Lyric-Based Music Mood Recognition

Emil Ian V. Ascalon, Rafael Cabredo
De La Salle University
Manila, Philippines
emil.ascalon@yahoo.com, rafael.cabredo@dlsu.edu.ph

Abstract: *In psychology, emotion is described as a conscious reaction caused by certain stimuli. Some of these stimuli come in the form of music and the songs we listen to. Mood Recognition in Songs is an increasingly popular topic among researchers. Among the studies, many of the works are devoted to the study of features and the advantages and strengths of a particular feature set. While most researches employ audio alone or using two or more sources of features for their study, most agree that there is some relevant information found in lyrics. To address this problem, this research focused on mood recognition of OPM songs using lyrics. Word level features such as TF-IDF and keyGraph keyword generation algorithm were experimented on, using different thresholds and parameters to determine how well these methods worked. Two approaches of labeling the mood of the songs were studied as well: the manual annotation of songs and an automated approach using arousal and valence. Results having manual annotation performing quite well while for the automatic approach there is still a need for improvement. Using keywords extracted from the lyrics that were manually labeled shows a lot of promise. Especially with keyGraph feature extraction where an 80% average accuracy rating experimented on two different classification models were achieved. Through this, more information was learned about the relationship of the content of lyrics with the determination of the mood.*

Keywords: Mood Recognition: OPM songs: Text Processing

1. Introduction

Sentiment analysis has been the topic of various studies around the world. It refers to the use of natural language processing and text analysis to identify what the author meant or wanted to convey in source materials. With the rise of social media, interest in sentiment analysis has grown, particularly among companies as a way for them to monitor people's opinion on different items and especially their products. As businesses look to automate this, the interest to this area of research rises as well. Sentiment analysis can be applied to different forms of media such as documents, blogs, or videos. One particular media where interest has been growing is music.

Music mood recognition is the process wherein the emotions of a musical piece are identified through various means which includes the analysis of audio and lyrical text. People may want to listen based on their mood, and multiple websites in the Internet tries to address this need of users. Sites such as LastFM, is a site used to play music and discover other types music through tags generated by the users some of which are moods. Musicovery is another music player websites that utilizes moods of songs as a way to discover or create a user's playlist. There is also the Echonest dataset, which is utilized by the recommendation engines of Spotify, Rdio, and other music streaming sites.

There are studies that have been conducted in the area of mood classification based on various combinations of lyric and audio features. Only a few though focuses solely on lyrics as the main feature set. In this study, the focus is on using word level features for mood classification.

This paper is organized as follows: Section 2 reviews related work on mood classification. Section 3 discusses the framework of the research; Section 4 contains the results and analysis; and Section 5 presents the conclusions.

2. Review of Related Works

Researches on the topic of mood classification of music tend to vary to the types of features the use to conduct the experiments. These features may be extracted from different types of sources. Most commonly used in researches are features extracted from audio files such wav and mp3. Researches use tools to extract different kinds features from audio tracks. These features have been heavily studied which features perform the best for mood classification. Such as studies of Chin et al. (2013) and Laurier et al. (2008), whose features for audio classification were studied from the Music Information Retrieval Evaluation eXchange. There are also researches that use lyrics as the main source of features. This approach tends to be more difficult from in terms of the lack of information and structure of lyrics. And finally a multi modal approach wherein the use of multiple modalities such as lyrics, audio, or statistical data. A combination of features seems to perform the best for classifying mood of songs. Such is the approach of Su et al. (2013), who combined lyric and audio features into a single feature space whose results were better when compared with the results of classification done with just one source of feature.

Using lyrics as the main source of features for classification can be achieved in various ways. One common way is using a bag of words containing words usually found in a specific mood. Each mood will have it's own set of words. These bags, which can be created by using certain databases or lexicons, will be then used to classify the lyrics depending on which words appear in the lyrics. There are databases with sentiment information such as SentiWordNet (sentiwordnet.isti.cnr.it). SentiWordNet is a lexical resource that assigns numerical scores extending WordNet, scores which relates to the sentiment it expresses. Kumar and Oh used this approach of using affective databases. The group of Kumar and Minz (Kumar & Minz, 2013) performed the examination of the use of SentiWordNet as a tool for classification. First the researchers collected lyrics as the data to be processed taken from websites and collected a total of 185 English songs. The lyrics undergo preprocessing steps, which include text cases, tokenizing, and removal of stopwords. SentiWordNet is then used to provide the polarity of a song by computing the features from positive and negative score of the word in the lyrics. The features used included the *TF-IDF* and three sentiment features of lyrics and are added to a document vector. Feature selection may be used to further improve the classification. Three classifiers were used for comparison of results namely Naive Bayes, K-Nearest Neighbor, and Support Vector Machine. The group experimented on the use of different statistical feature selection techniques such as Principal Component Analysis, Latent Symentic Analysis, Chi-Square, among others to determine to rank the features and constructed variants of the dataset for the classification. Results show that SVM show promise in classifying dataset of small documents such as lyrics.

The group of Oh et al. (2013) on the other hand used the Affect Norms for English Words (ANEW) database for training samples, which uses Valence, Arousal, and Dominance. The group further added to normal approaches by experimenting on whether location of the words is important. As such, the focus of the work is on the introduction and refrain parts of lyrics as opposed to the whole song. It utilized only these parts because the intro part supposedly includes the information to create the atmosphere of a song, and the refrain contains the most important keywords of the song. The team believed that this selection would enhance the accuracy of its classification by reducing the amount of words that may be meaningless. In their experiment they employed SVM as their classifier, and performed it on a hundred songs.

In this study, the implementation of word level features will be explored. The performance of using word level features in mood classification is compared with each other to determine which may perform the best.

3. Research Framework

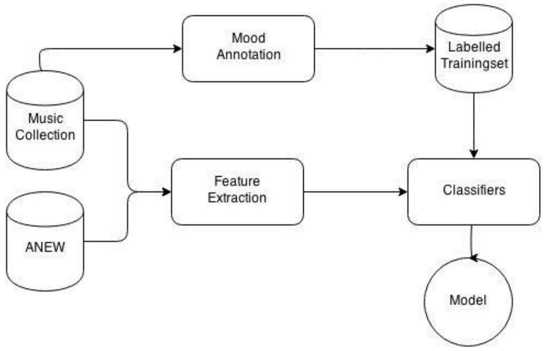


Figure 1: Architectural design for the study

3.1 Data Collection

In order to create a model that will serve as basis for comparing the automatic determination of moods in OPM songs, manual annotation by actual people was needed to determine the labels to be used for supervised learning. A sample of 200 song titles with artists and composers were collected through the online Philippine Music Registry set up by the National Commission for Culture and the Arts and the Organisasyon ng Pilipinong Mang-aawit. Their respective lyrics, on the other hands, were collected from lyric collections available in other websites online. The two-source rule in verifying if the songs matched the lyrics was employed. The lyrics were taken from different websites such as metrolyrics(www.metrolyrics.com) and lyric007 (www.lyric007.com). The lyrics are important to any song and may contain important information that can contribute to the type of mood the song is trying to convey.

3.2 Lyric Preprocessing

Lyrics were taken from websites that are user generated. These submissions are not monitored and may contain mistakes or errors. As such the lyrics were checked for correctness such as the spelling of the words. Lyrics consist of sections such as intro, refrain, chorus, and other parts. Repetitions of these parts happen in most lyrics. Users that submit lyrics may place instruction, such as [Chorus x4], instead of placing the whole song. In most researches, these instructions were replaced by the exact lyrics. This approach was followed as well for this study. The instructions were removed and replaced by the exact lyrics. The lyrics must also be found in both websites and must have the same words found. If not, the song was manually verified by reading the lyrics while listening to the audio to determine if it is correct.

3.3 Mood Annotation

In mood classification, each study tends to have different set of moods that are being identified in songs. They number from two to eighteen, even more. One way of representing moods is by using categories. Moods are divided into categories with closely related moods being group into one category. Moods can be also represented by plotting them in a dimensional space where moods lie in a XY plane. One emotion model that is commonly used for mood classification is Russell's model (Figure 2) wherein emotions lie in a two-dimensional plane of arousal and valence. (Kim, 2010)

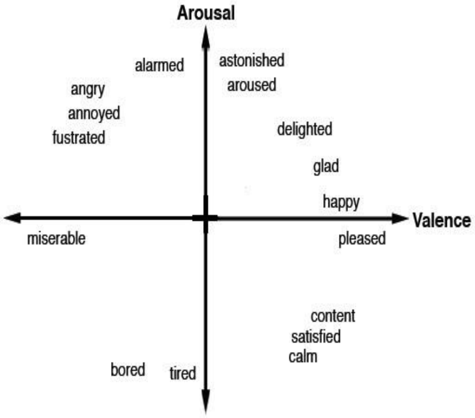
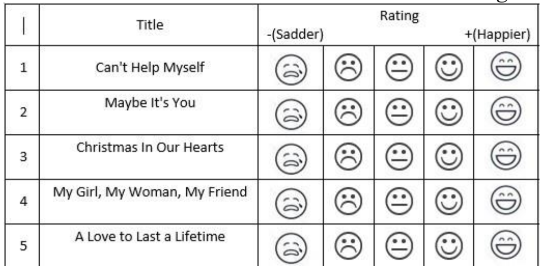


Figure 2: Russells Circumplex model of emotion

For this study, only two moods, *Happy* or *Sad*, were identified for a song. These are common themes of songs and can be identified by people at an easier rate. Two approaches for mood annotations were experimented in this study. An automated way using arousal and valence values and second through the use of manual annotation of people. For manual annotations, at least five people were be ask to rate the song which were used as labels for the songs. The annotators were asked to answer a survey to establish musical preferences and exposure. They were then given an answer sheet that contains the song titles to be rated, a scale of moods that utilizes emoticons that ranges from a face that is crying to one that is laughing. (Figure 3) Familiarity is also asked wherein a scale of one to five where one indicates that it was the first time that song was seen to five where the annotators really knows the song. After all the annotations were finished, they were collected into a single collection. The range of emoticons was converted to a numerical representation where in the *Saddest* emoticon is represented as a -2 and the happiest emoticon is represented as 2. All five annotations for the each song were averaged; the result value will become the annotated mood for that song.



|	Title	Rating -(Sadder)				+(Happier)
1	Can't Help Myself	😢	☹	😐	☺	😄
2	Maybe It's You	😢	☹	😐	☺	😄
3	Christmas In Our Hearts	😢	☹	😐	☺	😄
4	My Girl, My Woman, My Friend	😢	☹	😐	☺	😄
5	A Love to Last a Lifetime	😢	☹	😐	☺	😄

Figure 3: Sample "Happiness Scale"

Songs that have a value above 0 are considered as songs having a *Happy* mood, likewise songs whose values are lower than 0 are considered as song that are *Sadder* in nature. Songs with value that ended up as 0 will not be used in the study as no definitive mood was found.

For the automatic annotation, arousal and valence values were utilized. For this approach, the lyrics would first go undergo tagging in which the words in the lyrics are compared with the word list of ANEW and if the words from the lyrics were found it will be tagged. After the words are tagged, word segmentation will be done to output a tagged word list. That word list is filtered for those words that are not tagged meaning it was not found. The arousal and valence values for the word in the word list were added to the list. Two ways determining the label were experimented on. First method would entail the use of two cluster using K means to signify the two moods. The cluster containing the most words would become the label and if the cluster were equal the label would be considered as neutral. The other method would use distance formulas to determine if a word is closer to word *Happy* or *Sad* in terms of their respective arousal and valence. If song would have more *Happy* words, the song will be considered *Happy*, the same with *Sad* lyrics.

3.4 Datasets

Three sets of features were created for the study for comparison of results determining which features perform better. As stated before, the study was on the utilization of lyrics as sole features for mood recognition. As such only word level features were used in study and would not consider phrases in the classification. First set of features is words derived using TF-IDF or the term frequency multiplied by inverse document frequency of the words. This will prioritize words that appear frequently for a song but not in the whole music collection. The words in these methods were converted into a document vector wherein they will be used as the attributes. The number of words used, as attributes will depend on the thresholds set on TF-IDF scores. This served to determine if a word is important to the whole music collection and used as weights to improve the performance of the model. Different percentages of the total words derived from the lyrics were experimented on

Second set of features used were words derived from the keyGraph method. Words taken in this method differs from the previous feature since it does not only consider the frequency of a word, but also considers the probability of two words showing up together as a pair. The algorithm used is statistical measure where a probability is computed whether a certain word or pair of words would most likely be found in a *Happy* or *Sad* song. The number of words to be taken from the lyrics may vary and can be experimented to which number gives the best result.

The third set of features are the ANEW scores. The word sets derived from the previous features sets were used. Same with previous methods, the words with ANEW words will become the attributes with their corresponding values of arousal and valence. The arousal and valence values would be used separately.

4. Experiments

4.1 Experiment Design

This study aimed to study models created using features derived from lyrics that would be efficient enough to recognize moods. For the music collection, 200 song lyrics were collected from Internet sources. This study experimented on two approaches for labeling the song with it's perceived mood. First would be the manual annotation wherein individuals are ask to rate the songs using an emoticon scale which were converted into numerical values. Each song was annotated five times and these collected into one training set to get the average of the five to determine the dominant mood of the song. The second approach is an automated approach wherein arousal and valence values of are used. Different methods of determining the dominant mood of the song were experimented on.

This study focused on word level features alone, which were TF-IDF, keyGraph keywords, and a combination of feature sets derived from TF-IDF or keyGraph with ANEW. The features were extracted twice from the music collection, once as single dataset and using all the words extracted by these features. Another time where in the lyrics were separated by mood and extracted features separately extracting two sets of word list separated by mood. From these two-word lists, words that appear in both lists are removed so that words that would remain would not have any conflict and could have an impact to the results of experiments.

Feature selection must also be done to figure out how much words should be used to have the best results. Models using TF-IDF, the values derived can be used as weights to filter out less important words. The values derived from TF-IDF ranges from 0.001 to 0.700; as such thresholds were utilized to filter down the amount of words. The thresholds start from a range of 0.7 - 0.5, then we decrement by 0.2 to get words with values that ranges from 0.7 - 0.3, 0.7 - 0.1. Lastly, all the words found in the lyrics were also included as a baseline for this method. For models using keyGraph, a parameter is change to determine the number of words to be taken. It starts with 5 and will increment by 5 until 20 words per lyric is taken.

Table 1. Summary of Features Used

	TF-IDF	keyGraph	Combination
Feature Extracted	term frequency x inverse document frequency	keyword co-occurence	TF-IDF/keyGraph features with ANEW
Possible Values	values ranges from 0.001 to 0.7	binary values 0 or 1	AV values 1.25 - 8.85
Feature Selection	values as weights for the filter	set amount of words per lyric	words from TF-IDF/ keyGraph found only in ANEW
Thresholds Used	0.7-0.5, 0.7-0.3, 0.7-0.1, 0.7-0.001	20, 15, 10, 5 words per lyric	All words
Total Words Taken	170-2093 Words	310-1149 Words	5-8 Words

Using these features summarize in **Table 1**, together with the labels taken from annotation portion of the study, were fed to three different classifiers for the creation of the model. 10-fold cross validation was performed to determine the validity of these approaches and features.

4.2 Results

4.2.1 Results of the Annotation Approaches

The annotation approaches are divided into two the manual approach and the automatic approach.

Manual Annotation

As the manual annotation approach utilizes individuals for the labeling of songs, 50 people randomly selected were asked to rate the songs for the study. Having each person annotate 20 songs each, this would give each song an annotation of 5 times which should be sufficient enough to create a quality labeled training set. After collecting the responses and computing the average, the resulting mood label distribution follows 100 were labeled as *Sad*, 87 as *Happy*, and 13 as neutral. The total number of songs located in the labeled training set was 187, this is due to the lyrics that garnered a neutral label and were discarded.

Arousal & Valence Annotation

With manual annotation using people for labeling of songs there could be problems such as finding people to perform the annotations as well as the qualification needed for the task, as such an automatic approach is suggested using arousal and valence of words to label the songs. The following data show the results from the mood annotation module using arousal and valence of words found in lyrics. Two methods of labeling the songs were experimented and their findings are shown in **Table 2.**

Model[Threshold] (Total words taken)	% of Correctly Classified Instances	Kappa Statistics	Weighted Average of F-Measure
TF-IDF NB [0.001-0.7] (2093 Attributes)	64.7%	0.29	0.64
keyGraph NB [20 words] (1149 Attributes)	68.9%	0.37	0.69
Combination NB Arousal (508 Attributes)	69.5%	0.39	0.69
Combination NB Valence (508 Attributes)	69.5%	0.39	0.67

Table 2: Distribution of labels of songs.

	Clustering	Distance
Happy	151	158
Sad	36	32
Neutral	13	10

As shown above, the distribution of mood labels were biased towards the *Happy* label. Most of the lyrics were labeled as *Happy* even if it should have been labeled as *Sad*. Multiple variations of clustering and distance formulas were used with the best results are shown above. As this approach used arousal and valence of the words alone, there might not be enough information to have correct labeling of songs. Seeing as how one sided the labels are, the creation of models using the labels derived the automatic approach was not performed as it may perform in a bias manner towards the *Happy* label as well. A comparison and further analysis of both approaches is further discussed in the part.

4.2.2. Results of Experiment A & B

Shown here is the discussion of the results of the models using the features taken from the lyrics and the labels taken from the manual annotation approach. The experiments would be divided into two. **Experiment A** would have all the words extracted into a single feature set for each type of extraction method, and for **Experiment B** would first have the words separated by the mood of the lyrics from which they were extracted from and then have the conflicting words removed from both lists.

Experiment A

Experiment A contains the results of models using features extracted using TF-IDF, keyGraph, and the Combination of ANEW with TF-IDF or keyGraph. It does not consider the mood of words from which the lyrics are taken from and extracted as a whole.

Table 3: Best Results of Experiment A

As seen in the **Table 3**, the classifiers' performances were not that excellent in any of the various variations of filters used using the TF-IDF values as weights. The highest accuracy achieved was 64.7% using Naive Bayes with all the words found in the lyrics as the threshold are values that range from 0.001 to 0.7, which was essentially the same as having no filters. The keyGraph features performed better than that of TF-IDF as a whole compared to its counterpart in Naive Bayes without the removal of duplicate terms. Naive Bayes performed quite well with respect to the other classifiers. Achieving a 75% accuracy, the highest of all the results with having taken 15 words per lyrics as the filter used for this variation. It was able to equally identify both moods.

Table 4: Best Results of Experiment B

Model[Threshold] (Total words taken)	% of Correctly Classified Instances	Kappa Statistics	Weighted Average of F-Measure
TF-IDF NB[0.1-0.7] (534 Attributes)	67.1%	0.39	0.67
TF-IDF NB[0.001-0.7] (1048 Attributes)	90.5%	0.80	0.90
keyGraph NB[15 words] (637 Attributes)	81.2%	0.60	0.80
keyGraph NB[20 words] (775 Attributes)	84.2%	0.65	0.83

Experiment B

Experiment B contains the results of models using features extracted using TF-IDF or keyGraph and considers the moods of the lyrics from which the words are extracted. The words found in both moods are removed.

In the **Table 4**, the accuracy of the models generated indicated generally increased of accuracy when the duplicate words are removed. In some of the cases, the increase was really high having a 90% accuracy rating and could be considered as a well-performing model. But as the words that were not important were filtered out using thresholds, the accuracy greatly suffered going back down to the 60% accuracy rating. As the researcher further studied the results, some explanation was derived that could answer why the results happened. Such as since the best performing variation did not use a filter, all the words extracted from the lyrics were used. This would over fit the feature set to the training set and may be able to accept new unclassified data.

For keyGraph, similar with previous classifier the features underwent an additional step wherein the duplicate words in both list were removed. The results greatly improved for most of the classifiers. The accuracy of most of the variations was above 80% with 89% as the highest among all the results. This feature set showed better results than that of TF-IDF, as the results were more consistent in the results. Although it was not able to achieve an accuracy rating as high of that of TF-IDF, keyGraph does not suffer from the problems that TF-IDF encountered.

5. Conclusions

After seeing the results generated by Weka by both methods of annotation using only word level features, the researcher believes that the methods and processes utilized as well as the features used were not enough to produce a quality model for mood recognition in lyrics, although it showed some promise particularly the model using keyGraph features with 20 words per lyric which garnered an accuracy of 84%. Not only did the said model generate relatively good results from Weka in terms of accuracy and other statistical measures, the model was rather consistent in the recognizing of moods. As for the classifiers, Naive Bayes and SVM performed quite well and were equal in terms of performance. There is quite a difference between models' results when the duplicate terms in the lyrics between the moods are removed. However there are still some improvements needed in this method as it only had improved the accuracy for *Sad* lyrics. For the automated annotation using arousal and valence values, there is still much study and experimentation to be done for it to be a viable option for labeling songs.

Future studies may be improved by:

a. Further experimentation on upgrading the procedures in the manual annotation's data gathering.
b. Adding more songs to the collection as well as widening the scope of moods or emotion recognized in the study.
c. Utilizing phrase-based methods to counter some problems experienced by word-level features, and
d. Include the locations of the words as weights for important parts of the song.

6. References

Bradley, M.M. & Lang, P.J. (1999). Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical Report C-2. University of Florida, Gainesville, Fl.

Chin, Y.H., Lin, C.H., Siahaan, E., Wang, I.C., & Wang, J.C. (2013). Music Emotion Classification Using Double-Layer support Vector Machines. In

2013 International Conference on Orange Technologies (ICOT) (pp. 193-196). Piscataway: IEEE.

Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, O.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D. (2010 August 9-13). Music Emotion

Kumar, V., & Minz, S. (2013). Mood Classifiaction of Lyrics Using SentiWordNet. In *2013 International Conference on Computer Communication and Informatics January 04-06, 2013 : Coimbatore, India* (pp. 1-5). Piscataway, N. J.: IEEE.

Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal Music Mood Classification Using Audio and Lyrics. In *ICMLA 2008 Seventh International Conference on Machine Learning and Applications : Proceedings : 11-13 Dec. 2008, San Diego, California* (pp. 688-693). Los Alamitos, Calif.: IEEE Computer Society.

Oh, S., Hahn, M., & Kim, J. (2013). Music Mood Classification Using Intro and Refrain Parts of Lyrics. In *2013 International Conference on Information Science and Applications (ICISA 2013): Suwon, South Korea, 24-26 June 2013.* (pp. 1-3). Piscataway, NJ: IEEE.

Recognition: A State of the Art Review. Paper Presented at 11th International Society for Music Information Retrieval Conference, Utrecht, Netherlands. ISMIR.

Su, D., Fung, P., & Auguin, N. (2013). Multimodal Music Emotion Classification using AdaBoost With Decision Stumps. In 2013 International Conference on Acoustics, Speech and Signal Processing (ICASSP): Vancouver, BC, 26-31 May 2013. (3447-3451). IEEE.