Machine Learning Engineer Nanodegree

Capstone Proposal

Francisco Nogueira

30th March, 2021

Proposal

Domain Background

Ever since I was a child, I was always passionate about music and even learned how to play guitar at a young age. As time passed by the guitar practice had slowly abandoned my hobbies but listening to music never disappeared and currently, I still listen to songs based on my mood and find artists that represent the same mood I am having. The Music field has been analyzed extensively and several studies appear showing the benefits in the subjective well-being of humans, making it a major reason for me to deep dive more in the area.

As new sources of data appear and are made publicly available what was once quite a subjective field, it is becoming more and more quantitative as the years pass by. The topics regarding the music industry diverge from predicting sales of music albums by using Time Series to creation of Recommendation systems for users with similar musical tastes, such as what is being used by Spotify.

Another area of research within the music industry, Natural language Processing (NLP) has made it possible to make sense of unstructured data and create features that help in several problems, such as sentiment analysis. A good overview on sentiment analysis within the music industry can be found in the project made by Buhrer.K , Johns.M, and Stephens.S[1]. Tackling a more specific problem as to create a "mood predictor" based on music lyrics, Raschka.S did a great project predicting if a song was happy or sad based on the Million Song Dataset[2].

Problem Statement

The problem I would like to propose is if I am able to predict weather a song is positive or not. The solution will be to develop an NLP model that will receive as inputs the music lyrics and will output a prediction stating if the song is positive or not, making it a quantifiable problem. The problem can be measurable, in the sense that we are able to extract the level of positiveness of a song through publicly available websites, as described in the section of the data. Finally, the problem can be reproduced with the same type of inputs making it a replicable problem.

[1] https://michaeljohns.github.io/lyrics-lab/
[2] https://github.com/rasbt/musicmood

The complement of this solution will be a web app using AWS services as learned throughout this nanodegree program.

Datasets and Inputs

Data is not available as easily as I thought to at the beginning. Even though I had a huge benchmark dataset (the 1Million Song Dataset), this dataset does not provide the labels that I needed. To get the level of positiveness of a song, I would have need to use the Spotify API to extract the valence feature algorithmic estimation[3].

With that in mind, I searched for a dataset that on one hand, would have a relevant size and on the other would provide the information I need. Even though I could not find data regarding the lyrics of the song, I plan to extract those as part of the data extraction stage of the project. In summary, my main inputs will be the following:

- Kaggle Dataset with songs from 1921 to 2021
- Lyric extraction using lyrics-extractor

The first will be to get the relevant song names, artists, and level of positiveness (my target) and the second will be to extract lyrics which will be my main model inputs. The main dataset has over 170K songs ranging from 1921 sings up until 2021 from a lot of different genres so it is quite a diverse dataset.

Solution Statement

The solution will be to develop an NLP model, more specifically a Recurrent Neural Network deep learning model to predict weather a song is positive or not using solely as inputs the music lyrics with the appropriate pre-processing. There could be other options to try out and I may end up going through a logistic regression model depending on the final dataset I will have but, at the time of writing this proposal, I will leave a RNN as my main model.

Benchmark Model

My benchmark model will be a simpler model than a Neural Network, as to compare classical approaches with newer ones. After some investigation, I found out that the first models to show up regarding NLP are the ones with Bayesian statistics, namely a Bayesian classifier approach. [4]

This model can be directly compared against my main model through the AUC (Area Under the Curve metric) on a cross validated set, as well as on the same test set.

Evaluation metric

The evaluation metric I propose is the AUC. AUC stands for "Area under the ROC Curve." AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. In our case, the positive examples would be the

[3] https://developer.spotify.com/documentation/web-api/reference/
[4] https://underthehood.meltwater.com/blog/2019/08/22/deep-learning-models-for-sentiment-analysis/

predictions for labeling a song as positive. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.[5]

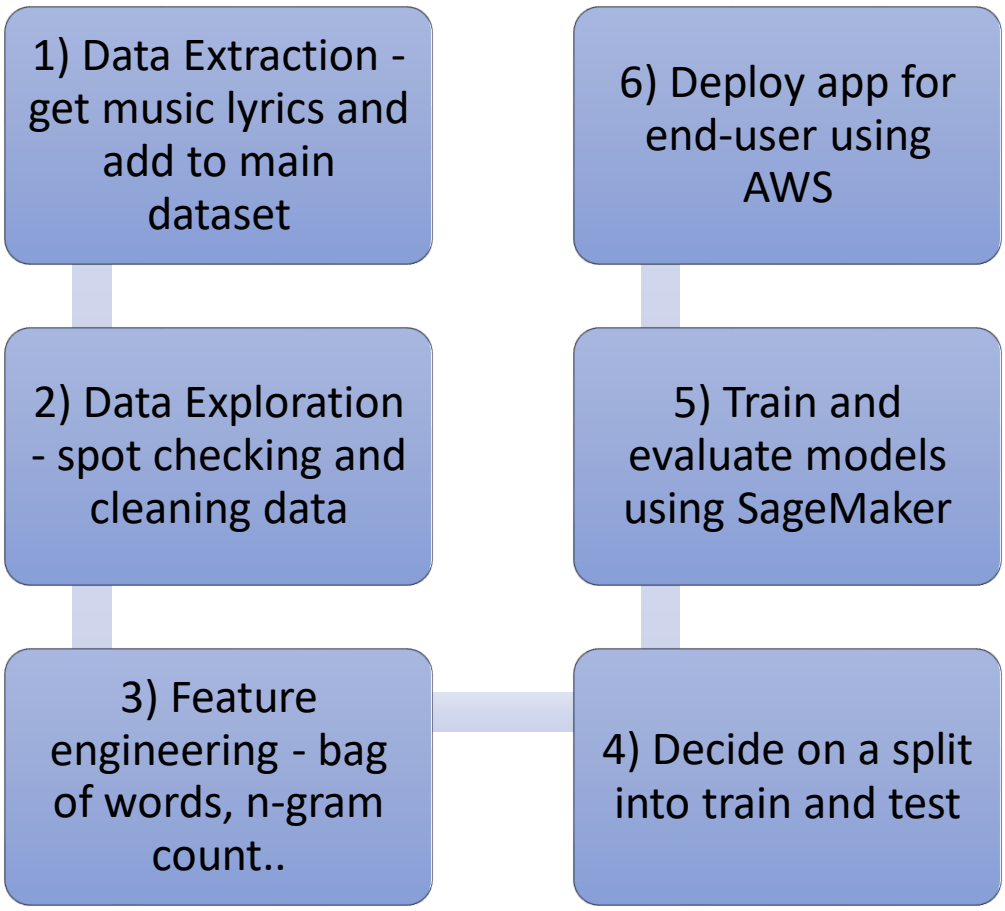
Project Design

I would like to propose the following activities for the project:

1) Data Extraction – main activity here will be to merge information from the two main datasets and extract Song Lyrics data using the lyrics extractor package.
2) Data Exploration/Cleaning – analyze the data and select what to bring for next stages. Also do some necessary cleanups that might be needed such as pre-processing the text data. In this stage I will also create my target variable.
3) Feature engineering – create features using the lyrics data such as the n-grams counts.
4) Split data – divide the data into two partitions: train and testing.
5) Train and evaluate models – train the base and benchmark model and evaluate results while providing a comparative analysis.
6) Deploy app – final stage is to use the AWS services to deploy an app that will take as input a user defined lyric and return a prediction for weather the lyrics are related to a positive or a negative song.

The proposed workflow is the following:

[5] https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc


1) Data Extraction - get music lyrics and add to main dataset
2) Data Exploration - spot checking and cleaning data
3) Feature engineering - bag of words, n-gram count..
4) Decide on a split into train and test
5) Train and evaluate models using SageMaker
6) Deploy app for end-user using AWS