**Analyzing Song Lyrics by Genre**

***BSAN-6200-Section-01****: Text Mining & Social Media Analytics*

Author(s):

**Mfolozi Dlamini, Kyle Guillen, Monami Mukherjee and Sarah Olsen**

# **Background**

*Objective*

## In this project, we observed trends and sentiments found in song lyrics across five music genres: country, rap, rock, pop, and R&B. We applied natural language processing and text mining tools to extract the information and create insights which led to business recommendations to benefit the music and podcast streaming service, Spotify.

## *Problem and Business overview/project summary*

It is no secret that music is a powerful tool that has the ability to affect our emotions and general mood, and no service capitalizes on this as well as Spotify does. One reason Spotify is one of the most popular music and podcast streaming services in the world is its playlist recommendation features. As a Spotify user, a subscriber can curate personalized playlists, share playlists with other subscribers, and choose from one of the several original Spotify playlists to listen to. These Spotify original playlists can range from mood-based playlists such as happy or sad, to genre-based playlists such as country or rap. Our goal as an entertainment analytics consulting firm was to make recommendations for Spotify to improve their playlist recommendations models and in turn improve a user’s listening experience.

**Data Processing**

*Data Collection*

Genius is “the world’s biggest collection of song lyrics and musical knowledge” on the web. Using their platform, we extracted information about the top 100 most viewed song lyrics in each of the following music genres: Rap, Pop, Country, and Rock.

Important Note: We are analyzing this data under the assumption that the more views a song lyrics webpage has, the more compelling the lyrics are to the listener compared to the other songs in the genre.

The following columns were copied from <https://genius.com/#top-songs>:

* Song
* Artist
* Views
* Genre Dummies (Rap, Pop, Country, R&B, and Rock)

The lyrics were extracted using the Genius API. We first signed up for an API account on the Genius developer website. Then we were able to find an open access python library made specifically to work with this Genius API, which made it very easy to simply enter the name of the song, artist, and our unique Genius API key and the lyrics of the song were returned instantly.

*Data Cleaning*

Using a combination of Pandas and Excel, we cleaned the Song Information data. Our first step was to transform the information into a tabular format. The data copied and pasted from the web is read into the csv as one long column with values stacked on top of each other, so we had to transpose the data into its appropriate rows and columns. We did so using the transpose function in Excel.

During the data collection process, we collected the top 100 most viewed songs from each genre. In all five of the song information tables we copied, we had to manually create a genre column. At this point, we had five tables, one for each genre. We did a union to stack the tables on top of each other to form one large data table. From there we created dummy variables for each genre, with the genre name as the column name and where 1 means yes and 0 means no. We removed all duplicate records and our song information data was ready to read in python.

The next challenge was to use python and Genius’s API to generate a column of song lyrics to match up with the song information in our table. Extracting lyrics from the API was easy; we fed the API the song title and artist and it returns the song lyrics. Unfortunately, the lyrics came through the API very messy. There was a prefix and a suffix added to the beginning and end of every set of lyrics and there were unprintable characters throughout. Using a series of regular expressions, the extra characters and words were removed from the lyrics. When the data table was complete with all song information and clean lyrics, it was using UTF-16 encoding. In order to avoid song lyrics including unprintable characters, specify the encoding type UTF-16 when exporting or importing the data in or out of python. That concludes our data cleaning process.

*Data analysis and Findings*

\*Note: The following were used to perform the analyses in Python: sklearn.linear model, sklearn.naive\_bayes model, sklearn.feature\_extraction.text, sklearn.model\_selection

First, we performed a sentiment analysis on our dataset. We used the Python library TextBlob to calculate the polarity scores of each of the songs. A song was classified as positive or negative depending on if it had a 1 or 0 polarity score. The dataset was relatively balanced with an overall 60% positive sentiment and 40% negative sentiment. (\*Note: the size of our dataset was about 420 songs so under sampling or oversampling was not appropriate). After using TextBlob’s polarity function, we grouped the sentiment by genre. Our results can be seen in **Figure 1** on the following page.

Chart

Description automatically generated

**Figure 1**

Rap was the only genre with a negative polarity score (-0.006340) while pop was deemed to be the most positive genre with a polarity score of 0.082111. Country, R&B, and rock had polarity

scores of 0.072947,0.070789, and 0.040430 respectively.

We also ran a logistic regression (**Figure 2**) and a simple Naïve Bayes model on our data. The

AUC of 67.9% gave us an indication that the model was *fairly* accurate at classifying our positive and negative outcomes.

Chart

Description automatically generated

**Figure 2**

**Chart, line chart

Description automatically generated**

**Figure 3**

Next, we performed Latent Dirichlet Allocation (LDA) topic modeling on the dataset. In essence, topic modeling allows us to analyze text data automatically to determine potential word clusters from the given information. For this project, we were intrigued to see what words would be grouped together in our dataset spanning across the 5 genres. If we could identify topic clusters and observe coherence scores, the topic clusters could be used in creating automated blended genre and mood playlists based on topic on Spotify.

Before creating the LDA model we had to ensure the data set was balanced. From the pie chart below (**Figure 4**), we see that country, rap, and rock songs were relatively the same in count with approximately 100 songs per genre. Pop and R&B have relatively lower counts with about 60 songs per genre. However, we still felt confident in our samples as this was not a significant gap.

Chart, pie chart

Description automatically generated

**Figure 4**

We can see the top terms by cluster across the 5 music genres below.

Graphical user interface, table, Excel

Description automatically generated

**Figure 5**

Our coherence score is shown below. We can use the score in our topic model as a metric for how interpretable our song topics are to humans. In this case, the topics are represented as the top N words with the highest probability of belonging to that specific topic. The coherence score essentially measures how similar these words are to one another. Our goal was to maximize intra-topic instances and minimize intertopic similarity. We implied the cosine similarity between words with word2vec embedding. As we can see, there is a steep decline in the coherence score when the number of topics increase, a result that is justified due to the pre-chosen nature of the genres.

Chart

Description automatically generated

**Figure 6**

The intertopic distance map is a visualization of the song topics in a two-dimensional space. The topic circles below have areas proportional to the number of words that belong to each of those topics across the dictionary. We used the map to determine which genres had topics in common so we could suggest genres with similar topics for the Spotify original playlist curation. Country, pop and R&B were shown to have similar topic clusters.

Chart, bubble chart

Description automatically generated

**Figure 7**

**Recommendations**

The UX design team and product managers can use this information to create Spotify blend playlists for genres and moods. The application feature will be added to the Mood section (**Figure 8a**) in Spotify and include automatically updated weekly playlist that gives the user the most personalized listening experience based on their mood of choice and preferred music genres. For example, a user may choose to enjoy a personalized playlist of a mix of country and pop songs that would be found under the happy playlists section. Implementing these niche playlist features can improve user listening experience and general service satisfaction. If these personalized mood and genre playlists are included with Spotify Premium, it incentives no-cost subscribers with another appealing application feature that requires them to purchase Spotify Premium.

Graphical user interface, application

Description automatically generated

**Figure 8a**

Graphical user interface, website

Description automatically generated

**Figure 8b**

**Bonus Model**

Can We Use Song Lyrics to Predict a Song’s Genre?

Out of curiosity, we attempted to use the BERT algorithm to predict songs’ genres. We ran BERT five times, one for each genre column. As we mentioned previously, the genre columns are dummies. For example, in the Pop column, a value of 1 represents “yes, this song is Pop” and a value of 0 represents “no this song is not Pop”. The Lyrics column is the input variable. The target variable is the genre column, so we ran BERT five times - one time for each genre as the target variable.

BERT was not able to reliably predict any of the song’s genres. There are a couple of reasons why. First, we did not have enough data. With only roughly 500 records, there’s not enough to compare. Also, the target classes are imbalanced for every genre. There are only roughly 100 records in each genre and a total of 500 records. Also, we hypothesized that genre might not be the correct target variable to look at. Genre blending is becoming more popular in mainstream music, so it makes sense that BERT has a hard time deciphering which songs belong to each genre. This idea aligns with our findings in LDA and sentiment analysis, where we found that all five genres have some overlapping topics and sentiments.

References

Catchpole, Dan. “How Spotify and Amazon Use A.I. to Learn Your Preferences.” *Fortune*, Fortune, 10 Nov. 2021, https://fortune.com/2021/11/09/spotify-amazon-ai-alexa-artificial-intelligence-preferences-mood/.