Music Emotion Recognition for Arabic Tracks Using Extreme Gradient Boosting

Mhd Shadi Hasan

Abstract

Intelligent information retrieval is a requirement with today's massive data available online. Researchers and private-sector companies are investing time into developing methods to progress information retrieval in many areas; including music. Music Emotion Recognition allows content providers to automatically recognize the emotion in music audio tracks and use this knowledge to improve user experience. Extreme Gradient Boosting Classifier was used in this research to classify Arabic music tracks recommended by listeners, on Arousal and Valence labels both separately and jointly, based on five acoustic features. Four evaluation metrics were used to evaluate the model's performance: Accuracy, Precision, Recall, and F1-Score. Results show that the model was able to predict Arousal labels better than it did with Valence ratings or Arousal/Valence labels jointly.

Keywords: Classification, Music Emotion Recognition, Valence, Arousal, Extreme Gradient Boosting

Introduction

Music is important in almost every culture. It is considered as a part of the identity of many peoples around the world and therefore great interest is given to preserving and presenting music as a mean of communication that crosses boarders. With the fourth industrial revolution and the integration of the internet, it has never been easier to listen to music anywhere through dozens of applications and websites. In fact, one can become overwhelmed by the amount of available options.

The exponentially-growing size of information on the internet (including music) requires focus on intelligent information retrieval techniques. This includes advanced search and filtering and personalized recommendations to offer a better user experience. Retrieving music audios based on emotion is of interest to many researchers. It has potential applications in Context-Aware Recommendation Systems (CARS) and can enhance the user experience for many music applications and websites. It requires an automated Music Emotion Recognition (MER) process. Music Emotion Recognition is difficult because the perceived emotion from the same music piece can vary from one person to another not to mention how they would describe their perceived emotion [1].

Furthermore, Arabic music in particular has received very little attention in machine learning and data science literature. Researchers in [2] worked on classifying Arabic music based on acoustic features to one of four categories of cultural style. However, automatically recognizing the emotion invoked by an Arabic music piece has not been addressed before.

This research investigates an Extreme Gradient Boosting-based Classification System's ability to predict the emotion invoked by Arabic music pieces in listeners. In particular, the model predicts emotional categories based on Arousal/Valence labels as per Russell's circumplex model [3].

Background

A. Classification and XGBoost

Classification is a Supervised Machine Learning technique used to assign a label to instances in a dataset. It is a Supervised Learning method because the Classification model requires labeled input data in order to find relationships between the features of each instance and its label.

There are several Classification algorithms, but for this research we will use Extreme Gradient Boosting (XGBoost). XGBoost is an ensemble learning method that uses boosting to reduce variance in a group of weak learners. Boosting is a powerful technique for combining multiple classifiers; known as base classifiers, and generates a single strong learner [4].

B. Russell's Circumplex Model of Affect

Russell's Circumplex Model of Affection [3] is a method of mapping emotions to a 2D plane [Figure 1]. The horizontal axis represents Valence; which is the measure of positivity. The vertical axis on the other hand, represents Arousal; which is a measure of activation.

This model has been used in a plenty of MER research and has been used in [5], [1], [6], [7], and [8]. It provides a simple way of understanding and mapping the emotion in music pieces. Continuous and Categorical approaches have been used by researchers who addressed this problem. Researchers in [1], [6], and [7] viewed the plane from a continues perspective and followed a Regression approach in MER. whereas, a Classification approach was followed in [8].

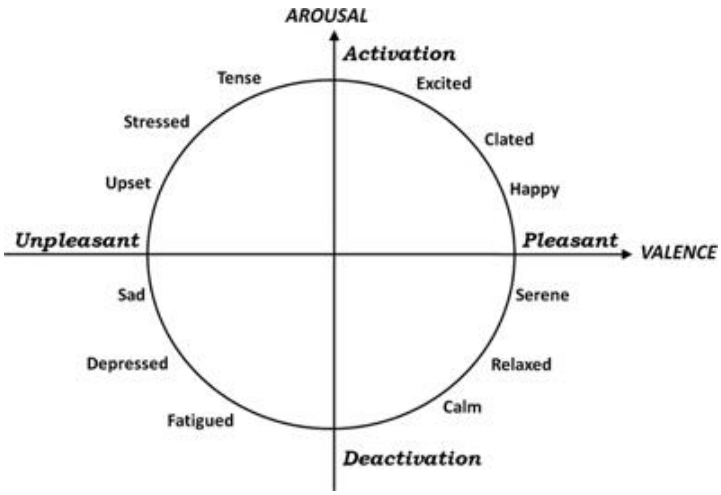

Figure 1: Russell's Circumplex Model of Affect [3]

Method

A. Data Collection

For the purpose of this research, a group of 24 males and females, with ages ranging from 20 to 28 years old, were asked about their favorite Arabic music piece or song. As a result, 24 audio tracks were collected with different durations (Mean= 551.125 sec, SD= 661.15).

B. Feature Extraction

Music audio files are a type of unstructured data. However, in order to train the model on data of tabular form, MIRToolbox [9] was used in MATLAB for acoustic feature extraction as it is simple, straightforward, and was used by many researchers for acoustic feature extraction.

Features related to Dynamics, Rhythm, Timber, and Tonality were extracted and organized in a data frame for later analysis and model training. In particular, seven features were extracted [Table 1] using MIRToolbox; namely Low Energy, Zero Crossing, Roll Off, Brightness, Centroid, Mode, and Pulse Clarity [10]. These features were found to have high correlation with Arousal and Valence ratings based on bivariate correlation analysis [5] [8].

Table 1: Summary of The Extracted Features

	lowEnergy	pulseClarity	zeroCross	rollOff	brightness	centroid	mode
count	24.000000	2.400000e+01	24.000000	24.000000	24.000000	24.000000	24.000000
mean	0.527851	3.403663e+05	1574.571742	6177.930587	0.477849	2853.042858	-0.068813
std	0.063069	3.382231e+05	2080.529341	2090.792539	0.117429	844.092503	0.142821
min	0.340570	1.097635e+04	488.822600	1739.158600	0.200570	1030.115300	-0.297740
25%	0.498947	1.638139e+05	851.946775	4731.641250	0.414660	2317.993050	-0.150358
50%	0.532205	2.747127e+05	1111.003250	6115.442300	0.494960	2894.001850	-0.070293
75%	0.550550	3.750679e+05	1561.542925	7704.912175	0.529075	3213.590100	0.015013
max	0.662000	1.626029e+06	11120.145600	10367.992400	0.686050	5073.971000	0.260500

Inter-correlation between the features was investigated numerically [Figure 2] and visually [Figure 3]. It was found that three features; namely Roll Off, Brightness, and Centroid, were highly correlated.

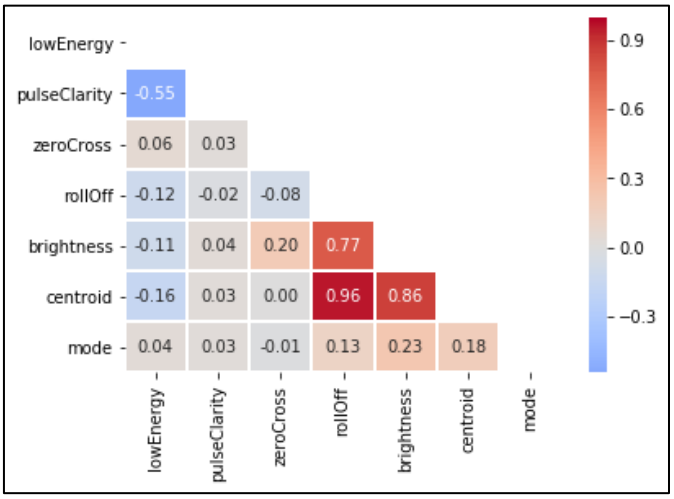

Figure 2: Numerical Features Inter-Correlation Analysis

Based on this analysis, two features; Centroid and Brightness, were dropped, leaving five acoustic features to be used to build the model.

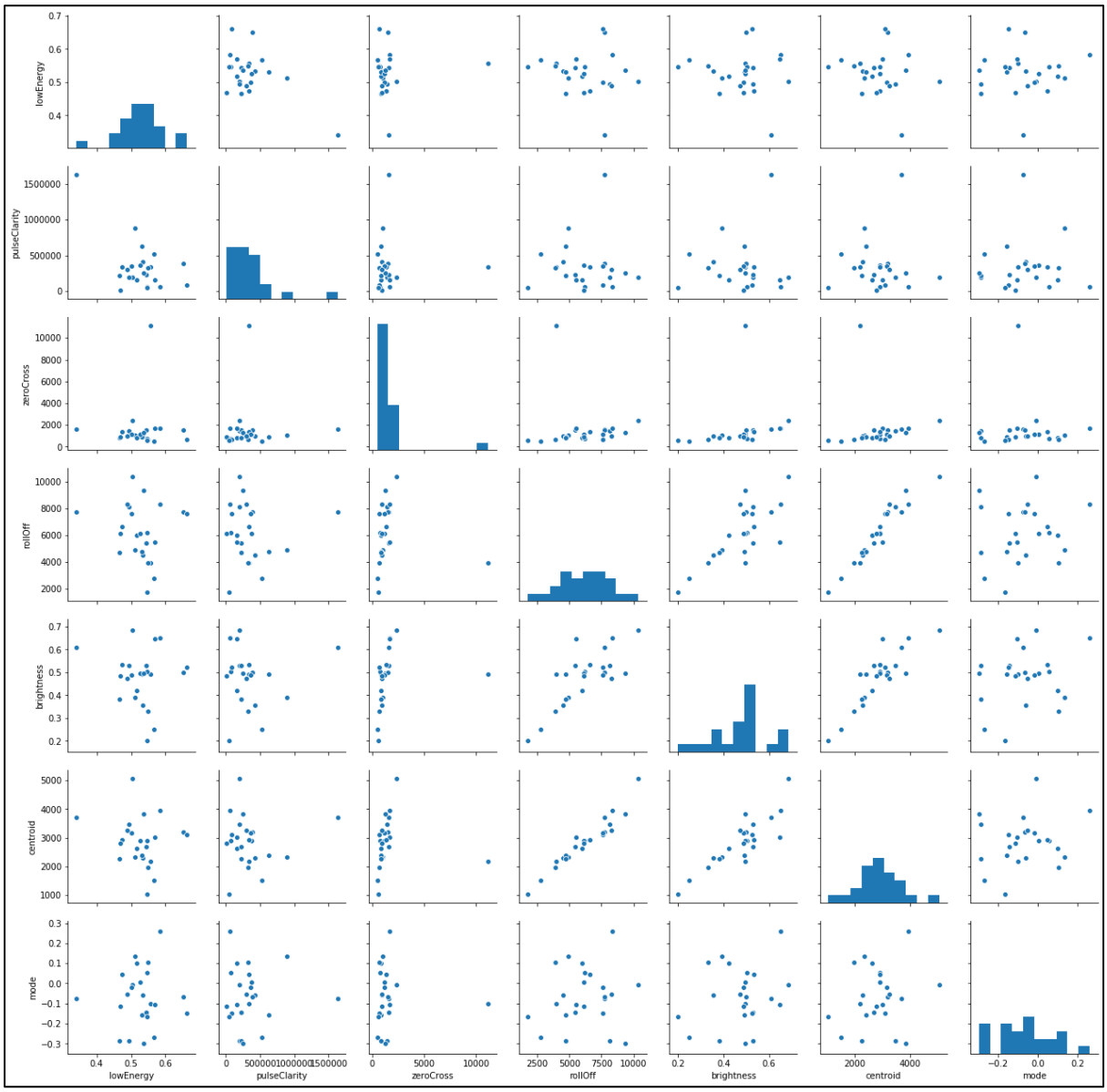

Figure 2: Visual Features Inter-Correlation Analysis

C. Target Labels

The labels used to train the classifier and predict music emotion were collected from the individuals who recommended the music tracks. Therefore, the emotion labels are based on listeners subjective rating. This method was used in [8], [5], [1] and others.

Participants labeled each track as High (*H*) or Low (*L*) for Valence as well as for Arousal. Consequently, each track's labeling to one of the four quadrants in Russell's Circumplex Model of Affect and their respective Arousal/Valence ratings [Table 2] is the input that the model will train on. Therefore, there are three target variables that the model will be tested against; *emotionQuadrant (1, 2, 3, or 4)*, *Valence (High 'H' or Low 'L')*, and *Arousal (High 'H' or Low 'L')*.

Table 2: Distribution of Tracks into Labels

Number of Tracks	emotionQuadrant	Valence	Arousal
8	1	H	H
2	2	L	H
8	3	L	L
6	4	H	L

D. Model Training

After standardizing the feature set, The XGBoost Classification model was trained on the five input features remaining after accounting for inter-corrolation among features. In fact, the model was trained and evaluated three times, each time a different target variable was taken among *emotionQuadrant, Arousal,* and *Valence*. To account for the small data size, five-fold cross-validation technique was used to evaluate the model performance.

Results, Discussion and Conclusion

Four evaluation metrics have been calculated each time the model was trained against a target variable. The evaluation metrics are:

1- Accuracy

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (1)$$

2- Precision

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2)$$

3- Recall

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3)$$

4- F1-Score

$$\boldsymbol{F1-Score} = \boldsymbol{2} \times \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision+Recall}} \quad (4)$$

The summary of the results [Table 3] shows that the model predicted the quadrent to which the track belongs very poorly with an accuracy of 17%. When tested against Valence and Arousal ratings independently, it had an accouracy of 67% in predicting Arousal labels, and 38% in predicting Valence labels.

It can be seen that the model was able to predict Arousal labels relatively well. Other features should be further considered to improve Valence predictions; possibly based on lyrics, which will require Natural Language Processing techniques to be applied. Furthermore, the model should be trained and tuned on larger datasets.

Table 3: Summary of Results

Target Variable	Accuracy	Class	Precision	Recall	F1-Score
emotionQuadrent	0.17	1	0.12	0.12	0.12
		2	0.00	0.00	0.00
		3	0.14	0.12	0.13
		4	0.22	0.33	0.27
Arousal	0.67	H	0.60	0.60	0.60
		L	0.71	0.71	0.71
Valence	0.38	H	0.47	0.57	0.52
		L	0.14	0.10	0.12

References

[1] Y. Yang, Y. Lin, Y. Su and H. H. Chen, "A Regression Approach to Music Emotion Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 448-457, Feb. 2008, doi: 10.1109/TASL.2007.911513.

[2] Soboh, Lama, Islam Elkabani, and Ziad Osman. "Arabic cultural style based music classification." In 2017 International Conference on New Trends in Computing Sciences (ICTCS), pp. 6-11. IEEE, 2017.

[3] Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39, no. 6 (1980): 1161.

[4] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[5] Vempala, Naresh N., and Frank A. Russo. "Predicting emotion from music audio features using neural networks." In Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), pp. 336-343. London, UK: Lecture Notes in Computer Science, 2012.

[6] Grekow, Jacek. "Music emotion maps in the arousal-valence space." In From content-based music emotion recognition to emotion maps of musical pieces, pp. 95-106. Springer, Cham, 2018.

[7] Bai, Junjie, Jun Peng, Jinliang Shi, Dedong Tang, Ying Wu, Jianqing Li, and Kan Luo. "Dimensional music emotion recognition by valence-arousal regression." In 2016 IEEE 15th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), pp. 42-49. IEEE, 2016.

[8] Nawaz, R., Nisar, H., Voon, Y.V. and Yee, T.P., 2018, July. Acoustic feature extraction from music songs to predict emotions using neural networks. In 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS) (pp. 166-170). IEEE.

[9] Olivier Lartillot, Petri Toiviainen, “A Matlab Toolbox for Musical Feature Extraction From Audio”, International Conference on Digital Audio Effects, Bordeaux, 2007.

[10] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, Jose Fornari, "Multi-feature modeling of pulse clarity: Design, validation, and optimization", International Conference on Music Information Retrieval, Philadelphia, 2008.