Natural Language Processing

Project

Ramiz Mammadli – 18011903

Lyrics-Based Genre Classification

Abstract

As people have access to increasingly large music data, song classification becomes critical in this industry. Especially, automatic genre classification is an important feature in song classification and has attracted much attention in recent years. As it can be guessed, to tell which genre the song belongs to, we need to listen to the music of the song, as well as, read its lyrics. However, according to the recent researches, there can be a predictable connection between the lyrics of the songs. This project aims to build a system that can identify the genre of a song based on its lyrics. By using various Natural Language Processing techniques and tools, we developed a system that guesses the genre of the English songs based on its lyrics. There are five major genres in the program to classify the lyrics into: Rock, Pop, Jazz, Indie and Hip-Hop.

1. Dataset

1.1. Dataset Retrieval

To start the project, we had to find the data, which include the lyrics of the songs and its matching genre. First, the raw dataset is found in Kaggle.com [1] which includes Artist, Song name, Genre, Language and Lyrics columns (Picture 1.1).

Artist	Song	Genre	Language	Lyrics
12 stones	world so cold	Rock	en	It starts with pain, followed by hate\nFueled ...
12 stones	broken	Rock	en	Freedom!\nAlone again again alone\nPatiently w...
12 stones	3 leaf loser	Rock	en	Biting the hand that feeds you, lying to the v...
12 stones	anthem for the underdog	Rock	en	You say you know just who I am\nBut you can't ...
12 stones	adrenaline	Rock	en	My heart is beating faster can't control these...

Picture 1.1

Naturally, there were some downsides of this dataset. For instance, it included 121,404 Rock, 108,714 Pop, 13,545 Jazz, 8,449 Indie and 2,240 Hip-Hop songs. Apparently, there is a huge imbalance in the dataset. Rock and Pop songs can be decreased to a reasonable number, which will eliminate the mentioned problem. But then, the dataset will be smaller than it is supposed to be to train and get a satisfying precision and accuracy rates, because of the little number of Jazz, Indie, and, most importantly, Hip-Hop songs. Hence, the second dataset is found, again via Kaggle.com [2]. This time, there were two separate datasets: the one with the artist information, genre, and the key column to represent the artist, and the other dataset with the song lyrics, language of the song and the key column to represent the artist whom the song belongs to. In this case, two datasets were needed to be joined according to their matching ‘key’ columns. In the end, the table shown in Picture 1.2 is got.

ALink	SName	SLink	Lyric	Idiom
/10000-maniacs/	More Than This	/10000-maniacs/more-than-this.html	I could feel at the time. There was no way of ...	ENGLISH
/10000-maniacs/	Because The Night	/10000-maniacs/because-the-night.html	Take me now, baby, here as I am. Hold me close...	ENGLISH
/10000-maniacs/	These Are Days	/10000-maniacs/these-are-days.html	These are. These are days you'll remember. Nev...	ENGLISH
/10000-maniacs/	A Campfire Song	/10000-maniacs/a-campfire-song.html	A lie to say, "O my mountain has coal veins an...	ENGLISH
/10000-maniacs/	Everyday Is Like Sunday	/10000-maniacs/everyday-is-like-sunday.html	Trudging slowly over wet sand. Back to the ben...	ENGLISH

Picture 1.2

After doing some preprocessing mentioned later on, there are 21,052 Rock, 8958 Pop, 6,610 Hip-Hop, and 3,095 Indie songs, all in English. There is no Jazz songs in this dataset, but it is not so limiting problem, as there were enough number of Jazz songs in the previous dataset.

Finally, two datasets are concatenated into together and one last data are prepared to go through preprocessing.

1.2. Data Preprocessing

The data is preprocessed in so many ways. First, in the final dataset we use, there are songs in a lot of languages. Since we developed the system for only English songs, the filtering had to be applied to get English songs only, by using the ‘Language’ column.

After preprocessing the data according to the language column, we will only need the Lyrics and Genre of the songs, for the use of the data in the project. Therefore, everything besides these two columns is dropped from the dataset.

There are more than 30 genres in the dataset, in total. In some of ‘indigenous’ genres, only a small amount of the songs is given. Training the models according to these genres is almost impossible and will give extremely inaccurate results. For this reason, we decided to take only 5 major genres into account, which are already mentioned before for several times.

Finally, punctuation removal process is applied to the dataset. The punctuations, such as comma and colons are dropped from the lyrics text of the songs. Also, question mark “ ? ”, brackets “ ( ) ”, box brackets “ [ ]”, and new line “\n” are removed from the lyrics by using Regular Expression specializations.

To conclude, after doing all these preprocessing steps, as it is mentioned in previous section, still there is an imbalance problem. There are 126,116 Rock and 97,525 Pop songs, which are far more than the rest of the genres. To eliminate this, the samples are taken out of Rock and Pop songs.

Eventually, we have 19,505 Pop, 18,016 Rock, 13,314 Jazz, 12,370 Indie and 11,081 Hip-Hop songs in the final dataset (Picture 1.3), which are not so small to train the model and have only a little imbalance.

Genre	Lyrics
Indie	leave now while you can 'cause growing old jus...
Indie	we've been poisoned tracked down herded to sta...
Indie	the needle's in hand but i cannot sew. my hear...
Indie	i catch every whisper surrounding your head yo...
Indie	get off of work come home pass out. my life's ...
...	...
Pop	rock me baby rock me rock me rock me baby rock...
Pop	it's sad to think we're not gonna make it and ...
Pop	tonight there's gonna be a whole lot of smoke ...
Pop	heaven help the soul that's severed from the p...
Pop	1 never felt like this before girl when you w...

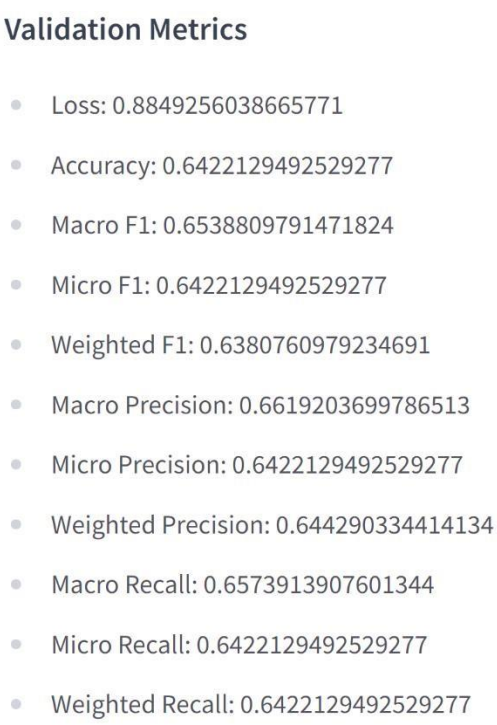
Picture 1.3

2. Natural Language Processing

In this section, the methods of Natural Language Processing applied on this project will be discussed.

Since there were more than two classes to break the data into, we concentrated on the multi-class text classification models. Instead of going for one or two models and trying to get the satisfying output, we discovered the new tool of Huggingface, called AutoNLP [3]. The preprocessed data are uploaded to the given interface and the best models are chosen for the purpose. The used models are more specific comparing to the well-known NLP models, such as Bert or Word2Vec. However, the performance and parameters are quite impressive.

Two models with the best output are used in the project, which are called EAGLE (Picture 2.1) and Barracuda (Picture 2.2).



Picture 2.2

Picture 2.1

As it is shown above, for example, the accuracy of Barracuda model is approximately 64.4% and EAGLE model is about 64.2%, which are reasonable numbers for multi-class text classification. Additionally, F1 scores and precisions are in a good level that let us experience satisfying outputs once the models are processed.

3. Code and User Interface

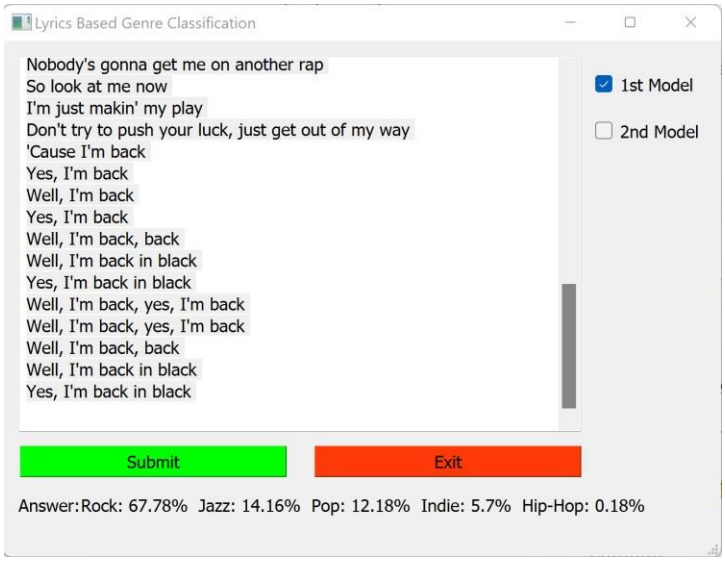
3.1. Code

A User Interface is developed to show the required outputs in a more effective way. To code the interface, PyQT5 library of Python is used (Picture 3.1).

Picture 3.1

As it is shown above, `setupUi()` is a PyQT5 method that prepares the window which the widgets will be shown. Then, `retranslateUi()` method is used to rename the widgets used in the user interface. In the `guess_genre()`, the JSON object is retrieved by sending POST request to HuggingFace API by using special token issued privately. In `print_func()` method, raw JSON object is used to be filtered and taken the required data out of it. `closeApp()` method serves to the termination of the application when the related button is clicked.

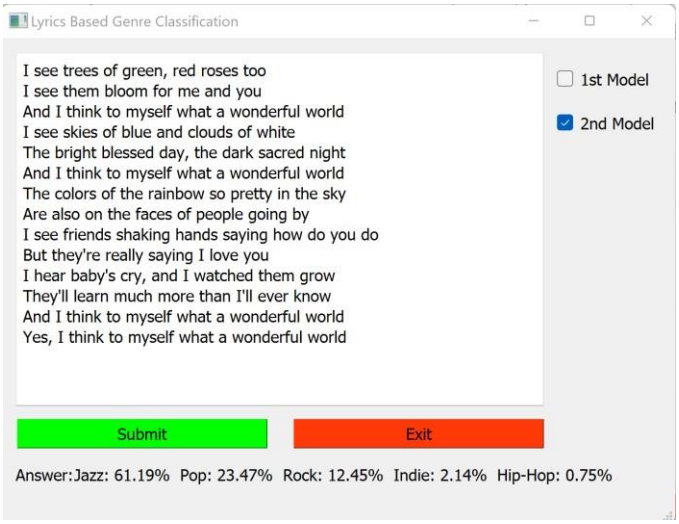
3.2. User Interface



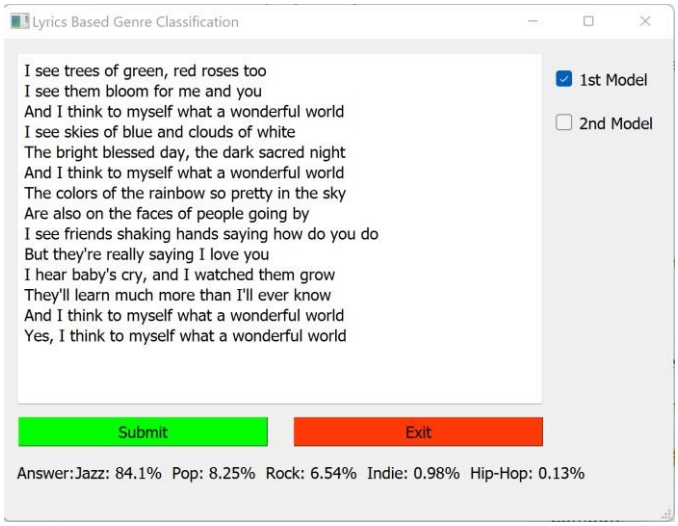
Picture 3.2

In the Screenshot above, we can see the lyrics of the song called ‘Back in Black’ by rock band AC/DC. Two buttons are used: Submit button to show the genre of input lyrics and Exit button to terminate the application. Below the buttons, we can see the result that shows the percentage in front of every genre to represent the genre of the given song. In the right side of the window, we can see two checkboxes: 1st Model represents the Barracuda, and the 2nd one stands for the EAGLE Model.

Other examples for both models (Picture 3.3 and 3.4) are shown below for the song called ‘What a Wonderful World’ by a famous jazz artist Louis Armstrong.



Picture 3.3



Picture 3.4

4. References

[1] Matei Bejan. (2021, January). Multi-Lingual Lyrics for Genre Classification, Version 1. Retrieved December 10, 2021, from https://www.kaggle.com/mateibejan/multilingual-lyrics-for-genre-classification/metadata

[2] Anderson Neisse. (2019,November). Song lyrics from 6 musical genres, Version 3. Retrieved December 14, 2021, from https://www.kaggle.com/mateibejan/multilingual-lyrics-for-genre-classification/metadata

[3] HuggingFace. (2021, March).HuggingFace: AutoNLP. Accessed December 14, 2021, https://huggingface.co/autonlp