Emotion Detection in Song Lyrics

Picascia Sergio
Data Science and Economics
University of Milan

Abstract

Emotions play a fundamental role in songs, in such a way that an entire branch of psychology is devoted to understand the relationships between human affect and music. Emotions are also the center of attention of numerous sentiment analysis studies, concentrated in particular to the identification of them inside texts or facial expressions. This research focuses on detecting from song lyrics which are the feelings that a listener can experience, starting from understanding what is an emotion and in which way it is possible to represent it. The EmoBank dataset will be used as the main source of knowledge in order to train different neural networks able to place sentences in the space drawn according to the VAD scheme. The analysis will then move to estimate a model for mapping values for valence, arousal and dominance, to categories, representing emotions. Eventually, it will be possible to generate playlists based on the current listener's mood and its music preferences.

Introduction

Emotion detection is a central topic in the field of sentiment analysis and numerous researches have been conducted about it. This project, in particular, focuses on the application of these methods on song lyrics, in order to study how different music genres relate to emotions. Before going in details about the methodologies that can be applied to accomplish this task, it is extremely important to define what is an emotion and how it is possible to represent it.

According to the literary definition, an emotion is a strong feeling deriving from one's circumstances, mood, or relationships with others. Over the years, several psychologists have given their own interpretations of models of emotions and, to date, there are two main methods for representing them, which are widely used in the literature: categorical and dimensional.

The categorical representation puts emotions in a discrete space, treating each of them as a class. This approach results in data having a binary nature (presence, non-presence) or presenting continuous values indicating the intensity of the emotions. If, on one hand, this view is extremely useful since it allows to concretely assign observations (documents, expressions) to precise emotions, on the other hand, it is quite limited in terms of applications, because different tasks may require the identification of different emotions and, therefore, the same model could not be applied to two distinct problems. There are several classification models, but the most widely acknowledged is the one drafted by Paul Ekman, which identifies six basic emotions: anger, disgust, fear, happiness, sadness and surprise ("An Argument for Basic Emotions", 1992).

The dimensional representation, instead, focuses on three aspects: valence (pleasant/unpleasant), arousal (energized/soporific) and dominance (controlling/controlled). Each dimension has an equal range of values which is usually identified as being between -1 and 1. This view allows to apply the same model to multiple tasks, even removing one of the dimensions (as often happens with dominance), and has been gaining much importance in this field during the last years. The model is referred to as VAD, while the first instance of this idea is the PAD model (Mehrabian & Russell, "An approach to environmental psychology", 1974), with pleasure used in place of valence. In this study, both methods will be used, in order to get the best out of both, using a mapping in order to pass from one representation to the other.

Research Question and Methodology

The aim of this project is to build a model able to detect emotions in song lyrics, identify a possible correlation with music genres and, eventually, create playlists based on someone's favorite genre and current mood. In order to accomplish this task, different steps must be performed: learn a model that is able to predict VAD values given a sentence; map VAD values to categories representing emotions; apply these methods to a dataset containing song lyrics; perform the remaining analysis on the resulting data.

The first step consists in building a model capable of estimating the values for valence, arousal and dominance from a list of words. After having applied all the necessary text preprocessing operations over sentences, in order to normalize and tokenize each word, the approach that has been followed is similar to the one explained in "Word Emotion Induction for Multiple Languages as a Deep Multi-Task Learning Problem", Buechel & Hahn 2018 (Figure 1). Firstly, sentences have been embedded in order to be used as input for the neural network, using three different methods: an average of the vectors of each word, resulting from the Word2Vec model built on the Google News corpus; a Doc2Vec model learnt on the EmoBank dataset; the same Doc2Vec model, this time trained without including in the vocabulary words considered neutral according to the SentiWordNet score. Each of these methods allowed to avoid any issues regarding out of vocabulary words when applying the model to a new dataset. Then, three feed-forward neural network have been built, with the same structure: two hidden layers and an output layer composed by three nodes, each of them predicting one values in the VAD model. In fact, in the previously cited paper, it has been demonstrated that building a single neural network and predicting the three variables altogether returns better results than building three separate models. The Leaky ReLU has been used as activation function in the

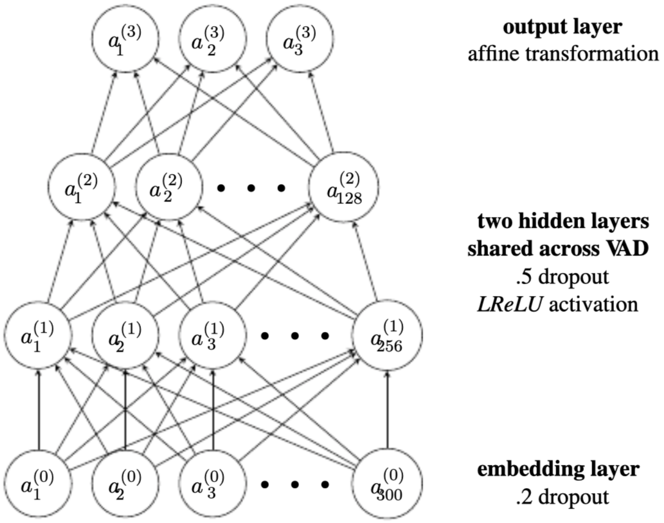

Figure 1: Structure of the neural network used by Buechel & Hahn.

hidden layers, Adam as optimizer and the mean squared error as loss function; a dropout layer has been put after each hidden layer, with a rate of 0.5.

The second crucial step of the analysis regards the mapping from the VAD model to the categorical representation of the emotions. This task has been accomplished in different ways in literature, using both deterministic and stochastic approaches. Starting from the latter, the most common procedure involves the estimation of a classification model using common machine learning techniques, such as decision trees or support vector machines. On the other hand, another idea is to determine the position of emotions in the VAD space, as done by Mehrabian and Russell in "Evidence for a Three-Factor Theory of Emotions", 1977 (Figure 2). This approach is extremely flexible since it can be applied to research questions having different target emotions: the classes are identified and located in the space; their locations are used as reference for the observations, in such a way that each point in the dataset is directly assigned to the emotion closer to it.

Experimental Results

This research has been conducted on two datasets, one used for the learning phase and the other for the application one. During the first part of the analysis, the EmoBank dataset has been chosen in order to learn the model. The choice was made given the distinction in the data between the emotion perceived by the reader and the emotion the writer wanted to transmit. This differentiation is necessary in this kind of analysis, especially when the model will then

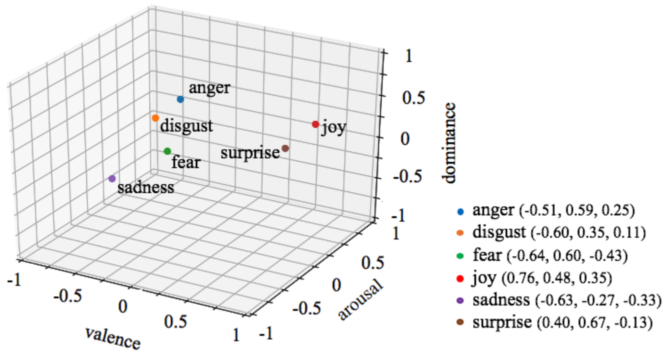

Figure 2: Representation of the Ekman's six basic emotions in the VAD space.

be applied to music: it can happen that the perspective of the listener differs from the one of the artist and, since the goal is to create playlist based also on the mood of the listener himself, the 'reader' dataset has been chosen over the 'writer' one. The data consists in more than ten thousand observations, each of it representing a sentence and the corresponding values for the VAD scheme; a subset of it has also been manually annotated with the Ekman's six basic emotions.

Firstly, the values for valence, arousal and dominance have been mapped to the more common range [-1, 1] from the original [1, 5]. After having preprocessed the text and embedded the sentences according to the methodologies described in the previous section, the resulting data has been used as input for the neural networks in order to train it. A k-fold cross validation has been adopted in order to estimate the distributions of the mean square error; in particular, after adopting a number of fold equal to 5, the results yielded mean losses of: 0.026, 0.03 and 0.031. The models have then been trained over the entire dataset.

Due to the presence of a subset of data having also information about Ekman's emotions, with values representing intensity ranging from 0 to 100, it has been possible to apply both the approaches previously described in order to map VAD values to specific emotions. The first approach consists in building a model capable of predicting the intensity or the presence of each of the six basic emotions given the values for the VAD scheme. This resulted in both a regression and a classification problem: the first having as targets the intensities of the emotions, the second having as independent variables the presence or absence of each emotion; in both cases a decision tree has been used as model. Unfortunately, either of them returned sufficient results: the regression tree performed awfully on the test set, giving a negative R-squared and a very high MSE; the decision tree classifier, instead, obtained high levels of accuracy, but low performances in terms of precision and recall since it was able to predict only negative examples but not positive ones.

Given the poor performances of the previous approach, the method that has been followed is the one consisting in positioning emotions in the VAD space and assigning observations to them according to distance of each point from them. An important assertion must be done with regards to those sentences which are located near the center and, therefore, can be considered neutral, not expressing any emotion. Those observation having a small distance from the center have been excluded from the evaluation and assigned to no emotion at all, while all the others have been labeled with the nearest emotion.

The second part of this research involves applying the discussed methods to song lyrics. Because of the unavailability of a public dataset containing this kind of text, due to copyright issues, the data has been retrieved using the Genius API from genius.com, a popular website collecting insights about music. For each of the five major music genres identified by the site (Rap, Pop, Country, R&B, Rock), the most popular 100 songs of all time have been retrieved, together with their respective lyrics. Each text has been divided into sentences and undergone the same procedures of tokenization, normalization and embedding that have been applied to the EmoBank dataset. Finally, both the models for predicting VAD values and the one for mapping these onto categorical emotions were applied to the lyrics dataset.

Concluding Remarks

The results were far from satisfactory when it came to applying the models on the lyrics dataset. The last model, in particular, was not able to capture any emotion at all, probably due to the strictness of the condition of non-neutrality imposed over the words before learning the model. The other two neural networks seemed to return values very close to the center of the VAD space for the majority of the sentences and, in turn, the emotion that appears most frequently, especially applying the second model, is 'disgust', which happens to be the one with the lower distance from the origin. The first model, the one learnt with the Word2Vec embedding, seemed to be slightly more balanced with respect to the second, build on the Doc2Vec embeddings, given its ability to assign songs also to the emotion of joy. However, given the inability of detecting all the other emotions, it seems useless to discuss about the relationships between emotions themselves and music genres.

The method used for predicting VAD values has been proved to be effective in other researches, therefore some probable explanation can be given to the poor performances in this case. The dataset is not sufficiently large and does not contain enough examples for a neural network. Using the SentiWordNet score in order to exclude neutral words is too restrictive; another method that gives less importance to these words should be explored. The mapping used seems to be inadequate since the most neutral sentences will be assigned to the emotion closer to the center in the majority of cases; the idea of building a classification model seems to be more appropriate to accomplish this task, perhaps training the model itself on a larger dataset.