Are song lyrics' emotions related to music genre?

Maria Giuseppina Brunelli - Mat 960160

Università degli Studi di Milano - La Statale
mariagiuseppina.brunelli@studenti.unimi.it
https://github.com/giusi07/Sentiment-Analysis-Emotion-Detection

Abstract. Emotions can be expressed in songs through music or lyrics, the goal of this project is to classify songs with emotions and to find out if there is a relation between emotions and music genre. The research methodology involved the training of an emotion classifier, then applied to a song lyrics data-set. Then, the use of Pearson's chi-square test and Cramer's V test allowed to analyze the dependence between the two categorical variables, which finally resulted to be very weak.

Keywords: Text classification · Song lyrics · Emotions.

1 Introduction

Nowadays, with the widespread of music streaming platforms, like Apple Music or Spotify millions of people have easy access to music, that has become a preponderant part of our lives. However, music has always been a powerful "tool" that allows artists to pour their emotions into a song to share with the public and arouse emotions for them. Emotions are expressed in songs through the melody or through the lyrics. In the first case the emotion expressed changes based on the instruments used, the tempo, the beat and on many other factors. In the second case, instead, the way in which words are laid down in the text and their semantics contributes to convey different emotions. However, emotions are subjective, thus different people might feel differently when listening to the same song. Moreover, sometimes the emotion that the artist is trying to convey is completely different from the one actually perceived by the public. This subjective nature of emotions, makes the task of emotion detection a non trivial one. Emotion detection in song lyrics and generally in text, aims to infer the underlying emotions influencing the author by studying their input texts. This is based on the premise that if a person is happy, it influences them to use positive words. Likewise, if a person is sad, frustrated or angry, the kind of words they use can allow us to infer their underlying negative emotion. Currently there are three main techniques used to detect emotions from text and they use features selected from syntactic and semantic data. They are: [1]

Keyword based approach: based on certain predefined keywords. These words are classified into categories such as disgusted, sad, happy, angry, fearful, surprised. It is easy to implement, intuitive and straight forward since it involves

identifying words to search for in text. However, it is domain specific and relies on the presence of keywords for accurate results, thus results cannot be transferred to other domains, making this approach not flexible.

Learning based approach: uses a previously trained classifier to classify input text into emotion classes. It is easier and faster to adapt to domain changes since it can quickly learn new features from corpora by supplying a large training set to a machine learning algorithm for building a classification model. However, acquiring large corpora may not always be feasible. Nonthetheless, this facilitates easy implementation of classifiers by novices who can then apply the learned model to new instances

Hybrid based approach: consists of a combination of the keyword based implementation and learning based implementation. The main advantage of this approach is that it can yield higher accuracy results from training a combination of classifers and adding knowledge-rich linguistic information from dictionaries and thesauri.

The aim of this project is to detect emotions expressed in music, exploiting song lyrics and then find out if there is a correlation between the music genre and expressed emotions. This paper is organized as follows: Section 2 contains an explicit definition of the research question and an overview of the used methodology. Section 3 describes the two data-sets used in this research, the metrics used to evaluate the performance and the experimental methodology along with the results. Finally section 4 contains a discussion of the obtained results, and ideas for future work.

2 Research question and methodology

The main goal of this project is to understand if emotions expressed in songs through lyrics depend on the music genre to which the song belongs, e.g. to understand if "Rock" songs are mostly sad or joyful songs. Therefore the research question is:
RQ: Are song lyrics' emotions related to music genre?
In order to test this dependence there is the need of song lyrics classified by emotions and music genre. The task of associating song lyrics to music genre is not too complex, since most of the time singers already belong to specific music genres (sometimes even more than one) and so do their songs. Hence, there are different music sources available to scrape data from, to link song lyrics to their authors and to the relative music genre (e.g. Genius.com or Vagalume.com). The problem, however is obtaining a corpus of song lyrics annotated with emotion. Indeed, since emotions are subjective, this task would require many manual annotators to classify songs and then take a summary of the different annotations, which in practice is quite hard to accomplish and extremely costly. To solve this problem, the proposed methodology of this research builds upon the use of a

"Learning Based Approach" to detect emotions. (defined in the previous section). In practice, the method unfolds into two steps: the first step is to exploit an already available corpus of documents annotated with emotions, to train a classifier capable of capturing specific patterns related to emotions. The next step, implies the use of a song lyrics data-set, with songs already classified by music genre, and further classify songs with emotions by exploiting the classifier trained in the previous step. Finally, a Pearson's chi-square test is employed to test the dependence between the two categorical variables (emotion and genre) and the Cramer's V test to assess the strength of their association.

3 Experimental Results

3.1 Datasets

As described above this research project unfolds into two steps and correspondingly two data-sets were used. The first data-set provided by WASSA-2017 [2] is made of Tweets annotated with four primary emotions i.e., anger, fear, joy and sadness, moreover, each tweet is associated with a measure of intensity of the emotion. The second dataset used for this task, is made available in Kaggle [3]. This data-set contains information about song lyrics and artists belonging to six different music genres. Lyrics are written in many different languages, however for this project only English ones were kept since the emotion detection classifier was trained on English documents. This data-set was made by scraping the Vagalume website. Some songs in this data-set appeared more than once, because classified with different music genres, e.g. "Pop" and "Rock". Given the goal of this research, those songs were reclassified with a unique multi-label e.g. "Pop, Rock", instead of removing one of the two occurrencies which would have introduced a bias.

3.2 Experimental methodology

The first data-set with emotion annotated tweets was used to train a classifier. To do so, common preprocessing techniques were applied to clean the text and after different classifiers have been tested. In order to evaluate the performances of the different classifiers, the metric considered was the f1 weighted score (since the dataset was slightly unbalanced). The train and development set provided by WASSA were merged together into a single training set used to train classifiers then evaluated using 5 fold cross-validation and to get the overall estimate of each classifiers over unseen data the test set has been used. First a baseline model was defined by using the Multinomial Naive Bayes algorithm using as features the Tf-Idf score. Whereas, the actual classifiers compared against the baseline were obtained by combining two algorithms and three different feature extraction techniques. In particular, the algorithms used were Support Vector Machine and XGBoost, the first one, is a supervised machine learning algorithm which can be used for both classification or regression, and aims at finding an

hyper-plane that differentiate the classes. In this context it was applied to a multi-class classification task and with a radial kernel. The second algorithm is also a supervised machine learning algorithm and is a decision-tree-based ensemble, that is an optimized gradient boosting algorithm through parallel processing, tree-pruning and regularization to avoid overfitting. The feature extraction techniques involved Tf-Idf score, word embedding and document embedding. The Tf-Idf score is a frequency score assigned to each word in the corpus that works by penalising the common words by assigning them lower weights while giving importance to words which are rare in the entire corpus but appear in good numbers in few documents. Word embedding, instead, works by creating for each word a real-valued vector that encodes its meaning such that words that are closer in the vector space are expected to be similar in meaning. Obtained vectors are then merged together, to embed the entire document. Vectors were merged in two different ways, once by taking a simple average of all the vectors of words that appear in the document, while the second way exploited words' Tf-Idf score to get a weighted average of the vectors. Finally, document embedding is practically an extension of word embedding and as such defines a vector for the entire document. Both techniques, word and document embedding practically require the training of a neural network that can be done by exploiting the Gensim library. However there is also the possibility of exploiting pre-trained models. Indeed, a pre-trained word embedding model, over a Twitter corpus was also used. In total, by combining different algorithms, different embedding techniques and different averaging ones, twelve different classifiers have been compared. The best classifier obtained was then applied to the second data-set and finally a cross table was produced to register the co-occurrences of each music genre with each emotion. Finally, after a visual analysis of the table, exploiting plots, statistical tests were used to better define the relation between the two variables. In particular the Pearson's Chi-Square test of independence assesses whether observations consisting of measures on two variables, expressed in a cross table, are independent of each other and the Cramer's V test that returns a number between 0 and 1 that indicates how strongly two categorical variables are associated and it is based on Pearson's chi-squared statistic.

3.3 Experimental Results

Figure 1 contains the F1 weighted score results of the 5-fold cross validation for each one of the twelve classifiers, as well as the test score. From the table it is clear that the best classifier is given by the combination of the XGBoost algorithm and Tf-Idf feature extraction technique with a test score of 0.82, which performs better than the baseline model that has a test score equal to 0.71. The XGBoost algorithm performs even better than SVM, whose f1 weighted score is 0.78. However SVM is less affected by the imbalance of the dataset, unlike XGBoost. Indeed as mentioned previously, the WASSA data-set is slightly imbalances since the class "fear" is the most represented one. XGBoost therefore, makes more mistakes than SVM by classifying more documents with the majority class (fear). The second embedding technique used, i.e. document embedding

created by training a model, instead, yields very bad results with both algorithms indeed, f1 score is 0.25 for SVM and 0.27 for XGB. Moreover, in both cases most instances are misclassifed with the label "fear". This is most likely due to the fact that document embedding captured more the pattern of documents classified with the emotion "fear" since it is the class that is most represented in the data-set, indeed the class "sadness" which is the least represented is the worst predicted one. This suggests that the document embedding is not well suited for this type of data-set, since it is imbalanced (and therefore some patters are easier to spot than others) and that is too small (indeed there are less than 4000 documents) to train the model appropriately. As described above the third embedding technique used is word embedding, once with a custom trained model and once with a pre-trained model. The custom trained model resulting vectors were used once by taking a simple average and once by taking a Tf-idf weighted average in order to embedd documents. The simple average gave better results than the weighted average for both algorithms as shown in figure 1, however in both cases results were worse than the baseline. This is most likely due to the fact that the data-set used to train the word embedding model is too small, and therefore it is incapable of capturing most of the relationships between words, ending up with poor performances, which were then emphasized even more by the Tf-idf score. To overcome the problem of the small data-set to train a word embedding model, a pre-trained model over a corpus of Tweets provided by Glove has been used. This model yields slightly better results compared to the custom trained model, with simple average of vectors. However f1 weighted score is still lower than the baseline one, even worse when Tf-idf weighted average is considered. Therefore, although the pre-trained model works better than the self-trained one is not domain specific, in this case emotions, so it does not improve performances compared to the baseline model.

f1-weighted score	FEATURES						
MODEL	Cv score / Test score	Tf-Idf	Doc2Vec	Custom Word2Vec Average	Custom Word2Vec Tf-Idf	Pre-Trained Word2Vec Average	Pre-Trained Word2Vec Tf-Idf
	Baseline (Multinomial NB)	0.74 0.71	-	-	-	-	-
	Svm	0.81 0.78	0.29 0.25	0.53 0.53	0.21 0.20	0.63 0.63	0.21 0.20
	Xgboost	**0.82** **0.82**	0.27 0.27	0.53 0.46	0.26 0.25	0.56 0.53	0.24 0.23

Fig. 1. 5-fold cross validated and test F1 weighted score of classifiers

Using the best classifier, XGB with Tf-Idf score, song lyrics from the second data-set were classified. Thus every song has been associated with one of the four labels that defines the emotion expressed in the song itself. Finally a cross-table was produced to explore the relation between emotion and music genre. From the plot of the cross table that contains all music genres (see Fig.2) and the stacked bar plot (see Fig.3) it looks like the number of songs of each class of emotion are proportional to the number of songs of each music genre. However since there are, in the plots, some music genres with few occurrences, they create "noise" in the graph that makes this proportionality harder to see.

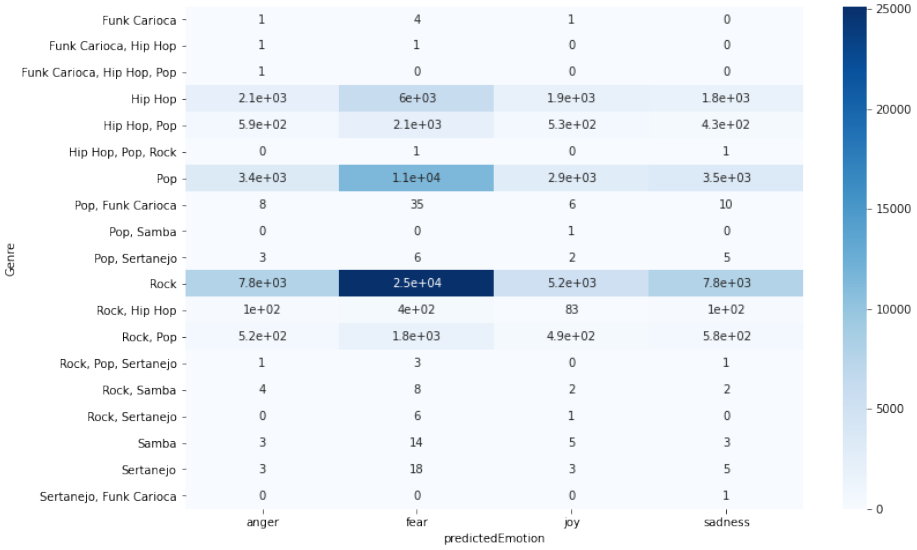

Fig. 2. Cross table: Emotion - Genre (all genres)

However when removing those genres with few occurrences and keeping only Rock, Pop and Hip Hop which have most occurrences, the plots are more clear and it is evident that there is not a particular relation between music genre and emotions. (see figure 4 and 5).

To confirm those results the Pearson's Chi-Square test was used which rejects the null hypothesis, thus suggesting that there might be a dependence between the two variables. Thus, suggesting the contrary of what the plots shows, however this could be due to the fact that a weak association in a large sample size may also result in $p = 0.000$. Therefore, in order to test how strong this dependence is the Cramér's V test was used, which returns a number between 0 and 1 that indicates how strongly two categorical variables are associated. In this case the test score is equal to 0.035, meaning that there is an extremely weak association between the two variables, thus confirming the intuition of the visual analysis.

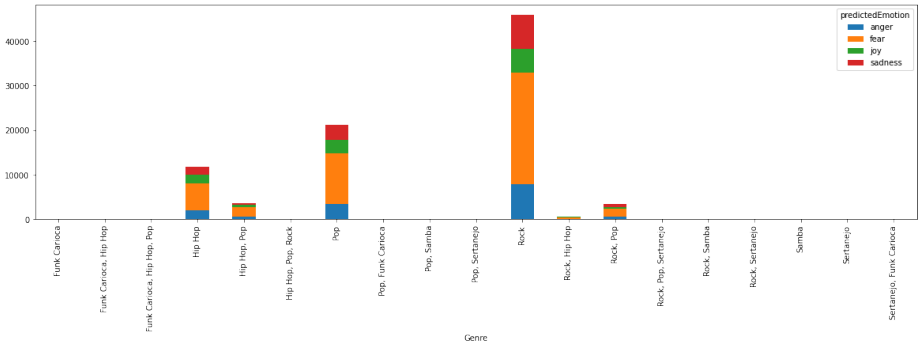

Fig. 3. Stacked bar plot (all genres)

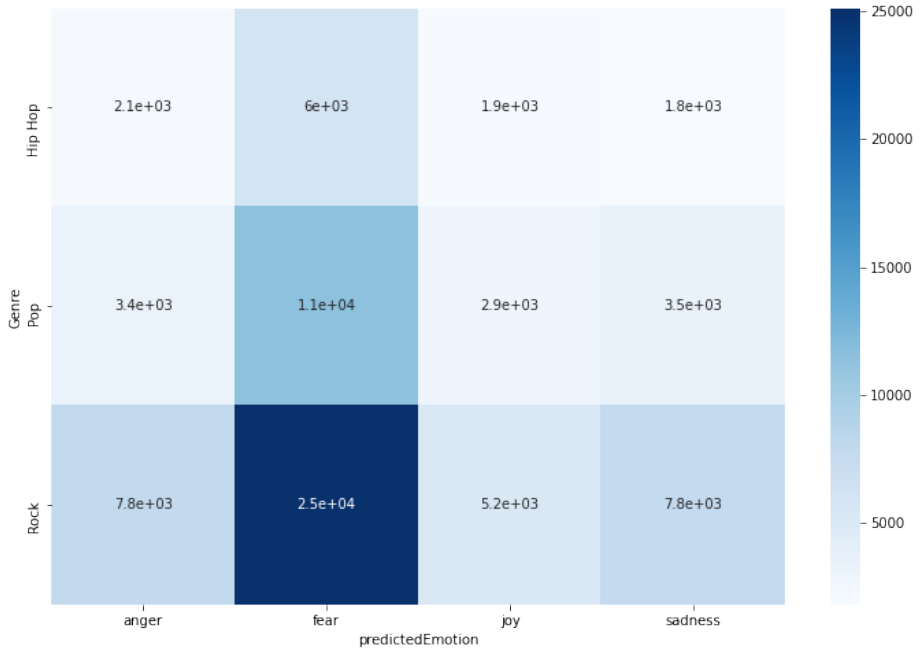

Fig. 4. Cross table: Emotion - Genre (Rock, Pop and Hip Hop)

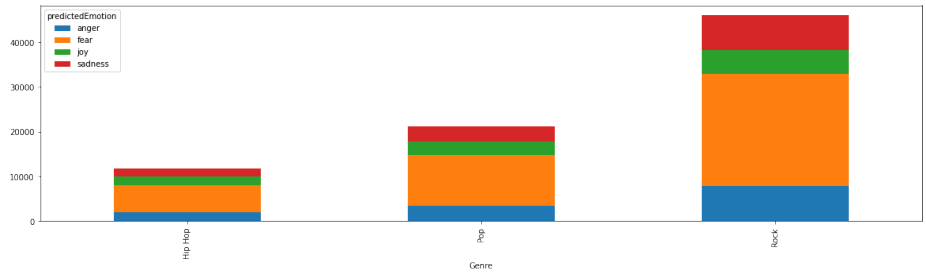

Fig. 5. Stacked bar plot (Rock, Pop and Hip Hop)

4 Concluding remarks

The goal of this research was to find out if emotions are related to music genres and the challenges posed by emotion detection were overcome by using a "Learning Based approach", thus by using a corpus previously annotated with emotions to train a classifier and then used to classify song lyrics. In this case the data-set employed was made of tweets and it contained less than 4000 documents. The small size of the data-set did not allow exploitation of word and document embedding techniques because models could not be properly trained. Moreover, results show that music genre and emotion are not related, however this result is not definitive and should be explored further. For future works, indeed, it could be helpful to introduce also audio features together with the lyrics to get the full picture and understand how genre and emotions are relates, especially because most of the time music genre is mainly defined by the audio characteristics of the song rather than simply the lyric. Furthermore, it could be interesting the idea of extending the tweets data-set or even better to create a big corpus of manually annotated song lyrics, and train a classifiers on it, instead of using tweets. If the corpus is big enough word and document embedding techniques could work better and in addition deep learning algorithms could be used to train a classifier.

References

1. Haji Binali, Chen Wu,Vidyasagar Potdar: Computational Approaches for Emotion Detection in Text. 4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2010)
2. WASSA-2017, http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html.
3. Song lyrics from 6 musical genres, https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres?select=artists-data.csv.