University of Warsaw

Faculty of Economic Sciences

Ewa Włodarczyk Karolina Sierocka Tymoteusz Mętrak

Album N°: 432571 Album N°: Album N°:

**Which type of song is the most popular: happy, sad or angry? (mood classification of song lyrics from Billboard Hot 100)**

Project for

Text Mining and Social Media Mining course

Project written under the supervision of

Dr. Hab. Jacek Lewkowicz

from Department of Political Economy

WNE UW

Warsaw, January 2024

# TABLE OF CONTENTS

|  |  |
| --- | --- |
| INTRODUCTION................................................................................................................. | 2 |
| CHAPTER I. Background............................................................................................ | 2 |
| CHAPTER II. DATAset..................................................................................................... | 3 |
| CHAPTER III. Modeling................................................................................................ | 5 |
| CHAPTER IV. results.................................................................................................... |  |
| CONCLUSION..................................................................................................................... |  |
| BIBLIOGRAPHY................................................................................................................. |  |

**INTRODUCTION**

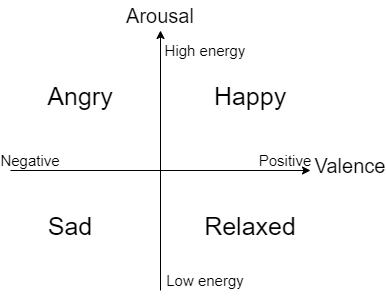
Listening to music can influence our mood, which can then in turn impact our behavior. But what kind of music are we listening to? The sentiment analysis of popular songs’ lyrics done by Napier and Shamir (2018) shows a significant increase in anger and sadness and a decrease in joy in the tone of popular music from 1951-2016. Still, in the same study, while analyzing tone expressed in lyrics, most popular genres, apart from rap, scored higher in joy than in sadness and anger. Additionally, as pointed out by Schellenberg and Von Scheve (2012) popular music has changed over the years to involve more mixed emotional cues.

The purpose of the project is to find out which type of music is the most popular: happy, sad or angry based on mood classification of song lyrics from the Billboard Hot 100. Our main assumption, taking into account the fact that most popular genres are: pop music, which mood is usually happy, and rap/hip-hop, which mood is rather angry, would be that most popular types of music have happy or angry lyrics. However, even looking at the title of the current (written on the 8th of November 2023) number 1 song from the Billboard Hot 100 - “Is it over now?”, which clearly signals a sad tone, we can see that deciding which type of music is the most popular will pose a greater challenge.

The organization of the paper is as follows. The related work and background of the research is discussed in Chapter I, description of the data used is in Chapter II, proposed methodology is presented in Chapter III, discussion about the results, model evaluation and comparison is showed in Chapter IV, which is then followed by conclusion.

**CHAPTER I. BACKGROUND**

Recently, several studies have been concentrated on a lyric-based mood classification through natural language processing techniques. In most of them mood categories are derived from the Russell’s (1980) circumplex model, which maps emotions across two dimensions: valence and arousal, as shown in Figure 1. In this study, we will also use Russel's model, but in a simplified version, classifying song lyrics into four mood categories: Happy, Sad, Angry and Relaxed. As there are more than two classes and each data point can only be assigned to one class this will be a multi-class classification.

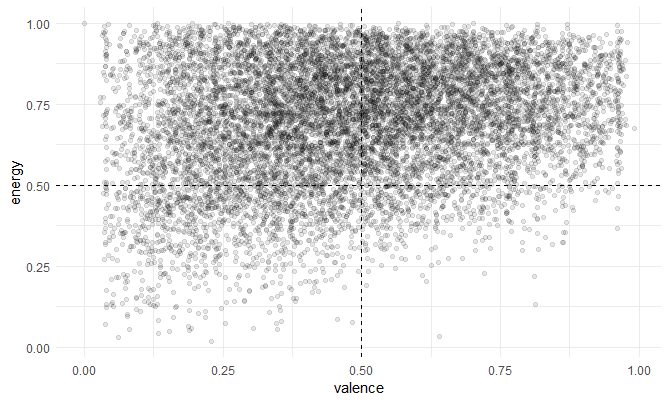
Fig. 1. Simplified Russel's circumplex emotion model.

*Source:* Own study, based on: Russell J. (1980) A circumplex model of affect, Journal ofPersonality and Social Psychology, vol. 39, 1161–1178.

Mood classification is performed with varying success using different machine learning processes and algorithms. Akella and Moh (2019) achieved an accuracy of 71% using Convolutional Neural Network. Similarly, Abdillah, Asror and Wibowo (2020) created a Recurrent Neural Network, which can produce an accuracy of 91%. Siriket, Sa-ing and Khonthapagdee (2021) however, instead of using deep learning techniques, used a boosting algorithm with 89% accuracy. This proves that multi-class classification can be successfully performed using not only neural networks, but also simpler machine learning techniques, which will be used in this study.

**CHAPTER II. DATASET**

Two distinct datasets were used for this analysis. The first one is from Spotify and it contains data of around 10 thousand songs released from 1999 to 2019. Apart from lyrics, it also consists of important song features: energy and valance, which determine the mood label in accordance with Russel's model. For both traits, their values range from 0.0 to 1.0. However, songs are not evenly distributed across those variables as it is shown on Figure 2, making mood labeling much more complicated. We want to train our model on songs in which emotional states are clearly expressed, with enough examples of lyrics for each category. That’s why we apply mood labels only to songs with attributes in the 1st or 3rd quartile.

Fig. 2. The top 10 of the most common terms in each mood category

*Source:* Own study, based on data from Spotify

After applying mood labels there are 757 happy, 526 angry and 706 sad songs. Unfortunately, as there are not enough relaxed songs, relaxed mood will not be included in the model. This dataset will then be used for training, evaluating and choosing the best mood classification model.

The second dataset is from Billboard Hot 100 and it holds data of around 5 thousand songs that have charted from 1999 to 2019. Apart from lyrics, it also consists of songs’ peak positions and the number of weeks they have been on the chart. We have selected only the songs which peaked at number 1, limiting our data to only 245 songs. On this dataset, after mood labels are applied to songs by the best mood classification model, analysis will be performed to determine which type of song is the most popular.

Additional data preparation on both datasets included text preprocessing: removing special characters, changing contraction to their multi-word forms, replacing informal writing with known semantic replacements and reducing whitespace. After that, stop words were removed and Porter Stemming Algorithm was applied to reduce words to their root form. Figure 3 shows the most common terms used in each mood category. A lot of words are common across all categories, for example: “can”, “will”, “know”, “love”, ”now”, ”just” or ”like”. However, we can see that words like “never” and “want’ are present more frequently in songs with angry mood. Also “feel” is more common in categories associated with negative moods and “baby” with happy mood.

Fig. 3. The top 10 of the most common terms in each mood category

*Source:* Own study, based on data from Spotify

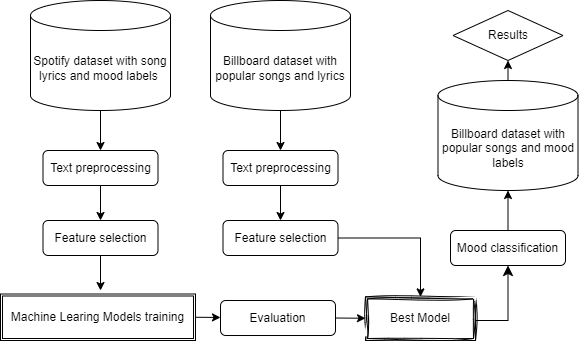
Additionally, most common words in each category are very interesting, namely: “can” and “will” for Angry, “like” and “love” for Happy and “love” and “know” for Sad. Nevertheless, the words which were common across all of the categories were removed, treating them as extra stop words.

**CHAPTER III. MODELING**

After text preprocessing, lyrics data was converted into a Document Term Matrix (DTM), where words were tokenized, changed into n-grams of one to three words and scores were assigned using Term Frequency - Inverse Document (TF-IDF) technique. This gives terms a higher score if they appear frequently in a particular document, but rarely across the entire corpus and it helps to balance off terms’ frequency and importance. In this case, some of the terms with the highest TF-IDF score were: “oh oh”, “without”, “let”, “need” or “night”. However, before applying any machine learning models, an additional feature selection was need, as there were more than 330 thousand terms in the matrix.

The first feature selection was filtering our DTM and keeping only terms, which had TF-IDF scores in the top 0.5 percentile. This limits our testing dataset to about 2 thousand variables.

Later, several traditional machine learning algorithms such as Naive Bayes, Decision tree and Random forest were used for text classification with varying results. The full workflow of this study can be seen on Figure 4.

Fig.4. The overall process of the proposed research

*Source:* Own study

**CHAPTER IV. results**

**conclusion**

**BIBLIOGRAPHY**

Abdillah, J., Asror, I., Wibowo, Y. (2020). Emotion Classification of Song Lyrics using Bidirectional LSTM Method with GloVe Word Representation Weighting. Journal RESTI 4. 723-729.

Akella, R., Moh, T. (2019). Mood Classification with Lyrics and ConvNets. Proceedings - 18th IEEE International Conference on Machine Learning and Applications. 511-514.

Napier, K., Shamir, L. (2018). Quantitative Sentiment Analysis of Lyrics in Popular Music. Journal of Popular Music Studies, 30(4), 161–176.

Russell J. (1980) A circumplex model of affect, Journal ofPersonality and Social Psychology, vol. 39, 1161–1178.

Schellenberg, E., von Scheve, Christian. (2012). Emotional Cues in American Popular Music: Five Decades of the Top 40. Psychology of Aesthetics Creativity and the Arts. 6. 196-203.

Siriket K., Sa-ing V., Khonthapagdee S. (2021) Mood classification from Song Lyric using Machine Learning. 9th International Electrical Engineering Congress. 476-478.