Billboard Hot 100 Music Analysis

MeiLi Charles, Lydia San George, Regan Moreno

***Abstract*—This project had two goals: build a classifier trained on the Billboard Hot 100 and investigate commonalities in lyrics from tracks that have debuted at #1 on the same chart. Features for the classifier were retrieved using *Spotify*'s Web API. Lyric analysis was done taking into consideration word frequency, repetition, and weight.**

I. Introduction

As of September 1, 2020, BTS debuted at #1 on the Billboard Hot 100 with their first all-English single, ”Dynamite.” In addition to being the first South Korean act to top the chart, the band retained their #1 ranking for two consecutive weeks. The single itself is a disco pop song, with cheerful and uplifting lyrics, accompanied by 70's style synths, layered vocals, and horns to create an upbeat sound. The previous top ranking for the previous 3 weeks was Harry Styles' Watermelon Sugar.

The Billboard Hot 100 serves as the music industry's standard for a track's popularity in the United States. The chart encompasses all genres and rank is determined by a combination of U.S.-based streaming, airplay, and digital or physical sales, where the weight of each of those elements decrease respectively.

There have been only 43 songs that have ever debuted at #1 in the history of Billboard Hot 100 since its launch in 1958 [1]. Some examples of these songs include

- You Are Not Alone by Michael Jackson (1995)
- Honey by Mariah Carey (1997)
- Candle in the Wind 1997/Something About The Way You Look Tonight by Elton John (1997)
- My Heart Will Go On by Celine Dion (1998)
- Hold It Against Me by Britney Spears (2011)
- Shake It Off by Taylor Swift (2014)
- What Do You Mean? by Justin Bieber (2015)
- 7 Rings by Ariana Grande (2019)

II. Retrieving Data

Song data was retrieved using Spotify's web API. The popular music-streaming platform hosts about 286 million active users according to the company's 2020 report [2]. *Spotify for Developers* offers an easy way to retrieve individual track data. Given a unique track ID, the *GET* command will return a variety of feature values for the track as a .json. For this project, all audio features listed in the following table were used for popularity classification. Value descriptions can be found on the *Spotify for Developers* website.

Further data processing was done using pandas, where the retrieved information could easily be imported into a dataframe.

key	value type
acousticness	float
danceability	float
duration_ms	int
energy	float
instrumentalness	float
key	int
liveness	float
loudness	float
mode	int
speechiness	float
tempo	float
time_signature	int
valence	float
popularity	float

Two audio playlists from Spotify were collected to build our classifier: the Billboard Hot 100 chart as of November 1, 2020 and Playlist Hits 2020.

III. Predicting a Song's Popularity

We tried a variety of different machine learning algorithms using scikit-learn's packages. We used its train-test split function to split our dataset in half. We set the average popularity of a track from the current Billboard 100 chart as the criteria for whether a song is ”popular” or not. For the train dataset, if its popularity was equal to or greater than the average popularity value calculated from the Billboard 100, its popularity value was set as 1, otherwise 0.

IV. Classification Accuracy

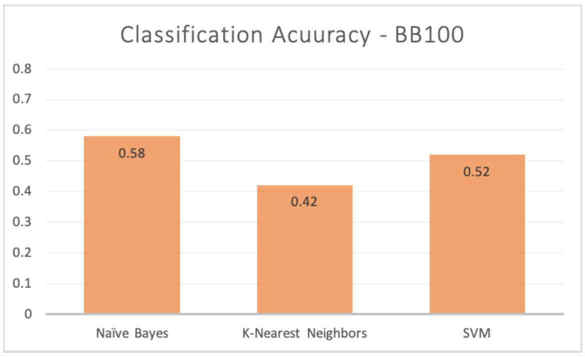

Fig. 1. Classification accuracy using the current Billboard Hot 100 chart as of November 1, 2020

Using the Hot 100 playlist, we can see the Support Vector Machine classifier performed the best. From these results, we may be able to infer that even extremely popular songs may not measure up the the threshold of ”average popularity.”

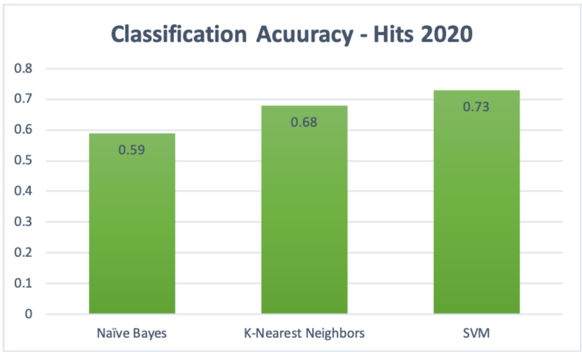

Fig. 2. Classification accuracy using the Playlist Hits 2020 chart

V. Lyric Analysis

Regan handled lyric extraction and cleaning. They used Beautiful Soup and the Genius.com API to scrape lyrics from the internet and store them within individual text files. The text files were then cleaned of punctuation and extraneous data. Those text files were then utilized by both Lydia and Regan to create varying lyric comparisons to attempt to view repetition of words, phrases, or slang.

Lydia was in charge of the visualization of the lyrics from the top billboard 100 songs. The songs were first formatted into clean text files to be evaluated based on frequency and repetition. The text files for each song were then run through the monkeylearn API to create CSV files containing each word's frequency, repetition, and weight. From the CVS files, word clouds were able to be created to visualize the song lyrics that stood out the most. By looking at the word clouds you could draw conclusions about the slang used, the song topic, the most catchy lyrics, and the complexity of the song's writing. These visuals were then used in the presentation to show off the success of the overall lyrics analysis and the areas that had room for improvement.

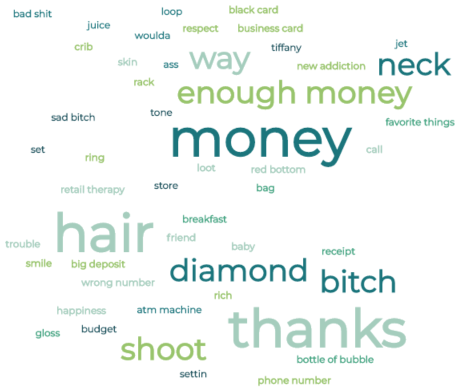

Fig. 3. 7 Rings Word Cloud

The figure above is from the song 7 Rings by Ariana Grande. This word cloud shows off the success of the lyrics analysis. The largest words shown are the size they are because of their repetition within the song. The words "money", "hair", "thanks", and "diamond" are the main focus of this song and of this word cloud. By looking at the figure you can tell the topic of the song and the tone.

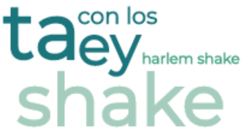

Fig. 4. Harlem Shake Word Cloud

The figure above shows how small the word clouds could end up. When given a song without a lot of lyrics, the visual turns out containing all of the words. This song is the Harlem Shake, which was a song for a popular trend online. One special quality of this word cloud is that it contains the full title of the song.

VI. Project Tasks and Schedule

Potential tasks to be completed include:

1) Gather data sets: determine dates for chart rankings.
2) Investigate Pandora's Music Genome project for song attributes.
3) Data exploration/Identify common factors among songs. This could include musical structure/composition, lyrics, origin, etc.
4) Classifier

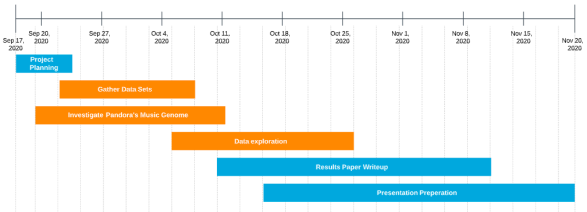

Fig. 5. Schedule of Tasks

VII. Team Member Responsibilities

MeiLi Charles: Senior Project Head

- Gathering Data Sets
- Investigate Pandora's Music Genome
- Data Exploration
- Results Paper
- Final Presentation

Regan Moreno: Vice Project Head

- Gathering Data Sets
- Investigate Pandora's Music Genome
- Data Cleaning
- Data Exploration
- Results Paper
- Final Presentation

Lydia San George: Project Team Member

- Investigate Pandora's Music Genome
- Data Exploration
- Results Paper
- Final Presentation

VIII. Conclusion

Our goal with this project is to uncover answers about the modern American's music taste. Although music is often considered subjective, there may be identifiable, core elements of a "hit" song. Similar chord progressions and lyrical content may be more easily accessible and relatable to the general public.

From classification, we saw that we may be able to use audio features to predict a popular song. Future improvements to increase classification accuracy could include further data investigation. Some features may not be needed for classification. Futhermore, the popularity threshold may have been set too high in this particular instance. A more comprehensive study, considering significant cultural events as well as the general public's mindset may be a good indicator as to how individuals may connect with a specific song. Given more time, it may be useful to create a popularity value predictor, that predicts a value between 0 and 1 (as opposed to this project's binary classification system).

Additionally, we may gain insight into whether identified trends are cyclical in nature or if each major hit marks a new musical trend. Is there a musical "sweet spot" of blending musical familiarity with newer, fresher ideas? Out project aims to provide answers and insights into this area of musical consumption

IX. Contact Information

MeiLi Charles: jcharl12@vols.utk.edu
Lydia San George: lsangeor@vols.utk.edu
Regan Moreno: rholmber@vols.utk.edu

References

[1] Callie Ahlgrim. Only 46 songs have debuted at no. 1 in the history of the billboard hot 100 — here they all are - insider, Nov 2020.

[2] Mansoor Iqbal. Spotify usage and revenue statistics (2020), Dec 2018.