MoodyLyrics: A Sentiment Annotated Lyrics Dataset

Erion Çano
Polytechnic University of Turin
Duca degli Abruzzi, Turin, Italy
+393478353047
erion.cano@polito.it

Maurizio Morisio
Polytechnic University of Turin
Duca degli Abruzzi, Turin, Italy
+390110907033
maurizio.morisio@polito.it

ABSTRACT
Music emotion recognition and recommendations today are changing the way people find and listen to their preferred musical tracks. Emotion recognition of songs is mostly based on feature extraction and learning from available datasets. In this work we take a different approach utilizing content words of lyrics and their valence and arousal norms in affect lexicons only. We use this method to annotate each song with one of the four emotion categories of Russell's model, and also to construct MoodyLyrics, a large dataset of lyrics that will be available for public use. For evaluation we utilized another lyrics dataset as ground truth and achieved an accuracy of 74.25 %. Our results confirm that valence is a better discriminator of mood than arousal. The results also prove that music mood recognition or annotation can be achieved with good accuracy even without subjective human feedback or user tags, when they are not available.

CCS Concepts
• Applied Computing→Arts and humanities • Applied Computing→Document management and text processing

Keywords
Intelligent Music Recommendation; Lyrics Sentiment Analysis; Music Dataset Construction; Lyrics Mood Annotations

1. INTRODUCTION
Today with the expansion of community networks, music listening and appraisal is changing; It is becoming more social and collective. Search and selection of songs that was once performed on the basis of Title, Artist or Genre, now also uses mood as a new and important attribute of music. In this context, there is a growing interest for automatic tools that perform Music Emotion Recognition, or Recommendation Engines that exploit users' context to provide them better music recommendations. Recent emotion recognition tools are mostly based on intelligent models that learn from data. To train such models datasets annotated with emotion or mood categories are required. Manual and professional annotation of song emotions is labor intensive. As a result most of existing works utilize datasets that consist of less than 1000 songs [33]. Also many datasets that are collected by researchers are utilized to evaluate their results only and are not rendered public. To solve the problem of emotion recognition in music, researchers base their methods or approaches in subjectively annotated song datasets (typically smaller than 1000 pieces) or user tags of songs, extraction of features (typically audio, text, or both) and supervised learning algorithms for classification (e.g., SVM) [34, 13, 12]. In this work we take an opposite approach. We employ a method that is based on content words of lyrics and generic lexicons of emotions only, avoiding any subjective judgment in the process of song emotion recognition. This method does not require any dataset or extraction of textual features (like unigrams, bigrams etc.). Our idea is to use this method for creating a larger mood dataset and then employing feature extraction and advanced learning algorithms for possible better results in sentiment analysis of songs. Russell's Valence-Arousal model with 4 mood categories is employed for the annotation process [27]. Valence and Arousal values of songs are computed adding the corresponding values of each word of lyrics that is found in a lexicon we build by combining ANEW (Affect Norm of English Words), WordNet and WordNet-Affect. An important output of this work is MoodyLyrics, a relatively big dataset of song lyrics labeled with four mood categories, Happy, Angry, Sad and Relaxed using the same method. To validate the quality of the method and MoodyLyrics, we used a lyrics dataset annotated by subjective human judgment and user tags [23] as a comparison basis. The evaluation process reveals an achieved accuracy of 74.25 %, which is comparable with results of similar works [12, 34]. The evaluation results also show that in general, valence appears to be a better emotion discriminator than arousal. On the other hand, even though slightly disbalanced (more Happy and fewer Angry or Relaxed songs), MoodyLyrics is bigger than most of the current publicly available datasets, consisting of 2595 song lyrics. A more comprehensive evaluation with bigger and better ground truth benchmark dataset would provide better insights about its annotation quality. The contribution of this work is thus twofold:

- First, we create and provide for public use MoodyLyrics, a relatively large sized dataset of lyrics classified in 4 emotion categories.
- Second, we investigate to what extent do objective sentiment annotations based solely on lyrics and lexicons agree with user tag or subjective human annotations of music.

MoodyLyrics corpus of songs and annotations can be downloaded from http://softeng.polito.it/erion/MoodyLyrics.zip. There is a slight difference between mood and emotion from a psychological point of view. Usually the term mood refers to a psychological state that lasts longer in time than other certain states of emotion [7]. Nevertheless in this paper we use this two terms interchangeably. The rest of this paper is organized as follows: Section 2 provides recent related works about the different mood

ISMSI '17, March 25-27, 2017, Hong Kong, Hong Kong

DOI: http://dx.doi.org/10.1145/3059336.3059340

annotation methods of songs, most popular models of music emotions and the use of lexicons for sentiment analysis problems. Section 3 illustrates the collection and textual processing of lyrics, describes the lexicons we use and explains in details the method we involve for the annotation process. Section 4 presents the evaluation results we obtained by comparing our dataset with a similar lyrics dataset that was manually annotated by experts and user tags. Finally, section 5 concludes and presents possible future uses of MoodyLyrics.

2. BACKGROUND

2.1 Creation of Ground Truth Datasets

In order to train and test a classifier, a dataset with assigned mood labels or emotion categories from an emotion music model is required. This so-called ground truth is difficult to obtain [8] because of the inherently subjective emotional perception and annotations of music [33]. The perception of music pieces and their emotions is influenced by various factors like age, gender, social context or professional background, and thus it is quiet difficult to reach cross assessor agreements on music mood labels. Furthermore the annotation or labeling of moods to music pieces is a time consuming and labor-intensive process, as it requires a heavy cognitive involvement of the subjects [33, 20]. These difficulties lead to small datasets that are usually annotated by less than five musical experts and show varying quality in practice. In different studies like [29, 19, 28], authors report the above problems and make use of crowdsourcing mechanisms for the annotation process. In [19] Mechanical Turk annotations are compared with those collected from MIREX [1] campaign. The authors show that the distribution of mood clusters and agreement rates from MIREX and Mechanical Turk are comparable, and conclude that Mechanical Turk can serve as a practical alternative for music mood ground truth collection. Similarly in [28] a high number of persons is crowdsourced, selected and involved (at least 10 annotators per song) to create a high quality dataset. Nevertheless the resulting dataset contains 1000 songs only. Actually most of the similar datasets that can be found are not any bigger. Another recent approach that attempts to facilitate song labeling process is picking up mood tags provided by users of music listening websites such as last.fm. However, considerable amount of preprocessing work is needed to clean and cluster the synonymous tags. Additional challenges like polysemy of tags and absence of a common and widely agreed vocabulary haven't been properly addressed yet, and lead to quality weaknesses of resulting datasets [29, 19, 18]. [16] is one of the first survey works about social tags and their use in music information retrieval. Tags are defined as unstructured and unrestricted labels assigned to a resource (in this case a song) to describe it. In that study of 2008, the author reports that in the domain of music, 68 % of tags are about genre and only 5 % about mood. Other researchers make use of last.fm tags to create ground truth datasets for their own experimentations. For textual feature experimentation, authors in [13] utilize last.fm tags to build a large ground truth dataset of 5585 songs and 18 mood categories. They use WordNet-Affect[2] lexicon and human expertise to clean up tags and cluster together synonyms. However they do not publish or evaluate the quality of the dataset they created. In [18], the authors utilize last.fm community tags to create a semantic mood space of four clusters, namely ***Angry***, ***Sad***, ***Tender*** and ***Happy***. They compare it with existing expert representations (e.g., clusters from MIREX AMC task) and report consistency, confirming the relevancy of social tag folksonomies for mood classification tasks. Furthermore their 4 clusters can also be interpreted as representations of the 4 quadrants in the Valence-Arousal plane of Russell. Several researchers have even designed games to collect mood annotations of musical pieces from online users. Annotation games try to employ the "Human Computation" by making the annotation task more entertaining. In [24] the authors present a web game that collects categorical labels of songs by asking players to describe short excerpts. In [15] the authors go one step further developing MoodSwings, a game that not only collects song mood labels form players, but also records the mood variability of each musical piece. They utilize the 2-dimentional Arousal-Valence model and ask each user to give feedback about five 30-seconds clips. Players are partnered to verify each others' results and thus produce more credible labels. Also, in [30] the authors compare effectiveness of MoodSwings annotations with those obtained from crowdsourced single paid subjects hired through Amazon Mechanical Turk. They report strong agreement between MoodSwings and MTurk data, but however advise that complexity and quality control of crowdsourcing methods should be carefully arranged.

2.2 Models of Music Emotions

Same as with dataset construction, the subjective nature of music perception is a serious difficulty for creating standard mood categories or models as well. The psychological models of emotion are necessary abstract constructs that help to reduce the mood space into a manageable set of categories. These models are usually either categorical or dimensional. Categorical models describe emotions of music by means of labels or descriptors. The synonymous descriptors are usually clustered together in one mood category. On the other hand dimensional models are based on few parameters or dimensions like ***Valence*** which can be positive or negative, ***Arousal*** which can be high or low, ***Stance*** which can be open or closed etc. All possible combinations the model is based on, create the different mood classes of that model. A comprehensive and detailed discussion about music emotion states and models can be found at [6]. In the resent years several music emotion models have been proposed by psychologists and used by researchers. Yet none of them is considered as "Universal" or fully acceptable. Nevertheless there are few music emotion models that have gained popularity in the community of researchers.

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 1. Mirex five mood clusters

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

[2] http://wndomains.fbk.eu/wnaffect.html

A popular categorical model that was proposed in [10] organizes mood descriptors in 5 clusters as shown in Figure 1. This model has been used in MIREX AMC[3] task since 2007. A problem of this model is the semantic overlap between cluster 2 and cluster 4 as reported in [17]. Another earlier categorical model was proposed by Hevner in [9]. It uses 66 descriptors categorized in 8 groups. There are obviously many other categorical models of affect presented in various studies. They are usually derived from user tags clustered in synonymous groups and describe mood categories of song datasets. On the other hand, one of most popular dimensional models is the planar model of Russell [27] shown in Figure 2. This model is based on two dimensions: Valence (pleasant-unpleasant) and Arousal (aroused-sleepy) which the author considers as the most basic and important emotion dimensions.

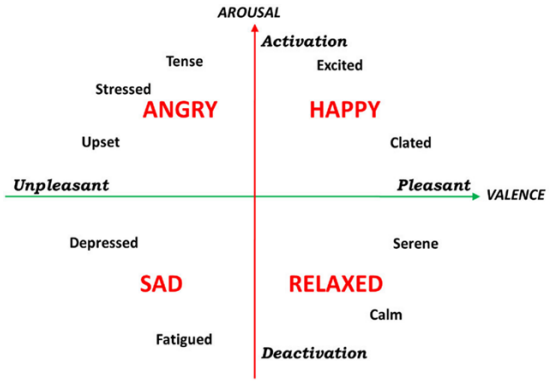

Figure 2. Circumplex model of emotions

Valence represents the positive or negative intensity of an emotion whereas Arousal indicates how strongly or rhythmically the emotion is felt. A 3-dimensional model named PAD (Pleasure-Arousal-Dominance) is based on the model of Russell. It adds dominance-submissiveness, a dimension related to music potency. PAD emotion model is described in [1].

2.3 Use of ANEW and other Lexicons

ANEW lexicon and its Valence, Arousal and Dominance word norms have been used in several sentiment analysis research works in the recent years. In [26] its words are used as a source for training sample words. The authors build a classifier using intro and refrain parts of each lyrics. In [34] the authors utilize both word-based and global lyrics features to build a mood-based song classifier. They conclude that tf-idf can be effectively used to identify moody words of lyrics and that the lingual part of music reveals useful mood information. A similar approach is presented in [14] where ANCW (Chinese version of ANEW) is created by translation of ANEW terms and used for building a mood classifier of Chinese songs. The authors preprocess the sentences of each lyric and extract the words appearing in ANCW which they call Emotion Units. They compute Valence and Arousal of each EU and afterwards of the entire sentence. Finally they make use of fuzzy clustering and Vector Space model to integrate the emotion values of all the sentences and find out the emotion label of the entire song. In [12] authors perform music feature analysis by comparing various textual features with audio features. They mix together various feature types like n-grams of content words, stylistic features and also features based on General Inquire, ANEW and WordNet. General Inquirer [31] is one of the first psycholinguistic lexicons created, containing 8315 unique English words organized in 182 psychological categories. We describe ANEW and WordNet in the next section where we also present the way we combined them for our purpose.

[3] http://www.music-ir.org/mirex/wiki/2007:AMC

3. CONSTRUCTION OF MOODYLYRICS

In this section we describe the steps that were followed for the annotation method setup and dataset construction. We first motivate the use of lyrics and describe corpus collection and textual preprocessing. Later on we explain the combined use of the 3 lexicons we chose. Finally we describe the annotation process and resulting dataset.

3.1 Collection and Preprocessing

In this work we chose to use lyrics of songs for several reasons. First, contrary to audio that is usually copyrighted and restricted, it is easier to find and retrieve lyrics freely from the Internet. Some websites like lyrics.wikia.com provide free services for searching, downloading or publishing lyrics. It is also easier to work with lyrics than audio which requires certain expertise in signal processing. Lyrics is rich in high level semantic features contrary to audio which offers low level features and suffers the resulting semantic gap [4]. Nevertheless, lyrics are different from other text documents (newspapers, books etc.) and pose some difficulties. They are usually shorter and often created from a small vocabulary. Furthermore, their metrical and poem-like style with metaphoric expressions can cause ambiguity and hamper mood identification. For our purpose, we first found public sources from where to get song titles and authors. The major part of our corpus was constructed from Playlist[4] collection which is a list of songs and tags of listeners crawled from Last.fm API. The construction of Playlist dataset is further described in [5]. It is good to have diversified songs in terms of genre or epoch. For this reason we tried to selected songs of different genres (Rock, Pop, Blues etc.) and from different periods ranging from the sixties (e.g., Beatles, Rolling Stones etc.) to few years ago. We thus added other song sources like MillionSongSubset[5], Cal500[6], and TheBeatles[7]. Further information about public music (and other) source datasets can be found at [3]. We downloaded song lyrics from lyrics.wikia.com using Lyrics[8], a Python script that finds and downloads lyrics of songs given song title and artist. Collected texts were first preprocessed removing empty or duplicate songs. Also English language filter was applied to remove any text not in English. We cleared out punctuation symbols, tokenized into words and removed stopwords as well. Part-of-speech tagging was not necessary whereas stemming was not performed as it could create problems when indexing words in the lexicon. At this point we removed entries with less than 100 words, as it would probably be impossible to correctly classify them. Finally year and genre information was added when available and the resulting corpus was saved in CSV format.

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

[5] http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset

[6] http://labrosa.ee.columbia.edu/millionsong/sites/default/files/cal500HDF5.tar.gz

[7] http://labrosa.ee.columbia.edu/millionsong/sites/default/files/AdditionalFiles/TheBeatlesHDF5.tar.gz

[8] https://github.com/tremby/py-lyrics

3.2 Construction of the Lexicon

The basic lexicon we used for sentiment analysis of lyrics is ANEW (Affective Norms for English Words) which provides a set of normative emotional ratings for 1034 unique English words [2]. The words were rated in terms of Valence, Arousal and Dominance dimensions by numerous human subjects that participated in the psycholinguistic experiments. Besides the average rate, the standard deviation of each dimension is also provided. WordNet is a much bigger and more generic lexicon of English language [25]. It contains more than 166000 (word, sense) pairs, where sense is an element from a given set of meanings. The basic relation of words in WordNet is Synonymy and word senses are actually sets of synonyms (called synsets). WordNet-Affect is a smaller lexicon obtained from WordNet synsets and represents affective concepts [32]. The corpus was marked with affect terms (called a-labels) representing different types of affective concepts (e.g., Emotion, Attitude, Sensation etc.). For our purpose none of the above 3 lexicons could be used separately. ANEW is small and not entirely focused on mood or emotions. WordNet is huge but is very generic and does not provide any Valence or Arousal rates. WordNet-Affect is enough relevant but it is small. As a result we combined the 3 lexicons in the following way: First we started from ANEW words. For each of them we checked the synsets of WordNet that include that word and extended with the resulting synonyms, marking the new words with same Arousal and Valence values (Dominance is not used at all) of ANEW source word. Afterwards we kept only words that belong to synsets of WordNet-Affect labeled as ***Emotion***, ***Mood*** or ***Sensation***, dropping out every other word. The final set is composed of 2162 words, each with an Arousal and Valence score. ANEW was extended in a similar way in [11] where the authors experiment with heterogeneous featuresets and SVM algorithm to increase mood classification accuracy.

3.3 Mood Annotation of Lyrics

The process of mood annotation starts by computing the aggregate Valence and Arousal values of each song, based on the corresponding values of words in that song that are found in the mixed lexicon we constructed. Lyrics words that are not part of the lexicon are not considered. Valence and Arousal values of each indexed word were added to the total Valence and Arousal of that song. Meanwhile lyrics with less than 10 words in the lexicon were discarded. At the end the aggregate affect values of each song were normalized to fall in [-1, 1] interval. For several reasons we decided to adopt a categorical version of Russell's model to represent the emotion categories of lyrics. First the model of Russell is simple and very popular. It is based on the two most fundamental "sources" of music affect, namely ***Valence*** and ***Arousal***. Furthermore it is easy to conceive or represent it geometrically (see Fig. 1). Each of the 4 mood categories namely ***Happy***, ***Angry***, ***Sad***, ***Relaxed*** represent one of the 4 quadrants in a 2-dimensional Euclidean plane (see Figure. 3). This representation seems a very good tradeoff between oversimplification (few categories, e.g. only positive vs. negative) and ambiguity (many categories which probably overlap with each other). We utilize the above model of music mood and put each song in one of the 4 quadrants if it has normalized Valence and Arousal values that are "distinctly" positive or negative. By "distinctly" we mean grater or lower than certain threshold Vt and At values. This threshold values are necessary in order to have high confidence for the categorization process and a polarized resulting dataset. To this end, we classified each lyrics as shown in Table 1. The subset of songs falling inside the rectangular zone [(Vt,At), (-Vt,At), (-Vt, -At), (Vt,-At)] were removed as they do not carry a high classification confidence. For certain sentiment analysis applications it might be necessary to have only positive or negative lyrics. For this reason, we also derived a version of the dataset with this 2 mood categories, using the same logic and based on Valence only, as shown in Figure 4. The songs are considered ***Positive*** if they have V>Vt and ***Negative*** if V<-Vt (see Figure. 4). To decide about Vt and At values we considered results of various comparisons with another dataset, as explained in Evaluation section.

Table 1. Classification of lyrics

V and A values	Mood
A>At and V>Vt	Happy
A>At and V<-Vt	Angry
A<-At and V<-Vt	Sad
A<-At and V>Vt	Relaxed

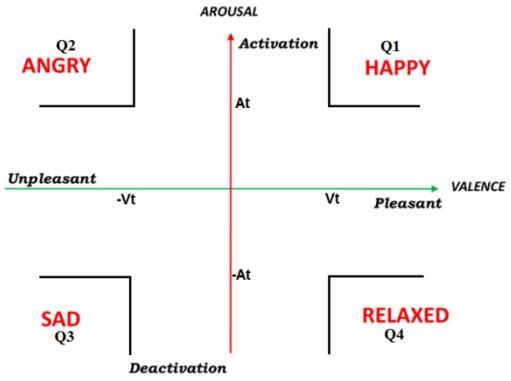

Figure 3. Dataset with 4 classes

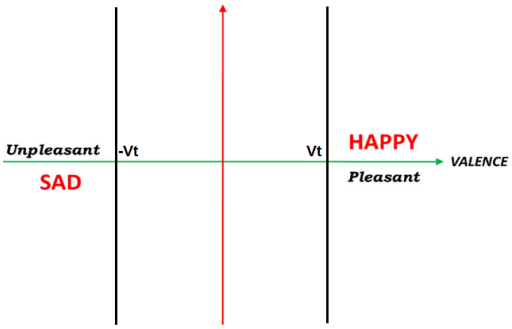

Figure 4. Dataset with 2 classes

4. STATISTICS AND EVALUATION

In this section we present and discuss some characteristics of MoodyLyrics. Quality assessment results of our method and dataset are also provided. Predictions of the method were compared with a lyrics dataset we used as benchmark.

4.1 Corpus Statistics

The current version of MoodyLyrics consists of 2595 song lyrics and the corresponding mood label for each. Table 2 provides the

distribution of songs according to the mood category they belong to. There is a slight disbalance of the clusters which is somehow inevitable; Today it is much easier to find happy songs rather than angry or relaxing ones. The version with two mood categories only (positive vs. negative) is a corpus of 1416 positive and 1179 negative lyrics. In Table 3 we summarize some statistics of the entire corpus whereas in Table 4 we list the absolute and relative frequency of the top 15 words. As expected the most frequent word among all songs is ***love***; Song lyrics are mostly about love and sentiments. It appears on average about 5 times in each song.

Table 2. Songs per mood category

Mood	Songs
Happy	819
Angry	575
Sad	604
Relaxed	597

Table 3. Corpus Statistics

Statistic	Value
Number of unique songs	2595
Number of unique artists	1666
Average songs per artist	1.558
Total words with stopwords	597933
Average words per song	230
Total words no stopwords	347851
Vocabulary size no stopwords	15329

Table 4. Most frequent words

Rank	Word	Freq	Avg. Freq
1	love	12229	4.712
2	im	4364	1.681
3	dont	3170	1.221
4	know	3064	1.180
5	baby	2658	1.024
6	like	2518	0.970
7	oh	2518	0.970
8	youre	2239	0.862
9	got	2037	0.784
10	na	2017	0.777
11	one	1757	0.677
12	want	1719	0.662
13	cant	1677	0.646
14	time	1661	0.640
15	come	1650	0.635

Figure 5 below shows the word cloud image of Moodylyrics.

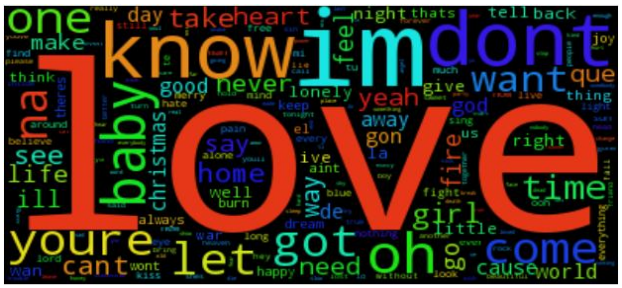

Figure 5. Word cloud of MoodyLyrics

4.2 Evaluation Results

To have an idea about the quality of our annotation method and the resulting dataset, we compared it with a similar lyrics dataset. Our goal was to explore to what extent do objective text based mood annotations of music, agree with subjective annotations performed by humans or obtained from subjective user tags. To have a direct basis of comparison we searched for datasets that are based on the same emotion model and categories. One such dataset is described in [28] and contains 1000 songs. The songs were annotated using Amazon Mechanical Turk on the basis of Valence and Arousal by a minimum of 10 subjects each. Unfortunately most of the songs in that dataset are instrumental (with few exceptions) making them unusable for our purpose. Other similar datasets are described in [23, 22, 21]. For our purpose we chose the lyrics dataset described in [23]. It is based on the same affect model and annotated using both human evaluators and user tags. The corpus consists of 771 (211 Happy or Q1, 205 Angry or Q2, 205 Sad or Q3, 150 Relaxed or Q4) song lyrics collected from AllMusic[9]. Each song has tags of AllMusic users which were considered by the authors for the first phase of annotation process. Later, 3 subjects were involved to provide feedback about each song. A song was set to one of the 4 quadrants if at least one of the annotators agreed with AllMusic tags. The authors use this dataset themselves for validating textual feature experiments they perform. We first collected the lyrics of the benchmark dataset. Afterwards, our method was applied in the lyrics generating the mood labels. Finally the mood labels generated by our method were compared with the original mood labels and an accordance rate was obtained. Initially we used Vt=0.25 and At=0.25 for which agreement between the two datasets was low. We increased threshold values of Valence and Arousal raising the polarization of our dataset and classification confidence of the songs. Nevertheless, using high threshold values reduces the size of the resulting dataset. Many more lyrics fall inside the "unknown" rectangular zone [(Vt,At), (-Vt,At), (-Vt,-At), (Vt,-At)] and are therefore discarded. Furthermore the 4 clusters of songs become disproportional, with many "happy" songs and few "relaxed" ones. We stopped at At=0.34 and Vt=0.34 values.

Table 5. Confusion Matrix

True\Pred	Happy	Angry	Sad	Relaxed
Happy	**68.57**	4.28	2.85	24.28
Angry	5.88	**81.18**	12.94	0
Sad	7.27	16.36	**74.54**	1.82
Relaxed	18.18	0	9.1	**72.72**

Additional increase of At or Vt would excessively shrink the size of MoodyLyrics and the benchmark dataset. In Table 5 we present the confusion matrix. We can see that Angry (Q2) songs are the best predictable. On the others hand Happy (Q1) songs have low prediction accuracy and are often confused with Relaxed (Q4) songs. They both have high valence and obviously it is not easy to discriminate based on their arousal values. There is also relatively high confusion between Angry and Sad (Q3) songs. On the other hand there is low confusion between Sad and Relaxed songs. Obviously it is more difficult to discriminate between high and low arousal than between high and low valence. Same results are reported in [23] where mood classifications based on various

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

textual features are evaluated. Higher accuracy is reported using valence and lower accuracy when using arousal. The overall accuracy of the method is 74.25 %, which is similar to that of other studies that are based on text feature learning and classification. In [12] they use features based on content words, ANEW and other lexicons, and text stylistics features to classify lyrics in 18 mood categories. Their reported accuracy ranges from 53.33 % for ***exciting*** songs to 79.66 % for ***aggressive***. This results are in line with our results, as ***exciting*** and ***aggressive*** are close synonyms with ***happy*** and ***angry***. Likewise, in [34] the authors report a maximal accuracy of 77.23 % when combining global features like word count, character count, and line count with word-based features (unigrams and bigrams) and classifying with SVM. An accuracy of 74.25 % is certainly not very good for a dataset to be considered as ground truth. For this reason we pushed Valence and Arousal threshold up to At=0.4 and Vt=0.4 believing that this way it has higher accuracy and quality at least to a certain scale. Unfortunately at this point the benchmark dataset shrunk to few lyrics, as mood of most of the songs was considered "unknown". This small corpus of lyrics couldn't be used as a valid and credible comparison set, leaving our dataset not fully validated. The corpus we used for validation is certainly not the best possible ground truth datasets. Comparing with higher quality ground truth datasets could give us better insights about the quality of the method and MoodyLyrics as well. Nevertheless we believe that MoodyLyrics is enough accurate to be used for several tasks, especially for text feature extraction and analysis.

5. CONCLUSIONS AND FUTURE WORK

In this work we presented an objective sentiment annotation method of song lyrics that is based on affect norm of content words. We used a lexicon that mixes together words from WordNet, WordNet-Affect and ANEW and exploited Valence and Arousal norms of the latter to find the quadrant each song belongs to, based on 2-dimensional model of Russell. We wanted to explore to what extent can lyrics mood annotations based on content words and lexicons mood annotation method agree with subjective, manual or tag based annotations. We also created and presented MoodyLyrics, a large and polarized dataset of mood annotated lyrics which will be available for public use. The accuracy of our method compared with a lyrics dataset annotated by means of user tags and human subjects was 74.25 % which is rather good in the domain of music. This result proves at a certain level, that mood annotation of musical pieces is a problem that can be also solved without any subjective feedback, when it is not available. Evaluation process also revealed that in general, valence appears to be a better discriminator of mood than arousal. A possible extension of this work could be combining more affect lexicons for mood prediction. Reconsidering threshold values of valence and arousal base on more careful empirical observations could also raise the accuracy of the method. A possible extended version of MoodyLyrics should also be followed by a more comprehensive evaluation to prove or disprove its validity as a ground truth music mood dataset and also provide insights about its possible practical uses.

6. ACKNOWLEDGMENTS

This work was supported by a fellowship from TIM[10].

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

7. REFERENCES

[1] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon. Pleasure, arousal, dominance: Mahrabian and russell revisited. *Current Psychology*, 33(3):405-421, 2014.

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

[3] E. Çano and M. Morisio. Characterization of public datasets for recommender systems. *In Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2015 IEEE 1st International Forum on*, pages 249-257, Sept 2015.

[4] O. Celma. Foafing the music bridging the semantic gap in music recommendation. In *5th International Semantic Web Conference (ISWC)*, Athens, GA, USA, 2016.

[5] Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[6] J. J. Dent, C. H. C. Leung, A. Milani, and L. Chen. Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation. *ACM Trans. Interact. Intell. Syst.*, 5(1):4:1-4:36, Mar. 2015.

[7] P. Ekkekakis. *Measurement in Sport and Exercise Psychology*, chapter Affect, Mood, and Emotion. Human Kinetics, 2012.

[8] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A Survey of Audio-Based Music Classification and Annotation. *Multimedia, IEEE Transactions on*, 13(2):303-319, Apr. 2011.

[9] K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48:246-268, 1936.

[10] X. Hu and J. S. Downie. Exploring mood metadata: Relatonships 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.

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

[12] X. Hu and J. S. Downie. When lyrics outperform audio for music mood classification: A feature analysis. In *proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 619-624, Utrecht, The Netherlands, August 9-13 2010.

[13] X. Hu and J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. In *Proceedings of the 10th International Society for Music Information REtrieval Conference*, pages 411-416, Kobe, Japan, October 26-30 2009.

[14] Y. Hu, X. Chen, and D. Yang. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, pages 123-128, 2009.

[15] Y. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. In Proceedings of the 9th International Conference on Music

Information Retrieval, pages 231-236, Philadelphia, USA, September 14-18 2008.

[16] P. Lamere. Social tagging and music information retrieval. Journal of New Music Research, 37(2):101-114, 2008.

[17] C. Laurier and P. Herrera. Audio music mood classification using support vector machine. In *International Society for Music Information Research Conference (ISMIR)*, 2007.

[18] C. Laurier, M. Sordo, J. Serr_a, and P. Herrera. Music mood representations from social tags. In *International Society for Music Information Retrieval (ISMIR) Conference*, pages 381-386, Kobe, Japan, 26/10/2009 2009.

[19] 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.

[20] Y.-C. Lin, Y.-H. Yang, and H. H. Chen. Exploiting online music tags for music emotion classification. ACM Trans. Multimedia Comput. Commun. Appl., 7S(1):26:1-26:16, Nov. 2011.

[21] R. Malheiro, R. Panda, P. Gomes, and R. Paiva. Music emotion recognition from lyrics: A comparative study. In *6th International Workshop on Machine Learning and Music*, n/a, 2013.

[22] R. Malheiro, R. Panda, P. Gomes, and R. Paiva. Bi-modal music emotion recognition: Novel lyrical features and dataset. In *9th International Workshop on Music and Machine Learning MML2016 in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML/PKDD 2016*, n/a, 2016.

[23] R. Malheiro, R. Panda, P. Gomes, and R. P. Paiva. Emotionally-relevant features for classi_cation and regression of music lyrics. *IEEE Transactions on Affective Computing*, PP(99):1-1, 2016.

[24] M. I. Mandel and D. P. W. Ellis. A web-based game for collecting music metadata. In S. Dixon, D. Bainbridge, and R. Typke, editors, *Proceedings of the International Society for Music Information Retrieval conference*, pages 365{366, Sept. 2007.

[25] G. A. Miller. Wordnet: A lexical database for english. Commun. ACM, 38(11):39-41, Nov. 1995.

[26] S. Oh, M. Hahn, and J. Kim. Music mood classi_cation using intro and refrain parts of lyrics. In 2013 International Conference on Information Science and Applications (ICISA), pages 1-3. IEEE, 2013.

[27] J. Russell. A circumplex model of a_ect. *Journal of personality and social psychology*, 39(6):1161-1178, 1980.

[28] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2Nd ACM International Workshop on Crowdsourcing for Multimedia*, CrowdMM '13, pages 1-6, New York, NY, USA, 2013. ACM.

[29] 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.

[30] J. A. Speck, E. M. Schmidt, B. G. Morton, and Y. E. Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 549-554, Miami (Florida), USA, October 24-28 2011.

[31] P. J. Stone and E. B. Hunt. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241-256, New York, NY, USA, 1963. ACM.

[32] C. Strapparava and A. Valitutti. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083-1086. ELRA, 2004.

[33] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.*, 3(3):40:1-40:30, May 2012.

[34] M. v. Zaanen and P. Kanters. Automatic mood classi_cation using tf*idf based on lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 75-80, Utrecht, The Netherlands, August 9-13 2010.