Springboard

Using NLP to Predict Song Genres through their Lyrics

Andrés García Molina, PhD

The Problem

A 2019 report published by Music Business Worldwide asserts that nearly 40,000 tracks are uploaded to Spotify every day. Given such a volume, manual classification is an expensive and slow task that is also error-prone. While in theory those uploading bear the burden of providing meaningful and accurate metadata, automating a process for sanity checks is well worth the investment. This notebook explores some ways to implement a genre-tagging model that is based on neural network training of labeled lyrics datasets.

In an era of digital circulation and streaming, music metadata is crucial in the process of accurately tagging and classifying music to aid its discoverability. How might we leverage NLP techniques in order to correctly assign genre labels to songs? This project combines NLP techniques with neural networks in order to predict genre labels on a corpus of song lyrics.

Stakeholders

- Artists who upload music to various online platforms and hope to maximize their searchability and findability
- The platforms themselves, who hope to optimize music discovery and recommendation processes
- Audiences who want to find music according to genre labels and also discover new acts that might be of interest

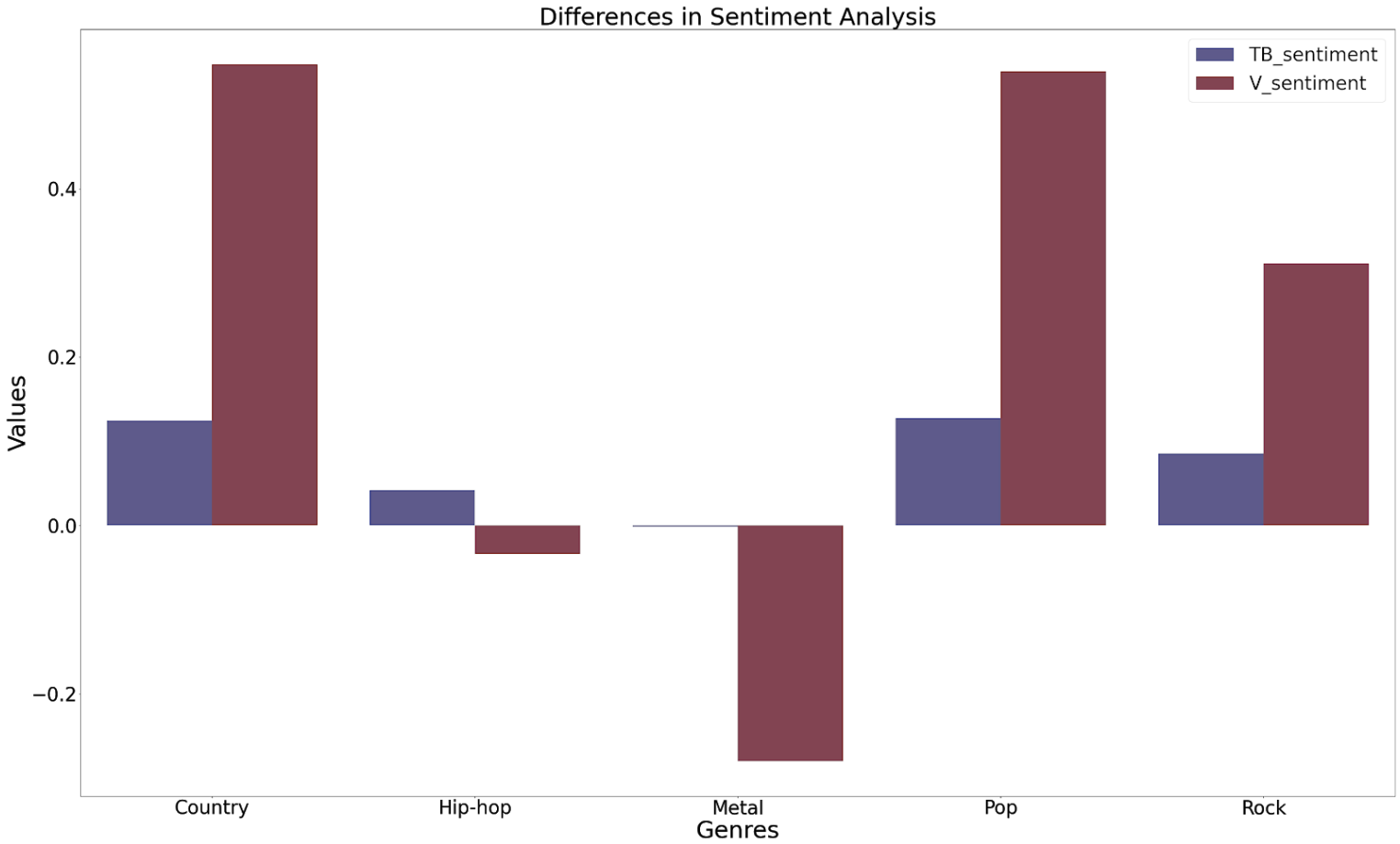
Data

Perhaps the biggest challenge in this project concerns acquiring good, labeled data that pairs song lyrics with genres. For the task at hand, I analyzed several already-available datasets that have performed similar work, including that of Bajwa et al, Sianipar et al, Kovachev et al, and Ram and Salz.

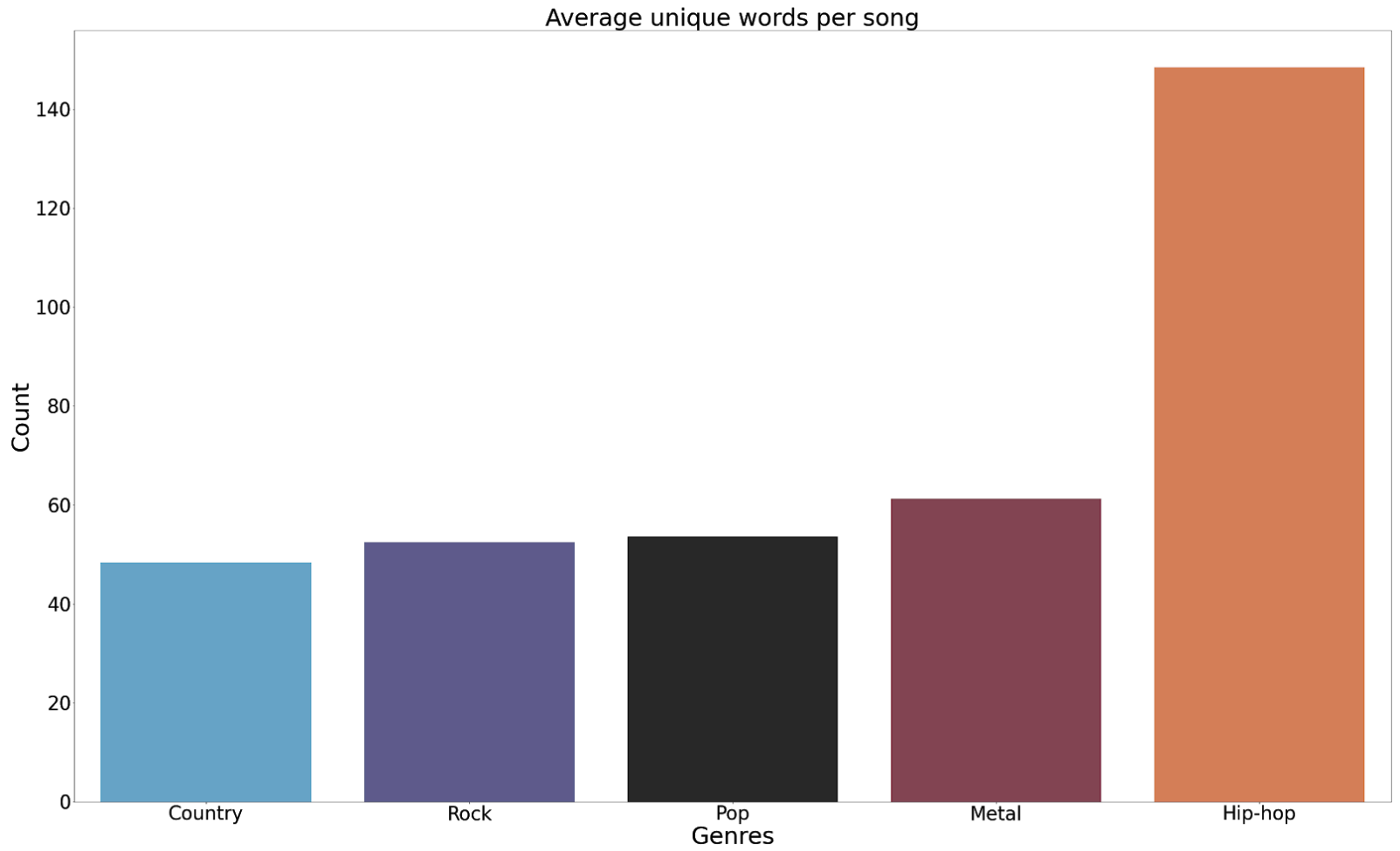
12 genres, 250,000…imbalanced…92900…Metal, Rock, Country, Pop, Hip-hop balanced

Exploratory Data Analysis

Some techniques we applied include sentiment analysis and word frequency per genre. Below, differences in analysis assessment are presented according to two different models:

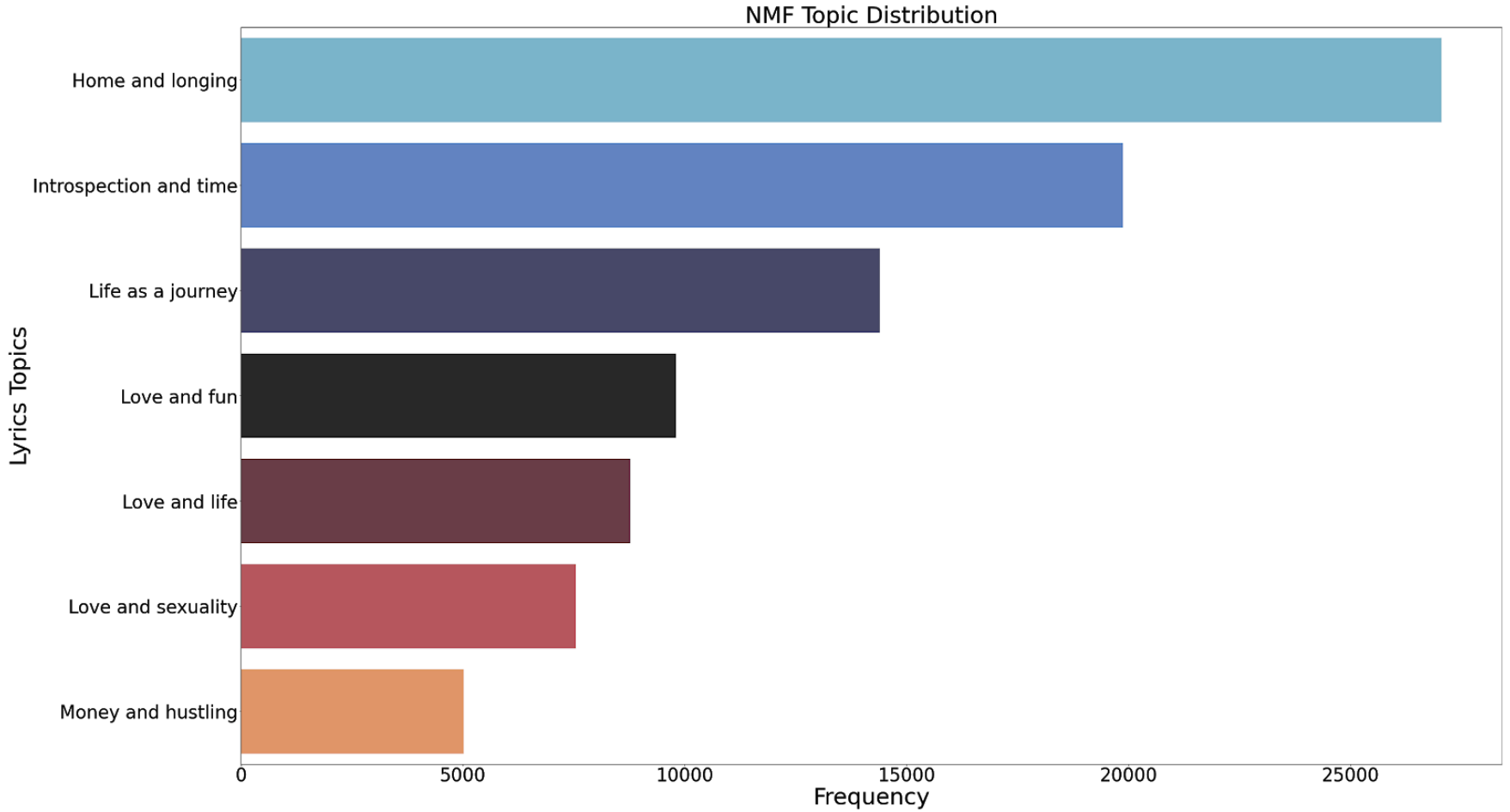


How many unique words are used in each genre?



Topic modeling and distributions

LDA (Latent Dirichlet Allocation) and NMF (Non-Negative Matrix Factorization) topic modeling techniques were used. NMF was favored due to its crisper model output. Below is a general view of the topics covered in our sample.



Modeling

- Prior to training and testing models, we use a label encoder for our genres target variable.
- We also follow the conventional splitting of our data into training and test sets.
- For the purposes of training a neural network we use a “bag of words” (BOW) representation of our lyrics, which is a vectorized count transformation.
- We also normalize the BOW representation so that all values in our matrix are between 0 and 1; scaled data allows neural networks to achieve better results.

Models

Various models were tested, including:

Bagged Decision Trees with Random Undersampling, Standard Random Forest, Random Forest With Class Weighting, Random Forest With Bootstrap Class Weighting, Random Forest With Random Undersampling, Easy Ensemble Classifier, Weighted Logistic Regression, Linear Support Vector Classifier, Extreme Gradient Boosting

Model performance

Several models were tested, including:

- **Baseline:** Logistic Regression, Random Forest Classifier, Multinomial Naive Bayes
- **Neural Networks:** Simple Sequential, Sequential with an Embedding Layer, Sequential with Embedding and Global Max Pooling, Sequential with pre-trained embedding (GloVe), Convolutional Neural Network.

The best performing model is Sequential with Embedding and Global Max Pooling, achieving 64% testing accuracy. Baseline models reach 61% testing accuracy and neural networks stagnate between 62 and 63%.

Conclusions and Recommendations

- Baseline models appear to produce similar results as neural networks. If this application were to be deployed, neural networks would not be favored as an appropriate solution, as the yielded results do not improve significantly on simpler models, but are much more computationally expensive.

- The next steps would require further testing with hyperparameter tuning using simple models and one of the less computationally expensive neural networks. Additional neural networks could be experimented with, such as a Long Short-Term Memory (LSTM) neural network, Recurrent Neural Networks (RNN), Gated Recurrent Unit (GRU), Hierarchical Attention Networks, Recurrent Convolutional Neural Networks (RCNN), Random Multimodel Deep Learning (RMDL), and Hierarchical Deep Learning for Text (HDLTex).

- Ultimately, the question might lie in the quality of the data, as the acquired datasets come from multiple sources and data quality is certainly an issue. This project ultimately demonstrates that one of the more crucial steps in the data science lifecycle concerns data acquisition and quality. In this case, generating a tidy lyrics dataset would produce high pre-processing costs.