Moodify - Emotion Detection in Music: A Multimodal Approach

Yonathan Ron Shiri Gilboa

March 2024

Abstract

Music is a powerful medium that can evoke a wide range of emotions in listeners. However, the perception of emotion in music is subjective and can be influenced by various factors, including lyrics, melody, and the listener's personal experiences. In this study, we present a novel approach to detect emotions in music tracks using a combination of audio features, lyrics analysis, and deep learning models. We leverage the Spotify API to gather audio features, employ the Roberta-based language model `SamLowe/roberta-base-go_emotions` for sentiment analysis of lyrics, and utilize both fully connected and convolutional neural network (CNN) models to analyze audio embeddings generated from the OpenL3 model. Additionally, we explore the application of audio embeddings to classic machine learning models to assess their performance . Our dataset comprises songs from Spotify playlists, each representing a specific emotion. This paper details our methodology for dataset creation and presents preliminary results of our emotion detection models. Our research aims to shed light on the question of what aspects of a song contribute most to its emotional impact, whether it be the lyrics, the melody, or other factors.

1 Dataset Creation

To create a diverse and representative dataset for our study, we manually selected Spotify playlists that are indicative of specific emotions, such as sadness, happiness, love, and anger. The playlists and their corresponding emotions are as follows:

For each playlist, we extracted the audio features using the Spotify API, downloaded the audio files using the YouTube API, and retrieved the lyrics using the Lyrics API (`https://api.lyrics.ovh/v1/`). We then processed the lyrics through the `SamLowe/roberta-base-go_emotions` language model to obtain emotion labels. Additionally, we converted the audio files into waveform files using FFmpeg, which were then used as input for the openl3 model, which later was the input for our CNN , fully connected network and the ML models. In our dataset, we gathered a total of 739 unique songs. Out of these, lyrics features were successfully extracted using the language model for 467 songs. To further analyze the audio content, we created a total of 2,217 waveforms by generating three different segments from each unique song.

Emotion	Playlist ID	Playlist Name	Author
sad	6kIofAY27hecJAuDln8t53	saddest songs ever	elliefarroww
sad	6nxPNnmSE0d5WlplUsa5L3	Sad Songs	Spaak Records
sad	3p0pyQmJi6h3xUn25403WH	missing someone who passed away	HAILEY
happy	37i9dQZF1DX84kJlLdo9vT	Happy Days	Spotify
happy	0RH319xCjeU8VyTSqCF6M4	Happy songs everyone knows	skye
happy	37i9dQZF1EIgG2NEOhqsD7	Feel Good Happy Mix	Spotify
in-love	6oNsYDhN95gkENsdFcAwTh	100 BEST LOVE SONGS OF ALL TIME	Juan Ceñal
in-love	37i9dQZF1DX4pAtJteyweQ	Valentine's Day Love	Spotify
angry	0jbaEzUwLTOlIOp42B5pXV	loud songs to listen to when you are angry and mad	rubenrxcket
angry	32jdsAx2HOE3I8cDKXynlK	angry songs for angry days	morinka-us
angry	4Tdgj7NprP4Ou3qzul2WLX	Songs For anger and rage	Heilong the Dragon

Table 1: Spotify Playlists Used for Dataset Creation

1.1 Spotify Features

The Spotify audio features extracted for each track include:

- **Acousticness** (float): A measure from 0.0 to 1.0 indicating the confidence that the track is acoustic.
- **Danceability** (float): A measure from 0.0 to 1.0 indicating how suitable a track is for dancing.
- **Energy** (float): A measure from 0.0 to 1.0 representing the intensity and activity of the track.
- **Instrumentalness** (float): A measure indicating the likelihood that the track contains no vocals.
- **Liveness** (float): A measure indicating the presence of an audience in the recording.
- **Loudness** (float): The overall loudness of the track in decibels (dB).
- **Mode** (integer): Indicates the modality (major or minor) of the track.
- **Speechiness** (float): A measure indicating the presence of spoken words in the track.
- **Tempo** (float): The overall estimated tempo of the track in beats per minute (BPM).

1.2 Language Model: SamLowe/roberta-base-go_emotions

The `SamLowe/roberta-base-go_emotions` language model is based on Reddit data and is designed for multi-label classification, with 28 emotion labels including 'sadness', 'neutral', 'disappointment', 'anger', 'joy', 'love', and others. For any given input text, the model outputs a probability for each label, typically using a threshold of 0.5 to determine the presence of an emotion.

This model utilizes the RoBERTa (Robustly Optimized BERT Pretraining Approach) tokenized model, which is an optimized version of the BERT (Bidirectional Encoder Representations from Transformers) model. RoBERTa improves on BERT's pretraining methodology by using dynamic masking, larger batch sizes, and longer training times. The tokenized nature of the model allows for efficient processing of text data, making it suitable for sentiment analysis and emotion detection tasks. By leveraging the RoBERTa framework, the `SamLowe/roberta-base-go_emotions` model is able to capture the nuances of emotional expression in text, providing a powerful tool for analyzing the sentiment of song lyrics.

1.3 Audio Embeddings: OpenL3 Model

OpenL3 is an audio embedding model based on the Look, Listen, and Learn (L3) approach, which leverages deep learning techniques to learn a joint audio-visual representation of the world. The original L3 model was trained using a large dataset of audio-visual pairs, with the aim of learning audio features that are predictive of visual content and vice versa. This training approach encourages the model to capture high-level semantic information from audio signals.

The OpenL3 model is a variant of the L3 model that focuses solely on the audio modality. It uses a convolutional neural network (CNN) architecture to process audio input and produce a fixed-size embedding vector. The CNN consists of multiple convolutional layers followed by pooling layers, which progressively extract and condense information from the audio signal.

In our study, we utilized OpenL3 to extract 512-dimensional embeddings from 20-second audio clips. These embeddings capture a rich set of audio features, including timbral, harmonic, and rhythmic characteristics, which are crucial for tasks like emotion detection in music. By leveraging the deep learning-based approach of OpenL3, we were able to obtain a compact yet informative representation of each audio clip, facilitating the classification of songs into emotional categories.

2 Methodology

2.1 Classic Machine Learning Models

In our initial approach to emotion detection in music, we explored the use of classic machine learning models to classify songs based on their Spotify audio features. We trained and evaluated several models, including Decision Tree, K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Logistic Regression, and Gradient Boosting and more...

Each model was trained on a dataset consisting of songs labeled with their corresponding emotions, using features such as danceability, energy, acoustic-

ness, and tempo extracted from the Spotify API. We performed cross-validation to assess the accuracy of each model.

The results showed that the Gradient Boosting model, specifically using the Random Forest algorithm, achieved the best performance among the models tested. The accuracy of the Gradient Boosting model was measured at approximately 68.29%, indicating its effectiveness in classifying emotions based on Spotify audio features.

These findings suggest that classic machine learning models can provide a viable baseline for emotion detection in music. However, further exploration is needed to improve accuracy and explore the potential of more advanced techniques.

2.2 Lyrics Features Analysis

In addition to the audio features, we explored the impact of lyrics on emotion detection. Lyrics can be a powerful indicator of a song's emotional content, as they often convey the songwriter's emotions and intentions. To analyze the lyrics, we employed the Roberta-based language model `SamLowe/roberta-base-go_emotions`, which is designed for multi-label classification of emotions.

We extracted emotion-related features from the lyrics of each song and used these features as input for the same set of classic machine learning models used in the previous analysis. The goal was to evaluate whether the lyrics alone could predict the emotion of a song effectively.

The average accuracy across all models was approximately 43.85%. The best-performing model was the Support Vector Machine, with an accuracy of 50.00%. These findings suggest that while lyrics are an important aspect of a song's emotional expression, they may not be sufficient on their own to accurately predict the emotion of a song. Combining lyrics features with audio features could potentially yield better results.

2.3 Combination of Audio Features and Lyrics

Building on the insights from our previous experiments, we hypothesized that a combination of Spotify audio features and lyrics features would provide a more comprehensive representation of a song's emotional content. For instance, the rhythm and tempo of a happy song might be captured well by audio features, while the emotional depth of a sad or love song might be more accurately reflected in its lyrics.

To test this hypothesis, we created a combined dataset that included both the audio features extracted from the Spotify API and the emotion-related features obtained from the lyrics using the `SamLowe/roberta-base-go_emotions` language model. We then applied the same set of machine learning models to this combined dataset.

The results were promising, with a noticeable improvement in accuracy compared to using either audio features or lyrics features alone. The best-performing model was the Cross Gradient Booster, which achieved an accuracy of approximately 74.39%. Logistic Regression also performed well, with an accuracy of 71.95%.

The classification report for the Cross Gradient Booster model is as follows:

	Precision	Recall	F1-score	Support
Happy	0.71	0.71	0.71	14
Sad	0.71	0.86	0.77	28
Angry	0.84	0.80	0.82	20
In-love	0.73	0.55	0.63	20
Accuracy			0.74	82
Macro avg	0.75	0.73	0.73	82
Weighted avg	0.75	0.74	0.74	82

Table 2: Classification report for Cross Gradient Booster model

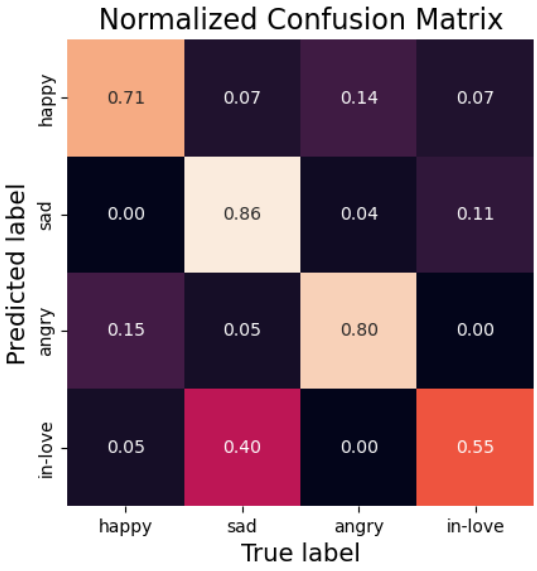

Figure 1: Confusion matrix for Cross Gradient Booster model

These results suggest that the combination of audio features and lyrics features provides a more holistic representation of a song's emotional content, leading to improved accuracy in emotion detection. This finding underscores the importance of considering multiple aspects of a song when analyzing its emotional impact.

2.4 Audio Embeddings and Neural Network Models

In our final experiment, we explored the use of audio embeddings as input to various neural network models for emotion detection. Audio embeddings provide a compact representation of the audio content, capturing its essential characteristics. We utilized the OpenL3 model, a deep learning-based audio embedding model, to extract 512-dimensional embeddings from 20-second audio clips taken from three different segments of each song.

We then fed these embeddings into two types of neural network models: a convolutional neural network (CNN) and a fully connected model. The CNN consisted of three convolutional layers followed by max-pooling layers, a flatten-

ing layer, and two dense layers, with the final layer using a softmax activation function for classification. The fully connected model, on the other hand, took the mean of the embeddings as input and consisted of two dense layers with softmax activation for classification.

Additionally, we applied the audio embeddings to classic machine learning models to assess their performance in emotion detection. The Cross Gradient Booster model achieved an accuracy of 75.16%, which was higher than both the CNN and fully connected approaches.

The models were trained to classify songs into four emotional categories: happy, sad, angry, and in-love. The confusion matrices for the models are shown below:

Fully Connected Model:

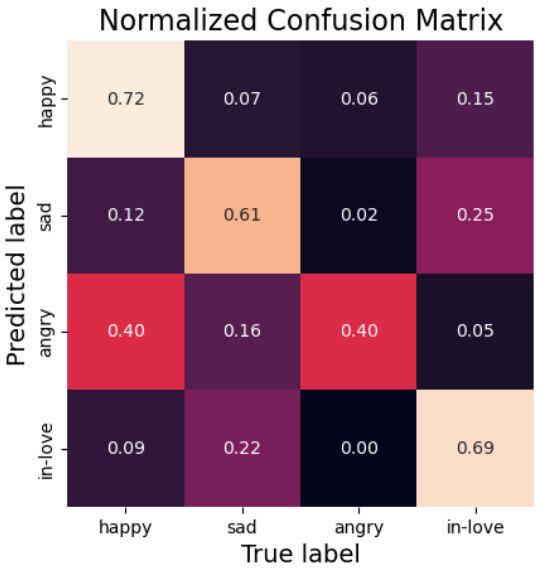

Figure 2: Confusion matrix for the fully connected model.

	Precision	Recall	F1-score	Support
Happy	0.53	0.72	0.61	81
Sad	0.60	0.61	0.61	93
Angry	0.82	0.40	0.53	81
In-love	0.61	0.69	0.65	87
Accuracy			0.61	342
Macro avg	0.64	0.60	0.60	342
Weighted avg	0.64	0.61	0.60	342

Table 3: Classification report for Fully Connected Model

CNN Model:

Classic ML Model with Audio Embeddings:

Although the results showed some improvement in emotion detection accuracy, the best performance was achieved with the classic ML model applied

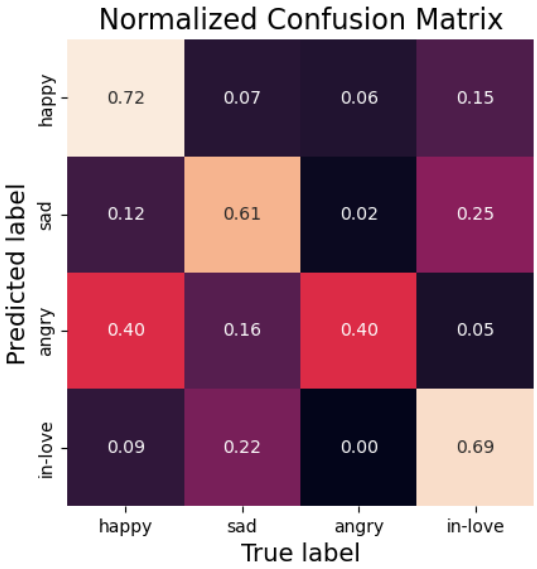

Figure 3: Confusion matrix for the CNN model.

	Precision	Recall	F1-score	Support
Happy	0.47	0.82	0.60	83
Sad	0.64	0.60	0.62	91
Angry	0.76	0.41	0.53	78
In-love	0.70	0.54	0.61	89
Accuracy			0.60	341
Macro avg	0.64	0.59	0.59	341
Weighted avg	0.64	0.60	0.59	341

Table 4: Classification report for CNN Model

to audio embeddings. This suggests that while audio embeddings and neural network models are powerful tools for audio analysis, classic ML models can also effectively leverage the rich information contained in audio embeddings for emotion detection.

3 Conclusion

[h!] In this paper, we presented a comprehensive study aimed to classifying songs based on the following emotions: Happy, Sad, Angry, and In-Love. Our methodology involved gathering a unique dataset comprising Spotify audio features generated by Spotify's algorithms, emotional tags from a language model based on Roberta using the songs lyrics, and feature embeddings extracted from the OpenL3 model. We employed a variety of classic machine learning models, including those from the Scikit-learn and XGBoost frameworks, to evaluate the effectiveness of our approach through precision, recall, and F1 score metrics.

Our project encountered significant challenges, particularly in acquiring high-

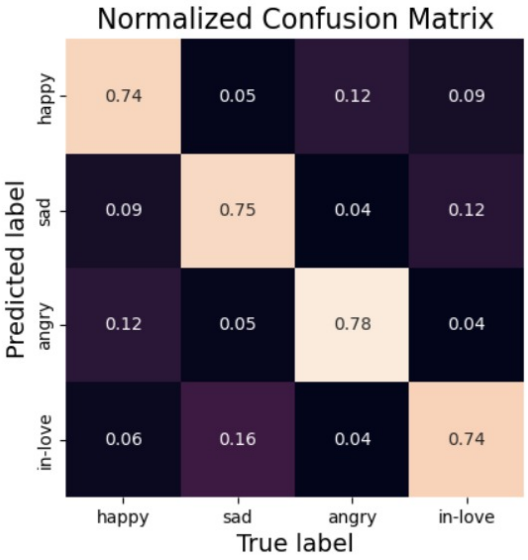

Figure 4: Confusion matrix for the classic ML model with audio embeddings using Cross Gradient Booster.

quality data. One of the challenges we faced in our project was the reliance on playlist names as a source for labeled emotions, which introduced some uncertainty in the emotional classification. The subjective nature of how individuals perceive emotions in music adds to this complexity. While using playlist names was a practical approach for gathering a large dataset, it's important to note that emotional classification in music is inherently subjective, and different listeners may interpret the emotions of the same song differently.

The integration of features from disparate sources—Spotify, Roberta, and OpenL3—proved to be a promising approach for enhancing classification performance. Notably, XGBoost emerged as the most effective model across most of our experiments, demonstrating the potential of ensemble learning methods in handling complex, feature-rich datasets.

Looking forward, we see several opportunities for further research. First, we could explore different types of neural networks to see if they can improve our ability to classify emotions in songs by its embeddings. This is especially promising since our previous use of XGBoost with feature embeddings showed good results. Also, combining features from Spotify and Roberta with OpenL3 embeddings in a more advanced neural network could help us make even better classifications. Another important area for future work is improving how we label the emotions of the songs. We plan to start an initiative to get better quality labels, where groups of people will listen to songs and tag them with one of the four emotions we're focusing on. We'll then use the labels that most people agree on.

Our work lays the groundwork for future investigations into the nuanced interplay between music and emotion, highlighting the importance of leveraging multimodal data and advanced machine learning techniques to decode the emotional content of songs.

4 References

References

[1] Cramer, J., Wu, H.-H., Salamon, J., and Bello, J. P. (2019). *Look, Listen, and Learn More: Design Choices for Deep Audio Embeddings*. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3852–3856.

[2] Dang, T., Shirai, K., and Khoa, N. (2009). *Music Emotion Recognition Using Support Vector Machines*. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, pages 1-4.

[3] Schindler, A. (2020). *Predicting the Music Mood of a Song with Deep Learning*. Towards Data Science. Available at: `https://towardsdatascience.com/predicting-the-music-mood-of-a-song-with-deep-learning-c3ac2b45229e`

[4] Lowe, S. (2021). *roberta-base-go_emotions*. Hugging Face. Available at: `https://huggingface.co/SamLowe/roberta-base-go_emotions`

[5] Spotify. (2021). *Spotify Web API*. Available at: `https://developer.spotify.com/documentation/web-api`

[6] Gomez-Cano, L. and Morisio, M. (2021). *Music and Emotions: A Bibliometric Analysis of the Scientific Literature*. Frontiers in Psychology. Available at: `https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.760060/full`