605.744: Information Retrieval
Emotion Extraction From Lyrics

Sabbir Ahmed

December 12, 2022

Contents

Abstract

Mood classification in music has become more prevalent with the growing streaming industry. Categorizing songs based on the perceived emotions allow music streaming services to improve their recommendation systems and automatic playlist generation. Services such as Spotify use audio features such as duration, energy, tempo, etc. and other combinations of audio features such as danceability and instrumentalness to group similar tracks. These classifications focus solely on the audio production of the songs while ignoring the lyrical content. Emotion extraction from text can be a difficult task, due to the subjectivity in quantifying or discretely categorizing emotions. In this project, 2 of the popular models of emotions, the Plutchik's Wheel and the VAD model, have been used to attempt emotion extraction. The tracks were reclassified using their emotion intensities extracted from their lyrical content.

1 Introduction

This paper discusses the methods used to extract emotion intensities from lyrics and using them to reclassify randomly selected songs into emotion-specific playlists.

2 Background

Most popular streaming services contain manually curated playlists of songs based on their perceived emotions they invoke on the listeners. For example, Spotify has many official mood specific playlists. The apparent moods they invoke range from highly energetic or angry songs to play at the gym to soft or sad songs to play when it is raining. These playlists are constantly updated to incorporate new releases, but are done so manually. This manual curation can be heavily influences from staff bias and subjectivity.

On Spotify, automatic playlist creation is achieved via audio features such as duration, energy, tempo, etc. and other combinations of audio features such as danceability and instrumentalness [1]. These features are evaluated from audio analyses of the tracks which do not include contexts from their lyrical contents.

Lyrics can be just as relevant when classifying songs into moods. For example, the production of *Pumped Up Kicks* by the indie pop/neo-psychedelia group Foster the People consists of pop chords, upbeat melodies and a relatively cheerful tone. However, the song describes the homicidal fantasies of a young boy named Richard who has gained access to his father's gun.

He found a six-shooter gun
In his dad's closet, and with a box of fun things
I don't even know what
But he's coming for you, yeah, he's coming for you
All the other kids with the pumped up kicks
You better run, better run outrun my gun
All the other kids with the pumped up kicks
You better run, better run faster than my bullet

Listing 1: Excerpt of *Pumped Up Kicks*

Relying solely on the audio features of this song will classify it as a joyous or calming song, with relatively positive sentiment.

2.1 Natural Language Processing on Lyrics

Natural language processing on English song lyrics assume additional restrictions. There are no standards set for preprocessing lyrics, but the following lists the exceptions address for this project:

- Songs may be entirely composed of stopwords.
- Repetition is considered significant.
- Lyrics may contain songwriting directions, such as "[gang vocals]", "[instrumental]", "[hook]", "[Speaker A]", etc.
- Lyrics may contain adlibs. For this project, adlibs are not considered part of the dictionary.

2.2 Sentiment Analysis

Sentiment scores are widely used to categorize documents as invoking positive or negative perceptions. However, the scores are limited as they are one-dimensional, ranging between [-1, 1]. This score lacks the depth required to distinguish negative-sentiment emotions such as anger from sadness or positive-sentiment emotions such as calm and excited.

2.3 Emotion Classification

The field of emotion classification is highly subjective. It is often difficult to quantify with absolute values or discretely classify emotions. There have been several psychological models developed over the decades. This project utilizes 2 of the popular models, the Plutchik's Wheel of Emotions and the VAD Emotional State Model.

2.3.1 Plutchik's Wheel of Emotions Model

Psychologist Robert Plutchik proposed a model composed of eight primary emotions: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy [2]. The model, pictured in Figure 1, consists of several concentric circles, with the outer circles being combinations of emotions from the inner circles.

Several categories described by the model do not translate well over text. For example, a human reviewer may find it difficult to extract emotions of *trust* or *anticipation* from a document without explicit usages of synonyms of such emotions.

2.3.2 VAD Emotional State Model

The VAD (Valence-Arousal-Dominance) Emotional State Model was proposed by psychologist Albert Mehrabian. The model plots emotional states across these 3 dimensions of emotion. Valence measures how pleasant or unpleasant an emotion is, arousal determines the energy of the emotion, and dominance refers to the sense of control over the particular emotion. The model, pictured in 2, implies a more granular approach to categorizing emotions.

The third dimension of the model can be disregarded to describe the more popular Valence-Arousal Emotional State Model (also known as the Circumplex Model), developed by psychologist James A. Russell. The two dimensions of this model allows for emotions to be categorized into quadrants which are sufficient in determining the general sentiment of the emotion. The four quadrants of the model, pictured in 3, can be labeled as:

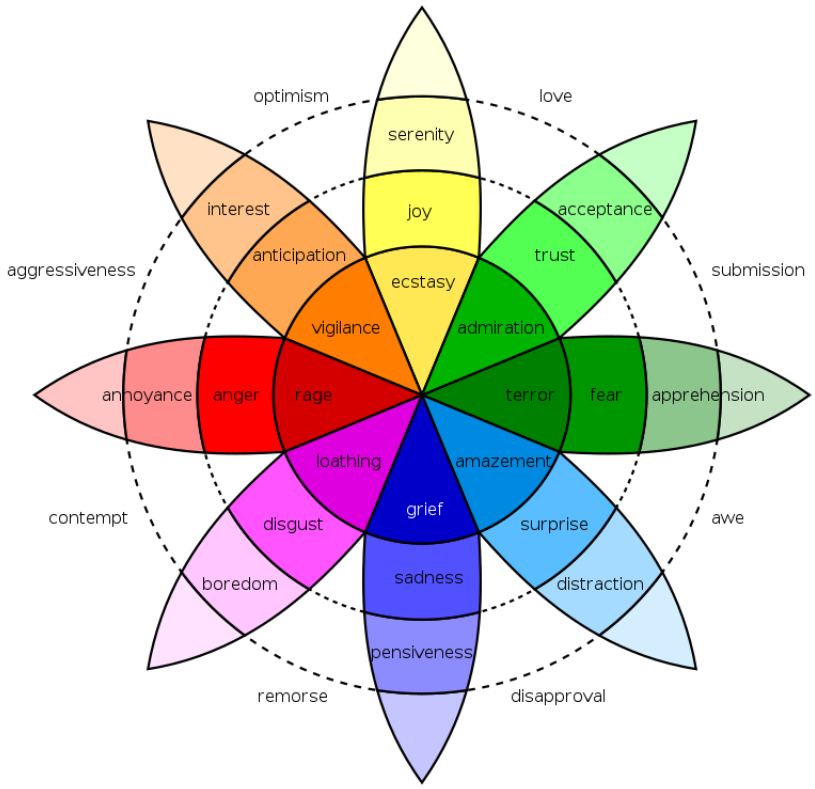

Figure 1: Plutchik's Wheel of Emotions Model

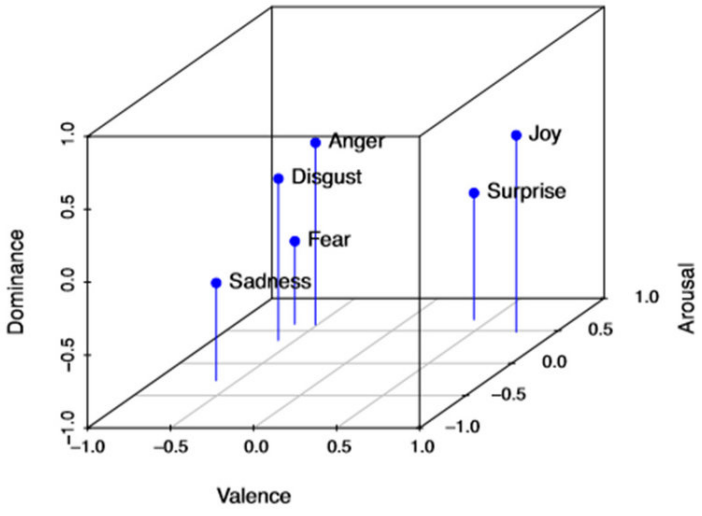

Figure 2: VAD Emotional State Model

- **Quadrant I**: High-arousal/positive-valence, "joy"
- **Quadrant II**: High-arousal/negative-valence, "anger"

- **Quadrant III**: Low-arousal/negative-valence, "sadness"
- **Quadrant IV**: Low-arousal/positive-valence, "calm"

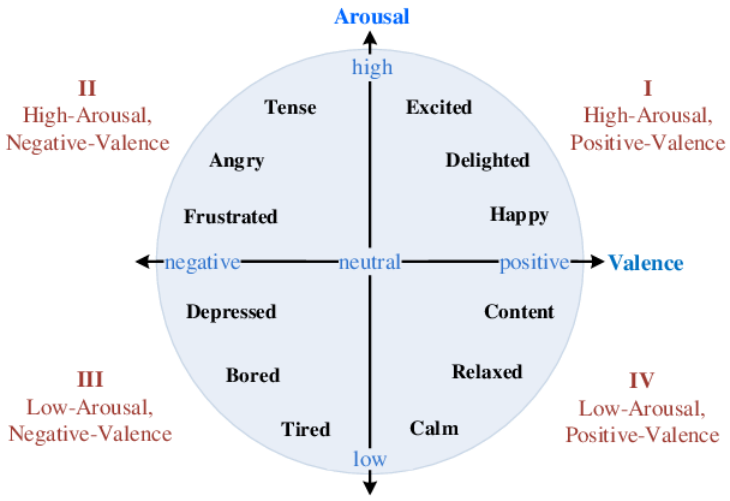

Figure 3: VA Emotional State Model

3 Source Datasets

The datasets for this project were obtained from various sources.

3.1 Playlist Datasets

Playlist datasets were obtained from Spotify via their web application programming interface (API) [1]. Top playlists curated by the Spotify staff were randomly selected, and the API was used to gather playlist names, lists of tracks, and the apparent targeted moods.

3.2 Lyrics

The lyrics of all of the tracks from the playlist datasets were obtained via scraping from various web sources. In total, 748 documents were obtained for analysis. The lyrics are tokenized as 1-grams and therefore lose context. The lyrics are tokenized, stemmed and preprocessed in 3 different ways:

1. With the frequencies preserved, represented as bags-of-words
2. With the frequencies normalized, represented as dictionaries with equal weights
3. With the frequencies normalized and stopwords removed, represented as dictionaries with equal weights

3.3 NRC Emotion Lexicons

The word associations, emotion intensities and the VAD lexicons for the emotion models were obtained from the National Research Council Canada [3, 4, 5]. The dictionaries for the positive and negative terms required to compute the sentiment scores were also extracted from these datasets.

4 Scores

The scores generated are based on sums of the emotion intensities and frequencies of word associations.

4.1 Emotion Intensity

Emotion intensity scores are generated for the 8 emotions identified by the Plutchik's Wheel Model. The scores are simply the sums of the corresponding emotion intensity values present in the NRC datasets. To normalize the intensity scores, each of the 8 emotions also have their corresponding intensity ratio scores, which are calculated by computing the ratio of the emotion intensities over the total sum of intensities. Equation 1 defines the ratio for all emotions e in the 8 emotions identified by the Plutchik's Wheel Model, E_{wheel}.

$$e_{ratio} = \frac{\sum_{i=0} e_i}{\sum_{e \in E_{wheel}} \sum_{i=0} e_i} \tag{1}$$

4.2 VAD Scores

Similar to the emotion intensity scores, the valence, arousal and dominance scores were evaluated by summing their intensities and averaging them over the number of terms identified in the word associations. Since every term in the dataset are associated with a 3-tuple of valence, arousal and dominance values, their sums are averaged by the same value. Equation 2 defines the ratios.

$$v = \frac{\sum_{i=0}^{N} v_i}{N}, a = \frac{\sum_{i=0}^{N} a_i}{N}, d = \frac{\sum_{i=0}^{N} d_i}{N} \tag{2}$$

4.3 Sentiment Scores

The sentiment scores are instead evaluated as the ratio of the positive terms (p) and negative terms (n) over the total number of terms. Terms are labeled positive or negative if they are found in their corresponding datasets.

$$S = \frac{|p| - |n|}{N} \tag{3}$$

If a document contains more negative terms, then the sentiment score will naturally be negative as well, and vice versa for a positive score.

5 Exploratory Analysis

5.1 Emotion Intensities

Table 1 show the mean of the distributions of the emotion intensities over the different types of normalization. The distributions of the intensities over the unnormalized lyrics and the normalized lyrics with the stopwords removed appear to be similar.

Table 2 shows the distributions of the 8 emotion ratios.

Table 1: Means of Emotion Intensities Over Different Normalization Processes

Emotion	No normalization	Frequency normalization	Frequency normalization with stopping
anger	1.436	3.609	1.888
anticipation	1.488	3.320	1.002
disgust	0.950	2.336	1.354
fear	1.626	3.437	1.825
joy	2.131	5.796	2.876
sadness	1.567	3.489	1.486
surprise	0.717	1.559	0.624
trust	2.145	6.043	2.199

Table 2: Distribution of Emotion Ratios

Emotion	mean	median	max
anger_ratio	0.110	0.106	0.676
anticipation_ratio	0.147	0.132	0.624
disgust_ratio	0.073	0.064	0.287
fear_ratio	0.131	0.123	0.506
joy_ratio	0.226	0.204	0.806
sadness_ratio	0.139	0.131	0.687
surprise_ratio	0.064	0.060	0.256
trust_ratio	0.221	0.207	0.706

The aforementioned song, *Pumped Up Kicks*, was also evaluated for its emotion ratios. Table 3 shows the prevalent emotions found from the lyrics of the song.

Table 3: Prevelant Emotions Extracted from *Pumped Up Kicks*

Emotion Ratio	Score
fear	0.378
anger	0.306

5.2 Valence, Arousal, and Dominance

All of the tracks had their valence, arousal, and dominance scores evaluated and displayed over the scatter plot in Figure 4.

The data points appear to cluster on the valence-arousal plane, with the exception of some outliers in the dominance axis. When translated into a valence-arousal graph, pictured in Figure 5, the data points appear to cluster in Quadrant IV. There does not appear to be any correlations among the playlists.

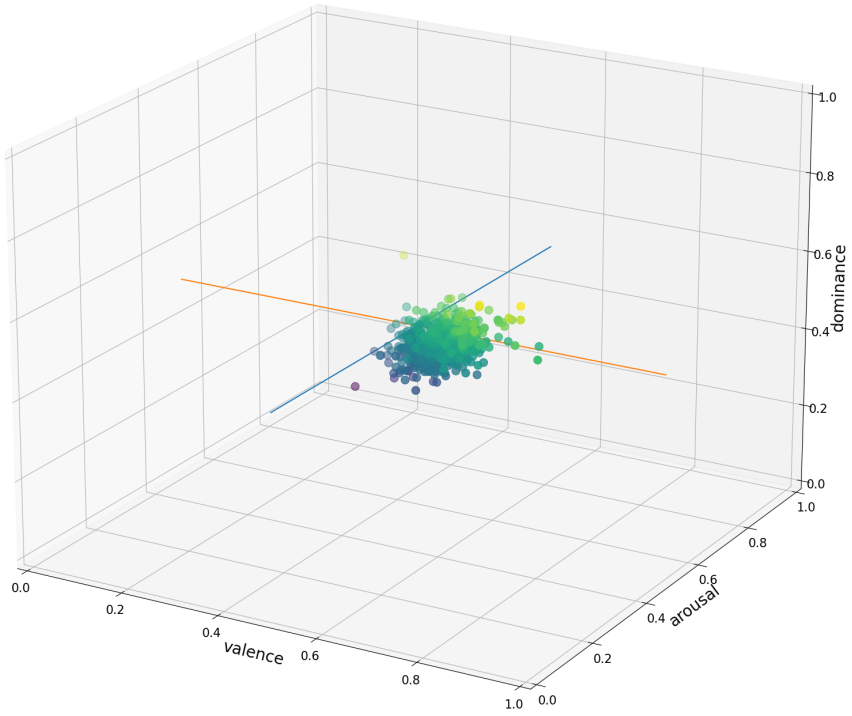

Figure 4: Valence, Arousal, and Dominance of All Playlists

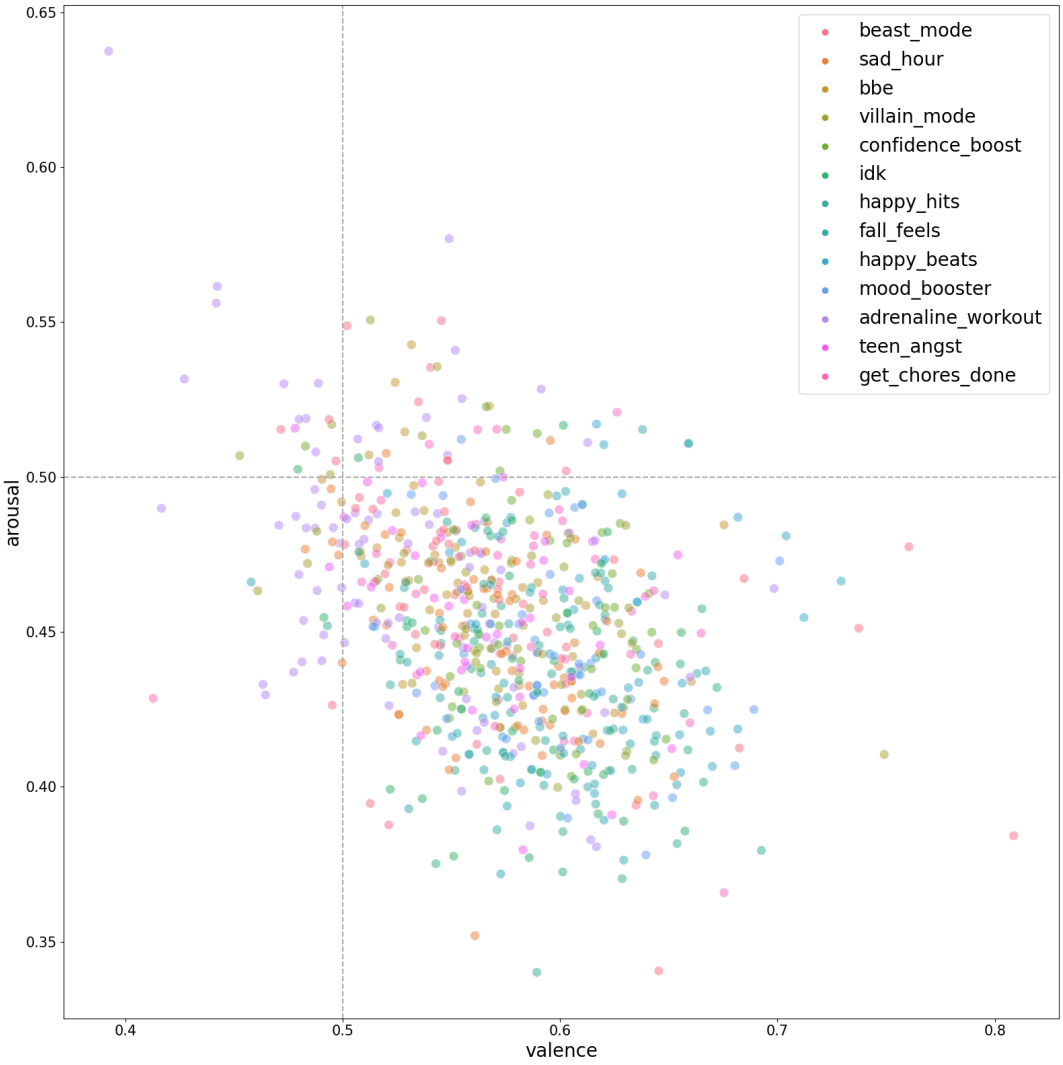

Figure 5: Valence and Arousal of All Playlists

6 Generated Playlist

The emotion ratio values were used to classify the tracks into generated playlists, Emotion Playlist and Quadrant Playlist.

6.1 Emotion Playlist

Emotion Playlists were generated by grouping tracks if the corresponding emotion ratios exceeded a threshold. The threshold was determined by evaluating a distribution measure, such as the median or mean. The best distribution measure was determined by comparing the accuracies. Equation 4 defines the formula used to calculate accuracy.

$$acc = \frac{\sum_{e^+ \in E^+} \frac{\sum_i [e_i^+ > 0]}{|e^+|} + \sum_{e^- \in E^-} \frac{\sum_i [e_i^- < 0]}{|e^-|}}{|E|} \tag{4}$$

Table 4 shows the accuracies and losses evaluated with the different distribution measures. The loss (Equation 5) is calculated as the ratio of uncategorized tracks over the total number of tracks.

$$loss = \frac{t_{uncat}}{t_{cat} + t_{uncat}} \tag{5}$$

Table 4: Accuracies of Emotion Playlist with Different Measures

Metric	Loss	Accuracy
mean	0.13%	0.673
25% quantile	0.13%	0.544
50% quantile	0.13%	0.643
75% quantile	2.27%	0.748

The best accuracy was achieved with the 75% quantile. The tracks were reclassified into the corresponding Emotion Playlists using this measure. Their valence-arousal scores were plotted over the scatter plot pictured in Figure 6.

Some clustering can be observed with the Emotion Playlists. However, they do not seem to be discrete enough to be considered concrete.

The distribution of the emotion ratios were plotted against the Emotion Playlists, pictured in Figure 7

6.2 Quadrant Playlist

The Quadrant Playlists were generated by grouping pairs of Plutchik's 8 emotions into the quadrants of the valence-arousal graph. The following emotion quadrants were paired:

- **Quadrant I**: "joy" and "surprise"
- **Quadrant II**: "anger" and "disgust"
- **Quadrant III**: "sadness" and "fear"
- **Quadrant IV**: "trust" and "anticipation"

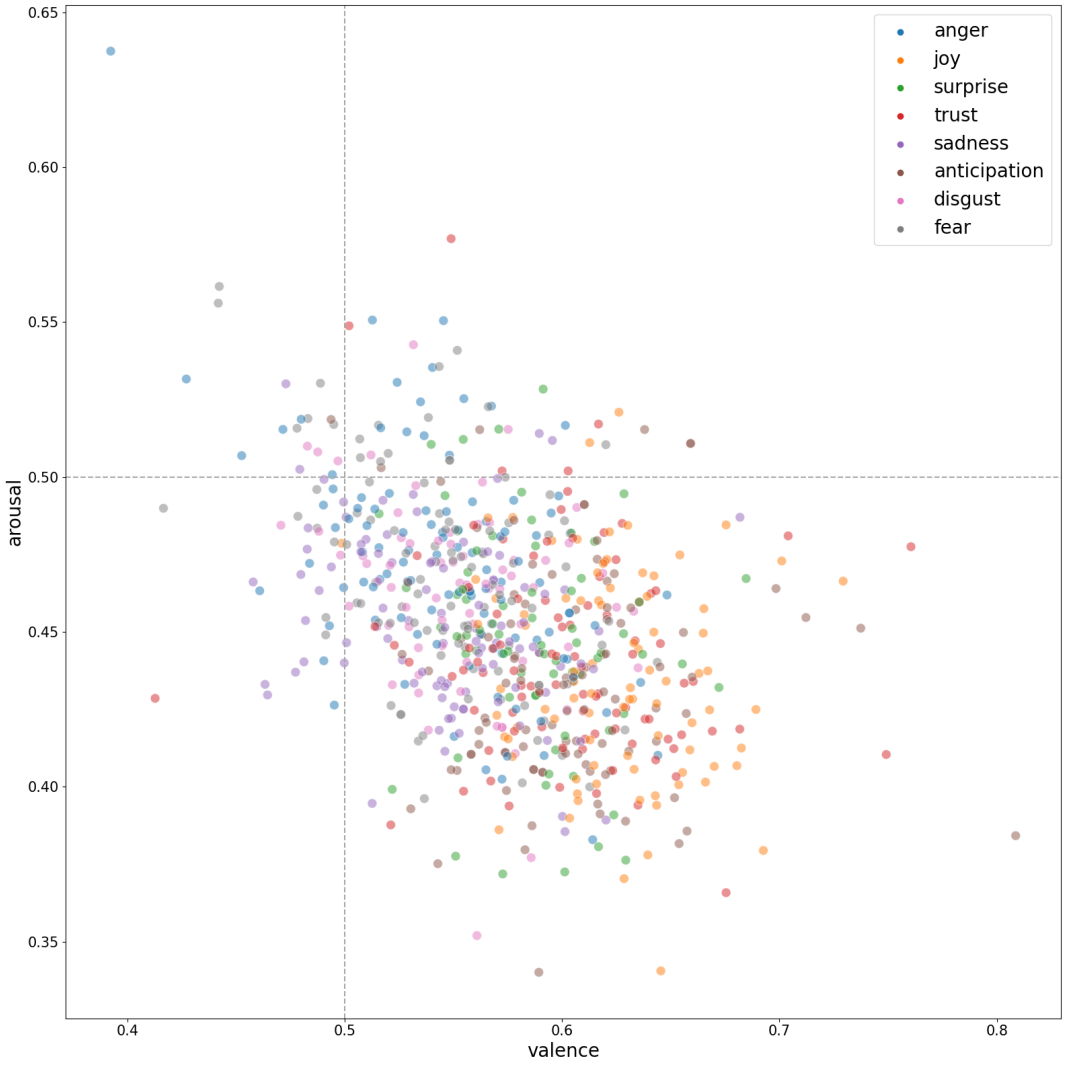

Figure 6: Valence and Arousal of Emotion Playlists

With these combinations, a track's quadrant scores were calculated using Equation 6.

$$e_q = \frac{e_{1,ratio} + e_{2,ratio}}{2} \tag{6}$$

Similar to the Emotion Playlists, the tracks were reclassified into Quadrant Playlists if their quadrant scores exceeded a threshold determined by the best accuracy with different distribution measures. The accuracy was calculated using Equation 7.

$$acc = \frac{\sum_{q^+ \in Q^+} \frac{\sum_i [q_i^+ > 0]}{|q^+|} + \sum_{q^- \in Q^-} \frac{\sum_i [q_i^- < 0]}{|q^-|}}{|Q|} \tag{7}$$

Table 5 shows the accuracies and losses evaluated with the different distribution measures.

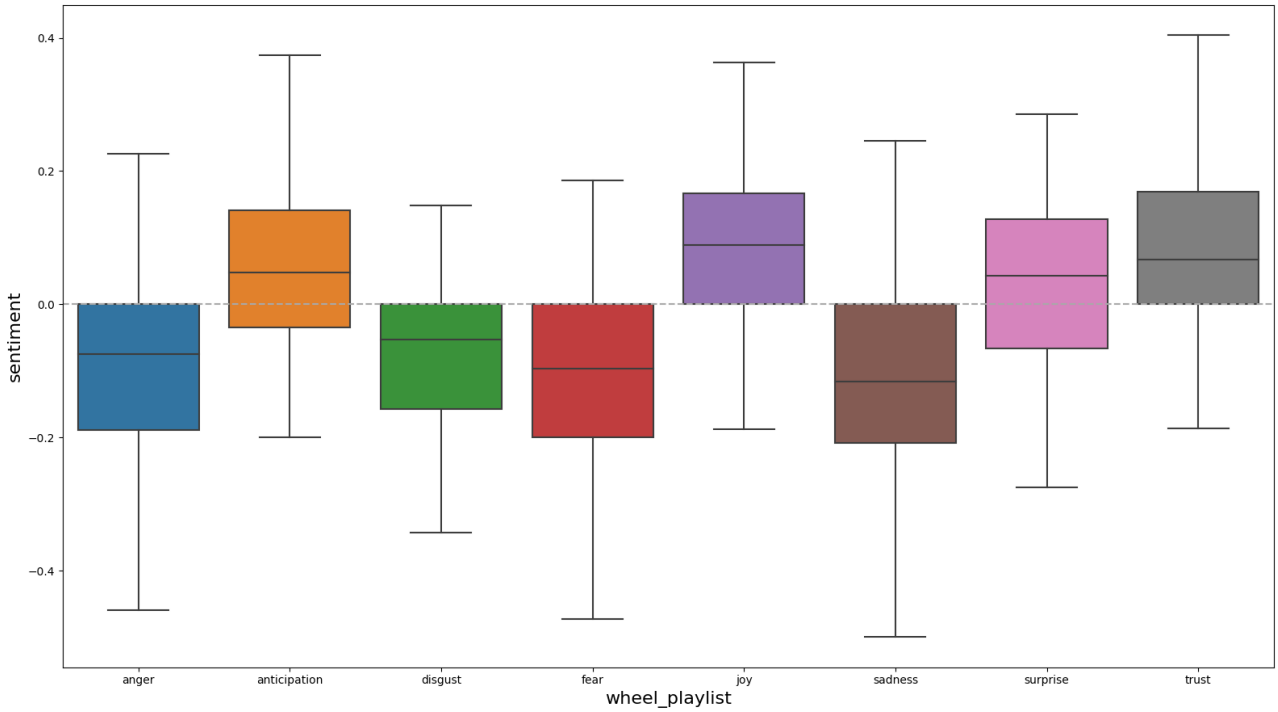

Figure 7: Distribution of Sentiment Scores Over The 8 Emotions

Table 5: Accuracies of Quadrant Playlist with Different Measures

Metric	Loss	Accuracy
mean	14.44%	0.713
25% quantile	16.31%	0.572
50% quantile	14.44%	0.719
75% quantile	18.45%	0.820

The best accuracy was achieved with the 75% quantile. The tracks were reclassified into the corresponding Quadrant Playlists using this measure. The distribution of the emotion ratios were plotted against the Emotion Playlists, pictured in Figure 8.

7 Conclusion

The tracks retrieved from Spotify playlists were automatically reclassified using emotions extracted from their lyrics. Using the emotion intensities, valence-arousal scores and sentiment scores, new playlists based on their prevalent emotions were created with decent sentiment score accuracy. With more tracks, it would be possible to increase the accuracy and even employ the use of machine learning classifiers.

References

[1] Spotify, “Features | spotify for developers.”

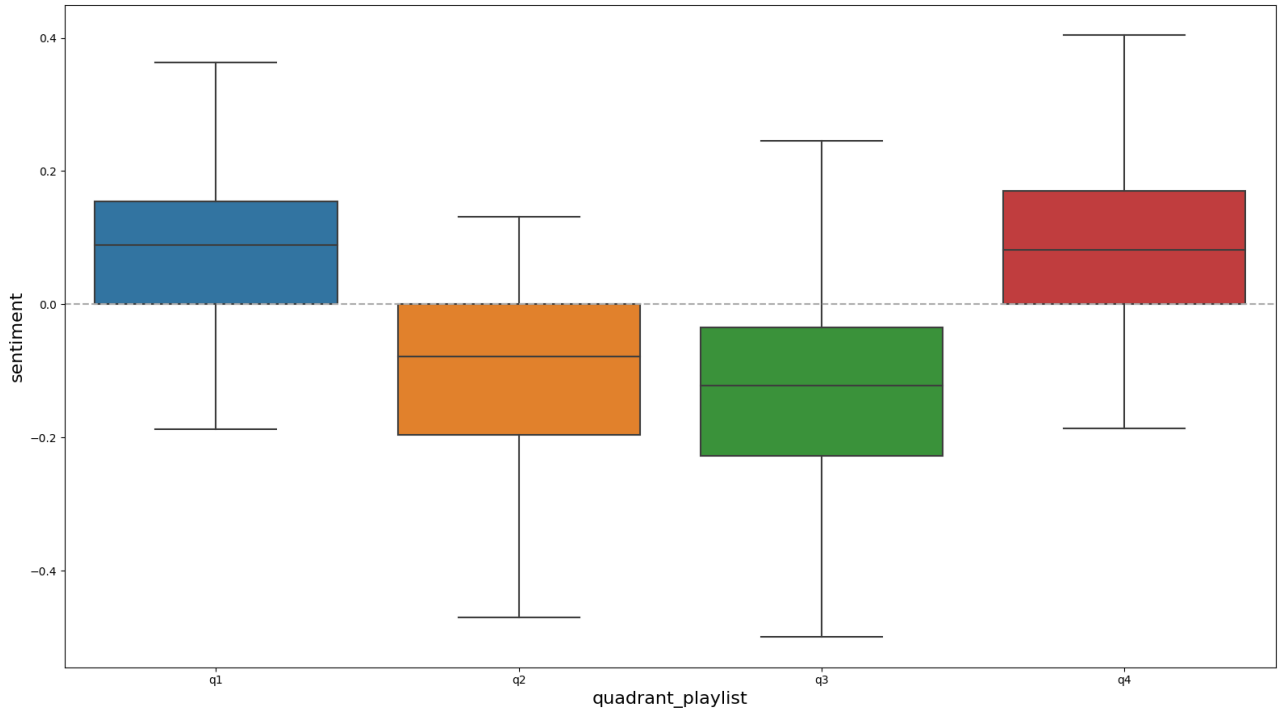

Figure 8: Distribution of Sentiment Scores Over The 4 Quadrants

[2] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[3] S. M. Mohammad, "Word affect intensities," in *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, (Miyazaki, Japan), 2018.

[4] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[5] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, (Melbourne, Australia), 2018.