Lyrics-Based Music Genre Classification

Raluca-Andreea GÎNGA

Faculty of Mathematics and Computer Science, Bucharest

January 23, 2022

Abstract

In this documentation, there are presented several solutions for detecting and classifying songs genres based on their lyrics. The approaches presented in the next few pages include methods like TFIDF to BERT and other pretrained models for Word Embeddings followed by the application of classical machine learning algorithms (some of them giving very good results), deep learning techniques used on some several datasets tha contain both the meta information extracted from the lyrics, such as number of syllables, rhymes, sentiment of the song, profanity checker, but the lyrics representation as well.

1 Introduction

Music is an important part of people's lives. Regardless of the type of music that people prefer, it influences them in different ways, in most cases without even realizing it. In Text Mining, the classification of song genres based on its lyrics is considered to be a difficult and challenging task. Developing an automatic software of music genre classification is a well researched topic in music information retrieval.

The present task has the main purpose to use Natural Language Processing and Text Mining techniques in order to assign the genre to a song based on its lyrics. Methods like Support Vector Machines, K Nearest Neighbors and Naive Bayes were used so far by various researchers in order to classify the lyrics into one of the 10 categories. Through this project, we decided to apply another methods considered "state of the art" in literatura and try to extract more information from the lyrics obtained.

The remainder of this document is organized as follows: section 2 describes the dataset and the exploratory data analysis and feature engineering parts. Section 3 describes the details of the implemented Machine Learning and Deep Learning methods, followed by their corresponding result on section 4. The last section (5) concludes the documentation about the implemented methods and comes with some further work that could be done in order to research more about this task.

2 Dataset and Data Analysis

2.1 Overview of the dataset

The dataset of this task is consisting in two .csv files:

- "Lyrics-Genre-Training" file containing the training dataset of approximately 18 thousands of songs

- "Lyrics-Genre-Test-GroundTruth" file containing the testing (validation) dataset of approximately 8 thousands of songs

For each song, the datasets contain the title, the artists, the year, the complete lyrics and finally, the genre that should be predicted.

An overview of the dataset format is shown in the figure from below.

	Song	Song year	Artist	Genre	Lyrics	Track_id
0	forest-enthroned	2007	catamenia	Metal	I am a night in to the darkness, only soul los...	18096
1	superhero	2010	aaron-smith	Hip-Hop	Yeah\nSometimes, i just wanna fly away.\nThey ...	22724
2	chicago-now	2007	fall	Metal	Do you work hard?\nDo you work hard?\nYou don'...	24760
3	the-secret	2007	geto-boys	Hip-Hop	You know what? I'm destined to be the last man...	24176
4	be-the-lake	2011	brad-paisley	Country	There ain't nothing that I would rather see\nT...	17260
...	...	...	...	...	...	...
18508	i-wish-he-didn-t-trust-me-so-much	2008	bobby-womack	R&B	I'm the best friend he's got\nI'd give him the...	12033
18509	i-totally-miss-you	2006	bad-boys-blue	Pop	Bad Boys Blue\n"I Totally Miss You"\nI did you...	15987
18510	sorry-for-love	2002	celine-dion	Pop	Forgive me for the things\nThat I never said t...	2722
18511	cure-for-aids	2008	dan-bern	Indie	The day they found a cure for AIDS\nThe day th...	10221
18512	iceberg-meadows	2015	crawdad-republic	Pop	Fourth of July has come, it's custom that we g...	13657

18513 rows × 6 columns

Figure 1: An overview of the dataset

2.2 Data Analysis & Feature Engineering

In order to understand more about the dataset, we decided to do some exploratory data analysis and create some new features as well to the dataset.

As we can see from the below figure (figure 2, the classes are not so well balanced, being present in our training dataset a lot of rock songs. We also noticed that almost 50% of the songs from training dataset were released in 2006 and 2007 years. We can also notice that there's an year (112) that is not like a year, so we can easily delete it since it's an outlier.

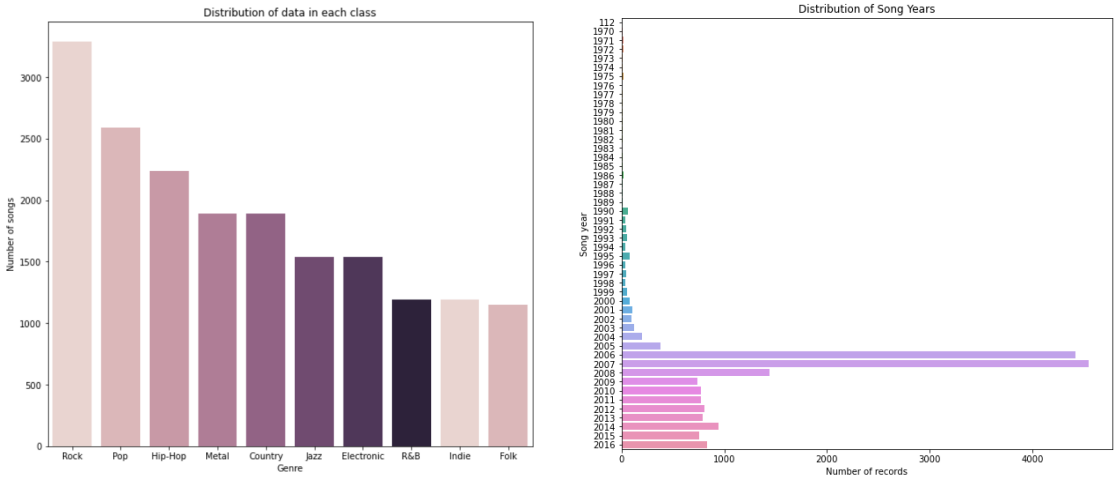

Figure 2: Distribution of songs regarding their musical genres and years

As we can see in figure 3, words like "love" are more predominant in almost all of the genres (excepting metal and hip-hop), whereas hip-hop contains more swearing words and metal contains words regarding time and life meditation.

Figure 3: Most common words per each genre

Another step that we took was to analyze how many songs contain swearings and profanity language. This thing was done using profanity check[1] library. As we can in the figure below, in the training dataset there were found 3519 songs with swearings, 43% of them being in the hip-hop genre and rock with metal with 28% (as noticed in figure number 4. The binary values generated by the profanity check were added further in the training and testing datasets in order to add new features.

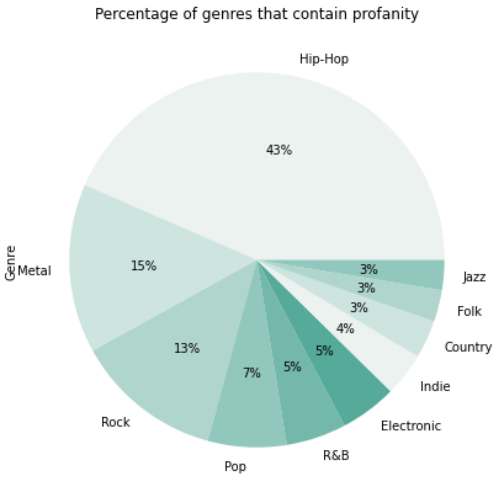

Figure 4: Genres that contain profanity

Another feature we decided to use was sentiment analysis and polarity of each song lyrics. For this, we used Flair[2] which predicts in a very professional and precised way the sentiment of each song lyrics. After the prediction, we noticed there were more negative songs than positive ones (as it can be seen in figure 5, but the balance is not that highly skewed, so we can say we don't have so many discrepances to worry about.

Another features that we've created were the number of syllables and number of rhymes.

As preprocessing step, we used the following techniques:

- text decoding
- special characters cleaning

[1]Profanity Check https://pypi.org/project/profanity-check/.
[2]Flair https://github.com/flairNLP/flair.

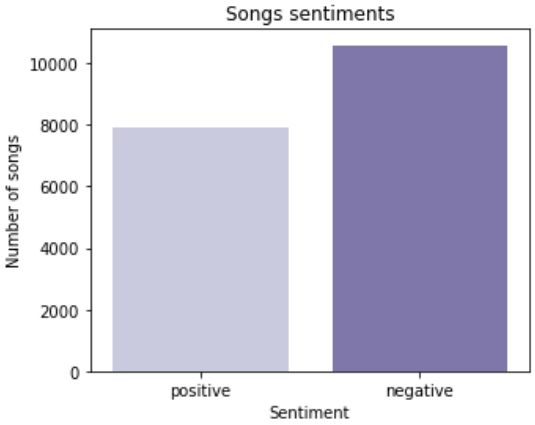

Figure 5: Songs Sentiments

- text lowercase
- tokenization
- stopwords removal
- removal of words shorter than 3 characters
- lemmatization

Thus, we created a final dataset (including the lyrics and meta information) having the following features:

- average of syllables
- number of rhymes
- number of words
- number of verses
- a binary feature representing if the song contains profanity or not
- a binary feature representing the sentiment of the songs (1 for positive and 0 for negative)
- the year continuous variable
- cleaned lyrics
- the genres encoded from 0 to 9 (labels):

0: Folk	5: Rock
1: Country	6. Metal
2: Jazz	7: Pop
3: Hip-Hop	8: Indie
4: R&B	9: Electronic

3 Approaches

3.1 Machine Learning models on three types of datasets

The first approach used was to apply different machine learning models on three types of datasets:

1. Meta-information (without lyrics) - this includes average of syllables, number of rhymes, number of words, verses, profanity, sentiment, year
2. Lyrics only - this includes only the lyrics, without any feature
3. Combination between the meta information about the lyrics and the lyrics as well

The Machine Learning models that were used for all of these 3 types of dataset representation were:

- Multinomial Naive Bayes
- Stochastic Gradient Descent
- Support Vector Machines
- Random Forest
- Light GBM
- XGBoost

We note that all of these models were supposed to cross validation in order to have a precise representation of the real accuracy.

For the lyrics dataset, we performed TF-IDF Vectorizer with 'word' analyzer and a features maximum of 5000 in order to obtain the numerical representations of the lyrics.

3.1.1 Hyperparameter Optimization

Because it can be noticed in results section that Light GBM provided very good results on combined and lyrics datasets, we decided to use Bayesian Optimization in order to find the proper parameters for Light GBM model. In the following tables, we could see the chosen parameters that provided the best results on a 5-fold cross validation.

	learning rate	number of leaves	feature fraction	bagging fraction	maximum depth
Combined Dataset	0.023	42	0.307	0.801	14
Lyrics Only Dataset	0.024	42	0.307	0.801	15

3.2 Deep Learning techniques

Because we've seen that the combination between meta information and lyrics have provided good results on Machine Learning algorithms, we decided to explore some Deep Learning techniques/algorithms in order to see if they provide better results. Also, some approaches used did also use the lyrics only representation of the songs.

1. 3-Layer Neural Networks

   This algorithm used 3 dense layers with 128, 64 and 32 neurons with 'relu' activation and a final layer with 'softmax' activation. The optimized used was Adam, the loss was sparse categorical crossentropy and accuracy used as main metric. We also provided to the neural network model a class weight because we noticed that imbalance regarding the genres of the lyrics and we wanted to have a more appropriate way to balance them and avoid the situations in which the network predicts only the majority class.

2. Universal Sentence Encoder as word embedding and 2-Layer Neural Network

   Universal Sentence Encoder is a model that is producing sentence embeddings that demonstrated over a lot of Natural Language Processing projects a good transfer to a number of other NLP tasks. Here, we used a Sequential Neural Network with Universal Sentence Encoder applied on lyrics, followed by 2 dense layers of 128 and 64 respectively neurons.

3. Bert Embeddings + Bert for Sequence Classifier

   In this approach, we used BertTokenizer for tokenizing the lyrics datasets and a classification BERT model - BertForSequenceClassification - that contains a single linear classification layer on top and that, surprinsingly, didn't provide the expected results, although this methods was used in a lot of Natural Language Processing competitions.

4. Sentence XLM-R + Bidirectional Gated Recurrent Unit

   For this technique, we used SentenceTransformers and provided a 'XLM-r-bert-base' model for embedding both the training and testing datasets. Then, as a deep learning architecture, we used 2 bidirectional GRU of 64 and 128 neurons, then a Dense layer of 256 neurons, a batch normalization, a dropout of 0.2, another dense layer of 64 neurons along with batch normalization and dropout, followed by a dense layer of 16 neurons and again, a batch normalization and concluding with the 10 classes. The model was trained on 50 epochs with Reduce LR On Plateau and Early Stopping in case the validation loss won't improve.

4 Results

4.1 Machine Learning models on 3 types of datasets

The results from the following table (1) contain the accuracy on 5-fold cross validation on the train dataset. As it can be seen from the below table, Light GBM seems to perform the best among all of the other Machine Learning algorithms. We can also notice that Multinomial Naive Bayes, Stochastic Gradient Descent, Support Vector Machines perform at their best only on lyrics TF-IDF representation, whereas Random Forest, Light GBM and XGBoost did a good job on all of the 3 dataset. Comparing the modeling time of each algorithm, Light GBM is a very fast model and it also provided very good results. We declared it as the winner for all of the datasets, both in computation time, but in best accuracy as well.

	Multinomial NB	SGD	SVM	Random Forest	Light GBM	XGBoost
Meta-Information	17.14 %	19.88 %	12.41 %	28.13 %	**32.24 %**	31.75 %
Lyrics	37.61 %	38.26 %	37.54 %	39.6 %	**41.1 %**	39.92 %
Meta + Lyrics	17.25 %	18.76 %	15.38 %	40.84 %	**44.1 %**	43.18 %

Table 1: Machine Learning results on 5-fold cross validation.

In table 2, there are presented the results of the algorithms on the ground-truth dataset (test dataset). Again, it can be seen that the best performing algorithm is Light GBM, providing even better results on test dataset than in the training dataset.

	Multinomial NB	SGD	SVM	Random Forest	Light GBM	XGBoost
Meta-Information	16.89 %	1 %	8.48 %	27.75 %	**32.54 %**	31.39 %
Lyrics	38.76 %	38.66 %	37.94 %	39.33 %	**41.88 %**	40 %
Meta + Lyrics	16.92 %	21.15 %	13 %	40.97 %	**44.55 %**	43.28 %

Table 2: Machine Learning results on test dataset.

In figure 6, it is shown the confusion matrix of Light GBM on both only lyrics dataset and combined lyrics with meta information.

4.2 Deep Learning techniques

In the table from below (table 3), there are presented the results given by the various Deep Learning techniques. As we can see, we tried to experiment more only with the lyrics given

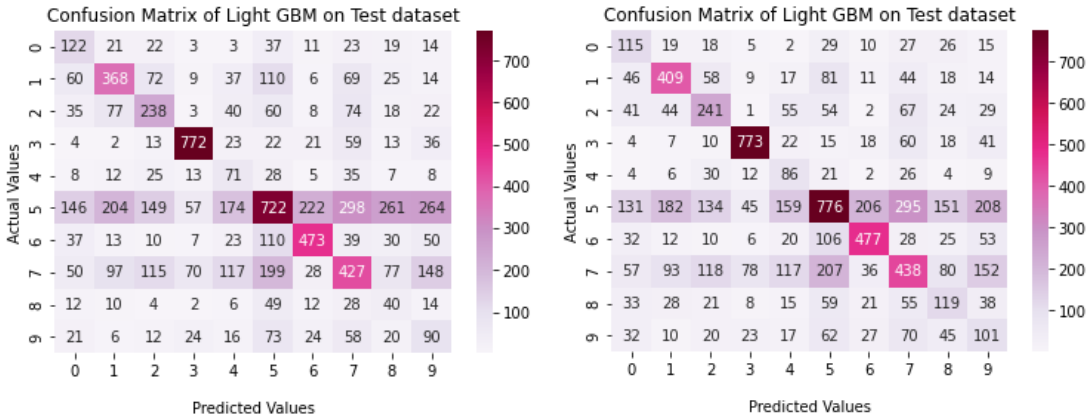

Figure 6: Confusion Matrix of Light GBM on Lyrics only dataset and Combined Meta Information + Lyrics datasets. (Left image - lyrics only and Right image - meta + lyrics)

	3-Layer NN		U.S.E. + 2-Layer NN		Bert Embeddings + Bert Classifier		XLM-R + BiGRU	
	Train	Val	Train	Val	Train	Val	Train	Val
Meta-Information	-	-	-	-	-	-	-	-
Lyrics	-	-	99.7%	37.61%	**42%**	**38%**	**48%**	**37.6%**
Meta + Lyrics	**52.22%**	**38.56%**	-	-	-	-	-	-

Table 3: Deep Learning techniques results.

in the dataset, obtaining the best results on Bert Embeddings and Bert for Sequence Classification technique (it is clearly from the differences between the train and test accuracies) and also, on XLM-R and that complex model with 2 bidirectional Gated Recurrent Units with Batch Normalization.

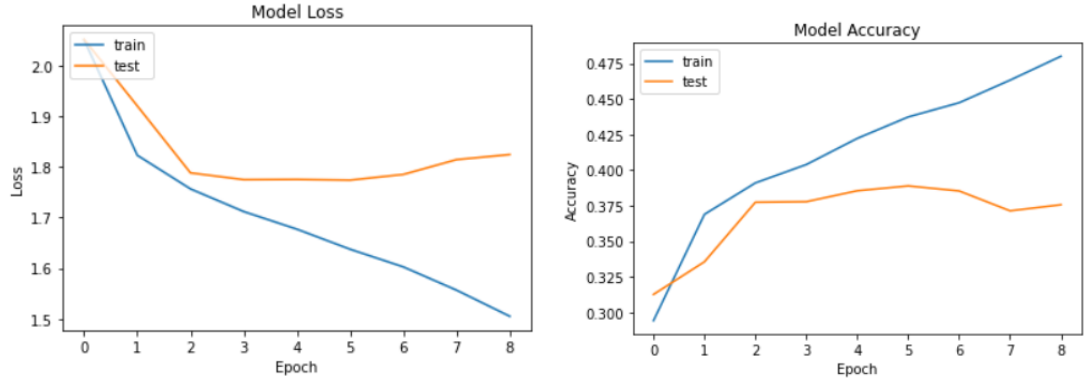

Figure 7: 3-layer NNs on Meta Information + Lyrics dataset combination. (Left: Loss on train - test. Right: Accuracy on train - test. - Early Stopping applied)

5 Conclusion and Further Work

In the current project, it was solved the problem posed by "Textract 2018 Machine Learning Hackathon: Task 1" competition. There were applied various methods, including the application of feature extraction for lyrics like TF-IDF, Bert Embeddings, XLM, Universal Sentence Encoders followed by classical Machine Learning models to Deep Learning architectures. The enormous potential was given by Light Gradient Boosting Machine on the combined meta in-

formation (number of syllables, number of rhymes, verses, song sentiment, profanity detection) with lyrics representations, obtaining 44.55% accuracy on test dataset. Deep Learning didn't seem to impress on this task, even trying state of the arts pretrained embeddings, like Bert Embeddings, XLM-R.

As a potential future work, we can experiment and try the following ideas:

- Data augmentation - adding new data containing all of those 10 classes
- Adding audio features
- Topic modeling
- Mixup strategies for text classification
- Hierarchical Attention Networks that are used for document classification

Bibliography

[1] Anthony Canicatti, *Song Genre Classification via Lyric Text Mining*, Int'l Conf. Data Mining | DMIN'16 |, 2016

Table of Contents