Automatic Valence and Arousal Recognition in Music

Damian Domela Nieuwenhuis Nyegaard

Leiden Institute of Advanced Computer Science, Leiden 2333 CA

Abstract. In this paper, an automatic music affect (emotion) recognition system is constructed based on lyrical and audio features. This system predicts which of the four quadrants of Russell's valence-arousal space a song belongs to. A Naive Bayes model based on lyrical features was used for valence classification, while an SVM based on audio features was used for arousal classification. Although this model ensemble did not outperform the state-of-the-art method in this area, the simplicity of the implementation could warrant its suboptimal performance.

Keywords: Affect recognition, Valence, Arousal, Music, Russell's model

1 Introduction

With the size of digital music collections rapidly increasing, the magnitude of manually annotating these datasets becomes more difficult everyday. Providing these music collections with annotation music tags allows for automatic search and recommendation of music based on music listening history. Starting from 2002, the standard annotation method was automatic genre classification proposed by Tzanetakis et al.[4]. Based on this research, automatic affect and emotion classification in music have developed significantly.

The field of affect recognition has significantly based its research on a model proposed by Russell et al.[1], in which a two-dimensional space represents valence on the x-axis and arousal on the y-axis. Valence in this context refers to *the pleasantness of a stimulus*, while arousal in this context refers to *the intensity of emotion provoked by a stimulus*[9]. In this model, all possible emotions can be represented using a linear combination of valence and arousal. As is visible in Fig. 1, these axes seperate four quadrants, with each quadrant representing low or high valence and low or high arousal.

In order to train a model to classify a set of songs towards a target, a representative quantification of these songs is required. This is done by extracting features from its corresponding data. Constructing the optimal audio feature set for this classification task has been researched by Grekow et al.[6]. Based on recent research by Raschka et al.[7], lyrical features are considered informative towards affect recognition. Thus, the corresponding lyrics are preprocessed and used as additional features.

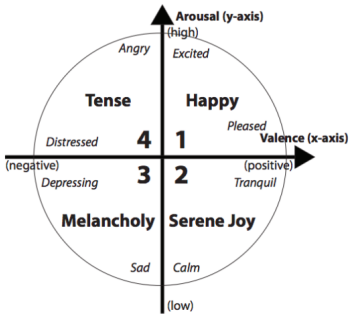


Fig. 1: Russell's Circumspect model of affect (adapted from Malheiro et al. [2])

In this paper, a model ensemble consisting of a Naive Bayes model, which is used to classify high or low valence in a song, and an SVM, which is used to classify high or low arousal in a song, is constructed based on the architecture proposed by Tan et al.[5]. These models were trained on lyrical and audio features respectively. By combining the output of these two binary classification models, a multi-class classification model is set up which maps a song to one of the four quadrants in Russell's model of affect.

2 Methodology

Section 2.1 discusses the bimodal dataset used to train and test the model ensemble, Section 2.2 outlines how the features were extracted from this dataset, Section 2.3 describes the structure of the used Naive Bayes model, Section 2.4 discusses the structure of the used SVM and Section 2.5 outlines the construction of the experimental setup. The source code for this project is available here[1].

2.1 Bimodal Dataset

The used dataset was proposed by Panda et al.[8]. It originally contained 180 lyric samples and 162 audio clips, with an overlap of 133 bimodal data points. The missing lyrical files were manually added to the dataset, creating a relatively larger overlap between lyrical and audio files of 162 bimodal data points. Each data point is labelled with one of the four quadrants in the Russell's model, representing the ground truth in this context. In Fig. 2, the class distribution of the improved dataset is visible. During the construction of the dataset, songs were selected which are diverse in genre and era of release. In addition to this, the dataset offers a proportional class distribution, preventing class imbalance complications with respect to fitting the model.

Despite the subtle increase in size for the dataset, it still is relatively small in size in comparison to datasets in other fields. The authors acknowledge this,

[1] https://github.com/TheBeast762/TheBeast762.github.io

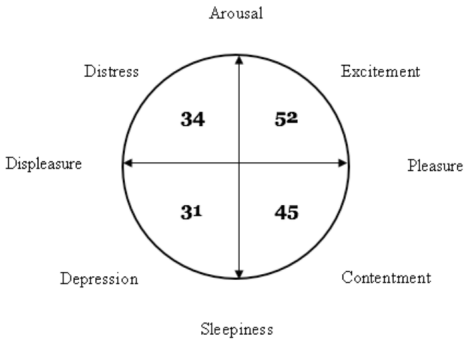


Fig. 2: Overview of the improved bimodal dataset its class distribution

but consider it large enough for experiments. It is challenging to further increase the size of the dataset, as the annotation was done manually by 39 people of different backgrounds, with an inter-annotator agreement of 0.87 and 0.82 for the valence and arousal classification respectively.

2.2 Feature Extraction

In order to extract the lyrical features, the raw lyrical text files were pre-processed by first removing stop words and punctuation and then lemmatizing each word. These lemmas were then matched with a valence dictionary proposed by Warriner et al.[9], in which 13,915 english lemmas are labeled with a valence score based on crowdsourced manual annotation. In this annotation process, each words is required to be labeled with a score ranging from 0-10. With 0 representing minimal valence and 10 representing maximal valence.
Tan et al.[5] made a cutoff based on the average lemma valence scores present in the lyrical files, providing a positive or negative average if the average valence score was above or below 5.0 respectively. These average valence scores for all lyrical files were represented as labels in a seperate concise dataset.

The structure of the audio feature set is based on research by Grekow et al.[6], in which varying combinations of audio features are tested for their performance towards music emotion recognition. This selected feature set is deemed fitting, as the research utilized the same target output of Russell's valence-arousal model. The Python library Librosa[11] was used to extract the Tonnetz feature, all other listed features were extracted using the pyAudioAnalysis library, proposed by Giannakopoulos et al.[10]. Table 1 shows the feature set that was used for arousal classification.

2.3 Naive Bayes model

Given an average valence label for each lyrical text file, a negative average valence label is considered low valence and a postive average valence label is considered

Audio Features	
Energy	Std. Dev. Energy
Entropy of Energy	Std. Dev. Entropy of Energy
Spectral Entropy	Std. Dev. Spectral Entropy
Beats per minute	Std. Dev. Beats per minute
Spectral Roll-off	Std. Dev. Spectral Roll-off
Spectral Flux	Std. Dev. Spectral Flux
(Mean) Tonnetz	

Table 1: **Arousal classification feature set**

high valence. A Naive Bayes model is then provided the pre-processed lyrics as input, given these lyrics with their corresponding frequency, the probabilities for high and low valence are learned. Once the training phase is completed, the model finds the largest probability for low or high valence given a pre-processed lyrics file.

2.4 SVM

Given the audio feature set, a Linear Support Vector Classifier (SVC) is trained with the regularization parameter C set to 150, searching for a relatively large-margined separating hyperplane as audio classification with a small dataset calls for high generalization. Seeing as audio classification is not linearly seperable, forming a linear hyperplane between low and high arousal classes based on audio features is guaranteed to be imperfect. The target labels were extracted from the bimodal dataset.

2.5 Experimental Setup

Since the used dataset is relatively small, k-fold Cross Validation is prone to produce skewed results due to extreme outliers. These extreme outliers are likely due to Naive Bayes being sensitive to imbalanced training set class distributions resulting from the small dataset. In order to circumvent this, the performance assessment is done by utilizing repeated random sub-sampling. This method allows for a large amount of iterations of a 10% test set and a 90% training set split, averaging out to more trustworthy performance results, despite the variance in results. This process is visible in Fig. 3. The results of this experiment are based on 30 iterations.

3 Results

Fig. 4a show the results of the model ensemble in terms of accuracy, Fig. 4b shows the results in terms of F1-score. The box & whisker plots clearly show a large amount of variance in performance across all iterations. This is a direct result of the relatively small dataset. The binary classification SVM and Naive

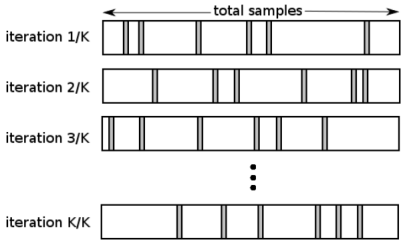


Fig. 3: Repeated random sub-sampling method (adapted from Chlis et al. [3])

Bayes model both achieved a mean accuracy(%) and F1-score of 78. The Naive Bayes and SVM model ensemble achieved a mean accuracy of 62% and a mean F1-score of 57. Given that the class distribution of the bimodal dataset is not exactly uniform, the lower F1-score relative to the accuracy illustrates that the model ensemble is prone to slightly overfit on the majority classes.

4 Discussion

State-of-the-art performance in this area was achieved by Catharin et al.[12], achieving an overall accuracy of approximately 73%. This research built upon the methods proposed by Tan et al.[5], but used a specific SVM structure with a custom-built feature set. Although this relatively simple implementation did not manage to rival this state-of-the-art performance, it is accurate enough to be used for song recommendations based on listening history.

In order for a human to manually assess the similarity of the model ensemble its prediction to the correct labels, as well as to human interpretation, a simple visual tool was developed, which is visible in Fig. 5. All four quadrants of the valence-arousal space are depicted with examples of corresponding emotions. The model ensemble its prediction and the correct label is visually represented, allowing for direct comparison while a 10 second clip of the associated song is played. This process is repeated for each song in the test set.

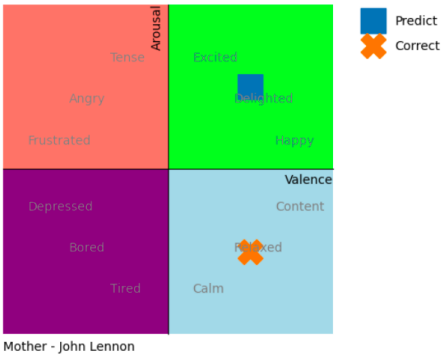


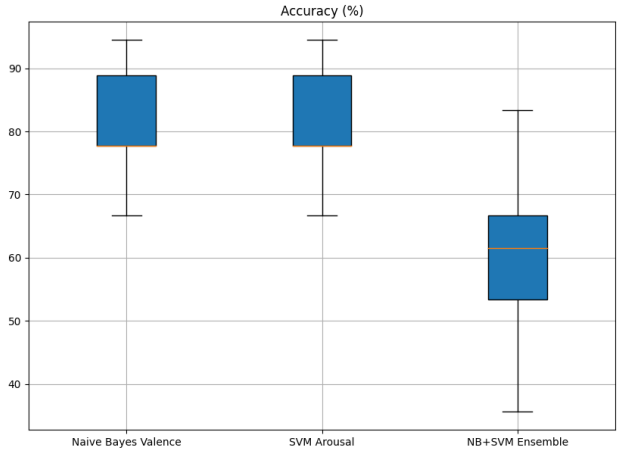
Fig. 5: Visual Tool to assess the model ensemble its performance

5 Conclusion

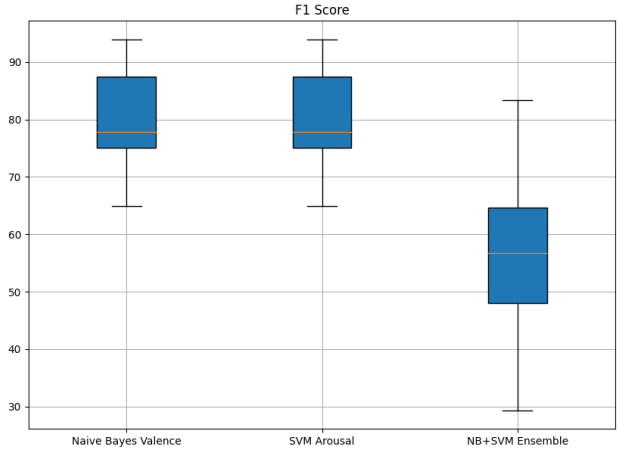
Previous research has shown that lyrical features are informative towards valence recognition in music, while audio features are informative towards arousal recognition in music. In this paper, an extensive explanation was given about the strengths and shortcomings of a reproduction of Tan et al.[5] their used model structure. The results for the model ensemble show that the individual models perform well on their associated feature sets, reaffirming that lyrics and audio are indicative towards valence and arousal recognition respectively. In addition to this, the results also show that the dataset its small size leads to a large variance in performance and that it slightly overfits on the majority classes.

References

1. Russell. J. A.: A Circumplex Model of Affect. In: Journal of Personality and Social Psychology 39(6). pp. 1161-1178 (1980)
2. Malheiro R., Panda R., Gomes P., & Paiva R. P: "Emotionally-relevant features for classification and regression of music lyrics," In: IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 240–254, (2016)
3. Chlis. N. K.: Machine Learning Methods for Genomic Signature Extraction. (2015)
4. Tzanetakis G., Cook P.: Musical genre classification of audio signals IEEE Trans. on Speech and Audio Process. 10(5) pp 293-302 doi:10.1109/tsa.2002.800560 (2002)
5. Tan K.R., Villarino M.L., Maderazo C.: Automatic music mood recognition using Russell's twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naive Bayes. In: IOP Conference Series: Materials Science and Engineering. pp. 12-19 (2019)
6. Grekow J.: Audio features dedicated to the detection and tracking of arousal and valence in musical compositions. In: J. of Inform. and Telecom. 1(12) doi:10.1080/24751839.2018.1463749 (2018)
7. Raschka S.: MusicMood: predicting the mood of music from lyrics using machine learning (Michigan: Michigan State University) Preprint arXiv:1611.00138 (2014)
8. Malheiro R., Panda R., Gomes P., Paiva R.: Bi-modal music emotion recognition: Novel lyrical features and dataset. In: 9^{th} Int. Work on Music and Machine Learning (2016)
9. Warriner A.B., Kuperman V. & Brysbaert M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. In: Behavior Research Methods 45, pp. 1191–1207 (2013)
10. Giannakopoulos T.: pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLoS ONE 10(12): e0144610. doi:10.1371/ journal.pone.0144610 (2015)
11. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8). (2015)
12. Catharin L. G., Ribeiro R. P., Silla C. N., Costa Y. M. G., Feltrim V. D.: Multimodal Classification of Emotions in Latin Music. In: 2020 IEEE International Symposium on Multimedia. (2020)



(a) Box & Whisker plot of achieved accuracies



(b) Box & Whisker plot of achieved F1-scores ($\times 100$)

Fig. 4: Results for model ensemble