering and Expected Maximization algorithm to the document-term matrix and report that the optimal number of term clusters is 4. Their 4 clusters of emotion terms are very similar to the 4 clusters of valence-arousal planar model of Russell (happy, angry, sad, relaxed). These results affirm that categorical mood models derived from user community mood tags are in agreement with the basic emotion models of psychologists and can be practically useful for sentiment analysis or music mood recognition tasks. Based on these literature observations, for our dataset we utilized a folksonomy of 4 categories that is very similar to the one described in [9]. We use ***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. This way we comply with the second requirement of the dataset. First we retrieved about 150 emotion terms from the studies cited above and also the current 289 mood terms of AllMusic portal. We conducted a manual process of selection, accepting only terms that clearly fall into one of the 4 clusters. For an accurate and objective selection of terms we consulted ANEW, a catalog of 1034 affect terms and their respective valence and arousal norms [20]. During this process we removed several terms which do not necessarily or clearly describe mood or emotion (e.g., *patriotic*, *technical* etc. from AllMusic). There was also ambiguity regarding different terms used in other studies which were also removed. For example, terms *intense*, *rousing* and *passionate* in [9] have been set into 'angry' cluster whereas in [10] they appear as synonyms of 'happy'. Same happens with *spooky*, *wry*, *boisterous*, *sentimental* and *confident* which also appear into different emotion categories. We also dropped out various terms that based on valence and arousal norms in ANEW, appear in the borders of neighbor clusters. For example, *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. A good music mood representation model must have high intra-cluster similarity of terms. To have a quantitative view of this synonymy of terms inside each cluster we make use of word embeddings trained with a 1.2 million terms Twitter corpus[8] which is rich in sentiment words and expressions. Word embeddings have been proved very effective in capturing semantic similarity between terms in text [29]. We tried to optimize the intra-cluster similarities by probing of a high number of term combinations inside each of the 4 clusters. The representation of Table 1 appeared to be the optimal one. That representation includes the 10 most appropriate emotion term in each cluster. Figure 3 shows the corresponding intra-cluster similarity values.

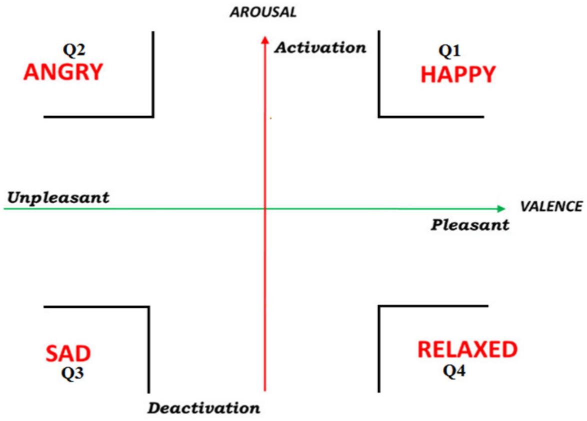

Figure 2. Mood classes in model of Russell

[8] http://nlp.stanford.edu/projects/glove/

4. DATA PROCESSING AND STATISTICS

To reach to a large final set and fulfill the third requirement, we chose a large collection of songs as a starting dataset. MSD is probably the largest set of research data in the domain of music [30]. Created with goal of providing a reference point for evaluating results, it also helps scaling MIR algorithms to commercial sizes. The dataset we used is the result of the partnership between MSD and last.fm, associating last.fm tags with MSD tracks. There are 943334 songs in the collection, making it a great source for analyzing human perception of music by means of user tags. Playlist dataset is a more recent collection of 75,262 songs crawled from yes.com, a website that provides radio playlists from hundreds of radio stations in the United States. The authors used the dataset to evaluate a method for automatic playlist generation they developed [32]. Merging the two above datasets we obtained a set of 1018596 songs, with some duplicates that were removed. We started data processing by removing songs with no tags obtaining 539702 songs with at least one tag. We also analyzed tag frequency and distribution. There were a total of 217768 unique tags, appearing 4711936 times. The distribution is highly imbalanced with top hundred summing up to1930923 entries, or 40.1% of the total. Top 200 tags appear in 2385356 entries which is more than half (50.6%) of the total. Also, 88109 or 40.46% of the tags appear only once. They are mostly typos or junk patterns like ”111111111”, ”zzzzzzzzz” etc. Most popular song is “Silence” of “Delerium” which has received 102 tags. There is an average of 9.8 tags for each song. Such uneven distribution of tags across tracks has previously been reported in

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

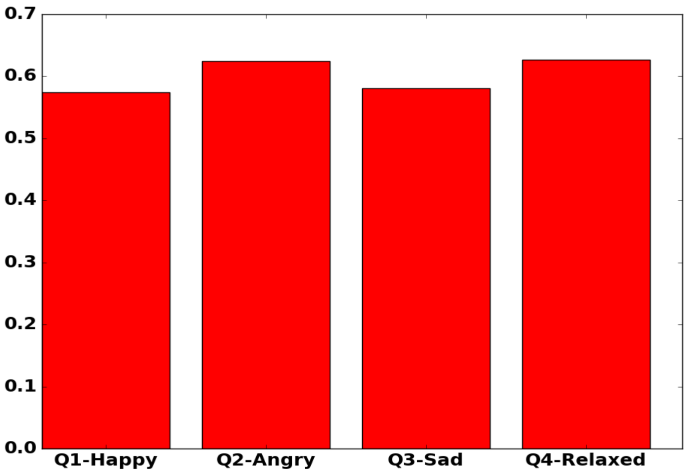

Figure 3. Synonymy rates for each cluster

[31] and [15]. Most frequent 30 tags are presented in Table 2. Top tag is obviously *rock* appearing 139295 times. From Table 2 we see that among top tags, those describing song genre are dominant. Same as in [7], we analyzed distribution of top 100 tags in different categories such as genre, mood, instrument, epoch, opinion etc. In that study of 2007 the author reports that mood tags make up 68% of the total, followed by locale, mood and opinion with 12, 5 and 4% respectively. 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 growth of opinion and mood that make up 16.2 and 14.4% respectively. Our interest here is in mood tags, most frequent of which are presented in Table 4. From the 40 terms shown in Table 1, only 11 appear in this list. There are however many other terms that are highly synonymous. We can also see that positive tags are distinctly 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 mood term is *mellow* appearing 26890 times. Obviously users are more predisposed to provide feedback when perceiving positive emotions in music. Word cloud of emotional tags is presented 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 authors in [16]. It is not possible to know if tag *love* means that the song is about love or that user loves the song. Similarly 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 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

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

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

5. Annotation Scheme and Results

At this point what's left is the identification of the tracks that can be distinctly fitted in one of the 4 mood clusters we defined, based on the tags they have received. To make use of as much tags as possible and reach to a large dataset (third requirement), we extend the basic 10 terms of each cluster with their related forms derived from lemmatization process. For example, is makes sense to assume that *relaxing*, *relax* and *relaxation* tags express the same opinion as *relaxed* which is part of cluster 4. We reached to a final set of 147 words that were the most meaningful from music emotion perspective. The next step was to identify and count those tags in each song. At the end of this step, for each song we had 4 counters, representing the number of tags from each mood cluster. 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:

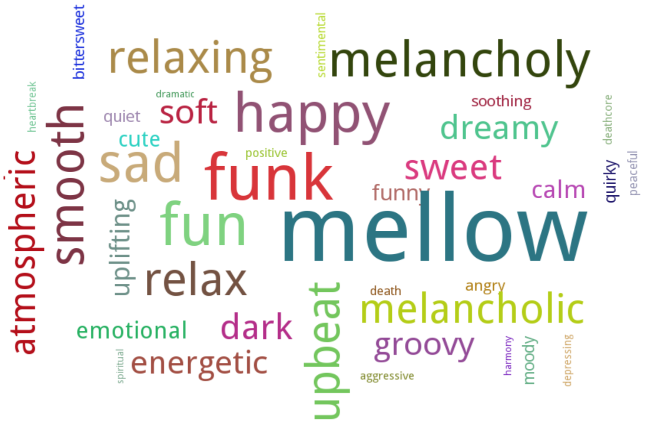

Figure 4. Cloud of most frequent affect tags

- 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 fewer tags or not fulfilling one of the above conditions were discarded. The remaining 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 reported in the previous section, tag distribution was imbalanced in the same way.

The pos-neg representation is clearly oversimplified and does not reveal much about song emotionality. Nevertheless, such datasets are usually highly polarized. Positive and negative terms are easier to distinguish. Same happens with several types of features that are often used for classification. The confidence of a binary categorization is usually higher not just in music but in other application domains as well. The pos-neg lyrics dataset we created here might be very useful for training and exercising many sentiment analysis or machine learning algorithms. 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. CONCLUSIONS

In this paper we presented the steps that we followed for the creation of two datasets of mood annotated lyrics based on last.fm user tags of each song. We started from two large and popular music data collections, Playlist and MSD. As music emotion model, we adopted a mood space of 4 term clusters, very similar to the popular model of Russell which has been proved effective in many studies. Analyzing last.fm tags of songs, we observed that despite the growth of opinion and mood tags, genre tags are still the most numerous. Within mood tags, those expressing positive emotions (happy and relaxed) are dominant. For the classification of songs we used a stringent scheme that annotates each track based on its tag counters, guaranteeing polarized clusters of songs. The two resulting datasets are imbalanced, containing higher number of positive songs and reflecting the bias of user tags that were provided. Both datasets will be available for public use. Any feedback regarding the annotation quality of the data is appreciated. Researchers are also invited to extend the datasets, especially the smaller clusters of songs.

ACKNOWLEDGEMENTS

This work was supported by a fellowship from TIM[9]. Computational resources were provided by HPC@POLITO[10], a project of Academic Computing within the Department of Control and Computer Engineering at Politecnico di Torino.

[9] https://www.tim.it/

[10] http://hpc.polito.it

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, pages 142–150, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[13] 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. http://ismir2010.ismir.net/proceedings/ismir2010-45.pdf.

[14] 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.

[15] 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.
http://ismir2007. ismir.net/proceedings/ISMIR2007_p067_hu.pdf.

[16] 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.

[17] 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.

[18] 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.

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

[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] 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.

[22] 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.

[23] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-label classification of music into emotions. In Proceedings of the 9th International Conference on Music Information Retrieval, pages 325–330, Philadelphia, USA, September 14-18 2008.
http://ismir2008.ismir.net/papers/ISMIR2008_275.pdf.

[24] A. Tellegen, D. Watson, and L. A. Clark. On the dimensional and hierarchical structure of affect. Psychological Science, 10(4):297–303, 1999.

[25] 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.

[26] 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.

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

[28] 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.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013.

[30] 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), 2011.

[31] Y.-C. Lin, Y.-H. Yang, and H. H. Chen. Exploiting online music tags for music emotion classification. TOMCCAP, 7(Supplement):26, 2011.

[32] 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.