Sentiment Analysis of Billboard Top 100 List Songs from 2000 - 2018

Amelia C. Teare

Northwest Missouri State University, Maryville MO 64468, USA
S555513@nwmissouri.edu

Abstract. The purpose of this paper is to investigate and search for patterns in the tone of songs placed on the Billboard Top 100 List from 2000 - 2018. The analysis of this project involved using Tableau to find trends among the polarity values of each song. A histogram of the polarity of each song on the list shows an even distribution with the bulk of songs having a polarity in the "neutral" zone (-0.1 to 0.1). In fact, taking the average polarity for each song in a given year placed the average yearly song polarity above 0.03 for each year. The year with the highest average polarity was 2000 with an average of .105 and the year with lowest average polarity was 2018 with an average of 0.037. While the average yearly polarity was consistently neutral-toned, there was a consistent decline in polarity over the course of those 19 years showing a rise in popularity for more negative songs. This project does not take into account the nuance of keys or notes when determining the overall tone of the song. Rather than looking at the song holistically, this project was only able to hone in one aspect of a song. That means a song with negative-seeming lyrics but in a happy, bouncy key will still present a polarity rating in the negative zones. The same can be said in the opposite direction. In the future, it would be necessary to account for the key and tempo of the song when determining the tone. Despite this limitation, this project provides key insights into the patterns and trends of songs most loved by American consumers.

1 Introduction

Music is able to bring people of all backgrounds together. It can tell a story of the personal history of the artist as well as provide a small peek into the history of its time. As history progresses and artists write from their own experiences, is there a type of music that consistently reigns supreme? Does the population trend towards positive music to escape from negative realities? This paper aimed to provide an answer to this question. From this answer, it may be possible to predict the popularity of a song based on the mood of the song, year it is released, and artist. The GitHub Project Repository can be found here. The Public Overleaf can be found here.

1.1 Related Works

Individuals have tackled this problem before. In one project, the lyrics for ten different songs were analyzed as having either a positive, a negative, or a neutral message [5]. In another project, sentiment analysis was done on posts that were uploaded to the website twitter [2].

2 Methodology

Data was collected and prepped using Microsoft power query. A few known songs were analyzed using the created Sentiment Analysis Model to determine the accuracy of the model. This was done using python. Following the validation of the model, the model was applied to the data file. The results were exported to a file and then uploaded into Tableau for visualization and pattern-identification. Much of the process followed the steps outlined in the "How to Perform Sentiment Analysis with Python?" online article [4].

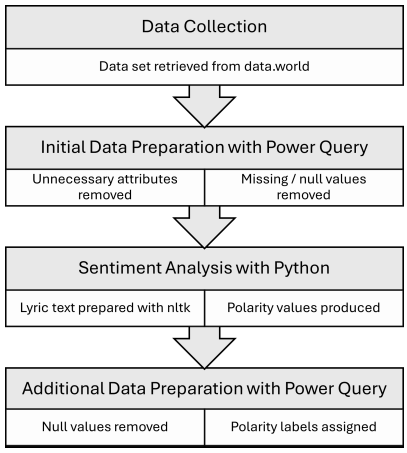


Fig. 1. Methodology. A. Teare 2024.

2.1 Data Sources and Limitations

The search for the data that was analyzed was done through Google Data. This file found on data.world containing the Billboard Hot-100 songs from 2000 - 2018 was ultimately chosen. The original file contained 7573 rows and 31 columns with information regarding a songs position on the list and analysis of the song like "accousticness", "danceability", and "tempo".

Due to the nature of sentiment analysis, there was limit on what could be identified from the lyrics of each song. Based on previous work completed [5][2], sentiment analysis was only able to give a result of negative, positive, or neutral.

2.2 Data Preparation

The raw data for this project was already in a simple tabular form and required very little cleaning. The biggest error found in the data were null values, missing lyrics, and unnecessary data fields. As this project revolved around the analysis of song lyrics, it was pertinent that only data with readable lyrics and only data that was needed for the analysis was included in the analyzed project data. Using Microsoft Power Query, the extra fields were removed and the data file was reduced to the fields that were necessary for this project. These fields were "date", "title", "artist", "peak pos", "rank", and "lyrics". A more in-depth data dictionary can be found in table 1 below.

Table 1. Data Dictionary

Field Name	Description	Format
date	date the song was on the list	mm/dd/yyyy
title	title of the song	string
artist	artist(s) who sang the song	string
peak pos	the highest position the song achieved on the list	integer
rank	overall position on the list	integer
lyrics	the complete written lyrics for the song	string

The original data file included a "year" field. While this field initially seemed useful to the purpose of this project, the rows contained mostly missing values and that information was also available in the "date" field. Therefore, the "year" field was also removed. In addition to the additional fields, the original file also included rows with missing data. It was found that the quantity of rows with missing data made up only 4.5 percent. This was deemed small enough to exclude from the data. To do this, Microsoft Power Query again was used to remove all rows where either the lyrics or data was missing. This left the file with 6 columns and 7229 rows.

2.3 Exploratory Data Analysis

Following the data collection and initial data preparation, the data underwent exploratory data analysis to identify any position trends prior to the sentiment analysis of the lyrics. Using Matplotlib from python through Jupyter Notebook, the data was read into the notebook and information was compiled pertaining to the data type of each field. After this, a code was carried out to determine the correlation between the two numeric fields and to complete additional summary statistics. Various visuals were then created using Matplotlib through Jupyter Notebook and through the use of Tableau.

Using the visualization tools of Tableau, initial patterns were discovered. An initial comparison of the quantity of songs that made the Billboard Top 100 List from 2000 - 2018 to the year the song made the list found that the years 2017,

```
data = pd.read_csv('clean_song_data.txt', encoding='MacRoman', delimiter='\t')
data.head()
data.shape
data.info()
data.describe()
data[['peak_pos','rank']].corr()
```

Fig. 2. Python code used to perform initial exploration of the data

2009, and 2011 had the highest number of songs with a total of 668, 524, and 489 respectively. The Billboard Top 100 List is updated weekly. This difference in values gives a sign that songs were more transient during those year than the others. A graph showing a further breakdown of this data can be seen in figure 3 below.

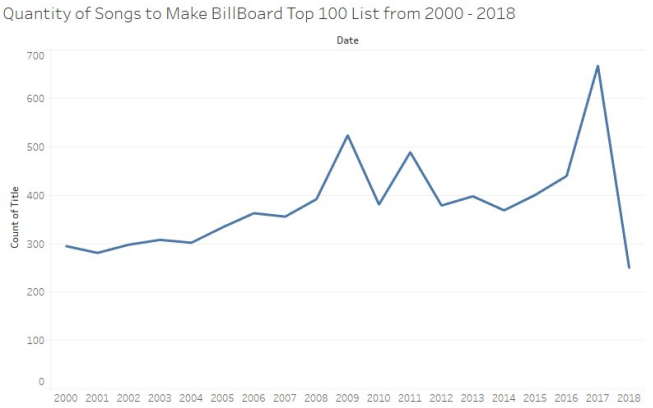


Fig. 3. Number of Songs to make the Billboard Top 100 List each year from 2000 - 2018. Smaller values indicate fewer additions/subtractions of songs from the list compared to higher values.

Additional analysis involved looking for the quantity of times an artist reached the peak position of 1 on the Billboard Top 100 List from 2000 - 2018. Rihanna reach this position the most times with a total of 11 times, Katy Perry was second with 9 times, and Usher was third with 7. A graph showing the top 10 artists to reach a peak position 1 on the Billboard Top 100 List from 2000 - 2018 can be seen in figure 4.

Finally, analysis was done to identify the artist to reach the overall top position on the Billboard Hot 100 List each year by looking for artists with a rank of 1. This list contained a few multiples and did not show an artist for each year. The table showing the Artists to reach a rank of 1 on the Billboard Hot 100 List can be seen below in figure 5.

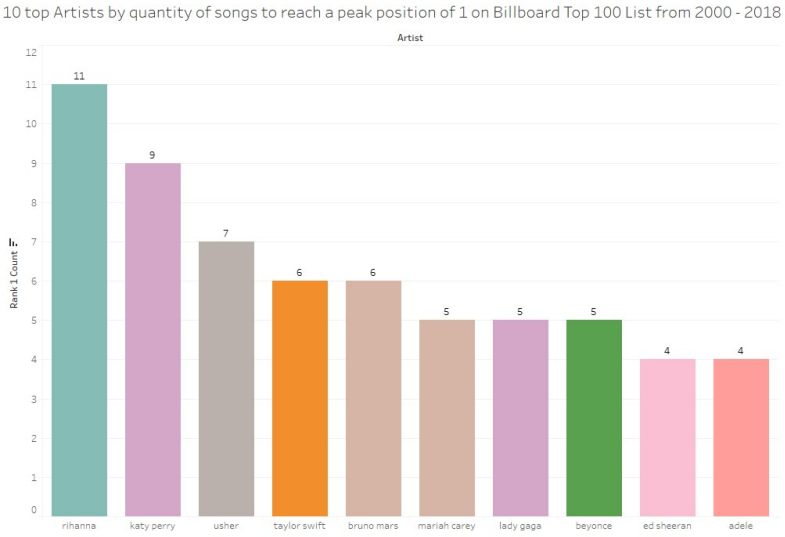


Fig. 4. Top 10 Artists with most songs to reach the top spot on the Billboard Top 100 List from 2000 - 2018

Artists to reach overall top position on Billboard Hot 100 List by year

Artist	Year of Date
Drake	2018
ed sheeran	2017
flo rida	2009
gotye	2012
jay-z + alicia keys	2009
maroon 5	2007
rihanna	2011
soulja boy	2007

Fig. 5. A list of each artist to reach the overall top position on the Billboard Hot 100 List for each year.

2.4 Sentiment Analysis Model

In creating the sentiment analysis model, it was chosen to use Python's NLTK Library [3][1]. In order to properly utilize this tool, necessary modules needed to be installed and imported. These modules included "nltk", "nltk.tokenzie", "nltk.corpus", and "nltk.stem"After compiling and cleaning the necessary data, preparation had to be done to the text that was to be analyzed. First, unnecessary white space and line breaks were removed. This was done using excel. After this, all numbers, punctuation, and special characters were removed. All text was converted to lower case and then tokenized; the splitting of words into indi-

vidual words or tokens. Then all stop words were removed; common words that do not have much meaning like "the", "is" and "and". The function was then programmed to observe negated words as negative. Words were then reduced to the root forms and tokens were joined back into a single string to return a cleaned text. This cleaning function was applied to text in the "lyrics" column and the cleaned text was analyzed for its polarity using sentiment analysis. That analysis was added as an additional column directly to the right of the lyrics. The code for this sentiment analysis function can be seen in figure 6.

```
def preprocess_text_column(data, column_name):
    # Define a function to preprocess a single text
    def preprocess_text(text):
        # remove all numbers, punctuation, and special characters.
        cleaned_text = re.sub(r'[^a-zA-Z\s]', '', text)

        # convert all text to lower case to ensure consistency with analysis
        cleaned_text = cleaned_text.lower()

        # Tokenization: splitting words into individual words or tokens
        tokens = word_tokenize(cleaned_text)

        # remove all common words that do not hold much meaning like "the", "is", and "and"
        stop_words = set(stopwords.words('english'))
        tokens = [word for word in tokens if word not in stop_words]

        negation_words = set(['not', 'no', 'never', 'none', 'nobody', 'nothing', 'neither', 'nor'])
        negated = False
        result = []
        for word in tokens:
            if word in negation_words:
                negated = not negated
            else:
                if negated:
                    word = "NOT_" + word
                result.append(word)
        tokens = result

        # reduce words to their root forms
        lemmatizer = WordNetLemmatizer()
        tokens = [lemmatizer.lemmatize(word) for word in tokens]

        # Join tokens back into a single string
        cleaned_text = ' '.join(tokens)

        return cleaned_text

    # Apply the preprocess_text function to the specified column
    data[column_name] = data[column_name].apply(preprocess_text)

    return data
```

Fig. 6. The function used to prepare the lyrics for sentiment analysis.

Following the execution of the sentiment analysis tool, the updated text file was opened in Microsoft Excel. This updated text file contained all the initial fields listed in Table 1, but also included an additional field with the polarity value found using sentiment analysis. This value was a numeric value between -1.0 and +1.0. The values closer to -1 indicated more negative tones, values closer to +1 indicated more positive tones, and values closer to 0 indicated neutrality. To provide additional fields for analysis, one more field was added to the file. This field placed the polarity values in range-based categories. These categories are defined in Table 2. Using these categories, further patterns and analysis were able to be completing using Tableau.

Table 2. Polarity Category Descriptions

Category	Polarity Range
Extremely Positive	0.75 to 1.0
Positive	0.5 to 0.75
Mostly Positive	0.3 to 0.5
Somewhat Positive	0.1 to 0.3
Neutral	-0.1 to 0.1
Somewhat Negative	-0.3 to -0.1
Mostly Negative	-0.5 to -0.3
Negative	-0.75 to -0.5
Extremely Negative	-1.0 to -0.75

3 Results and Analysis

The main objective of this project is to identify any patterns in the tone of songs and their placement on the Billboard Top 100 List. Immediate analysis of the polarity values for every song listed on the Billboard Top 100 List from 2000 - 2018 found with the sentiment analysis tool shows a normal distribution as seen in figure 7. The vast majority of songs on the list obtained a polarity value

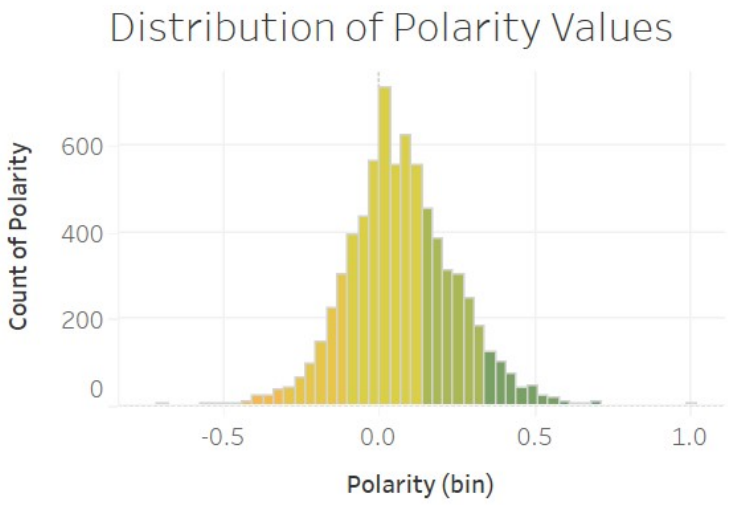


Fig. 7. The distribution of polarity values for every song listed on the Billboard Top 100 List from 2000 - 2018. This graph shows a normal distribution of polarities.

between -0.1 and 0.1. Beyond that, more songs ended up on the positive side of the histogram compared to the negative side, indicating that it is slightly more common for songs with a positive message to make the list than songs with a

negative message. Even more interesting, while three songs obtained a polarity value in the Extremely Positive range, not one song obtained a polarity value in the Extremely Negative range. A complete breakdown of song totals in each polarity category can be found in figure 8.

Quantity of Songs in Each Polarity Category

Polarity Label	
Extremely Positive	4
Positive	78
Mostly Positive	612
Somewhat Positive	2,258
Neutral	3,257
Somewhat Negative	901
Mostly Negative	102
Negative	16

Fig. 8. The quantity of songs in each polarity category. The vast majority fall within the Somewhat Positive, Neutral, and Somewhat Negative categories. Over half of all songs had a polarity value of neutral or above.

When comparing the average polarity of all songs on the list for each year, the polarity maintains in the neutral range as described above. However, the average polarity does trend in the negative direction showing a decrease in the polarity of songs as time moves forward. This relationship can be seen in figure 9. If each year is simplified down to show only those songs that obtained an overall rank between 1-10 for that year, more variety becomes apparent. While the majority of songs in that category fall within the neutral range, a few extremes come forward. In 2004, the song ranked fifth had a polarity of .2542, which falls in the Somewhat Positive range. Even higher, the song fourth in 2006 had a polarity of .3500 which places that song in the Mostly Positive range. On the other end of the spectrum, the song that placed fourth in 2012 had a polarity of -0.4155 which places that song in the mostly negative range. These metrics can be found in figure 10.

On the Artist level, the neutral theme continues. Focusing on the 10 artists with the most rank 1 songs, their polarities stay between 0.0247 and 0.1464. Rihanna has the most songs ranked number 1 on the Billboard Top 100 List and she has an average polarity of 0.1004. This value just barely pushes her over the line from the Neutral range into the Somewhat Positive range. Katy Perry has the second most songs ranked number 1 on the Billboard Top 100 List and she has an average polarity of 0.1093. Again, this value places her just over the line into the Somewhat Positive range.

The artist on this list with the lowest polarity value is tied for the least number of songs ranked number 1. This artist is Adele. She has four songs ranked number 1 on the Billboard Top 100 List and has a polarity of 0.0247 which places

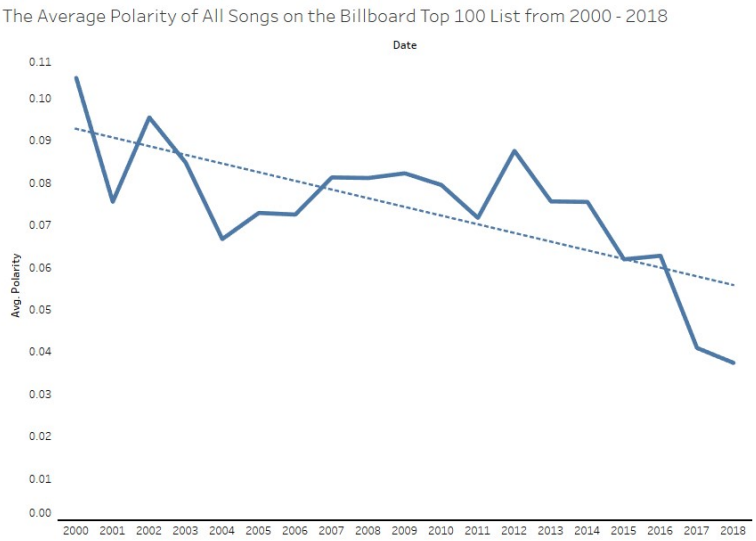
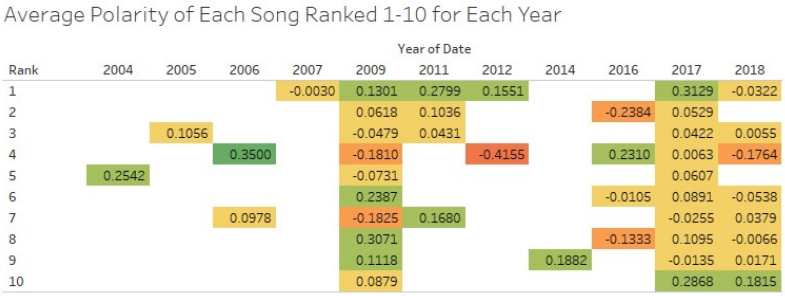


Fig. 9. The average polarity of every song placed on the Billboard Top 100 List each year. While the average polarity for each year is in the neutral range, there is a clear decrease in polarity as the years continue forward.



Rank	2004	2005	2006	2007	2009	2011	2012	2014	2016	2017	2018
1				-0.0030	0.1301	0.2799	0.1551			0.3129	-0.0322
2					0.0618	0.1036			-0.2384	0.0529	
3		0.1056			-0.0479	0.0431				0.0422	0.0055
4			0.3500		-0.1810		-0.4155		0.2310	0.0063	-0.1764
5	0.2542				-0.0731					0.0607	
6					0.2387				-0.0105	0.0891	-0.0538
7			0.0978		-0.1825	0.1680				-0.0255	0.0379
8					0.3071				-0.1333	0.1095	-0.0066
9					0.1118			0.1882		-0.0135	0.0171
10					0.0879					0.2868	0.1815

Fig. 10. The average polarity of every song placed on the Billboard Top 100 List each year. While the average polarity for each year is in the neutral range, there is a clear decrease in polarity as the years continue forward.

her comfortably in the neutral range. The artist with the highest polarity value is the other person tied with Adele for the least number of songs ranked number 1 in this figure. This artist is Ed Sheeran. He has four songs ranked number 1 and has a polarity of 0.1464 which places him in the Somewhat Positive range. This information can be found in figure 11. All figures shown in the Results and Analysis section can be further analysed here.

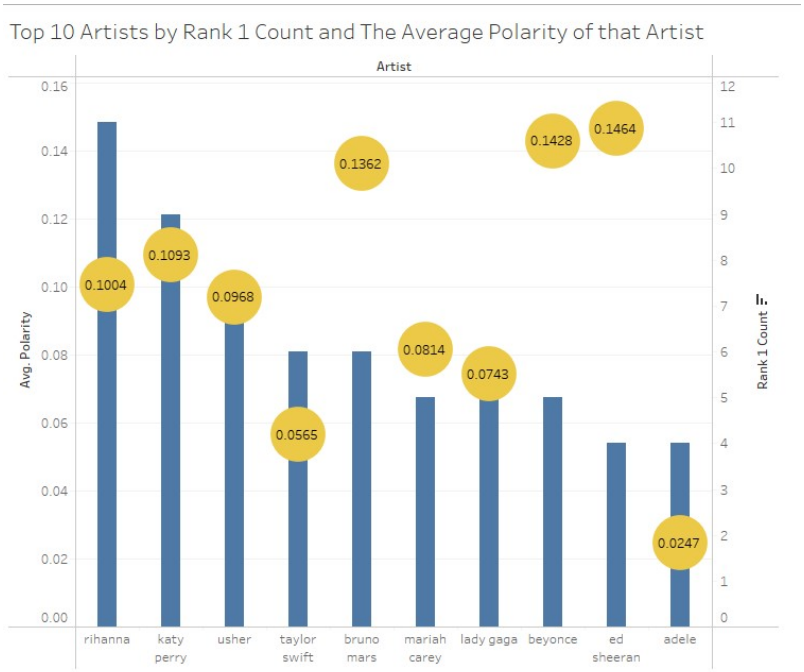


Fig. 11. The Top 10 artists with the most rank 1 songs on the Billboard Top 100 List and their average song polarity. The maximum polarity is 0.1454 from Ed Sheeran and the minimum polarity is 0.0247 from Adele.

4 Conclusion

This project set out to find a pattern or trend among the songs listed on the Billboard Top 100 List from 2000 - 2018. One thing that stood out was the overall neutrality of these songs. Very few songs with overtly positive or negative tones made the list during the 19 year span. This is further confirmed by the average polarity of the most common artists on that list being existing in either the Neutral or Somewhat Positive polarity ranges. One major limitation to this project and it's results is the focus on lyrics only. The mood of songs are defined by the combination of notes, key, and lyrics. While focusing on the lyrics provides a concrete evaluation of the words being sung, it does not provide a holistic view of the overall message. Further work would include all aspects of the song when determining its tone.

—- Bibliography —-

References

1. Cheng, R.: Text preprocessing with nltk, https://towardsdatascience.com/nlp-preprocessing-with-nltk-3c04ee00edc0

2. Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P.: Study of twitter sentiment analysis using machine learning algorithms on python. International Journal of Computer Applications **165**(9), 29–34 (2017)
3. Mogyorosi, M.: Sentiment analysis: First steps with python's nltk library, https://realpython.com/python-nltk-sentiment-analysis/#installing-and-importing
4. Selvaraj, N.: How to perform sentiment analysis with python?, https://365datascience.com/tutorials/python-tutorials/sentiment-analysis-with-python/
5. Singh, K.: Sentiment analysis on lyrics of popular music artists, https://kvsingh.github.io/lyrics-sentiment-analysis.html