Detecting Emotions in Lyrics

Diptanu Sarkar
Rochester Institute of Technology
New York, USA
ds9297@rit.edu

Abstract

Music stimulates strong human emotions and feelings. Past studies have shown people spend a significant amount of time listening to music. In the last decade, the way people consume music has also changed. Music platforms provide highly customized playlists to every user along with playlists based on moods. Emotions are subjective, and the subjective nature of emotions makes emotion detection a very challenging task when applied to music. Previously, music emotion detection solely relied on acoustic features. In recent studies, it's observed that using music lyrics features along with acoustic features significantly improves the classification result. To explore the challenges music emotion detection, we first annotated 1,160 lyrics using in 9 categories of Geneva Emotional Music Scales. Then, we developed single-label and multi-label classifiers to detect emotions in lyrics, which achieved 0.65 and 0.82 F_1 scores respectively.

1 Introduction

Music is far more powerful than language to stimulate strong emotions in humans (Sloboda and Juslin, 2001). It can arouse different emotions and feelings in us, which is not restricted by race, culture, nationality. The correlation between music and emotion has always remained an interesting research domain in Music Psychology. Neuroimaging studies have shown that music can activate the brain areas typically associated with emotions (Schaefer, 2017). Listening to music is a prominent component of human lives now, surpassing watching television, reading (Rentfrow and Gosling, 2003). One significant reason for music's universal appeal is the emotional rewards that it extends to its listeners (Zentner et al., 2008).

With the proliferation of the internet and online media platforms, there's a radical change in the way how we consume music now. Online music streaming services such as Spotify[1], Apple Music[2] automatically creates personalized playlists to the users. These providers also generate an ample number of playlists based on moods or emotions, which are not limited to only genres, artists, albums. Such playlists are mostly created by various automatic playlist generation algorithms (Bonnin and Jannach, 2014).

Automatic Music Information Retrieval (MIR), which detects emotion or mood dimensions of music is a growing research field. Human emotional states are not discrete, but continuous (Tettegah and Gartmeier, 2015). Emotions are quite subjective, and the subjective nature of emotions makes emotion detection a very challenging task when applied to music. Moreover, some music contains multiple strong emotions at the same time, on the other hand, some music is very hard to correlate with some emotions. Previously, music classification tasks have primarily relied on features such as track's audio signal or metadata (Hu et al., 2009; Mayer et al., 2008). Lyrics features are also an important attribute along with audio features for song emotion classifications (Hu et al., 2009). The multimodal systems classify emotions in music better than singular models (Hu and Downie, 2010).

The main goal of this project is to explore emotion detection in music using lyrics. Firstly, 1,160 song lyrics are hand-annotated using 9 categories of the Geneva Emotional Music Scales (GEMS) (Zentner et al., 2008) emotions. Using the generated dataset, we developed single-label and multi-label classifiers using unigram, bigram, term frequency-inverse document frequency (tf-idf) BOW features to detect emotions in lyrics, which achieved 0.65 and 0.82 F_1 scores respectively.

The rest of the paper is organized as follows. Section 2 presents previous related work in lyrics

[1] https://www.spotify.com/

[2] https://www.apple.com/music/

emotion detection. Section 3 details the data annotation task and datasets. In Section 4 the methods for experiments are described. Section 5 outlines the result, and Section 6 summaries the conclusion and direction for future work.

2 Related work

Hu et al. proposed three simple yet meaningful set mood categories for MIR systems, using the dataset from Last.fm [3] tags and the USPOP (Ellis, 2003) audio collection. However, because of domain oversimplification, it was not practiced by many researchers.

Zentner et al. proposed a domain-specific device to capture the richness of musically induced emotions - the Geneva Emotional Music Scale (GEMS). This model is designed after an extensive range of studies among music listeners. GEMS consists of 9 categories of musical emotions with 45 emotion labels. In our project, we used GEMS emotions to annotate the music lyrics.

Previously, to improve audio emotion classification, Yang and Lee employed lyric bag-of-words (BOW) text analysis. Even though the classification accuracy improved, due to the smaller dataset size (145 songs), the result was not convincing. Laurier et al. applied both Natural Langauge Processing (NLP) and MIR techniques. Firstly, using a difference between the language model they reported performance close to audio-based classifiers. Later, integrating this in a multimodal system (audio + text), they reported improved classification accuracies in all the four categories over 1,000 songs. Yang et al. proposed three different methods to infuse lyrics along with acoustic features for better performance. They assessed both unigram and bigram BOW lyric features on 1,240 songs in four categories. However, in these studies, the set of four emotion categories is apparent to oversimplify the problem.

Hu et al. analyzed the effect of lyrics text to improve music mood classification. The best performing lyric features are combined with audio features to classify moods in songs. Also, a large dataset with 5,585 songs annotated over 18 mood categories is used as a ground truth set. It is shown that combining lyrics and audio features can improve performance in many mood categories, but not all of them.

Zangerle et al. proposes ALF-200k, a reusable, high-quality publicly available datasets. It includes 176 audio and lyrics features of more than 200,000 songs. Employing a multimodal model they conclude that, while acoustic features are major to attribute tracks to playlist, lyrics features are also important.

Some earlier researches suggest solely relying on lyrical features of music for classification tasks. Fell and Sporleder proposes different dimensions of a song text, such as vocabulary, style, semantics, orientation towards the world, and song structure to model a classifier. Moreover, combining these features with n-gram features further improves the classification accuracy in different classification tasks such as genre detection, distinguishing the best and worst songs. Hu and Downie analyzed songs metadata, lyrics, psychological categories, contained sentiment and text-stylistic features for music emotion recognition. Then best of lyric features are combined with audio that significantly outperformed both the singular methods.

3 Data

For the dataset, over eleven hundred instances of song lyrics were collected from Genius.com[4] across four genres- rock, country, rap/hip-hop, and reggae. Then those lyrics are hand-annotated in 9 emotional categories of the GEMS (Zentner et al., 2008) using LightTag[5]. LightTag provides a set of powerful tools to the researchers to assign annotation tasks and also different statistical measures to asses the quality of data being produced. All subjects annotated about two hundred lyrics according to the emotion(s) they perceived after reading the lyrics. The annotation guidelines followed along with the list of emotions considered are presented in Table 1. Finally, two separate datasets were derived after completing the annotation task - single-label dataset, multi-label dataset.

3.1 Single-label dataset

For the single-label dataset, the label with the maximum number of annotations is selected as the default label for the lyrics. The final dataset consists of only one label per lyrics, along with other metadata such as artist, genre, title, album, year. One interesting fact about the single-label dataset is that it only consists of three classes - Sad-

[3] https://www.last.fm/

[4] https://www.genius.com/

[5] https://www.lighttag.io/

Emotional Category	Definition
Amazement	Feeling of wonder and happiness
Solemnity	Feeling of transcendence, inspiration. Thrills
Tenderness	Sensuality, affect, feeling of love
Nostalgia	Dreamy, melancholic, sentimental feelings
Calmness	Relaxation, serenity, meditativeness
Power	Feeling strong, heroic, triumphant, energetic
Joyful activation	Feels like dancing, bouncy feeling, animated, amused
Tension	Nervous, impatient, irritated
Sadness	Depressed, sorrowful

Table 1: The categories and definition for the annotation task.

Total	Count
Sadness	569
Tenderness	326
Tension	265
Total	***1160***

Table 2: Single-label dataset class distribution.

ness, Tension, Tenderness out of nine total classes. The dataset class distribution is shown in Table 2. From the distribution, it is evident that Sadness is the most represented class, followed by Tenderness and Tension. As our work is mainly focused on the single-label classification task, Word Cloud is visualized in Figure 1, 2 and 3 for the three classes. Word Clouds are visual representations of words that appear more frequently in the corpus.

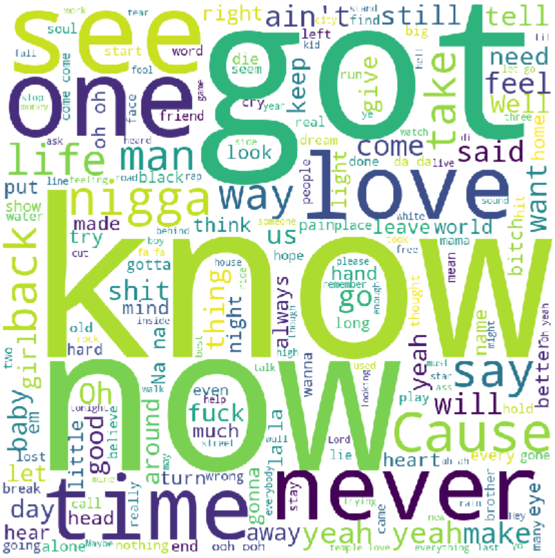

Figure 1: Word Cloud for the Sadness class in the single-label dataset.

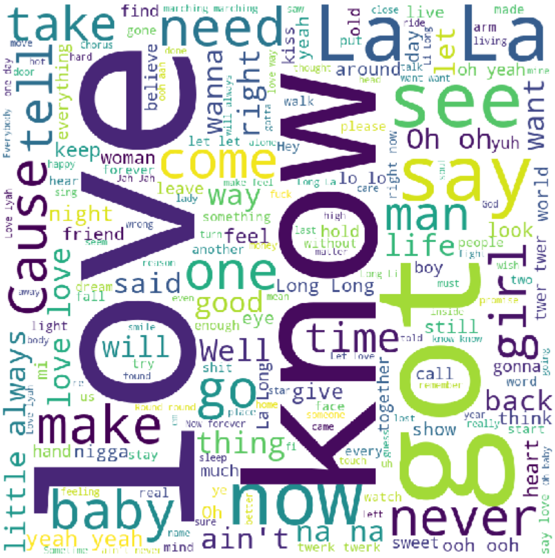

Figure 2: Word Cloud for the Tenderness class in the single-label dataset.

3.2 Multi-label dataset

For the multi-label dataset, all the labels with at least one annotation are included as a label for the lyrics. The final dataset consists of multiple labels per lyrics, along with other metadata such as artist, genre, title, album, year. The dataset class distribution is shown in Table 3. Interestingly, it is observed that the dataset has an average of 3 emotions per example with 1 and 7 minimum and maximum emotions per example respectively.

4 Methods

In this section, we discuss different features and classifiers employed for the lyrics classification task.

Class	Count
Sadness	574
Tension	553
Tenderness	498
Power	477
Nostalgia	438
Solemnity	377
Joyful Activation	349
Calmness	250
Amazement	239
Total	***3755***

Table 3: Multi-label dataset class distribution.

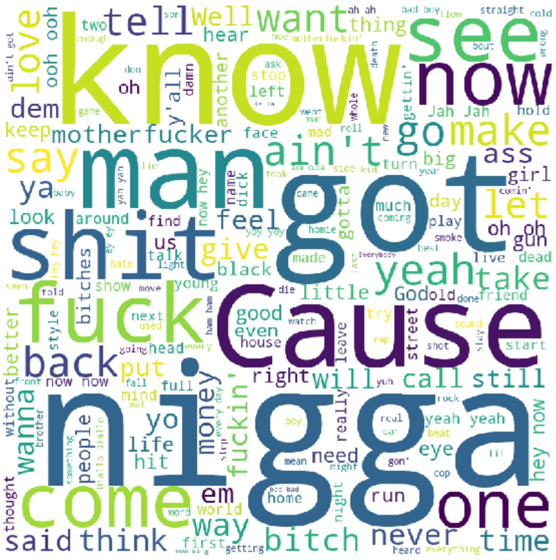

Figure 3: Word Cloud for the Tension class in the single-label dataset.

4.1 N-gram

In Natural Language Processing n-gram of texts is a set of n co-occurring words in the sample text. An n-gram of size 1,2 and 3 are called Unigram, Bigram and Trigram respectively. An n-gram model is a type of probabilistic language model for predicting the next item in a sequence. N-gram models are widely used in probaility, computational linguistics, sequence analysis, data compression. Another use of n-grams is to develop features for supervised machine learning models such as Support Vector Machines, Naive Bayes, etc.

4.2 Term Frequency–Inverse Document Frequency

Term Frequency–Inverse Document Frequency (tf-idf) is a statistical measure intended to reflect how important a word is to a document in a corpus. It is a term-weighting scheme and used in different applications such as information retrieval, text mining, and user modeling. The Term Frequency measures how frequently a term occurs in a document and Inverse Document Frequency measures how much information the word provides. The IDF factor decreases the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. Hence, the higher the tf-idf value of a particular word, the more the importance of the word in the collection or corpus.

4.3 *k*-Fold Cross Validation

Cross-validation is the process of re-sampling the dataset to evaluate how well machine learning models will perform in practice. The goal of cross-validation is to determine how well the predictive model generalizes an independent dataset that was not used to train the model. In *k*-fold cross-validation, the given data is grouped into k data samples. It helps to flag problems such as selection-bias, overfitting. In this project, we used *5*-fold cross-validation.

4.4 Part-of-speech Tagging

Part-of-speech tagging or POS tagging in linguistics is the process of marking a word in a sentence or corpus to a particular part of speech considering both the definition of the word and its context. Automatic POS tagging algorithms are generally of two types - rule-base and probabilistic. POS tagging is useful for building lemmatizers and parse trees. It is also employed in word sense disambiguation and sentiment analysis. E. Brill's tagger (Brill, 1992) is a rule-based POS tagger, which is most widely used in English. In this project, we employed POS tagger from NLTK (Loper and Bird, 2002) library.

4.5 Logistic Regression

In statistics, logistic regression is a classification algorithm to predict binary dependent variables. The simple logistic model can be extended to categories of multiple classes with a probability assigned to each class. It needs quite large sample sizes to perform better. However, logistic regression is widely used in various disciplines such as machine learning, medical fields, and social sciences.

4.6 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm widely used for tasks such as classification, regression. It is a discriminative classifier with an objective to determine a hyperplane in an N-dimensional space that precisely classifies the data point. SVM uses a subset of training points in the decision function called support vectors. It is simple, memory efficient yet highly effective in high dimensional space.

Linear Support Vector Classification (LinearSVC) is an implementation of Support Vector Classification with a linear kernel. With a large number of samples, the LinearSVC tends to converge faster and give good results.

4.7 Decision Tree

A decision tree is a simple and popular model for classification and prediction tasks. It is build based on a tree-like model of decisions and possible outcomes. A tree can be trained by splitting the source set into subsets based on an attribute value test. The model is trained by splitting the training set into subsets based on the target attribute value. The same process is recursively repeated on each derived subset to build the complete decision tree. Due to the ease of understandability and less computation power requirement, it is widely used in operations research and management.

4.8 Random Forest

Random Forest is an ensemble learning method for tasks such as classification, regression, etc. It consists of a high number of individual decision trees that operate as a collection. All the individual models predict the output and the class with the maximum number of support become the model's prediction. Random forest algorithm is a great fit for high dimensional data since it works only with subsets of data in each split. It tends to overfit hence it requires careful tuning on hyperparameters.

4.9 Multilayer Perceptron

A multilayer perceptron (MLP) is an Arti

facial Neural Network (ANN) that connects multiple neurons in the form of a network. It is a multi-layered model with three hidden layers followed by an output layer. In the feed-forward phase, the input layer multiplies the weights assigned to the values and sums them up. The relu activation function is used the hidden layers and 3-way softmax for the output layer. In the backpropagation stage, loss functions are calculated and minimized for better performance.

5 Experiments and Results

In this section, we discuss different experiments performed to classify emotions in lyrics using both single-label and multi-label datasets. A thorough analysis of the result is also illustrated.

5.1 Single-label Classification

For our experiments, considering the *lyrics* and *label* columns of the single-label dataset, we divided the dataset into train and test sets in 75% and 25% subsets. We analyzed tf-idf weighted POS, unigram, bigram, and trigram features. The tf-idf vectorizer transforms text to feature vectors that can be used as input to a classifier. Then, five different classifiers - Logistic regression, SVM, Decision Tree, Random Forest, Multi-layer Perceptron are trained on the features extracted in the previous step. 5-fold cross-validation is also performed on the trained models and the mean accuracies of the models are reported, along with test accuracies. Table 4 shows a detailed report of features, classifiers, and the test and 5-fold cross-validation accuracies.

It is observed that linear SVC and logistic regression models, in general, perform better than other employed classifiers. The highest accuracy reported is 0.65 using a logistic regression classifier with bigram and trigram as features.

In machine learning, a confusion matrix helps to visualize which classes the model is making more mistakes. Each row of the matrix represents the instances of predicted class while each column represents an actual class. The confusion matrix of the best performing model is shown in Figure 4. As we can observe, about 75% of total test data consists of *sadness* class and also contributes more to the total classification accuracy.

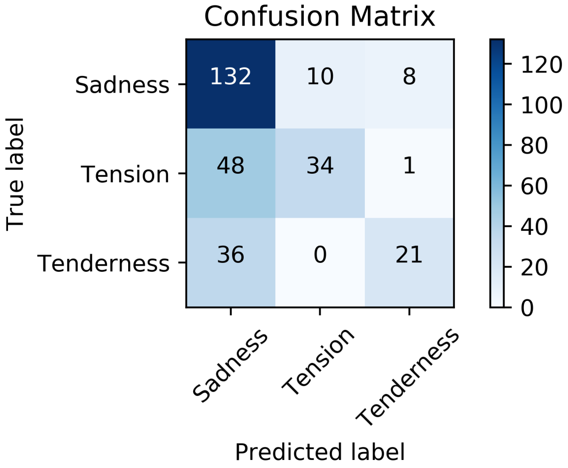

Figure 4: The confusion matrix of the logistic regression classifier with bigram and trigram as features.

5.2 Multi-label Classification

For the multi-label classification, we divided the data in the same way as for a single-label dataset. We used the tf-idf vectorizer with unigram features to transform the text into vectors. The model was trained using a logistic regressor. On the test set, the classifier has achieved an F_1 score of 0.82. F_1 score is a measure of test accuracy, it considers both the precision and the recall to compute the score.

6 Conclusion

Emotion detection in music is an interesting field of research as the music plays a prominent role in human lives. There are multiple approaches to classify the underneath emotion of music. In this work, we firstly annotated 1,160 song lyrics using 9 categories of the GEMS emotions and produced single-label and multi-label datasets. Then using the datasets, we explored a set of text features such as POS, unigram, bigram, and trigram combinations to model a classifier to detect the emotions of the lyrics. In the single-label dataset, the logistic regression classifier with bigram and trigram as features has outperformed other classifiers achieving 0.65 accuracy over the test set and 0.61 5-fold cross-validation accuracy. The multi-label logistic regression classifier has achieved a 0.82 F_1 score.

As future work, we plan to explore the acoustic features of music along with different metadata available to improvise the baseline classification accuracy.

Acknowledgments

We are thankful to Dr. Marcos Zampieri for collecting the lyrics and setting up the data annotation task. We also want to acknowledge all our colleagues for their contribution to the annotation task. Lastly, we are grateful to Mukunda Jajoo for suggesting multiple enhancements in the experiments and reporting.

References

Geoffray Bonnin and Dietmar Jannach. 2014. Automated generation of music playlists: Survey and experiments. *ACM Comput. Surv.*, 47(2):26:1–26:35.

Eric Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, pages 152–155. Association for Computational Linguistics.

Dan Ellis. 2003. The uspop2002 pop music data set. *http://www. ee. columbia. edu/% 7Edpwe/research/musicsim/uspop. html.*

Michael Fell and Caroline Sporleder. 2014. Lyrics-based analysis and classification of music. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 620–631.

Xiao Hu, Mert Bay, and J Stephen Downie. 2007. Creating a simplified music mood classification ground-truth set. In *ISMIR*, pages 309–310.

Xiao Hu and J. Stephen Downie. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 159–168, New York, NY, USA. ACM.

Xiao Hu, J Stephen Downie, and Andreas F Ehmann. 2009. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209.

Classifier	Features	Test Accuracy	CV Mean Accuracy
Linear SVC	Unigram	0.61	0.59 (+/- 0.07)
Linear SVC	Unigram + POS	0.60	0.60 (+/- 0.07)
Linear SVC	Bigram	0.53	0.51 (+/- 0.03)
Linear SVC	Bigram + POS	0.56	0.56 (+/- 0.05)
Linear SVC	Trigram	0.51	0.48 (+/- 0.05)
Linear SVC	Unigram + Bigram	0.61	0.57 (+/- 0.04)
Linear SVC	Bigram + Trigram	0.61	0.59 (+/- 0.07)
Linear SVC	Bigram + Trigram + POS	0.60	0.60 (+/- 0.07)
Linear SVC	Unigram + Bigram + Trigram	0.60	0.57 (+/- 0.03)
Logistic Regression	Unigram	0.64	0.61 (+/- 0.01)
Logistic Regression	Unigram + POS	0.64	0.60 (+/- 0.04)
Logistic Regression	Bigram	0.58	0.53 (+/- 0.03)
Logistic Regression	Bigram + POS	0.61	0.58 (+/- 0.02)
Logistic Regression	Trigram	0.51	0.48 (+/- 0.06)
Logistic Regression	Trigram + POS	0.54	0.54 (+/- 0.03)
Logistic Regression	Unigram + Bigram	0.64	0.59 (+/- 0.03)
Logistic Regression	Unigram + Bigram + POS	0.63	0.59 (+/- 0.03)
Logistic Regression	**Bigram + Trigram**	**0.65**	**0.61 (+/- 0.01)**
Logistic Regression	Bigram + Trigram + POS	0.64	0.60 (+/- 0.04)
Logistic Regression	Unigram + Bigram + Trigram	0.63	0.59 (+/- 0.03)
Logistic Regression	Unigram + Bigram + Trigram + POS	0.62	0.60 (+/- 0.02)
Decision Tree	Unigram	0.59	0.55 (+/- 0.03)
Decision Tree	Unigram + POS	0.58	0.52 (+/- 0.04)
Decision Tree	Bigram	0.54	0.51 (+/- 0.02)
Decision Tree	Bigram + POS	0.50	0.54 (+/- 0.03)
Decision Tree	Trigram	0.52	0.49 (+/- 0.01)
Decision Tree	Trigram + POS	0.55	0.51 (+/- 0.05)
Decision Tree	Unigram + Bigram	0.57	0.57 (+/- 0.04)
Decision Tree	Bigram + Trigram	0.60	0.56 (+/- 0.03)
Decision Tree	Unigram + Bigram + Trigram	0.59	0.54 (+/- 0.02)
Random Forest	Unigram	0.55	0.52 (+/- 0.02)
Random Forest	Unigram + POS	0.57	0.53 (+/- 0.03)
Random Forest	Bigram	0.51	0.49 (+/- 0.01)
Random Forest	Bigram + POS	0.53	0.51 (+/- 0.03)
Random Forest	Trigram	0.50	0.49 (+/- 0.00)
Random Forest	Trigram + POS	0.52	0.49 (+/- 0.02)
Random Forest	Unigram + Bigram	0.52	0.52 (+/- 0.02)
Random Forest	Unigram + Bigram + POS	0.57	0.54 (+/- 0.05)
Random Forest	Bigram + Trigram	0.54	0.52 (+/- 0.02)
Random Forest	Bigram + Trigram + POS	0.55	0.51 (+/- 0.03)
Random Forest	Unigram + Bigram + Trigram	0.55	0.52 (+/- 0.01)
Random Forest	Unigram + Bigram + Trigram + POS	0.54	0.54 (+/- 0.02)
Multilayer Perceptron	Unigram	0.58	0.54 (+/- 0.07)
Multilayer Perceptron	Bigram	0.52	0.48 (+/- 0.04)
Multilayer Perceptron	Bigram + POS	0.53	0.51 (+/- 0.09)
Multilayer Perceptron	Trigram	0.51	0.47 (+/- 0.06)
Multilayer Perceptron	Unigram + Bigram	0.57	0.54 (+/- 0.02)
Multilayer Perceptron	Bigram + Trigram	0.56	0.55 (+/- 0.07)
Multilayer Perceptron	Bigram + Trigram + POS	0.59	0.57 (+/- 0.07)
Multilayer Perceptron	Unigram + Bigram + Trigram	0.56	0.53 (+/- 0.03)
Multilayer Perceptron	Unigram + Bigram + Trigram + POS	0.59	0.54 (+/- 0.10)

Table 4: Detailed report of features, classifiers, and the test and 5-fold cross-validation mean accuracies.

Cyril Laurier, Jens Grivolla, and Perfecto Herrera. 2008. Multimodal music mood classification using audio and lyrics. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 688–693. IEEE.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. 2008. Combination of audio and lyrics features for genre classification in digital audio collections. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 159–168. ACM.

Peter Rentfrow and Samuel Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84:1236–56.

Hans-Eckhardt Schaefer. 2017. Music-evoked emotions—current studies. *Frontiers in Neuroscience*, 11:600.

John A Sloboda and Patrik N Juslin. 2001. Psychological perspectives on music and emotion. *Oxford University*.

Sharon Y Tettegah and Martin Gartmeier. 2015. *Emotions, technology, design, and learning*. Academic Press.

Dan Yang and Won-Sook Lee. 2004. Disambiguating music emotion using software agents. In *Music*.

Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen. 2008. Toward multi-modal music emotion classification. In *Pacific-Rim Conference on Multimedia*, pages 70–79. Springer.

Eva Zangerle, Michael Tschuggnall, Stefan Wurzinger, and Günther Specht. 2018. Alf-200k: Towards extensive multimodal analyses of music tracks and playlists. In *European Conference on Information Retrieval*, pages 584–590. Springer.

Marcel Zentner, Didier Grandjean, and Klaus R Scherer. 2008. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion*, 8(4):494.

A Datasets and Code

The datasets and the code of the project is publicly available at GitHub [6].

[6] https://github.com/imdiptanu/lyrics-emotion-detection