Song Emotion Detection Based on Arousal-Valence from Audio and Lyrics Using Rule Based Method

Fika Hastarita Rachman
Informatics Departement
InstitutTeknologiSepuluhNopember
Surabaya, Indonesia
Informatics Departement
University of Trunojoyo Madura
Bangkalan, Indonesia
hastarita.fika@gmail.com

RiyanartoSarno
Informatics Departement
InstitutTeknologiSepuluhNopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

ChastineFatichah
Informatics Departement
InstitutTeknologiSepuluhNopember
Surabaya, Indonesia
chastine.fatichah@gmail.com

***Abstract*—Arousal and Valence value represent of song emotions. Arousal is an emotional dimension of musically energy level, while Valence is an emotional dimension of the comfortable level of the listener.Label emotion of Thayer using Arousal and Valence dimension.This research proposed a rule base method for detecting song emotion using arousal and valence values, however many studies do not use this data. The datasets are audio and lyric features of the song structural segment chorus.Preprocessing of Audio and lyric data are uses Correlation Feature Selection (CFS) and preprocessing text.Audio feature extraction is using MIRToolbox. Stylistic and psycholinguistic are used for lyrics feature extraction. Rule based method is used to detect the emotions of the whole song by using the predictive feature of the arousal and valence values.The arousal and valence prediction values are representing withmatrices of frequencyfor audio and lyrics.From the analysis of testing data, it shows that the audio feature more represents the value of Valence while the lyrics feature more represents the Arousal value.There are seven (7) rule base models that used in this research, the best accuracy is 0.798.**

Keywords—song emotion detection,arousal,valence,rule based method, MIRToolbox, stylistic, pcycholinguistic, matrices of frequency.

I. INTRODUCTION

In Music Emotion Recognition there are categorical models and dimensional models that support each other. Thayer emotional label is one of the categorical emotional models based on two (2) dimensional values: arousal and valence[1]. Thayer's emotional labels are mapped on a 2-dimensional plane with Arousal and Valence as their coordinate axis.

Song documents have audio and lyrics that used for emotion detection.Generally, song emotion detection uses lyrics and audio features separately.Lyric features for detecting emotions has been done[2][3][4]. Audio features also have been used for it[5][6][7]. Several previous studies show an increase in accuracy with the use of both features[8][9][10]. This research used audio and lyrics for features of song emotion detection.

Song emotion detection model that is often used is combining audio and lyric features before classification method.In research [8], there are four (4) audio features including *inharmonicity, roughness, loudness and tempo.* Lyric features that are used are 13 psycholinguistics and 11 stylistics. The combined features are a combination of the number of audio and lyric features, which are 28 features. Similar with the research [9], The fusion process between audio and lyric features with hough forest is done before classification. The result of the classification process is the Thayer emotion label.This research uses the result prediction of classification for detection emotion of the whole of song.

Research of automatic chorus detection has conducted by other researchers. Research[11] is an example of a research that carries out automatic audio structural segmentation using boundary segments and structure detection. Research[12] and[13] also detecting of song structure with different methods. Research [12]using repetition in chord sequence for detection structure feature, while research[13] used the concept of clustering the beat of song audio. From research of detection structural, it can be seen that there is information in the form of structural segments of the audio of a song.

Research[14] conducted an automatic chorus detection with the colormap and MFCC features, then the result is used for detected the emotion of that segment. Detected emotion using the audio feature without lyrics. Detection of emotions using the chorus segment and audio features with Adaboost classifier produces an average F-measure of 92.63%. Seeing the great influence of chorus for emotional detection, we used the chorus segment of the combined bimodal dataset and Ep-dataset. But we use audio and lyric features to the emotional detection.

Research [15] doing research to detect emotion of song using audio features and lyrics features. But the audio feature used was extracted from an audio dataset with a duration of 30second from experts. The lyric features were extracted from all lyric in one song. This is an obstacle if a song doesn't have a sample of 30 second audio music.

In this study, we used structural segments that are owned by song with audio and lyrics features to detect the emotions of the whole of a song. Seeing the success of the research [14], we choose the chorus structural segments for a sample of song audio. Audio and lyric features are extracted from the duration

978-1-7281-4610-2/19/$31.00 ©2019 IEEE

of the chorus. This research uses audio and lyrics of chorus that synchronized using *.lrc* file format.

The contribution of this research is in the emotion detection model of songs with chorus segment.First, the arousal and valence area classification of a chorus segment is performed for each audio and lyric feature.By looking at the predicted area of arousal and valance values, the emotional labels also be known.The classification results, prediction of arousal and valence area labels are representing inmatrix of frequency for each audio and lyric.Then the process of detecting emotion Thayer based on matrix of frequency for audio and lyric is done using the rule base method.

II. Proposed Method

The dataset used in this study was 100 songs from Ep-Dataset and Bimodal Dataset. The part that will be processed is the chorus segment of each song.The overall scheme is seen in Fig.1. The audio and lyric data in this study are synchronous data.In one song there are one or more chorus segments. Each audio data chorus segment is extracted using MIRToolbox. Then the featuresareselection using CFS.The results of the selection feature are used in the process of classifying arousal and valence areas. The prediction of arousal and valence area which is the result of classification is representing in matrix of frequency for each audio and lyrics. Matrices of frequency are an input for rule base method to detecting song emotion.

A. Dataset

The dataset used is a combination of the Bimodal dataset[15] and Ep-Dataset [16]that has been expanding. Bimodal dataset is a song dataset which has a Thayer emotional label, while Ep-dataset is a dataset that has song structural data.From the merging of the two datasets, incomplete data are obtained. There are data that have emotional labels but do not have structural segment data and there is data that has structural segment data but does not have Thayer emotional labels.

Completeness of data that does not have an emotional label is assisted by music experts, while data that does not have a segment structure are equipped with .lrc file obtained from the official store of song. This dataset consists of 100 songs with 25 data on each Thayer emotional label. The entire data in the dataset have structural data of song and Thayer emotion label. But of the 100 songs that have chorus data are 89 songs. Soour research used 89 songs with 294 chorus audio data and 284 chorus lyric data.

In Fig.2 there is a classification of area arousal and valence for chorus audio and lyrics segments. Each of these classifications requires training data. To support it, Thayer label emotion in the dataset is further convert into arousal (high, low) and valence area (high, low) area labels. Thayer's emotional label consists of four (4) labels: Q1 (high arousal; high valence), Q2 (high arousal; low valence), Q3 (low arousal; low valence), and Q4 (high arousal; low valence). It shows in Fig.2.

B. Audio Features Extraction

The audio feature is extracted using waveform analysis using Fast Fourier Transform (FFT). In Matlab, there is a tool that uses this concept, namely MIRToolbox 1.6.1[17]. In MIRToolbox, audio featureshavefive (5) main features: dynamics, timbre, rhythm, tonality and pitch.The result of a subset of each feature are scalar data and signal data. Signal data requires further processing using statistical parameters. Our research usesthe statistical parameter, namely:mean, median and standard deviation, so that each data signal has these statistical parameters.

There are 54 subsets of audio chorus extraction features. The extraction results are further processed in feature selection. The method used for feature selection is Correlation Feature Selection (CFS)[18]. The result of CFS is 16 features, seen in Table 1.

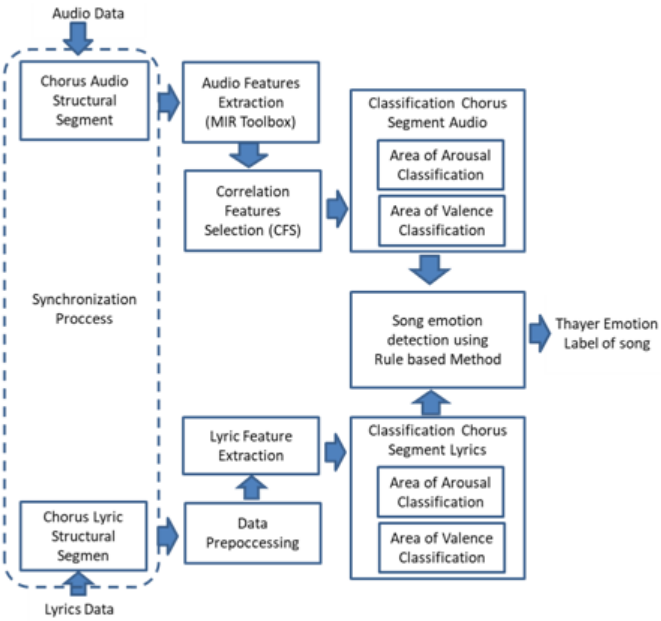

Fig.1. Proposed method scheme

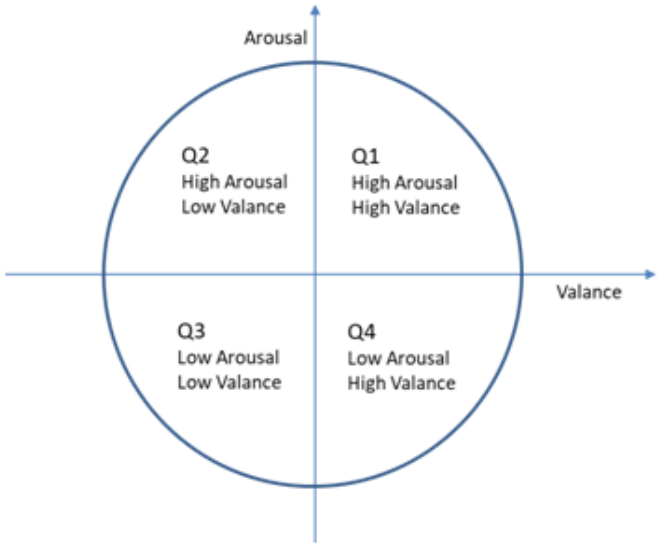

Fig.2. Label emotion of thayer

TABLE I. AUDIO DAN LYRIC FEATURES

Feature	Subset of features
Audio	Std-beat,eventdensity, pulseclarity, mean-attcktime, mean-decreaseslope, med-decreaseslope, std-decreaseslope, zerocross, kurtosis, mean-roughness, std-hcdf, mean-mfcc, std-mfcc, spectrum, chromagram, envelope-halfwavediff
Lyrics	Question_mark, Eclamation_mark, oh, yeah, ah, uh, woo, ha, la, oo, hemm, mm, ie, du, bye, hoo, shh, hey, wow, sum_v, sum_a, sum_d, avg_v, avg_a, avg_d, freq_nolabel, freq_q1, freq_q2, freq_q3,freq_q4

C. *Lyric Features Extraction*

Before extracting, the lyrics preprocessing are performed. The lyrics preprocessing process shows in Fig.3. Repair data for slang word, POS Tagging, Porter stemming and stopword removal are used for preprocessing data.

There are stylistic and psycholinguistic features extracted. Stylistic features are features in the form of unique words found in the lyrics and not found in the English dictionary [8]. Whereas psycholinguistic features are features obtained using an emotional corpus. In this case the corpus used is Corpus Based Emotion(CBE)[19] that expand according to Thayer label emotion. CBE previously was CBE with emotional labels according to the MIREX emotion label, this study developed it so that it became expand CBE (CBE-Ex). The total lyrics feature totaling 30 features, seen in Table 1.

D. *Classification of Chorus Segment*

The result of extraction process isaudio and lyric features. This research used Random Forest Method. The method was chosen, because in previous studies [8] compared to the SMO, Random Forest and Naive Bayes methods. That research was obtained that the Random Forest method was better.

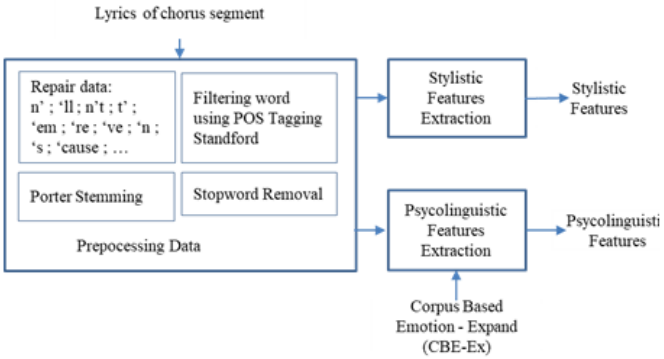

Fig.3. Preprocessing of chorus lyric

TABLE II. ACCURACY VALUE OF CHORUS SEGMENT CLASSIFICATION

Features	Prediction label		
	Arousal area	Valence area	Thayer label
Audio	0.69	0.76	0.66
Lirik	0.74	0.71	0.71

Classification of Chorus Segments produce predictive labels for the arousal area and the valence area for each audio and lyric feature separately. The F-Measure value for the chorus segment classification using the Random Forest 10 folds cross-validation method is shown in Table 2.

Analysis in table 2 shows that the accuracy of the classification results with the Thayer emotion prediction label is no better than the result of classification with prediction of the arousal and valence area.The audio feature is better used to predict the valence areathe lyrics feature is better used to predict arousal areas.

One song has one or more choruses, so at this step we have *n* arousal prediction, *n*valence prediction and *n* thayer prediction data.Then the data will be representing in the matrices of frequency for arousal, valence and thayer prediction. One song data have one vector of frequency.

Matrices of frequency for arousal and valence prediction, matrix *A*has two columns, there are column arousal prediction *"high"* and *"low"*. Matrices of frequency forthayer prediction, matrix *B*, has four (4) columns, there are column "*q1*", "*q2*", "*q3*", and "*q4*" predictions.

$$A = \begin{bmatrix} \sum a_{1,1} & \sum a_{1,2} \\ \sum a_{2,1} & \sum a_{2,2} \\ \sum a_{n,1} & \sum a_{n,2} \end{bmatrix} \quad (1)$$

$$B = \begin{bmatrix} \sum b_{1,1} & \sum b_{1,2} \sum b_{1,3} & \sum b_{1,4} \\ \sum b_{2,1} & \sum b_{2,2} \sum b_{2,1} & \sum b_{2,4} \\ \sum b_{n,1} & \sum b_{n,2} \sum b_{n,3} & \sum b_{n,4} \end{bmatrix} \quad (2)$$

E. *Rule Based Method*

Rule base method is a method created for the detection of emotions using arousal prediction data, valence area and thayer labels.There are 4 main rules that are used to detect Thayer based emotions on the arousal and valence area, there are:

- `Arousal(high) AND Valence(high)` → `Q1`
- `Arousal(high) AND Valence(low)` → `Q2`
- `Arousal(low) AND Valence(low)` → `Q3`
- `Arousal(low) AND Valence(high)` → `Q4`

The combining of audio and lyric features using the concept of adding matrices of frequency. The predicted value of the combined matrix of frequency is taken from the maximum element of a matrix. The main rule is used in the emotional detection model, there areseven (7) models. M1 is a model that uses arousal prediction values from audio and valence prediction from lyrics, shown in Fig. 5. M2 is a model that uses arousal prediction values from lyrics and valence prediction from audio. M3 is a model that uses arousal and valence prediction values from audio and lyrics, shown in Fig.6.Combining arousal and valence prediction using adding concept of matrix.

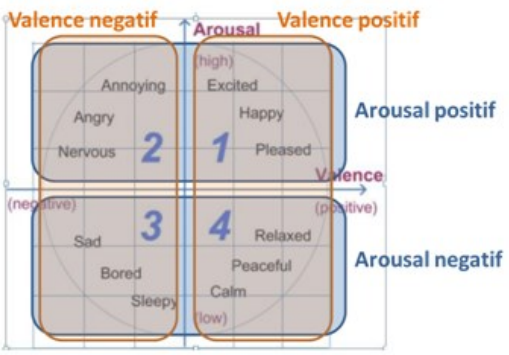

Fig.4. Mapping of arousal and valence areas in the thayer emotion model

M4 is model that uses thayer label prediction values from audio and lyrics. M5, M6, and M7 are a detection model that is made to overcome the existence of the same value in the elementsmatrix of frequency.M5 is a combined model of M4 and M1. First we detection the emotion using M4, when there is the same element value to get maximum value, then we use M1. M6is a combined model of M4 and M2. Then M7is a combined model of M4 and M3.

III. Result and Discussion

The accuracy values that used to measure the reliability of the emotional detection model are:

$$Accuracy = \frac{sum\ of\ true\ predicted}{total\ document}.100\% \qquad (3)$$

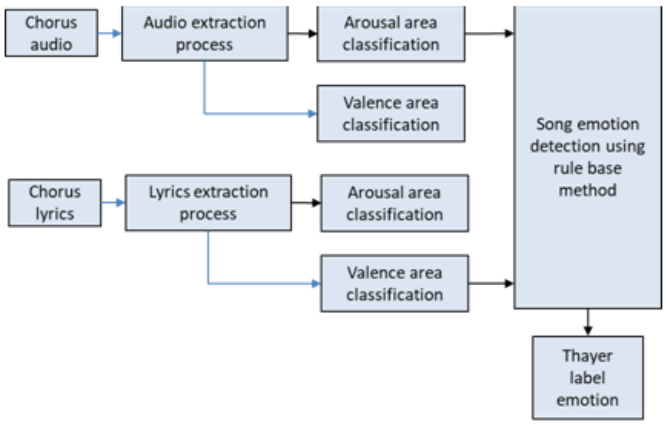

Fig. 5. Scheme of model 1 (M1) for song emotion detection

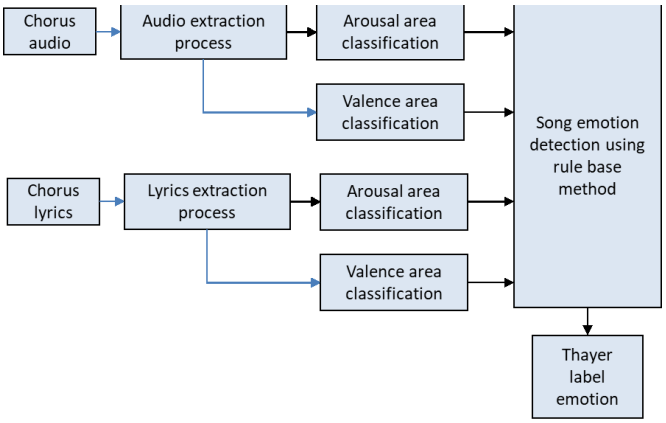

Fig. 6. Scheme of model 3 (M3) for song emotion detection

Table 3 shows the accuracy of different detection model. The best accuracy is 0.798 using detection model M6. M6is a combined model of M4 and M2. M4 is model that uses thayer label prediction values from audio and lyrics. M2 is a model that uses arousal prediction values from lyrics and valence prediction from audio.

IV. Conclusion

From the result of experiment shown that, Thayer's emotional label predictions are important for detection emotion. Furthermore, the predictive value of the arousal in the lyrics and the predictive value of the valence in the audio affect the emotional prediction of the whole of song.

In future work, another emotional detection model needs to be analyzed which can increase the value of accuracy. Uses more than one data song structural segment may also affect the accuracy results.

TABLE III. The Accuracy Value of Different Detection Model

Detection model	Accuracy
M1	0.517
M2	0.562
M3	0.517
M4	0.742
M5	0.764
M6	0.798
M7	0.787

REFERENCES

[1] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York: Oxford University Press, 1989.

[2] J. S. Downie, "When Lyrics Outerform Audio for Music Mood Classification : A Feature Analysis," no. Ismir, pp. 619–624, 2010.

[3] M. Kim and H. Kwon, "Lyrics-based Emotion Classification using Feature Selection by Partial Syntactic Analysis," in *International Conference on Tools with Artificial Intelligence*, 2011.

[4] Y. Hu, X. Chen, and D. Yang, "Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method," in *International Society for Music Information Retrieval Conference*, 2009, pp. 123–128.

[5] R. Panda and R. P. Paiva, *Automatic Mood Tracking in Audio Music*. Universidade de Coimbra, 2010.

[6] L. Lu, D. Liu, and H. Zhang, "Automatic Mood Detection and Tracking of Music Audio Signals," in *IEEE Transactions on Audio, Speech, and Language Processing (Volume: 14, Issue: 1, Jan. 2006)*, 2006, vol. 14, no. 1, pp. 5–18.

[7] J. A. Ridoean and R. Sarno, "Music Mood Classification Using Audio Power and Audio Harmonicity Based on MPEG-7 Audio Features and Support Vector Machine," in *International Conference on Science in Information Technology (ICSITech)*, 2017, pp. 72–77.

[8] F. H. Rachman, R. Sarno, and C. Fatichah, "Music emotion classification based on lyrics-audio using corpus based emotion," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 3, pp. 1720–1730, 2018.

[9] F. Xue, Hao ; Xue, Like ; Su, "Multimodal Music Mood Classification by Fusion of Audio and Lyrics," in *21st International Conference, MultiMedia Modeling*, 2015, pp. 26–37.

[10] C. Laurier, "Multimodal Music Mood Classification using Audio and Lyrics," in *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, 2008, pp. 688–693.

[11] E. Peiszer, T. Lidy, and A. Rauber, "Automatic Audio Segmentation : Segment Boundary and Structure Detection in Popular Music," *Proc of LSAS*, vol. 106, no. August, pp. 45–59, 2008.

[12] W. B. De Haas, A. Volk, and F. Wiering, "Structural segmentation of music based on repeated harmonies," *Proceedings - 2013 IEEE International Symposium on Multimedia, ISM 2013*, pp. 255–258, 2013.

[13] L. D. Quadros, "Automatic structural segmentation of music," 2015.

[14] C. Yeh *et al.*, "Popular music representation : chorus detection & emotion recognition," *Multimedia Tools Application*, vol. 73, pp. 2103–2128, 2014.

[15] R. Malheiro, R. Panda, P. Gomes, and R. Paiva, "Bi-modal music emotion recognition: Novel lyrical features and dataset," *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*, pp. 1–5, 2016.

[16] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," *Ismir*, no. Ismir, pp. 625–636, 2010.

[17] O. Lartillot, "MIRtoolbox 1.6.1," Denmark, 2014.

[18] M. Doshi, S. K. Chaturvedi, and D. Ph, "Correlation Based Feature Selection (CFS) Technique to Predict Student Performance," *International Journal of Computer Networks & Communications (IJCNC)*, vol. 6, no. 3, pp. 197–206, 2014.

[19] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE : Corpus-Based of Emotion for Emotion Detection in Text Document," in *ICITACEE*, 2016, pp. 331–335.