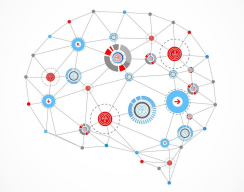
Mood Classification of Song Lyrics Using Deep Learning

Cyprian Gascoigne, Jack Workman, Yuchen Zhang
Instructor: Daniel Cer

Abstract



Mood classification of lyrics can help in the creation of automatic playlists, search engines, music libraries, and recommendation systems.

We measure the effectiveness and accuracy of different machine learning algorithms on mood classification, including Naive Bayes, SVM, and Convolutional Neural Network models.

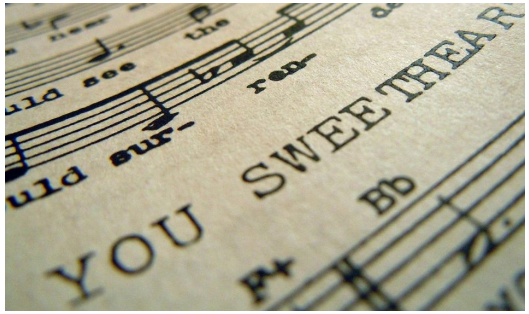
- Our results indicate optimal performance using our neural network vs the other models.

We use mood categories derived from Russell's model of affect

- Where mood is represented by a vector in a 2D valence-arousal space.

We will test the extensibility of our deep learning model using the Million Song Dataset (MSD), a freely available contemporary music track dataset.

Introduction

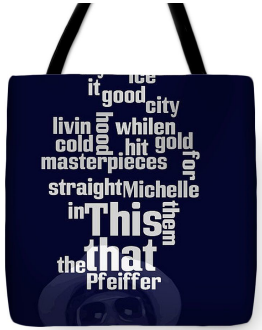


Mood and emotional classification is crucial to the entertainment value of songs, but much research on music classification is based on audio analysis versus lyric analysis (our focus).

- One reason may be that obtaining a large dataset legally is difficult (as lyrics are copyrighted material).

However, classifiers that incorporate textual features can outperform audio-only classifiers

Lit Review



Previous work reached contradictory conclusions and employed smaller datasets and simpler methodology like bag-of-words to derive unigrams, fuzzy clustering, and linear regression.

The approach of past research on musical mood classification can be grouped into three categories: classification using joint audio-lyric data, using only lyric data, and using only audio data.

Previous studies saw that lyrics alone can be used to classify music mood achieving an accuracy of up to 70% for a single mood with the MSD and SVM classification algorithms.

- Our choice of using a convolutional neural network is based on past research demonstrating that CNN-based systems are effective and successful for sentiment analysis (Kim, 2014).

Methodology

Data Acquisition

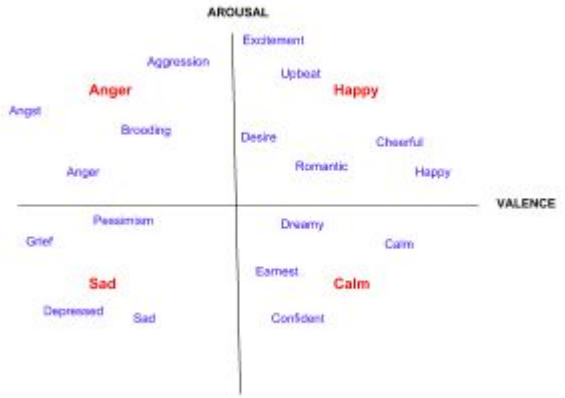
One initial challenge was in acquiring corresponding lyrics to song data.

To obtain the 63,803 lyrics in our dataset, we used the python package lyricsgenius (interfaces with www.genius.com) for retrieving lyrics

From the MSD, we matched 327,234 songs to lyrics.

- For our analysis, we filtered out non-english songs and those for which we could not match the mood.

Methodology (cont.)



Mood Annotation

We use the popular Russell's model of affect and focus on 18 mood classes.

- In Russell's model, mood is represented by a vector in the 2-D valence-arousal space where mood is an element of (valence, arousal): m M = (v, a)

Mapping Labels

To map each song to its mood, we make use of the sqlite db version of the last.fm dataset.

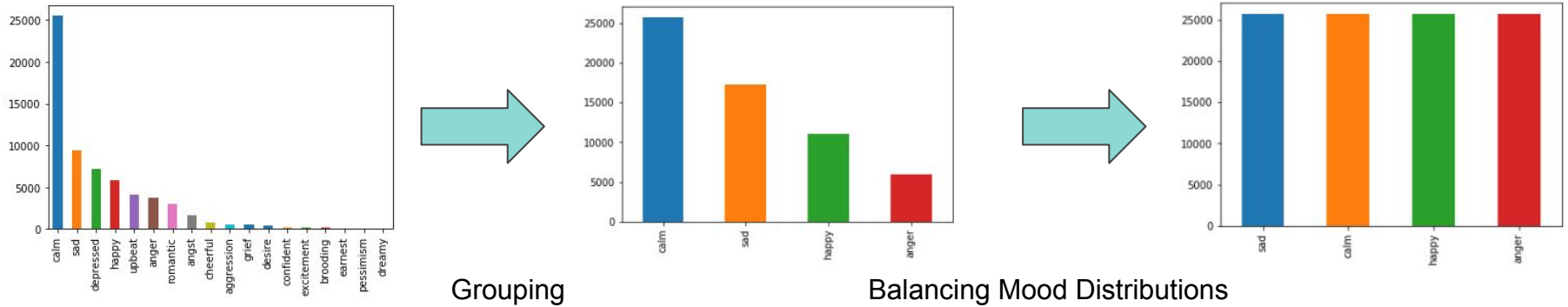
- We query for all associated tags of each song, then attempt to match tags against our mood categories with substring matching

Methodology (cont.)

Data Balancing

To balance the data:

- We group related-moods into four quadrants.
- Then oversample the underrepresented moods by copying and shuffling each song's lines and then appending them to the dataset



Grouping

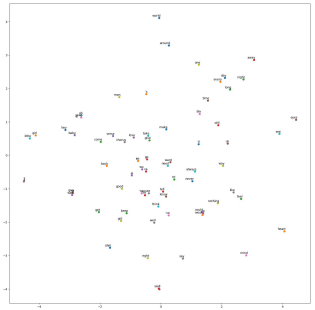
Balancing Mood Distributions

Methodology (cont.)

Word Embeddings

To form our word embeddings, we make use of the word2vec model and the implementation provided by TensorFlow.

- We use a vocabulary generated from the top most common 10,000 words of our full set of mood-labeled lyrics to train a skip-gram word2vec model



Visualization of Embeddings

Methodology (cont.)

Baseline Classification

To establish a baseline classification accuracy, we use the most-common-case (MCC) classification approach and two supervised machine learning algorithms, Multinomial Naive-Bayes (NB) and Support Vector Machines (SVM), both from sklearn.

Methodology (cont.)

Deep Learning Classification

For our deep learning classifier, we use a Convolutional Neural Network (CNN) implemented in TensorFlow.

- To prepare for training, we normalize all songs to a length equal to the 75th percentile of our dataset's song length, remove stopwords, remove punctuation, and tokenize lyrics

We experiment with two different embeddings inputs: random initialization with on-the-fly model training and word2vec.

Results

Our results show that the highest accuracy for each classifier is found with the Balanced Mood Quadrants dataset: 77.08% by our CNN w2v0 model.

- We attribute this to the embedding training that happened alongside the CNN's training.

The most performant machine learning classifier is NB also in the Balanced Mood Quadrants dataset with 55.19%.

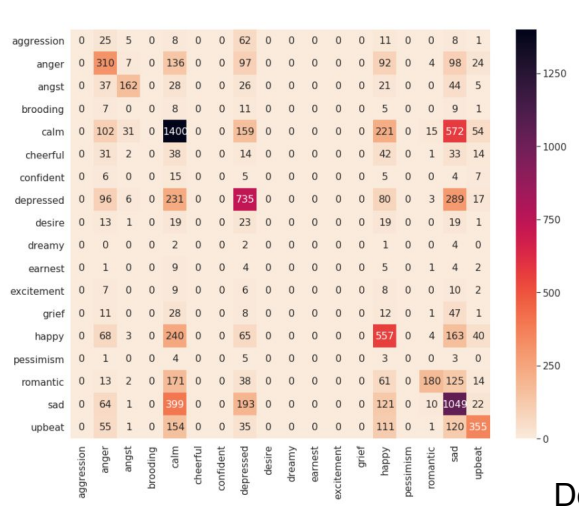
- NB performs better with balanced data as SVM outperforms NB in the unbalanced datasets.

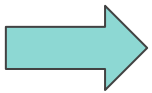
Model	Unbal. Mood	Unbal. Mood Quadrs.	Bal. Mood Quadrs.
MCC	39.81%	43.61%	25.34%
NB	39.93%	46.78%	55.19%
SVM	44.88%	50.95%	54.07%
CNN w2v0	**56.79%**	62.15%	**77.08%**
CNN w2v1	54.33%	**63.53%**	75.45%

Results

To review CNN results, we use a confusion matrix to visualize accuracy.

- Calm and sad are the two most confused quadrants





Detailed vs Consolidated Form

Conclusion

As more music moves online, intelligent and accurate music recommendation systems will be greatly enhanced by the use of lyrics to aid prediction.

- As larger datasets become more readily available, the real-life applicability of these algorithms will become easier to vet.

Our work and datasets are available online and open-sourced:
https://github.com/workmanjack/w266-group-project_lyric-mood-classification