

Multi-Modality in Music: Predicting Emotion in Music from High-Level Audio Features and Lyrics

Tibor Krols

mdl1163@alumni.ku.dk

Yana Nikolova

xjv550@alumni.ku.dk

Ninell Oldenburg

gxf740@alumni.ku.dk

Abstract

This paper aims at testing if a multi-modal approach for music emotion recognition (MER) performs better than a uni-modal one on high-level song features and lyrics. We use 11 song features retrieved from the Spotify API, combined lyrics features including sentiment, TF-IDF and Anew to predict valence and arousal (Russell, 1980) scores on the Deezer Mood Detection Dataset (DMDD) (Delbouys et al., 2018) with 4 different regression models. We find that out of the 11 high-level song features, mainly 5 contribute to the performance, that multi-modal features do better than audio alone when predicting valence.

This text contains 24906 characters.

1 Introduction (NO)

The emotional effect of music is a fact both well-studied (Hunter & Schellenberg, 2010; Balkwill & Thompson, 1999; Deng et al., 2015) and intimately familiar for most people. Furthermore, cross-cultural consistency in emotion recognition in music (Balkwill & Thompson, 1999; Fritz et al., 2009) indicates that fundamental cognitive capacities are at play in the process. With the rise of music digitization and affective computing technologies, there has been an increasing interest in extracting emotional information from music for research, therapy (Dingle et al., 2015; Bernatzky et al., 2011) or music recommendation (Deng et al., 2015; Park et al., 2015) purposes. This research falls under the term Music Emotion Recognition or MER (Chen et al., 2015). Nowadays, even music streaming platforms like Spotify come with an API that makes a wide range of features accessible and therefore open to the public.

But which features can actually predict the emotion of a song and how well perform Spotify’s

annotations? Building on existing literature presented in Section 2 we hypothesize *that a multi-modal approach combining high-level auditory and lyrics-extracted features performs better than a uni-modal one* (Y.-H. Yang, Lin, Cheng, et al., 2008; Hu & Downie, 2010b, 2010a). We introduce our MER model in Section 3 before presenting and discussing the results of our exploratory and regression experiments in Sections 4 and 5.

2 Related Work (NO)

The work on automatic information or prediction methods from music began with the rise of digital music. The annual Music Information Research Evaluation eXchange (MIREX)¹ included audio music mood as a classification task for the first time in 2007 (MIREX, 2007) already finding that ground-truth labeling and human mood judgment play a critical role on this task (Hu et al., 2008). The used range of features already varied from spectral, so-called low-level features (raw frequency signal, for example, spectral centroids and SD (X. Yang et al., 2018)) to higher-level features like “danceability” (Hu et al., 2008), rhythm, or timbre (X. Yang et al., 2018) that can be extracted from low-level features and are being described as closer to human perception (Panda et al., 2021). From there on, systems were consequently further being improved (Kim et al., 2010).

While the literature broadly distinguishes between three tasks of MER, i) classification into emotion categories, ii) emotion as a regression problem, and iii) variation detection throughout a song (Y.-H. Yang & Chen, 2012), we will focus on the second problem of MER as regression since it is seen as more accurate in regards to the represented emotion (Guan et al., 2012). For this, Russell (1980)’s two-dimensional model of emotion as a numerical value of valence (negative to

¹A “community-based formal evaluation framework” (Downie, 2006)

positive) and arousal (passive to active) is the most widely used metric (Huang et al., 2016; Trigeorgis et al., 2016; Wang et al., 2011). For example, anger is high arousal and low valence emotion. This framework is preferred for its simplicity and is used in the present study, although it does not capture the full range of emotions (Cowen et al., 2020), which is why other MER studies (Hu & Downie, 2010a; Hu et al., 2009) use a larger number of affective categories.

2.1 MER as Regression Task (NO)

The first approaches for this regression problem were done by Y.-H. Yang, Lin, Su, and Chen (2008) where they reached a valence-arousal (VA) R^2 score of up to 0.282 and 0.538 with low-level auditory features using a Support Vector Machine Regressor (SVR), similar to the SVR approach presented in Soleymani et al. (2014) that reached a VA score of 0.42 and 0.52 based on supra-segmental features. Later, Bai et al. (2016)’s Random Forest Regressor (RFR) reached a VA score of 0.293 and 0.625 on 548 dimensions of low- and higher-level audio features. Other approaches include a Multiple Linear Regression (MLR) (X. Yang et al., 2018; Vatolkin & Nagathil, 2019), or a Multilayer Perceptron (MLP) (Soleymani et al., 2014; Bhattarai & Lee, 2019).

The rise of fast and efficient deep learning (DL) models across a wide range of applications also impacted MER. One of the first DL bi-modal classification models was presented in (Jeon et al., 2017) using an end-to-end approach for lyrics and raw audio without any feature engineering, fusing the modalities mid-level. Their model outperforms uni-modal approaches at that time with an accuracy of 80%. They also find that lyrics predict the emotion more precisely than audio. A study by Delbouys et al. (2018), investigates the performance of a classical feature engineering approach vs. DL. On the Deezer Mood Detection Dataset (DMDD), they extract low-level auditory features and word embeddings from a list of VA-related words and compare the mid-level (second layer or later) and late-level (weighted outputs) fusion of both of these feature classes against an end-to-end DL approach. They can show that i) mid-level fusion on DL has comparable results to the feature engineering and that ii) the multi-modal mid-level fusion outperforms both uni-modal models (audio vs. textual).

2.2 Lyrics as Prediction Metric (YN)

D. Yang and Lee (2004), who are often seen as the first to combine audio and lyrics in MER modeling, use a Bag-of-Words (BoW) approach combined with the General Inquirer psychological lexical annotation to model lyrics.

Hu et al. (2009); Hu and Downie (2010a, 2010b); Hu et al. (2017); Laurier et al. (2008) expanded lyrics features to include n-gram counts, stylistic features, and ANEW sentiment annotation (Bradley & Lang, 1998) and found that lyrics tend to outperform auditory features on emotion classification and regression tasks, though the best results were achieved by combining the features. As mentioned above, the same pattern holds in Delbouys et al. (2018)’s DL model on word embeddings.

In their recent, Han, Kong, Han, and Wang (2022) report that BoW, TF-IDF, and word embeddings are the most used text representation methods in MER.

2.3 Higher-Level Features (TK)

For the relation of features to the two predicted categories, Gabrielsson and Lindström (2001) and Juthi et al. (2020) describe arousal as being related to tempo (fast/slow), pitch (high/low), loudness level (high/low), and timbre (bright/soft) as well as valence being related to mode (major/minor) and harmony (consonant/dissonant). Vatolkin and Nagathil (2019), however, state the importance of also rhythm for the prediction of valence.

Panda et al. (2021) were the first ones to assess the correlation of 12 high-level Spotify features on the prediction accuracy of valence and arousal. They found that only three (energy, valence and acousticness) were relevant.

2.4 Spotify (TK)

Spotify, with its 422 million active monthly users is the largest music streaming service on the market (Ruby, 2022). It provides easy access to 70+ million songs (Ruby, 2022) as well as an API that makes it possible to extract a customized dataset of annotated musical pieces². The annotation procedure of such is unknown to the public as this is one of Spotify’s major revenue streams (Hucker, 2021). However, the annotation technology seems automated was developed by *The Echo Nest* (acquired by Spotify in 2014) based

²<https://developer.spotify.com/documentation/web-api/>

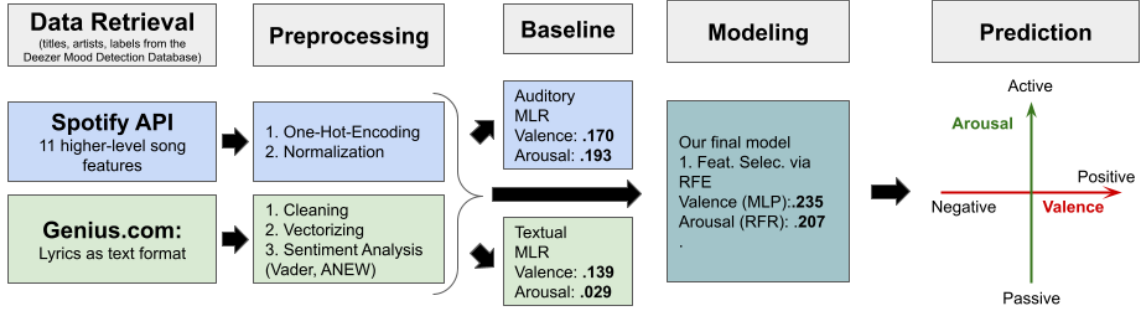


Figure 1: Pipeline of our model. Scores indicate R^2 on the test data.

on expert-annotated data. (*The Echo Nest Blog*, 2017).

3 Experimental Setup

3.1 Data (TK)

The song dataset used in the present study is the Deezer Mood Detection Dataset (DMDD) (Delbouys et al., 2018) which holds VA scores for 18,644 songs and is based on the Million Song Dataset (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011) as well as tags from LastFM³ that are related to mood (Hu & Downie, 2010b). The VA scores were obtained by applying an extended ANEW (XANEW) dataset with 14,000 words and their respective VA scores (Warriner et al., 2013; Bradley & Lang, 1998) to the tags from LastFM.

To represent the auditory modality, we retrieved the auditory features that X. Yang et al. (2018) describe as high-level features for all available songs from the DMDD via the Spotify API⁴, resulting in 13,445 songs with 11 auditory features each. Of these, the features acousticness, danceability, energy, instrumentality, liveness, loudness, speechiness, tempo and valence are continuous, while mode is binary, and key is categorical. We created dummy variables for key resulting in 23 auditory features overall. A detailed explanation of the features can be found in Appendix A. Spotify’s valence annotation is derived differently from our ground-truth valence, avoiding circularity, and is also used as a predictive feature for emotion in Panda et al. (2021).

Lastly, for the textual modality, we scraped the song lyrics for the available songs from genius.com as suggested by Li, Mou, and Chang (2018), further reducing the dataset to a final number of 12,471 songs, of which we used ca. 60% for

training, 20% for validation and 20% for testing.

3.1.1 Extracting Lyrics Features (YN)

To represent the lyrical information, we created three types of features.

1. **Sentiment information** consisting of positive, negative, neutral and compound scores was obtained with Vader sentiment analysis (Hutto & Gilbert, 2014).
2. **TF-IDF features**, which are widely used (Han et al., 2022) and aim to capture the general lexical features of the song texts relating to both valence and arousal. TF-IDF unigram counts were obtained after lemmatizing the lyrics texts and reduced to 100 principal components using PCA.
3. **Extended ANEW (XANEW) features** aimed to capture the lexical basis of valence and arousal. This data set was introduced in 3.1. We generated two count vectors for each pre-processed lyric text and multiplied the counts by the respective VA scores. These count features were also reduced with PCA to 100 dimensions. These features were not included in the final models (see 3.2).

3.2 Feature Selection (TK)

Based on these features, we selected feature subsets for the two different modalities (see Figure 2). Firstly, we tested for significance in an MLR (similar to our baseline, see Section 3.3.1) to get the optimal combination among the auditory features resulting in the five significant predictors danceability, energy, instrumentality, valence, and mode.

To select lyrics features, we calculated validation VA scores for all the combinations of Vader sentiment, XANEW and TF-IDF type features

³ <https://www.last.fm/>

⁴ <https://developer.spotify.com/console/>

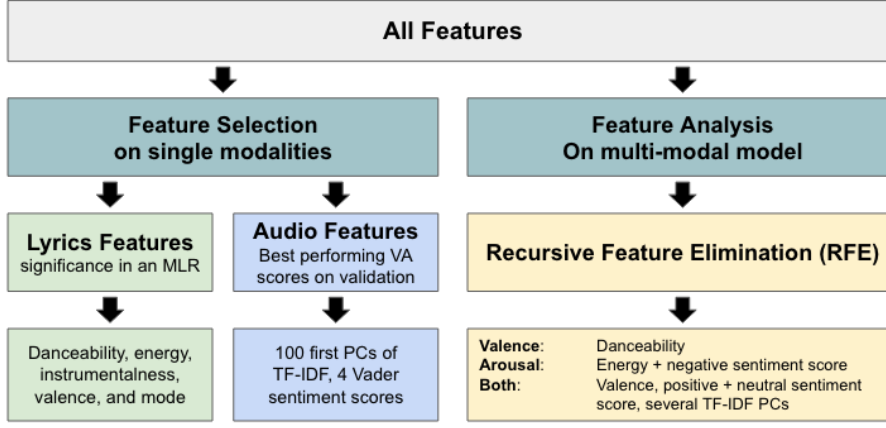


Figure 2: Feature Selection and Analysis Process

(see Appendix B). The combination of Vader and the first 100 PCs of the TF-IDF features had the overall best performance across regression models and was used in our final modelling.

Following Hu and Downie (2010a, 2010b) and especially Panda et al. (2021) for comparability and with the aim of leading to a true multi-modal feature set, we combined all data at feature level resulting in a data matrix $M \in \mathbb{R}^{N \times d}$ with $N = 12,471$ and $d = 23 + 4 + 100 = 127$.

3.3 Models (NO)

Based on the literature presented in Section 2, we compare different regression models to evaluate the performance of uni- and multi-modal approaches for MER on high-level auditory features and lyrics⁵. The regression learning algorithms used are all from the `scikit-learn` package (Pedregosa et al., 2011): `LinearRegression`, `RandomForestRegression`, `SupportVectorRegression`, `MultilayerPerceptronRegressor`, all trained on the optimal parameters we obtained over `GridSearchCV`.

3.3.1 Baseline (TK)

We created a Multiple Linear Regression (MLR) baseline for both uni- and multi-modal prediction of VA scores. For the audio baseline, we firstly we used all auditory features as input. Secondly, we performed the MLR on only statistically significant ($p < 0.05$) coefficients. Then for the lyrics baseline, we performed the MLR on the selected lyrics features (sentiment and TF-IDF). Lastly, we combined the significant audio features and the se-

lected lyrics features in MLR to form the multi-modal baseline.

3.4 Post-Hoc Feature Analysis (YN)

Additionally, for investigating the set of most significant features overall, we applied Recursive Feature Elimination (RFE) (Granitto et al., 2006) using the `sklearn.feature_selection.RFE` package (Pedregosa et al., 2011) on an MLR model. Furthermore, we created a MLR model with all audio features + compound sentiment scores included, for the purpose of analyzing coefficients (see Table 2).

4 Results (All)

We evaluate the predictive power of different regression models in terms of R^2 as applied in the aforementioned literature (e.g. Hu et al. (2008, 2009); Kim et al. (2010)).

4.1 Model Results

Table 4.1 provides a summary of test results for our regression models with different modalities. The highest scores overall are achieved with multi-modal features for both valence and arousal. This is more pronounced for valence: multi-modality significantly improves R^2 compared to both audio- and lyrics-only features. Within the uni-modal models however, audio features are slightly better at predicting valence than lyrics features.

The arousal results are more complicated. Although the highest arousal score is achieved with multi-modal RFR, all the other regression algorithms perform better at arousal prediction with uni-modal (audio) features. This suggests that

⁵ <https://github.com/Tibor-Krols/CogSci2-Spotify>

Mode	Model	Valence	Arousal
Audio	MLR	0.170	0.193
	RFR	0.171	0.204
	SVR	0.165	0.203
	MLP	0.176	0.203
Lyrics	MLR	0.139	0.029
	RFR	0.121	0.027
	SVR	0.042	-0.074
	MLP	0.117	0.020
Multi-modal	MLR	0.236	0.190
	RFR	0.224	0.207
	SVR	0.208	0.154
	MLP	0.235	0.196

Table 1: R^2 test scores for all uni- and multi-modal models based on selected feature subsets as described in 4.2.1.

adding lyrics features may interfere and compound error for arousal, because we also see that the lyrics-only arousal scores are practically 0.

No single regression algorithm emerged as the best, although SVR consistently performed worst. This might be due to insufficient hyperparameter optimization, as testing a larger number of parameter combinations was very time-consuming for this type of model.

Surprisingly, for lyrics-only valence and arousal as well as multi-modal valence, our more complex regression algorithms did not outperform the baseline MLR.

4.2 Feature Analysis

Based on RFE, the most important features were valence, the positive and neutral sentiment score and several PCs from the TF-IDF analysis for both target variables (see Figure 2) as well as danceability for valence and energy and the negative sentiment score for arousal.

Considering the p-values of coefficients in MLR, valence has 7 significant predictors, where arousal has 6 as can be seen in Table 2.

4.2.1 Selected Features Performance

Table 3 compares VA scores of the MLP with all features vs. selected features for each modality. The feature subsets are listed here:

1. $\text{all_features}_A = \{\text{Acousticness, Danceability, Energy, Instrumentalness, Key, Liveness, Loudness, Mode, Speechiness, Tempo, Valence}\}$

Feature	Valence	Arousal
Constant	-1.6885*	-0.9836*
Danceability	0.6915*	-0.3266*
Energy	0.6378*	1.4254*
Loudness	-0.0091	-0.0073
Speechiness	-0.1101	0.3952*
Acousticness	0.1649*	0.0207
Instrumentalness	0.0929	-0.3278*
Liveness	0.1916*	0.0207
Valence	1.0901*	0.5158*
Tempo	0.0005	0.0004
Mode	0.0977*	0.1272*
Compound sentiment	0.2275*	-0.0051

Table 2: Coefficients for MLR. *significant with $p < 0.05$

2. $\text{selected}_A = \{\text{Danceability, Energy, Instrumentalness, Valence, Mode}\}$
3. $\text{all_features}_L = \{200 \text{ PCs ANEW scores, 100 PCs TF-IDF, 4 Vader sentiment scores}\}$
4. $\text{selected}_L = \{100 \text{ PCs of TF-IDF, 4 Vader sentiment scores}\}$

We chose to illustrate these results using the MLP, as it has a consistent performance and takes little time to optimize and train. We see that using selected features gives a small but consistent boost to the R^2 for all modalities, so we trained our final models using these features.

Overall, we can see from Table 3 that all subsets of selected features perform always better in comparison to all the features.

	Feature set	Valence	Arousal
Audio	all_features_A	0.163	0.193
	selected_A	0.176	0.203
Lyrics	all_features_L	0.091	0.009
	selected_L	0.117	0.019
Multi	$\text{all_features}_A + \text{all_features}_L$	0.230	0.193
	$\text{selected}_A + \text{selected}_L$	0.235	0.196

Table 3: Comparison of MLP R^2 -scores for different feature subsets

5 Discussion (All)

Our main results can be briefly summarized in two notions. One, both uni-modal lyrics features and

uni-modal audio features reasonably predict valence, although a multi-modal approach outperforms either modality individually. And two, predicting arousal is hard to do with lyrics features, since audio features alone perform almost as well as the multi-modal approach.

5.1 Effect of Features (TK)

Valence. As seen in Table 2, danceability, energy, valence (Spotify), mode, acousticness, liveness and compound sentiment positively contribute to the valence score, while the first three are by far the strongest predictors.

Arousal. The features energy, speechiness, mode, and valence positively contribute, while danceability and instrumentalness have negative coefficients. Energy is by far the strongest predictor.

Both of these findings are consistent with Panda et al. (2021) in regards to energy, valence and acousticness. However, unlike them, we found that danceability, instrumentalness, mode and speechiness are also relevant for MER. Especially the difference in findings about the importance of danceability is interesting. Where in our baseline model danceability has one of the highest coefficients, Panda et al. (2021) found that danceability has the lowest feature weight.

Danceability captures tempo, rhythm and beat aspects of a song, where perceptual features such as perceived loudness and timbre contribute to energy according to Spotify’s descriptions (see appendix A). Danceability being a strong predictor of valence supports previous findings about the importance of rhythmic properties for predicting valence (Y.-H. Yang, Lin, Su, & Chen, 2008; Vattolkin & Nagathil, 2019). Our finding that energy is a strong predictor of arousal indirectly supports research stating timbre is related to arousal (Gabrielsson & Lindström, 2001; Y.-H. Yang, Lin, Su, & Chen, 2008). As energy also captures perceived loudness, the association between loudness and arousal (Gabrielsson & Lindström, 2001; Y.-H. Yang, Lin, Su, & Chen, 2008) is also supported by our results.

The finding that Spotify’s valence is the strongest predictor of valence is an indication that it is indeed somehow measured by the same underlying construct, despite being constructed in different ways. Furthermore, we find a positive coefficient of mode and predicted valence, which cor-

responds to traditional associations of major keys being related to positive emotions and minor keys to negative ones (Gabrielsson & Lindström, 2001; Panda, Malheiro, & Paiva, 2020).

The positive coefficient of sentiment score for predicting valence could be interpreted as that songs with more positive lyrics have higher valence scores and songs with more negative lyrics have more lower valence scores.

Surprisingly, XANew features were among the least predictive features based on RFE and VA validation scores, which is why they were removed from the final models. The ground-truth VA scores are based on tag analysis with XANew (Delbouys et al., 2018), and studies Hu and Downie (2010a) yielded good results with ANEW features. The main difference between XANew features and VA scores is that the former is based on tags from LastFM, where the latter is based on the lyrics. This suggests that words in lyrics of the songs substantially differ from the tags people give a song on LastFM. TF-IDF may have been more successful in finding distinguishing words because it takes into account the entire lyric vocabulary.

5.2 Multi-modal vs. Uni-modal (YN)

Valence. Our results for modality are generally consistent with the literature ((Y.-H. Yang, Lin, Cheng, et al., 2008; Hu & Downie, 2010a; Delbouys et al., 2018)), where multi-modal MER models outperform uni-modal ones. However, in our case this is only unambiguously true for valence, where lyrics and audio features appear to have complementary effects when combined. The fact that audio seems to contribute more to valence (based on uni-modal scores) is consistent with Ali and Peynircioğlu (2006), Delbouys et al. (2018) and Xue et al. (2015), as well as anecdotal evidence (Heinrich, 2019) that people pay more attention to music than lyrics, which would influence the listener annotations that our target variables are based on. On the other hand, older studies like Hu and Downie (2010a, 2010b) do report that lyrics can outperform audio for valence prediction. However they make use of more extensive feature engineering and more fine-grained emotion categories (predicted in binary classification tasks), which limits comparability.

Arousal. With regards to arousal prediction, the multi-modal advantage is not clear. Only one of our regression models performs better multi-

modally, while the other three perform better with only audio features. Lyrics features seem to either not contribute or actively interfere with audio features to give lower multi-modal performance. This is consistent with [Delbouys et al. \(2018\)](#)’s findings that arousal is mostly determined by audio features such as rhythm and tempo. [Malheiro et al. \(2018\)](#) succeed in classifying songs with regards to both arousal and valence exclusively from lyrics, but with a small dataset and very extensive feature engineering, especially including more stylistic and structural features, which we pre-processed away in our BoW TF-IDF approach. On the other hand, [Delbouys et al. \(2018\)](#), who do not report pre-processing lyrics but feed them directly into a variety of neural networks, achieve comparably low arousal scores for their lyrics-only models on the same dataset.

5.3 Summary and Future Directions (NO)

Lastly, we want to shed light on possible reasons for our model’s performance as well as limitations and give an outlook for future research direction.

Early Feature Fusion. As stated in 3.2, we chose to combine our features at feature level, following [Hu and Downie \(2010a, 2010b\)](#) and [Panda et al. \(2021\)](#). However, [Y.-H. Yang, Lin, Cheng, et al. \(2008\)](#) states that a potential difficulty of early fusion is achieving a truly common representation. Especially the combination of radically different feature types like TF-IDF vectors and auditory features could be responsible for our results. An exploratory investigation of late-fusion approaches (see e.g. [Y.-H. Yang, Lin, Cheng, et al. \(2008\)](#); [Delbouys et al. \(2018\)](#) for this dataset is left for future investigation. Another interesting direction for future research is investigating interaction effects of the features we used.

Deep Learning as State-of-the-Art. Interestingly, we find that our results align very well with the findings of [Delbouys et al. \(2018\)](#)’s comparison of classical feature engineering based approaches and DL models on the same dataset. We reach competitive scores on valence on the audio-only prediction with a MLP (0.176) compared to their convolutional network (0.179) while our multi-modal MLR (0.236) even outperforms their middle-fusion DL network (0.219) on valence. This suggests that using higher-level features makes our model more interpretable, while not compromising on performance.

Vague Annotation Standard. That we obtain similar results as [Delbouys et al. \(2018\)](#) indicates that part of the limitations may be related to the dataset itself. The fact that Spotify’s valence doesn’t correlate to the predicted valence very well ($R^2 = 0.35$) means it also may be hard to measure and evaluate accurately. The VA annotations obtained through LastFM tags in the [Delbouys et al. \(2018\)](#) were not checked or validated with human annotation, so their quality is unknown. Furthermore, the target VA variables assume that there is a single value for one whole song rather than displaying a possible variation of valence and arousal within the piece, while [X. Yang et al. \(2018\)](#) suggest that a song’s sentiment can vary throughout the piece, which falls under the MER subdomain of music emotion variation detection (MEVD) ([Han et al., 2022](#)).

Audio and Predictions. Even though we find some improvement to the overall scores when including the selected Spotify features, other approaches that directly source from the song’s spectral features (e.g. [Hu et al. \(2008\)](#); [Hu and Downie \(2010a\)](#); [Soleymani et al. \(2014\)](#)), seem to achieve better results at least on valence scores of up to 0.42 already in 2014. This leads to the conclusion in alignment with [Panda et al. \(2021\)](#) that there is still some room for improvement and clarity for Spotify features and that valence and arousal are still rather hard to predict from high-level features.

Lyrics and Predictions. Already the first music mood classification task, [MIREX \(2007\)](#), found that MER depends highly on the ground-truth labeling and human judgement, including how much people pay attention to the lyrics. An example for this is ”Hey-Ya” by OutKast, that sounds happy and upbeat (which corresponds to a high valence and arousal score in our dataset), whereas the artists themselves attribute a more negative meaning to it ([Archive-Corey-Moss, 2004](#)). Henceforth, the true, multi-modal valence of such songs is hard to determine, both for human judges as well as predictive models. Aggravating factors to this are the lyrics shortness and the use of rhetorical figures like metaphors, oxymorons, and irony that makes it even harder to capture the lyrics’ true sentiment. We suggest for future direction a more advanced sentiment analysis.

6 Conclusion (YN)

We have shown that we can combine audio and textual modalities to predict valence and arousal for a large dataset of songs, improving in particular valence prediction relative to uni-modal approaches. Predicting arousal from lyrics proved unsuccessful with the features we used and may require more extensive feature engineering or DL approaches. On average, audio-only models were better than multi-modal ones at predicting arousal based on auditory features such as energy, speechiness and mode.

References

- Ali, S. O., & Peynircioğlu, Z. F. (2006, October). Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, 34(4), 511–534. Retrieved 2022-05-30, from <https://doi.org/10.1177/0305735606067168> (Publisher: SAGE Publications Ltd) doi: 10.1177/0305735606067168 6
- Archive-Corey-Moss. (2004). *Road To The Grammys: The Making Of Outkast's 'Hey Ya!'*. Retrieved 2022-05-28, from <http://www.mtv.com/news/1484730/road-to-the-grammys-the-making-of-outkasts-hey-ya/> 7
- Bai, J., Peng, J., Shi, J., Tang, D., Wu, Y., Li, J., & Luo, K. (2016, August). Dimensional music emotion recognition by valence-arousal regression. In *2016 IEEE 15th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)* (pp. 42–49). doi: 10.1109/ICCI-CC.2016.7862063 2
- Balkwill, L.-L., & Thompson, W. F. (1999, October). A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues. *Music Perception*, 17(1), 43–64. Retrieved 2022-05-10, from <https://online.ucpress.edu/mp/article/17/1/43/62058/A-Cross-Cultural-Investigation-of-the-Perception> (Publisher: University of California Press) doi: 10.2307/40285811 1
- Bernatzky, G., Presch, M., Anderson, M., & Panksepp, J. (2011, October). Emotional foundations of music as a non-pharmacological pain management tool in modern medicine. *Neuroscience & Biobehavioral Reviews*, 35(9), 1989–1999. Retrieved 2022-05-10, from <https://www.sciencedirect.com/science/article/pii/S0149763411001151> doi: 10.1016/j.neubiorev.2011.06.005 1
- Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011, January). The Million Song Dataset. In (pp. 591–596). 3
- Bhattarai, B., & Lee, J. (2019, June). Automatic Music Mood Detection Using Transfer Learning and Multilayer Perceptron. *INTERNATIONAL JOURNAL of FUZZY LOGIC and INTELLIGENT SYSTEMS*, 19, 88–96. doi: 10.5391/IJFIS.2019.19.2.88 2
- Bradley, M. M., & Lang, P. J. (1998). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. , 49. 2, 3
- Chen, S.-H., Lee, Y.-S., Hsieh, W.-C., & Wang, J.-C. (2015, December). Music emotion recognition using deep Gaussian process. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 495–498). doi: 10.1109/APSIPA.2015.7415321 1
- Cowen, A., Fang, X., Sauter, D., & Keltner, D. (2020). What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117(4), 1924–1934. Retrieved 2022-05-11, from <https://www.pnas.org/doi/10.1073/pnas.1910704117> doi: 10.1073/pnas.1910704117 2
- Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., & Moussallam, M. (2018, September). Music Mood Detection Based On Audio And Lyrics With Deep Neural Net. *arXiv:1809.07276 [cs, stat]*. Retrieved 2022-05-03, from <http://arxiv.org/abs/1809.07276> (arXiv: 1809.07276) 1, 2, 3, 6, 7
- Deng, J. J., Leung, C. H. C., Milani, A., & Chen, L. (2015, March). Emotional States Associated with Music: Classification, Prediction of Changes, and Consideration in Recommendation. *ACM Transactions on Interactive Intelligent Systems*, 5(1), 1–36. Retrieved 2022-05-10,

- from <https://dl.acm.org/doi/10.1145/2723575> doi: 10.1145/2723575 1
- Dingle, G. A., Kelly, P. J., Flynn, L. M., & Baker, F. A. (2015, September). The influence of music on emotions and cravings in clients in addiction treatment: A study of two clinical samples. *The Arts in Psychotherapy*, 45, 18–25. Retrieved 2022-05-10, from <https://www.sciencedirect.com/science/article/pii/S019745561500043X> doi: 10.1016/j.aip.2015.05.005 1
- Downie, J. S. (2006, December). The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12). Retrieved 2022-05-31, from <http://www.dlib.org/dlib/december06/downie/12downie.html> doi: 10.1045/december2006-downie 1
- The Echo Nest Blog*. (2017, April). Retrieved 2022-05-31, from <https://web.archive.org/web/20170422195736/http://blog.echonest.com/post/66097438564/plotting-musics-emotional-valence-1950-2013> 3
- Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., ... Koelsch, S. (2009, April). Universal Recognition of Three Basic Emotions in Music. *Current Biology*, 19(7), 573–576. Retrieved 2022-05-10, from <https://www.sciencedirect.com/science/article/pii/S0960982209008136> doi: 10.1016/j.cub.2009.02.058 1
- Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression. In *Music and emotion: Theory and research* (pp. 223–248). New York, NY, US: Oxford University Press. 2, 6
- Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. , 8. 4
- Guan, D., Chen, X., & Yang, D. (2012). Music Emotion Regression based on Multi-modal Features. *Proc. International Symposium on Computer Music Modeling and Retrieval*, 70–77. 1
- Han, D., Kong, Y., Han, J., & Wang, G. (2022, December). A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6), 166335. Retrieved 2022-05-18, from <https://link.springer.com/10.1007/s11704-021-0569-4> doi: 10.1007/s11704-021-0569-4 2, 3, 7
- Heinrich, S. (2019, October). 22 "Fun" Songs That Actually Have Some Pretty Dark Lyrics. Retrieved 2022-06-01, from <https://www.buzzfeed.com/shelbyheinrich/catchy-songs-dark-lyrics> (Section: Rewind) 6
- Hu, X., Choi, K., & Downie, J. S. (2017). A framework for evaluating multimodal music mood classification. *Journal of the Association for Information Science and Technology*, 68(2), 273–285. doi: 10.1002/asi.23649 2
- Hu, X., & Downie, J. S. (2010a). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 159–168). Retrieved 2022-05-04, from <https://doi.org/10.1145/1816123.1816146> 1, 2, 4, 6, 7
- Hu, X., & Downie, J. S. (2010b, August). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. *ISMIR*, 619–624. 1, 2, 3, 4, 6, 7
- Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric Text Mining in Music Mood Classification. *Poster Session*, 6. 2, 4
- Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX Audio Mood Classification Task: Lessons Learned. , 6. 1, 4, 7
- Huang, M., Rong, W., Arjannikov, T., Jiang, N., & Xiong, Z. (2016). Bi-Modal Deep Boltzmann Machine Based Musical Emotion Classification. In A. E. Villa, P. Masulli, & A. J. Pons Rivero (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2016* (pp. 199–207). Cham: Springer International Publishing. doi: 10.1007/978-3-319-44781-0_24 2
- Hucker, M. (2021, November). *Uncovering How the Spotify Algorithm Works*. Retrieved 2022-05-31, from <https://towardsdatascience.com/>

- uncovering-how-the-spotify-algorithm-works-4d3c021ebc02
- Hunter, P. G., & Schellenberg, E. G. (2010). Music and Emotion. In M. Riess Jones, R. R. Fay, & A. N. Popper (Eds.), *Music Perception* (pp. 129–164). New York, NY: Springer. Retrieved 2022-05-11, from https://doi.org/10.1007/978-1-4419-6114-3_5 doi: 10.1007/978-1-4419-6114-3_5 1
- Hutto, C., & Gilbert, E. (2014, May). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. Retrieved 2022-05-10, from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550> (Number: 1) 3
- Jeon, B., Kim, C., Kim, A., Kim, D., Park, J., & Ha, J.-W. (2017). Music Emotion Recognition via End-to-End Multimodal Neural Networks. , *RecSys Posters*, 2. 2
- Juthi, J. H., Gomes, A., Bhuiyan, T., & Mahmud, I. (2020). Music Emotion Recognition with the Extraction of Audio Features Using Machine Learning Approaches. In P. K. Singh, B. K. Panigrahi, N. K. Suryadevara, S. K. Sharma, & A. P. Singh (Eds.), *Proceedings of ICETIT 2019* (pp. 318–329). Cham: Springer International Publishing. doi: 10.1007/978-3-030-30577-2_27 2
- Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., ... Turnbull, D. (2010). Music Emotion Recognition: A State Of The Art Review. , 12. 1, 4
- Laurier, C., Grivolla, J., & Herrera, P. (2008, December). Multimodal Music Mood Classification Using Audio and Lyrics. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 688–693). doi: 10.1109/ICMLA.2008.96 2
- Li, S., Mou, C., & Chang, C. (2018). Prediction of Genres and Emotions by Song Lyrics. *Deep Learning, Stanford University, CA*, 6. 3
- Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2018, April). Emotionally-Relevant Features for Classification and Regression of Music Lyrics. *IEEE Transactions on Affective Computing*, 9(2), 240–254. (Conference Name: IEEE Transactions on Affective Computing) doi: 10.1109/TAFFC.2016.2598569 7
- MIREX. (2007). *2007:Main Page - MIREX Wiki*. Retrieved 2022-05-31, from https://www.music-ir.org/mirex/wiki/2007:Main_Page 1, 7
- Panda, R., Malheiro, R. M., & Paiva, R. P. (2020). Audio Features for Music Emotion Recognition: a Survey. *IEEE Transactions on Affective Computing*, 1–1. (Conference Name: IEEE Transactions on Affective Computing) doi: 10.1109/TAFFC.2020.3032373 6
- Panda, R., Redinho, H., Gonçalves, C., Malheiro, R., & Paiva, R. P. (2021, July). How Does the Spotify API Compare to the Music Emotion Recognition State-of-the-Art? In *Proceedings of the 18th Sound and Music Computing Conference (SMC 2021)* (pp. 238–245). Axa sas/SMC Network. Retrieved 2022-05-26, from <https://estudogeral.sib.uc.pt/handle/10316/95161> (Accepted: 2021-07-04T16:31:54Z) doi: 10.5281/zenodo.5045100 1, 2, 3, 4, 6, 7
- Park, S.-H., Ihm, S.-Y., Jang, W.-I., Nasridinov, A., & Park, Y.-H. (2015). A Music Recommendation Method with Emotion Recognition Using Ranked Attributes. In J. J. J. H. Park, I. Stojmenovic, H. Y. Jeong, & G. Yi (Eds.), *Computer Science and its Applications* (pp. 1065–1070). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-45402-2_151 1
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830. Retrieved 2022-05-26, from <http://jmlr.org/papers/v12/pedregosa11a.html> 4
- Ruby, D. (2022, May). *Spotify Stats 2022 — (Facts, Data, Infographics)*. Retrieved 2022-05-31, from <https://www.demandsage.com/spotify-stats/> 2
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. (Place: US Publisher: American Psychological Association)

- ation) doi: 10.1037/h0077714 **1**
- Soleymani, M., Aljanaki, A., Yang, Y.-H., Caro, M. N., Eyben, F., Markov, K., ... Wiering, F. (2014, November). Emotional Analysis of Music: A Comparison of Methods. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 1161–1164). New York, NY, USA: Association for Computing Machinery. Retrieved 2022-05-26, from <https://doi.org/10.1145/2647868.2655019> doi: 10.1145/2647868.2655019 **2, 7**
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016, March). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5200–5204). (ISSN: 2379-190X) doi: 10.1109/ICASSP.2016.7472669 **2**
- Vatolkin, I., & Nagathil, A. (2019). Evaluation of Audio Feature Groups for the Prediction of Arousal and Valence in Music. In N. Bauer, K. Ickstadt, K. Lübke, G. Szepannek, H. Trautmann, & M. Vichi (Eds.), *Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement* (pp. 305–326). Cham: Springer International Publishing. Retrieved 2022-04-26, from https://doi.org/10.1007/978-3-030-25147-5_19 doi: 10.1007/978-3-030-25147-5_19 **2, 6**
- Wang, X., Chen, X., Yang, D., & Wu, Y. (2011). *Music Emotion Classification of Chinese Songs Based on Lyrics Using Tf*idf and Rhyme*. ISMIR. **2**
- Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013, December). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191–1207. Retrieved 2022-05-10, from <https://doi.org/10.3758/s13428-012-0314-x> doi: 10.3758/s13428-012-0314-x **3**
- Xue, H., Xue, L., & Su, F. (2015). Multi-modal Music Mood Classification by Fusion of Audio and Lyrics. In X. He, S. Luo, D. Tao, C. Xu, J. Yang, & M. A. Hasan (Eds.), *MultiMedia Modeling* (pp. 26–37). Cham: Springer International Publishing. doi: 10.1007/978-3-319-14442-9_3 **6**
- Yang, D., & Lee, W.-S. (2004). *Disambiguating Music Emotion Using Software Agents*. **2**
- Yang, X., Dong, Y., & Li, J. (2018, July). Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4), 365–389. Retrieved 2022-05-10, from <https://link.springer.com/article/10.1007/s00530-017-0559-4> (Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Springer Berlin Heidelberg) doi: 10.1007/s00530-017-0559-4 **1, 2, 3, 7**
- Yang, Y.-H., & Chen, H. H. (2012, May). Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 40:1–40:30. Retrieved 2022-05-10, from <https://doi.org/10.1145/2168752.2168754> doi: 10.1145/2168752.2168754 **1**
- Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., & Chen, H. H. (2008). Toward Multi-modal Music Emotion Classification. In Y.-M. R. Huang et al. (Eds.), *Advances in Multimedia Information Processing - PCM 2008* (pp. 70–79). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-89796-5_8 **1, 6, 7**
- Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*. doi: 10.1109/TASL.2007.911513 **2, 6**

A Feature description

Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. $0 \leq x \leq 1$
Danceability	Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
Instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Key	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C/Dx, 2 = D, and so on. If no key was detected, the value is -1. $-1 \leq x \leq 11$
Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
Loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
Mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). $0 \leq x \leq 1$

Table 4: Used features and description taken from the Spotify documentation⁶

B Performance of Lyrics Features

Features	LinReg		RF		SVR		MLP	
	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
tfidf	0.0705	0.0188	0.0631	0.0251	-0.0138	-0.0099	-0.8320	-2.0411
anew	-0.0146	-0.0138	0.0060	-0.0091	-0.0255	-0.0336	0.0121	-0.0026
vader	0.0478	-0.0025	-0.0049	-0.0765	0.0076	-0.0003	0.0536	0.0041
tfidf+anew	-0.0028	-0.0291	0.0427	-7.0790e-06	0.0765	-0.0056	0.0962	0.0167
tfidf+vader	0.0931	0.0159	0.0842	0.0267	0.0817	0.0044	0.0921	0.0216
anew+vader	0.0238	-0.0187	0.0536	-0.0056	0.0284	-0.0223	0.0518	-0.0038
tfidf+anew+vader	0.0304	-0.0251	0.0796	0.0019	0.0890	0.0041	0.0911	0.0089

Table 5: Validation R^2 scores for valence and arousal for different combinations of lyrics features. Tf-idf and Vader sentiment emerged as the best combination of features. The regression models were run on their default parameters.