On the Study of Thai Music Emotion Recognition Based on Western Music Model

N Satayarak, and C Benjangkaprasert*

School of Engineering, King Mongkut’s Institute of Technology Ladkrabang, Bangkok, THAILAND, 10520

nitjaree.sa@kmitl.ac.th, chawalit.be@kmitl.ac.th, *corresponding author

Abstract. The mood of the song could be identified by tracking the listener's emotion. The research in this area is growing significantly at the present. There are many research studies in western music, but a few in Thai music. Therefore, in this research, Thai songs were chosen because the Thai is a native language and Thai songs are quite popular in the region of research. This research is divided into 2 parts. First, Thai music was evaluated by the set of a system based on western music training settings. By using valence-arousal values, multiple linear regression, and k-nearest neighbors to represent the emotional annotations from the music. As a result, the highest f-measure of Thai music from multiple linear regression by ALL model was 41% and the f-measure of western music from multiple linear regression by No Tempo model was 51%, which was very different because ALL model in western music has lower efficiency than other models. Second, we measured the mood of 125 Thai popular songs and used valence-arousal (energy) values from Spotify API to investigate the results. In this research we used multiple linear regression (MLR) and support vector regression (SVR). Experimental results show that the multiple linear regression provides the highest accuracy of 61.29% with the precision of 65%, recall of 61%, and f-measure of 60% which is more than support vector regression.

1. Introduction
From the past, music is one of the most accessible works of art for humans, having various types of music to choose from. Nowadays, the music industry has brought technology to support and provide more convenience. Although music is based on art, the existence of music cause an effect to science, giving rise to new various scientific theories. In this research, the knowledge of engineering was applied to study music emotion recognition (MER) [1]. MER utilizes signal processing and machine learning applied with psychology and musicology to predict the emotion of the music. MER is becoming very popular nowadays, as a result of an application to listen to music called music streaming which has become the main channel for listening to music. Moreover, music streaming has offered a music recommendation system that allows listeners to choose songs according to their personal preferences or mood situation. MER will then help the provider understand the emotion of the music from the listener and use the analyzed emotion data to create a recommendation system to bring appropriate music to listeners, resulting in increased business values. However, such music emotion is highly subjective. Therefore, the development of a music emotion analysis system is a challenge for researchers today, especially research on non-western music since many research works...
are still based on western music. We found a few non-western works such as Sinhala (Sri Lanka) songs [2] and Indian (India) songs[3].

Although today’s popular songs around the world are based on the same genres (i.e., pop, rock etc.), the major difference is languages, which inevitably affects the model efficiency of machine learning. Possibly, the models trained by western music may not give accurate emotional results for non-Western music. The purpose of this research is to create a model to train western music emotion by using only the provided audio features that are not related to the language. Then, we bring Western and Thai songs as testing sets, to predict the emotion from that model. The results will prove whether the western music model could be applied on Thai songs with the same efficiency or not.

2. Hypothesis and Research Scope
The most efficient Western model should be able to apply to Thai songs. Considering removing language features from songs, the various musical structures should be very similar. The chosen music features in this research included pitch, dynamic, estimate tempo, tonality, and timber. Also, the significant key variable to determine emotion is rhythm, while the dynamic of the song seems to have the least influence that affects the emotion of the song.

3. Datasets Preparation

3.1. Western music dataset
The chosen western music dataset is MediaEval2013 [1] which is developed by MIREX consists of 744 songs from freemusicarchive.org, which were separated into a training set (619 songs) and testing sets (125 songs).

The dataset consists of 8 genres which include pop, rock, classical, jazz, folk, blues, country, and electronic. Each song has 45 seconds long, 16 bits, the sampling rate of 44.1 kHz with mono audio and MP3 files. The process of testing the music emotion of this dataset consisted of over 300 participants. Each participant took an online quiz by listening to a 45-second song and manually adjusting the emotion indicator which is called Valence-Arousal [4], [5] while listening to each song. According to Figure 1. where the most right represents the largest values and the most left represents the lowest value on the -1 scale.

![Annotating Valence](image)

Figure 1. MediaEval2013 valence emotion quiz example.

The authors of the dataset trim out the emotion values by the subject during the first 15 seconds because the mood at the beginning is unstable. The result is the average mood of each song in all 0.5 seconds. In this research, the emotion values were taken at every moment of each song to find the
average emotion and plotted on the diagram Valence-Arousal to find a mood value of each song according to Table 1 and Figure 2.

Table 1. The number of training and testing sets of MediaEval2013 dataset.

| Data     | Training | Testing | Total |
|----------|----------|---------|-------|
| Blues    | 82       | 18      | 100   |
| Classic  | 98       | 18      | 116   |
| County   | 90       | 15      | 105   |
| Electronic | 76      | 16      | 92    |
| Folk     | 62       | 9       | 71    |
| Jazz     | 84       | 21      | 105   |
| Pop      | 65       | 13      | 78    |
| Rock     | 62       | 15      | 77    |

Figure 2. Song emotion values of MediaEval2013 on Valence-Arousal diagram.

3.2. Thai music dataset
This research used songs from personal collections consisting of Thai popular songs. All songs were supported by copyright from iTunes and CD licensed. The quality of Thai music files was modified to the same standard as Western music which is an MP3 file of 45 seconds in length, 16 bits, with a sampling rate of 44.1 kHz and mono audio. The Spotify API [6] was chosen to extract data and determine mood by showing the Valence-Arousal values (Note that Arousal in Spotify API that we used in this research, called as Energy). The basic API operation system is shown in Figure 3, and the example of API data is shown in Figure 4. The emotional values from SpotifyAPI were plotted on the Valence-Arousal diagram as shown in Figure 5.

Figure 3. The basic functionality of the Spotify API.
Figure 4. Audio features from Spotify API.

Figure 5. Thai music emotions on the Valence-Arousal map.

Table 2. The number of emotion values in 125 Thai songs.

| Emotions of Spotify API | Happy | Excited | Sad | Peaceful |
|------------------------|-------|---------|-----|----------|
|                        | 64    | 43      | 15  | 3        |

As shown in Figure 5 and Table 2. The emotional values of popular Thai songs have high alertness including a mood of happiness (green) and excitement (red), while on the other hand, there were only a few songs with low alertness including a mood of sadness (black) and a mood of peace (blue).

4. Performance Evaluation
The obtained results of machine learning have to be measured in various ways. In this research, the outcome is divided into 4 fields including accuracy, precision, recall, and F-measure. Each field represents results that indicate the performance of each model which can be calculated from the equation as shown in equation (1), equation (2), equation (3), and equation (4), respectively. The performance evaluation also consists of the definition called true positive, true negative, false positive, and false-negative which are displayed in the confusion matrix as shown in Figure 6.
Figure 6. Example of the Confusion matrix.

Table 3. Confusion matrix with 4 emotions.

|        | Happy | Excited | Sad | Peaceful |
|--------|-------|---------|-----|----------|
| Happy  | 41    | 0       | 8   | 5        |
| Excited| 12    | 0       | 2   | 1        |
| Sad    | 5     | 1       | 21  | 2        |
| Peaceful | 9    | 0       | 10  | 8        |

The accuracy measurement formula is given by

\[
Accuracy = \frac{\text{True positive} + \text{True negative}}{\text{Total data}} \quad (1)
\]

As shown in equation (1), the Y-axis is the actual emotion and the X-axis is the predicted emotion. If the predictions match the actual emotions, it is True positive (green). If the prediction does not match the actual emotions, it is True negative (yellow). These two field values are used together to find the correct value. As in equation (1), the accuracy equation indicates the accuracy of the system compared to the total number of songs. However, the accuracy can be unreliable in the scenario that the True group data is not distributed.

Then, the precision measurement formula is given by

\[
Precision = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (2)
\]

As in equation (2), the precision equation indicates the performance of model prediction. By false positive (blue) from Figure 6 and Table 3. This shows that the prediction model was happy. However, when compared to the actual emotions, the returned results matched with other moods including excited, sad, and peaceful total of 26 songs. This indicates that it is effective in terms of quantity but not quality.

The recall measurement formula is given by

\[
Recall = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (3)
\]

As in equation (3), the recall equation indicates the performance of the model compared to actual emotion. This indicates that how much is the actual emotion matched the predicted emotion by false-negative value. From Figure 6 and Table 3, shows that the actual happy mood is misdirected to the other emotions for the total of 13 songs.

The F-measure formula is given by
\[ F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4} \]

As in equation (4), F-measure is the overall performance measurement which is more reliable than precision because the formula brings both precision and recall value to be calculated.

In summary, Table 3 illustrated that after the accuracy was calculated, the result was 56\% which on the emotional side is quite enough for reliability. However, when the actual results have been analyzed, it was found that the correct data did not spread because excited values were equal to 0. This means that there was no correct prediction for this excited emotion. The majority of correctness came from 56\% of happy emotion.

As shown in Figure 7, the ROC curve stands for receiver operating characteristic curve which will also indicate the performance values according to the area under the curve values as shown in Table 4. Also, this model has the highest rate of sad mood prediction with an average value of 0.76.

![Figure 7. ROC characteristic curve.](image)

| Table 4. ROC performance level indicator. |
|-------------------------------------------|
| AUC (Area Under the Curve)                |
| Excellent | Good | Fair | Poor | Fail | Worthless |
| 1-0.9     | 0.89-0.8 | 0.79-0.7 | 0.69-0.6 | 0.59-0.5 | <0.5 |

5. Results and Discussion

The results of regression analysis prediction on the Thai songs emotion will be shown in this section. We chose multiple linear regression (MLR) and support vector regression (SVR) \([7]-[8]\) because they were powerful models for western music emotion prediction. In this section, regression classification was selected which is a type of linear model.

The confusion matrix for MLR and SVR models were shown in Table 5 and Table 6 respectively. As seen in the confusion matrix for both models, there was only one difference in terms of the number of a song which is the one song with exciting emotion that appeared to be a true positive for MLR. However, there was no correct data for peaceful emotion which was zero hence only 3 peaceful songs are being predicted, so there was a low chance of making a correct prediction for peaceful mood songs.
Table 5. Confusion matrix for MLR model Thai songs.

| Emotions | Happy | Excited | Sad | Peaceful |
|----------|-------|---------|-----|----------|
| Happy    | 12    | 4       | 0   | 0        |
| Excited  | 4     | 6       | 0   | 0        |
| Sad      | 0     | 2       | 1   | 1        |
| Peaceful | 1     | 0       | 0   | 0        |

Table 6. Confusion matrix for SVR model Thai songs.

| Emotions | Happy | Excited | Sad | Peaceful |
|----------|-------|---------|-----|----------|
| Happy    | 12    | 3       | 1   | 0        |
| Excited  | 5     | 5       | 0   | 0        |
| Sad      | 0     | 2       | 1   | 1        |
| Peaceful | 1     | 0       | 0   | 0        |

The precision, recall, and F-measure of MLR and SVR are shown in Table 7 and Table 8, respectively. The precision for the MLR model for sad mood prediction was 100%. Although SVR predicted sad mood correctly for only one song, the reason that precision value for sad mood was only 50% is that there was one song from the prediction of sad mood matched with a happy mood, resulting in loss of efficiency in precision. Nevertheless, the precision value of Happy mood for both MLR and SVR are 71% and 67% respectively which is not much different.

The recall value of happy and sad mood for MLR and SVR was equivalent while in excited mood for MLR and SVR was 60% and 50%, respectively since SVR has lost its performance in one song while MLR took that difference as a valid data value.

The higher F-measure in MLR showed more effective result for every mood. When the average performance for both models has been calculated, the means for MLR performed better in all aspects as shown in Figure 8.

Table 7. Results of Precision, Recall, and F-measure values of MLR model prediction on Thai songs.

| Emotions | Precision | Recall | F-measure |
|----------|-----------|--------|-----------|
| Happy    | 71%       | 75%    | 73%       |
| Excited  | 50%       | 60%    | 55%       |
| Sad      | 100%      | 25%    | 40%       |
| Peaceful | 0%        | 0%     | 0%        |
| Average  | 65%       | 61%    | 60%       |

Table 8. Results of Precision, Recall, and F-measure values of SVR model prediction on Thai songs.

| Emotions | Precision | Recall | F-measure |
|----------|-----------|--------|-----------|
| Happy    | 67%       | 75%    | 71%       |
| Excited  | 50%       | 50%    | 50%       |
| Sad      | 50%       | 25%    | 33%       |
| Peaceful | 0%        | 0%     | 0%        |
| Average  | 57%       | 58%    | 57%       |
Figure 8. Comparison of Accuracy, Precision, Recall, and F-measure between MLR and SVR model prediction on Thai songs.

6. Conclusion
Based on the objective of selecting the most efficient emotional classification model that originally applies with the western songs model for predicting and classifying the emotion of Thai songs. The results of Thai song’s emotion prediction using MLR and SVR linear models turned out to be better than any other studies on western music mood. Moreover, the MLR model prediction outperformed the SVR model prediction in every aspect which contains the aspect of measuring accuracy, precision, recall, and F-measure values. The MLR model prediction results have an accuracy of 61.29%, precision of 65%, recall of 61%, and F-measure of 65% which are within the acceptable range in terms of mood classification.

All results from this research show that western songs and Thai songs are very different even though there was no language barrier involved since the lyrics had been cut out in the dataset preparation process. This could be that there are significant differences in audio attributes even in general, they are the same music genres on the same basis. Therefore, if one wants to create a model prediction based on the emotion in a desired country, the country-specific music model prediction should be newly generated according to each country’s audio tones.

In this research, the western music dataset MediaEval2013 and Thai music dataset Spotify API were used. However, the result of the simulation is not based on the actual situation or real scenarios. Therefore, in the future, the newly conducted experiment should be tested with actual Thai people to predict the mood of Thai songs most effectively. Also, it might be more effective if the Thai music dataset is specifically generated for the research’ or we could include alternative ways to improve the efficiency of analysis [10-12].

Acknowledgments
The authors wish to thank Mr. Soravitt Sangnark for his technical assistant.

7. References
[1] Y. E. Kim, Erik M. Schmidt and R. Migneco, 2010. “Music emotion recognition: A state of the art review,” Proc. of Int. Society for Music Information Retrieval Conference, pp. 255-266.
[2] M. G. Viraj and K. L. Jayaratne, 2016. “Melody analysis for prediction of the emotions conveyed by Sinhala songs,” Proc. of Int. Conference on Information and Automation for Sustainability, pp.1-6.

[3] S. H. Deshmukh and S. G. Bhirud, 2014. “North Indian classical music’s singer identification by timbre recognition using MIR toolbox,” Int. Journal of Computer Applications, vol. 91, No. 4, pp. 1-5.

[4] J. A. Russell, 1980. “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178.

[5] R. E. Thayer, 1989. The biopsychology of mood and arousal, Oxford University, New York, Inc.

[6] https://developer.spotify.com/documentation/web-api/  

[7] J. Bai, J. Peng and J. Shi, 2016. “Dimensional music emotion recognition by valence arousal regression,” Proc. of Int. Conference on Cognitive Informatics & Cognitive Computing, pp. 42-49.

[8] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, 2008. “A regression approach to music emotion recognition,” IEEE Trans. Audio, Speech & Language Processing, vol. 16, no. 2, pp. 448-457.

[9] P. M. F. Vale, 2017. “The role of artist and genre on music emotion recognition,” Universidade Nova de Lisboa.

[10] Y. An, S. Sun and S. Wang, 2017. “Naive Bayes classifiers for music emotion classification based on lyrics,” Proc. of Int. Conference on Computer and Information Science, pp. 635-638.

[11] D. Ayata, Y. Yaslan and M. E. Kamasak, 2018. “Emotion based music recommendation system using wearable physiological sensors,” IEEE Trans. on Consumer Electronics, vol. 64, no. 2, pp. 196-203.

[12] D. Chaudhary, N. Pratap and S. Singh, 2021. “Development of music emotion classification system using convolution neural network,” Int. Journal of Speech Technology, vol. 24, pp. 571–580.