Automatic music mood recognition using Russell's twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes

Automatic music mood recognition using Russell's two-dimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes

K R Tan[1], M L Villarino[2], C Maderazo[3]

Department of Computer and Information Sciences - University of San Carlos – Talamban Campus, Cebu City, Philippines

[1]ringoreigns@yahoo.com.ph
[2]mlvillarino@yahoo.com
[3]cvmaderazo@usc.edu.ph

Abstract. Automatic music mood recognition is still a new field of research that is gaining attention in the last decade. This study created a system that predicts which of the four quadrants of the valence-arousal space the song belongs to. The system used support-vector machine (SVM) for audio features while Naïve Bayes was used for lyrical features. audio classification achieved a high accuracy for arousal while lyrics classification achieved a high accuracy for valence.

1. Introduction

The music industry has been facing a problem on annotating music tags to songs since the creation of large digital music collections used in radio broadcasts or for streaming. Having many songs available, it becomes difficult and time consuming to have experts listen to a song and then decide on what tags to put on it. In 2002, Tzanetakis and Cook [1] tried to address this issue by creating an automatic genre classifier. Their method of using features extracted from audio files has since served as a standard procedure for future research on the field. By 2007, the new field had added automatic emotion and mood classification.

Feature extraction is a process which reduces raw data to a smaller set of features such as tone, timbre, and tempo. This enabled researchers to use the data without having to sort through the whole song. The process of extracting these features require tools [2] [3] and finding the optimal features to use for a study requires further research on these features [4]. Feature extraction was first focused solely on audio data but in the recent years, lyrics [5] [6] [7] also had been used to detect the mood of a song by using dictionaries containing valence and arousal values [8] [9]. Other studies that used both audio and lyric features [10] [11] had also shown that lyrics showed higher accuracy when compared with audio features and the combination of both features yielded greater results [10] [11].

Research on this new field saw the use of machine algorithms to automatically detect mood using decision trees [13], neural network [14], SVM [15] [16], and Naïve Bayes classifiers [5] [17]. Other methods such as the use of weighted graphs [12] and linear regression on ratings of words from word embedding [16] had also been used.

The model that was used as the basis for classification was Russell's circumspect model [18]. It is a two-dimensional plane with valence and arousal serving as the axes which divides the plane into four

quadrants. Figure 1 shows the emotional plane. Derivations of Russell's model have been created such as Thayer's model used in Bhat et al.'s study [14] [19] and Plutchick's model [7].

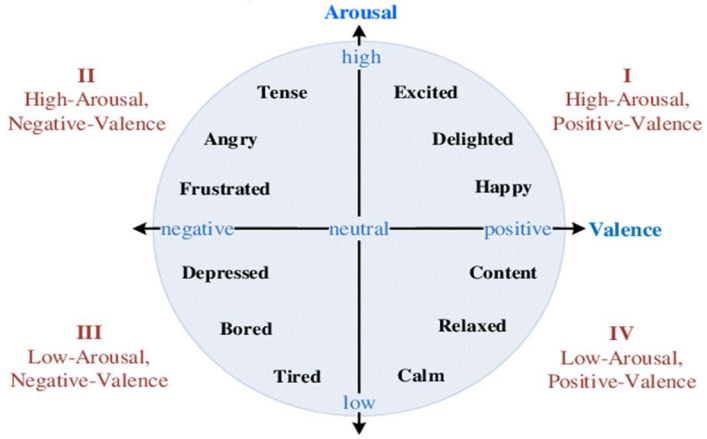

Figure 1. Two-dimensional valence-arousal space. [12]

For this study, the goal was to create a system that would be able to detect the mood of a song based on the four quadrants in Russell's model found at figure 1. To achieve this, the researchers used two classification algorithms SVM and Naïve Bayes to train separate classifier models for valence and arousal using selected audio features for SVM and lyrical features for Naïve Bayes. This process returns four trained models of valence and arousal for each algorithm. Valence is the positive or negative (pleasantness or unpleasantness) while arousal is the definition of how exciting or calm a person is towards a situation.

2. Design and Methodology

This section discusses the data processing and data handling.

2.1. Dataset

The dataset that was used for this study is the dataset used by Panda et al.'s [11] study. It contains annotations of valence and arousal and the quadrant it belongs to. The dataset consisted of 180 songs that had lyrics annotated while 162 songs had audio annotations. The number of songs that contained both lyrics and audio annotations totaled 133 songs.

2.1.1. Audio files. The dataset contained 158 mp3 files containing 30-second clips each. Each file was converted to wav format with a sample rate of 44100 Hz mono-channel which were used for this study. 126 songs were used for this study as some annotated songs did not have their audio data saved while some songs had low audio quality. The total number of songs in used each quadrant are set as: Quadrant 1 – 38, Quadrant 2 – 36, Quadrant 3 – 27, and Quadrant 4 – 26. The dataset was split into two parts: 66.66% for training and 33.33% for testing. Out of the 33.33% testing data, the test cases contained two parts: 26 songs (20.5%) which only contained audio ground-truth values and 16 songs (12.5%) which contained both audio and lyrical ground-truth data.

2.1.2. Lyrics. The dataset contained 180 songs with links to their lyrics made available. The lyrics were stored in text files and were pre-processed to fit the criteria of the dictionary that was used. The total number of songs in each quadrant are set as: quadrant 1 – 44, quadrant 2 – 41, quadrant 3 – 51 and quadrant 4 – 44. 66.67% out of the 180 lyrics were selected for training and 33.33% were for testing.

2.2. Feature extraction

The features of both audio and lyrical were extracted. Both single (Zero Crossing Rate) and multi-dimensional features (Mel-Frequency Cepstral Coefficient) were used.

2.2.1. Audio extraction. The features were extracted using Giannakopolous' pyAudioAnalysis [2] and an extra feature called tonnetz was extracted from Librosa [3] as tonal features. Table 1 shows the audio features that were used along with their standard deviation. The features selected were not the same for valence and arousal and instead were selected based on Grekow's [4] research. The following features that were used for arousal detection were Energy, Entropy of Energy, Spectral Energy, Spectral Flux, Spectral Roll-off, Beats per Minute, and their standard deviations. Audio features that were used for valence detection were Zero Crossing Rate (ZCR), Energy, Entropy of Energy, the three spectral features, MFCC, Chroma Vector, Chroma Deviation, and their standard deviations. Another feature called tonnetz was extracted from Librosa to have more tonal features present in valence detection.

Table 1. Audio features and their standard deviations were used.

Valence	**Arousal**
Zero Crossing Rate	Energy
Energy	Entropy of Energy
Entropy of Energy	Spectral Entropy
Spectral Energy	Spectral Flux
Spectral Flux	Spectral Roll-off
Spectral Roll-off	Beats Per Minute
MFCC (13)	
Chroma Vector (12)	
Chroma Deviation	
Tonnetz (6)	

2.2.2. Lyrics extraction. The dataset used links to redirect to the lyrics of each song. The lyrics were stored in text files (collected from the provided links in the dataset [11]. Some of the links provided were unavailable or that the lyrics were written wrong and so the researchers searched the lyrics from other links and corrected them. The resulting lyrics were then saved to text files and were pre-processed to fit the criteria of the dictionary [9] that was used. NLTK was also used to increase pre-processed screening. The lyrics were first stripped of their stop words and punctuations and were finally lemmatized down to its root word. Words with negative prefixes were preserved and words ending with "in'" were corrected as well. The results were saved in text files for the training and testing phase.

2.3. Machine-learning algorithms

The selected machine-learning algorithms, support-vector machine and Naïve Bayes, were used for audio and lyrical classification respectively.

2.3.1. Support-vector machine classifier. The available classifier in Python's sklearn was used with the following parameters. For arousal, the C parameter was set at 150 while valence's C parameter was set at 10^5. The training consists of two models, one for valence and another for arousal.

2.3.2. Naïve Bayes classifier. A modified naïve Bayes classifier that used Warriner et al.'s [20] and the NLTK library in Python to extract and use the pre-processed lyrics was used. Training and testing were split to 66.67% and 33.33%.

3. Results

The SVM classifier was tested with the following results shown on Table 2. Sixteen songs (12.5%) out of the 126 songs were used to test the accuracy of songs where its arousal is predicted with audio features and valence is predicted using lyrics while 26 songs (20.5%) were used for audio only detection. The

remaining 84 songs (66.66%) were used for training. Table 3 shows the precision, recall, and f1-score for Naïve Bayes. Tables 4 and 5 show the results of the tenfold cross-validation scores of the audio trained dataset's valence and arousal. The results of the 16 songs used for testing both accuracies for the two classification algorithms are found on Table 6.

Table 2. Results of SVM classifier on the training and testing data

	Training (84) (%) Valence	*Training (84) (%) Arousal*	Testing (26) (%) Valence	*Testing (26) (%) Arousal*	*Testing (16) (%) Valence*	*Testing (16) (%) Arousal*
Precision	88	90	58	100	45	94
Recall	88	89	58	100	44	94
F1-Score	88	89	57	100	44	94

Table 3. Results of Naïve Bayes classifier on the testing data

	Testing (16) (%) -Valence	*Testing (16) (%) +Valence*	Testing (16) (%) *-Arousal*	*Testing (16) (%) +Arousal*	Testing (60) (%) -Valence	*Testing (60) (%) +Valence*	Testing (60) (%) *-Arousal*	*Testing (60) (%) +Arousal*
Precision	90	100	57	78	80	100	76	82
Recall	100	86	67	70	100	62	88	67
F1-Score	95	92	62	74	89	77	82	73

Table 4. Valence Tenfold Cross-Validation Score (%)

Test Case	1	*2*	3	*4*	5	6	7	8	9	10
Precision	56	56	59	47	91	91	66	85	89	79
Recall	56	56	56	33	62	62	62	62	25	62
F1-Score	56	56	54	39	69	69	63	64	25	56

Table 5. Arousal Tenfold Cross-Validation Score (%)

Test Case	1	*2*	3	*4*	5	6	7	8	9	10
Precision	100	100	91	92	100	92	91	100	94	90
Recall	100	100	89	89	100	88	88	100	88	50
F1-Score	100	100	89	89	100	88	88	100	89	57

3.1. Audio detection

In extracting the audio features for training and testing, a total of 84 songs for the training result showed a (0.88 for the precision, 0.89 for the recall, and 0.88 for the f1-score) of valence while arousal showed (0.90 for precision, recall, and f1-score). Tenfold cross-validation was performed on the training data and showed poor results for valence but high accuracy for arousal. The 26 songs used for testing audio only detection showed (0.58 precision, 0.58 recall, and 0.57 f1-score) for valence and arousal achieved (1.00 precision, recall, and f1-score). The 16 songs that were used for testing resulted in (0.45 precision, 0.44 recall, and 0.44 f1-score) for valence while arousal resulted in (0.94 precision, 0.94 recall, and 0.94 f1-score). The cause of low accuracy on detecting valence has been discussed in the work of Yang, Dong, & Li [20]. It is considered that arousal can easily be distinguished between exciting or calm, with tempo being the key factor in this study. But valence is difficult to distinguish because it is ranked as either positive or negative (pleasant or unpleasant) and people have different opinions towards a song's pleasantness.

3.2. Lyrics Detection

Training data consisted of 120 songs and 60 were used for testing. The classifier achieved an 85% accuracy for valence (51 songs) and 75% accuracy for arousal (45 songs). The confusion matrix at Table 3 shows a high accuracy for valence but arousal predicted poorly this time. The trained model was tested on the 16 selected songs and achieved a high accuracy for valence detection as opposed to the low accuracy found on audio classification by SVM.

Table 6. SVM arousal and NB valence have higher accuracies.

	SVM Arousal Correct (%)	SVM Valence Correct (%)	*NB Arousal Correct (%)*	NB Valence Correct (%)
Quadrant 1	**100**	33.33	33.33	**100**
Quadrant 2	**100**	42.86	85.71	**100**
Quadrant 3	**100**	50	50	**100**
Quadrant 4	**75**	50	75	**72**

4. Conclusion

Arousal detection is highly accurate when used with audio features while valence detection is highly accurate when using lyrics. Arousal is easily distinguishable when listened to since its range would be from high to low. This study focused more on the use of tempo for arousal detection using extracted audio features. Valence detection using lyrics with Naïve Bayes resulted in higher accuracy than the use of audio because it is difficult to distinguish and analyze the tune and the positiveness or negativity of a word as they cannot be distinguished properly [20]. An example would be the difference between Quadrant 3 and Quadrant 4 – the tempo and tone of their songs are mostly alike and are confused with one another. Lyrics, however, uses the meaning of the words to signify the positivity or negativity of a word. This reveals a significant difference between the total positivity or negativity of a song, and can also identify songs representing a happy tone but the meaning of their lyrics is sad.

Future Works

For future tasks, a larger dataset is required for comparison of both audio and lyrical data. Further research also will be focused on valence detection using text classification and how it can be used to increase the accuracy of automatic mood classification problems.

References

[1] Tzanetakis G and Cook P 2002 Musical genre classification of audio signals *IEEE Trans. on Speech and Audio Process.* **10(5)** pp 293-302 doi:10.1109/tsa.2002.800560
[2] Giannakopoullos T 2015 pyAudioAnalysis: an open-source Python library for audio signal analysis *PloS ONE* **12(10)** doi:10.1371/journal.pone.0144610
[3] McFee B, Raffel C, Liang D, Ellis D P W, McVicar M, Battenberg E and Nieto O 2015 Librosa: v0.4.0. *Zenodo 2015* doi:10.5281/zenodo.18369
[4] Grekow J 2018 Audio features dedicated to the detection and tracking of arousal and valence in musical compositions *J. of Inform. and Telecom.* **1(12)** doi:10.1080/24751839.2018.1463749
[5] Raschka S 2014 MusicMood: predicting the mood of music from lyrics using machine learning (Michigan: Michigan State University) *Preprint* arXiv:1611.00138 [cs.LG]
[6] Ascalon E I V and Cabredo R 2015 Lyric-based music mood recognition *DSLU Res. Congress 2015 De La Salle University, Manila, Philippines*
[7] Oh S, Hahn M and Kim J 2013 Music mood classification using intro and refrain parts of lyrics *2013 Int. Conf. on Inform. Sci. and Appl. (ICISA)* pp 1-3 doi:10.1109/ICISA.2013.6579495
[8] Yu L, Lee L, Hao S, Wang J, He Y, Hu J, Lai K R and Zhang X 2016 Building Chinese affective resources in valence-arousal dimensions *Proc. of the 15th Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16) San Diego, CA, USA* doi:10.18653/v1/N16-1066
[9] Warriner A B, Kuperman V and Brysbaert M 2013 Norms of valence, arousal, and dominance for 13,915 English lemmas *Behavior Res. Methods* **45** pp 1191-207
[10] Hu X, Choi K and Downie J S 2016 A framework for evaluating multimodal music mood classification *J. of the Assoc. for Inform. Sci. and Technol.* **68(2)** pp 273-85 doi:10.1002/asi.23649
[11] Malheiro R, Panda R, Gomes P and Paiva R P 2016 Bi-modal music emotion recognition: novel

lyrical features and dataset *9th Int. Work. on Music and Machine Learning – MML'2016 – in conjunction with the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases – ECML/PKDD 2016 Riva del Garda Italy*

[12] Yu L, Wang J, Lai K R and Zhang X 2015 Predicting valence-arousal ratings of words using a weighted graph method *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing* pp 788-93 retrieved from http://www.aclweb.org/anthology/p15-2129

[13] Patra B G, Das D and Bandyopadhyay J 2013 Automatic music mood classification of Hindi songs *3rd Work. on Sentiment Analysis where AI meets Psychology (SAAIP 2013)*

[14] Bhat A S, Amith V S, Prasad N S and Mohan D M 2014 An efficient classification algorithm for music mood detection in Western and Hindi music using audio feature extraction *2013 Fifth Int. Conf. on Signal and Image Processing (ICSIP 2014)* pp 359-64 doi:10.1109/ICSIP.2014.63

[15] Ren J, Wu M and Jang J R 2015 Automatic music mood classification based on timbre and modulation features *IEEE Trans. on Affective Comput.* **6(3)** pp 236-46 doi:10.1009/TAFFC.2015.2427836

[16] Li M, Long Y and Lu Q 2016 A regression approach to valence-arousal ratings of words from word embedding *2016 Int. Conf. on Asian Language Processing (IALP)* pp 120-3 doi:10.1109/IALP.2016.7875949

[17] Romeu A U 2016 Emotion recognition based on the speech using a Naïve Bayes Classifier (Bachelor's thesis) retrieved from https://upcommons.upc.edu/bitstream/handle/2117/98368/Bachelors_thesis_ANGEL_URBANO.pdf?sequence=1&isAllowed=y

[18] Russell J A 1980 A circumplex model of affect *J. of Personality and Soc. Psych.* **39(6)** pp 1161-78 doi:10.1037/h0077714

[19] Nuzzolo M 2015 *Music Mood Classification* (Massachusetts, USA: Tufts University)

[20] Yang X, Dong Y and Li J 2017 Review of data features-based music emotion recognition methods *Multimedia Systems* pp 1-25 doi:10.1007/s00530-017-0559-4