BANA 277 MARCH 2021

Sentiment Analysis: Podcast Reviews

Ankit Jain, Yu Hsin (Kathy) Lee,
Kevin Cheung, Mira Daya, Site Bai

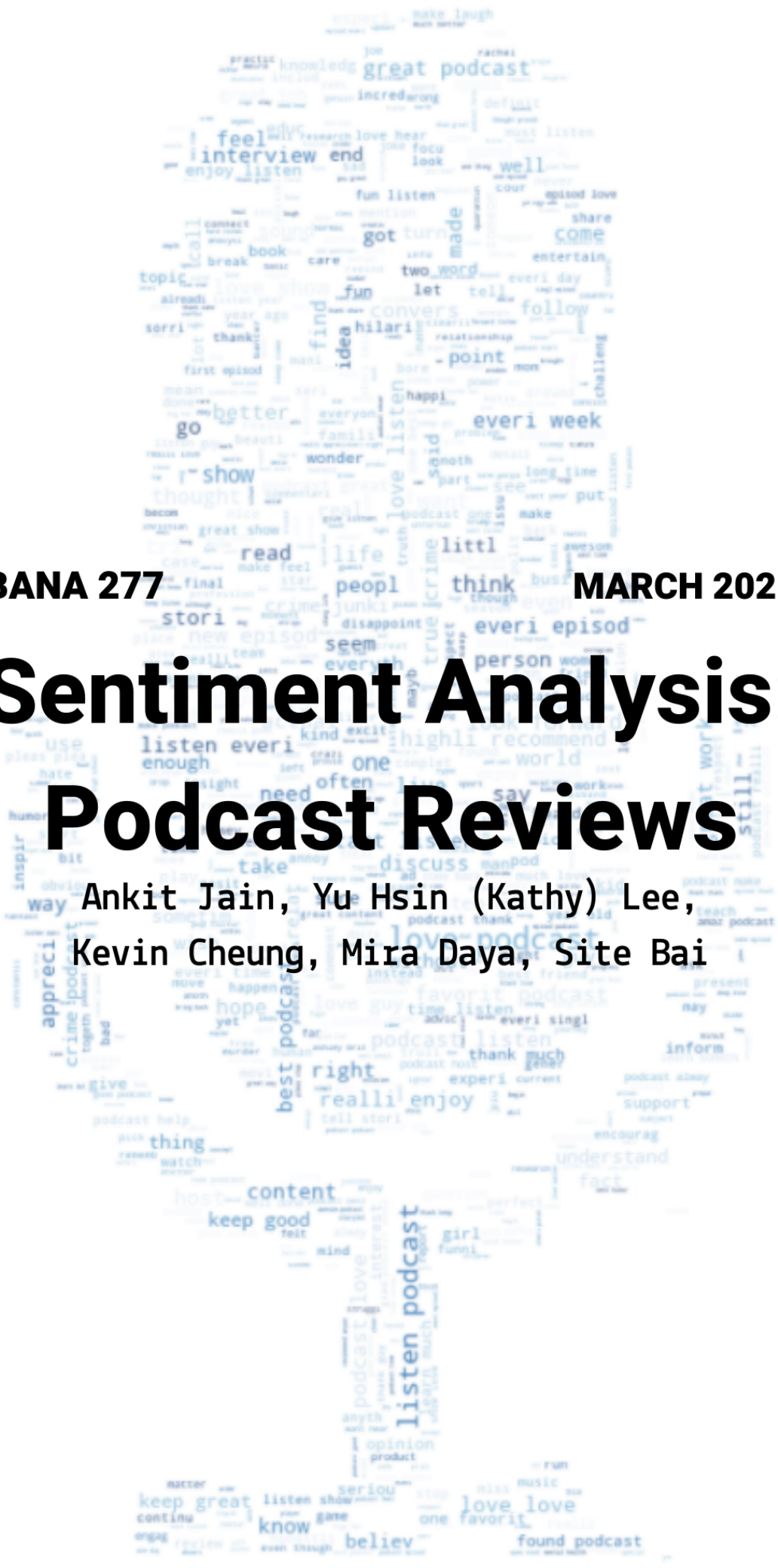

TABLE OF CONTENTS

I. Overview

Podcasts are becoming ever increasingly popular, ranging from a wide variety of topics such as the arts, society, business, career, religion, and so on. Specifically, Apple Podcasts currently is the only platform which offers reviews and ratings, whereas platforms like Spotify and Stitcher do not. We also have two teammates who have started a podcast on career and networking advice. We want to explore the business implications and the effects of reviews on the podcasts. By performing text, sentiment, and causal analysis on the Podcast dataset, we hope to determine:

1. What factors (time, category, and sentiment) influence the rating and review content?
2. How have podcast reviews changed over time in terms of sentiment and rating?
3. How certain historic events happening at the time might have affected the number of reviews and the ratings of the podcast? Do these events have a causal impact on reviews and ratings?

A. Business Implications

Ratings and Reviews are not only important for podcast creators in terms of receiving feedback and motivation, but they are also an indicator of how well a podcast is being received by audiences. These ratings are often used by potential advertisers and sponsors to gauge whether they should invest their time and funds in a podcast or not. Through this project, we hope to understand what factors (with a main focus on historic and current events, and time) may influence the rating and review of a podcast. We will provide recommendations to podcast creators about how they can possibly improve their rating and how they should approach their content creation process in the future.

II. Data

A. Data Acquisition

This Apple iTunes Podcast review dataset was collected from Kaggle, which consists of four tables: **Categories, Podcasts, Reviews, and Runs**. Table 1 below outlines a breakdown of the number of rows and columns in each data table. Table 2 outlines a breakdown of attributes in each data table.

Table 1: Table Dimensions

Data Tables	Total Rows	Total Columns
Categories	70952	2
Podcasts	46665	5
Reviews	1162840	5
Runs	9	3

Table 2: Tables & Attributes Overview

Data Tables	Attributes
Categories	podcast_id
	category
Podcasts	podcast_id
	itunes_id
	slug
	itunes_url
	title
Reviews	podcast_id
	title
	content
	rating
	created_at
Runs	run_at
	max_rowid
	reviews_added

Our main focus will be on the Reviews table, specifically the content variable, which has free text response reviews from users, and the rating variable, which rates the podcast on a scale from one to five (five being the best score). Furthermore, the category variable under the Categories table can provide some insight about which topics have the highest reviews.

B. Data Cleaning

Before performing any analysis, we wanted to make sure the dataset was appropriately cleaned for further exploration. We first checked for any NA values in all the tables, and found that there were no missing values. Furthermore, we proceeded to transform the *Created At* variable under the Reviews table to a DateTime object and created extra columns within the Reviews table, consisting of *Year*, *Month*, *Weekday*, and *Hour*. We also checked for any outliers within our *Rating* variable in the Reviews table, and found that all reviews were within the range of 1 through 5 (integers only). Additionally, we renamed *content* and *title* to *Review_Description* and *Review_Title* respectively.

C. Data Preprocessing

For textual analysis, we removed stopwords using nltk stopwords. We also removed additional stopwords relevant to our dataset included in Table 3 below. Additionally, we removed punctuation and white spaces from the reviews. In order to perform textual analysis, we needed to tokenize the reviews. We first tokenized each sentence into a list of sentences. Then, upon tokenizing each sentence, we split each sentence into a list of individual words or, in this case, tokens. We also converted any words with uppercase to lowercase characters and removed words with a length of less than 3 characters. After word tokenization, the tokens were stemmed using PorterStemmer, and the tokens were converted to stem words.

Table 3: Additional Stopwords Used

Additional Stopwords Added		
'Podcast'	'Show'	'Don'
'Listen'	‘Episodes'	'T'
'Podcasts'	'Used'	'M'
'Listening'	've'	‘&’
'Episode',	'Good'	'Re'
'Listened'	'Love'	'One'
'S'	'Listener'	'Like'
'Really'	'Will'	'People'

III. Exploratory Data Analysis

A. General EDA on Full Dataset

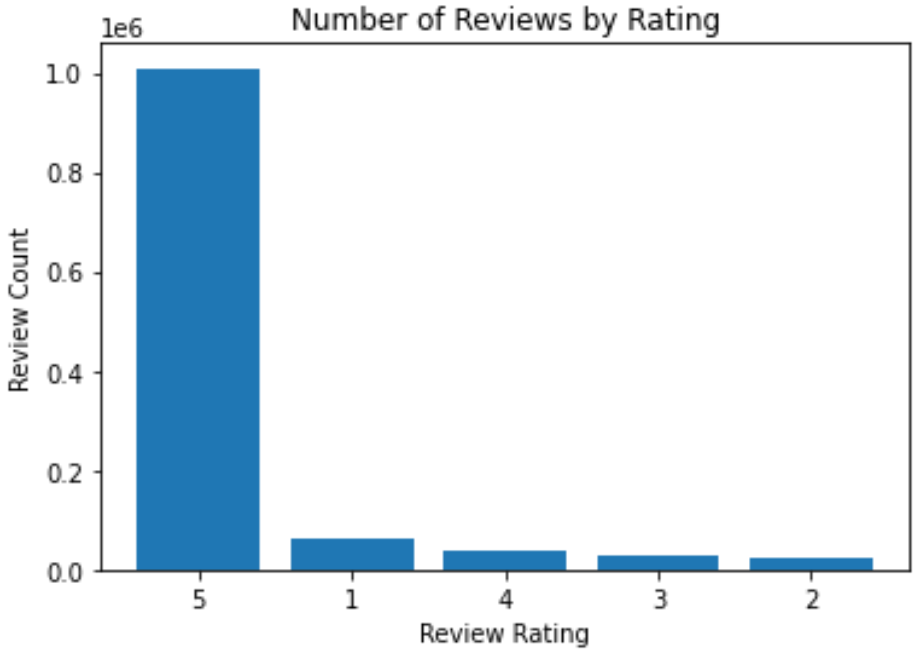

Figure 1

Figure 1 shows that rating 5 has the highest number of reviews and rating 1 has the second highest number rating. Mid-value ratings of 2 to 4 have lesser reviews compared to the extreme values. People seem to give more extreme value ratings, specifically more on opposite ends of the rating spectrum.

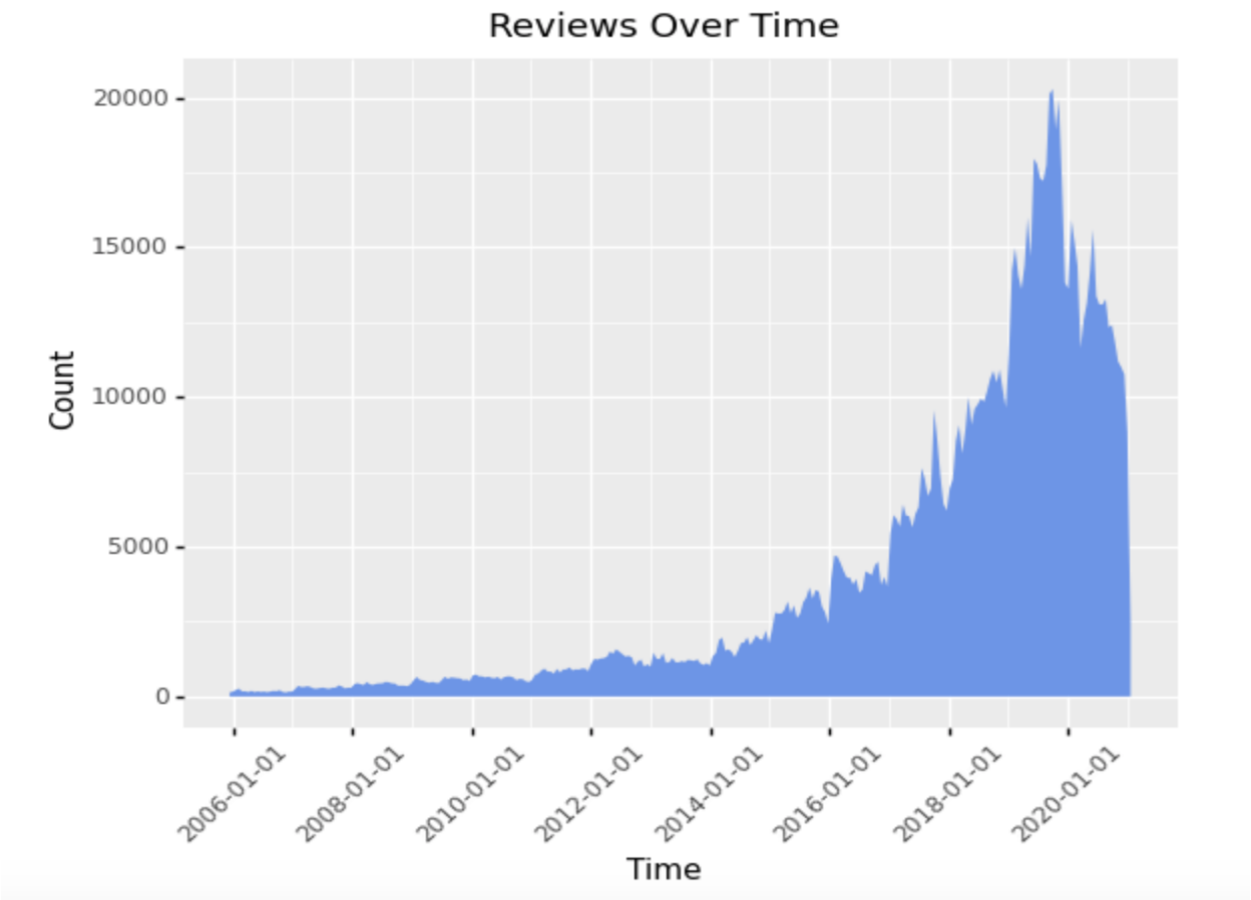

Figure 2

Figure 2 shows a dramatic increase in reviews over time. The quantity of reviews seem to increase drastically after 2014 and peak in 2019 with a drop going into 2020. We found this interesting as we expected a higher number of reviews in 2020, as more people were staying at home and may have had more time listening to podcasts.

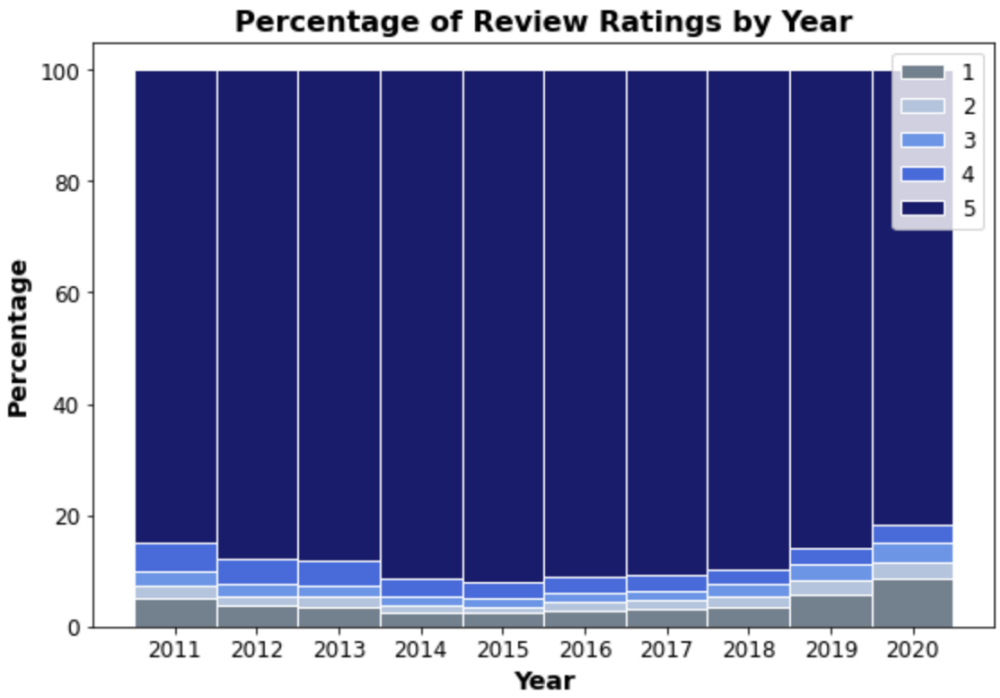

Figure 3

Figure 3 shows the percentage of review ratings by year. Starting in 2016, the percentage of reviews for rating 1 starts to increase, however starting in 2019, the percentage of reviews for rating 1 increased more than the previous few years. This pattern is also followed into 2020 where we can see that the percentage of 1-star reviews have increased the most. A possible reason for this increase can be due to the increasing popularity of podcasts themselves. Podcasts have become more popular in the past few years and people may become more critical and aware while writing reviews and giving ratings.

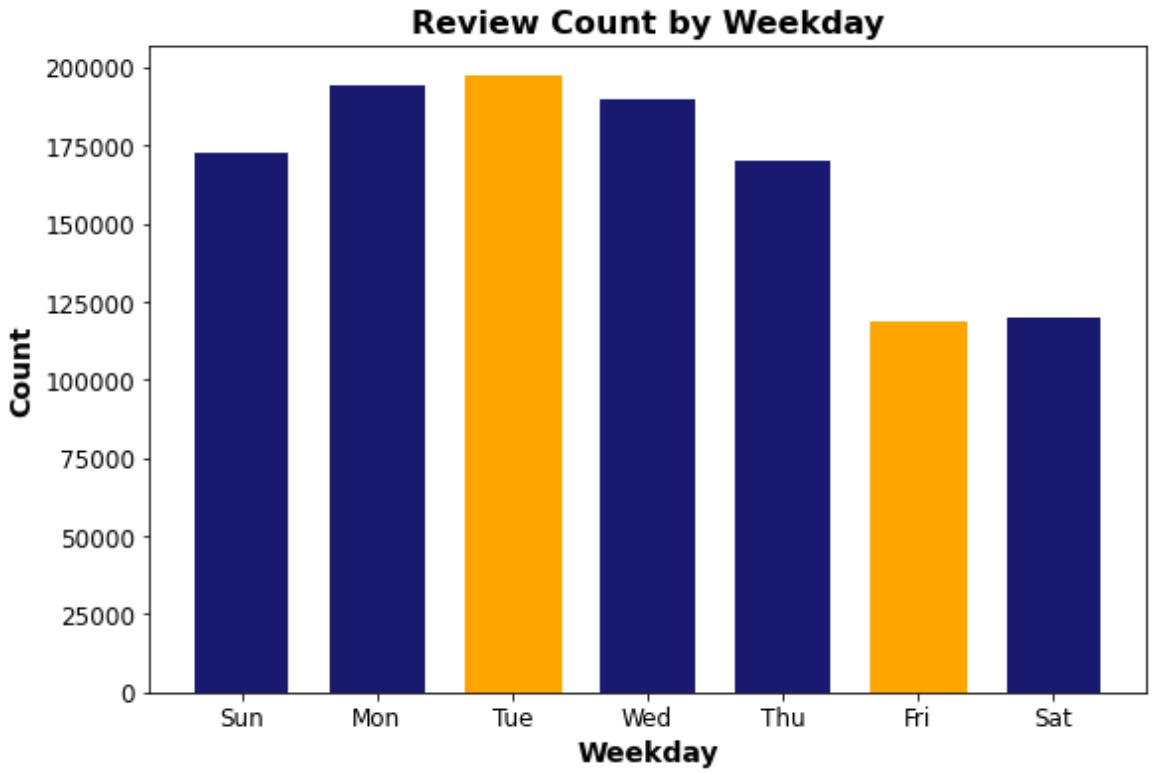

Figure 4

Figure 4 shows the distribution of reviews throughout the week. The most reviews occur in the early-mid point of the week from Monday to Wednesday with a peak on Tuesday. Additionally, reviews are at a minimum on Friday and also low on Saturday.

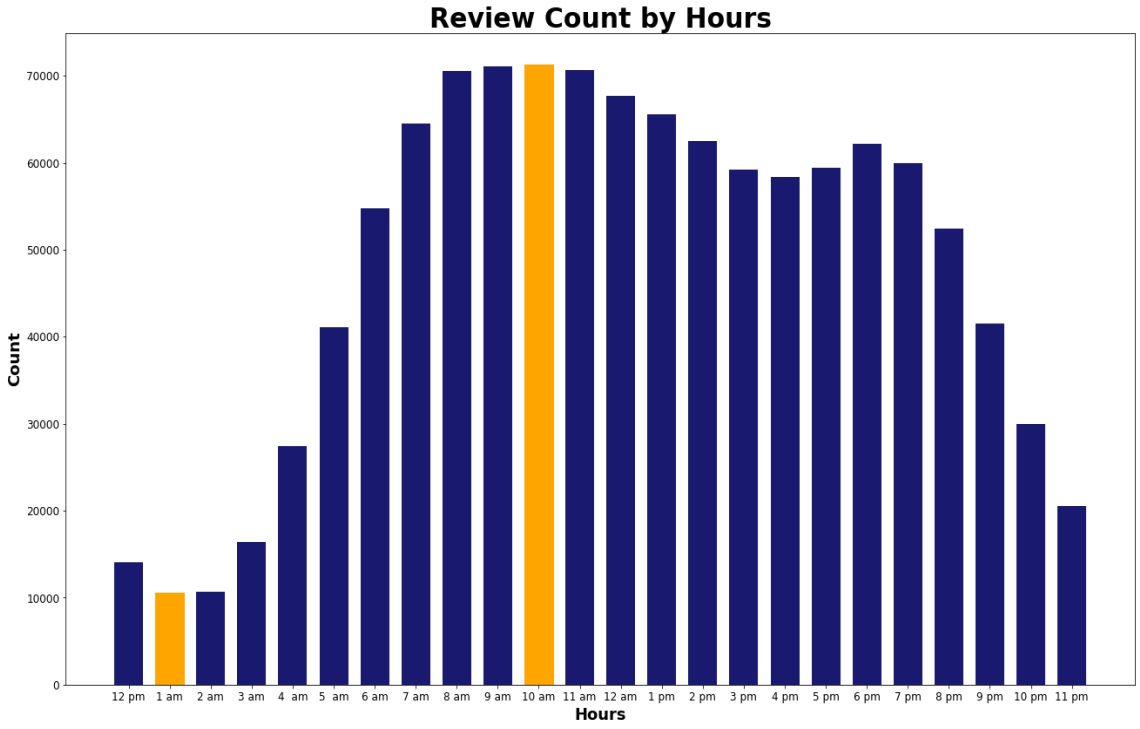

Figure 5

Figure 5 shows a similar graph but it is the distribution of reviews by hour. Notice the least amount of reviews are being left around 1AM while the most are around 10AM. Between the hours of 8AM to 11AM are the most active times for reviews being left by users.

Figure 4 and Figure 5 provide insights into telling us about when users are leaving reviews. Throughout our analysis, we are assuming that podcast listeners are leaving reviews right after they listen to a podcast. These figures can help us determine when the best time for a podcaster should release a new episode if their goal is to increase their number of reviews. This will be covered in Part VI, Key Takeaways & Conclusion.

B. Text Modeling

The next section of our exploratory analysis is on text modeling. Due to the large size of our dataset, from here on, we focus our analysis on data from 2020. In order to utilize review description data, we preprocessed the *Review Description* through tokenization and stemming as mentioned earlier. Additionally, we utilized a TF-IDF, Term Frequency-Inverse Document Frequency, vectorizer for feature extraction and NMF, non-negative matrix factorization, as our topic modeling technique. The TF part of TF-IDF, summarizes how often a given term appears within a document and the IDF part downscales terms that appear more frequently across the corpus, or collection of

documents. TF-IDF quantifies a word in documents, which computes a weight to each word, signifying the importance of the word in the document and corpus. The TfidfVectorizer will tokenize documents, learn the vocabulary and IDF weightings, and returns a matrix of TF-IDF features. The TfidfVectorizer first applies the CountVectorizer, then applies the TfidfTransformer, which takes the count matrix and normalizes it into vector formats with the corpus as length and features (tokens) as width with the weight under each token. NMF is based on linear algebra, which uses the original matrix (A) from the TfidfVectorizer and gives two matrices (W and H). W is the topics found and H is the weights for those topics. In other words, A is documents by words, H is documents by topics, and W is topics by words.

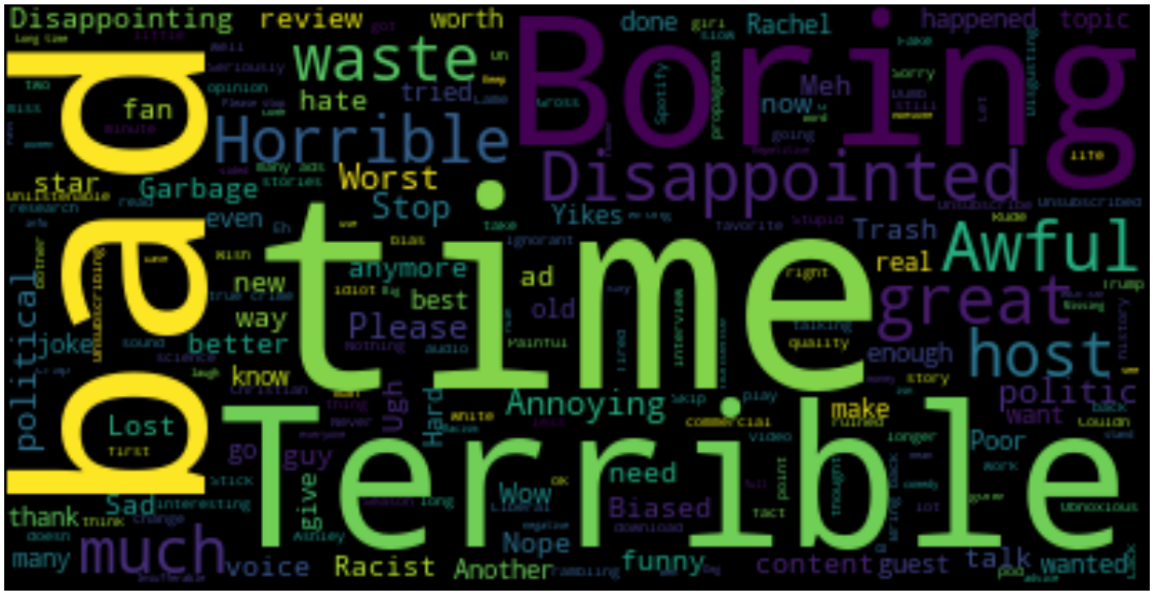

Figure 6A: Word cloud for 1-Star Ratings

Figure 6B: Word cloud for 5-Star Ratings

First, we created the word clouds in Figure 6A and 6B. These word clouds illustrate the most commonly occurring words for 1-star and 5-star ratings, respectively. We chose to focus our illustrations on these ratings because as we noticed in Figure 1 (from Part III.A), the highest frequency counts are for 5-star ratings followed by 1-star ratings. Notice for 1-star ratings, Figure 6A, words such as “bad”, “terrible”, “boring”, “waste”, and “time” show up. On the other hand, for 5-star ratings, Figure 6B, words like “Best”, “Great”, “Awesome”, and “Amazing” show up. This shows us how different 1-star ratings and 5-star ratings are with 1-star ratings being a lot more negative and 5-star ratings being very positive.

Table 4

Topic #0:love podcast absolut love love podcast much podcast much podcast love love love realli love podcast great year old podcast alway much love also love podcast thank podcast make podcast keep
Topic #1:listen podcast love listen podcast must listen enjoy listen time listen podcast everi podcast help podcast like stop listen learn much like listen podcast realli everi time podcast one podcast make
Topic #2:look forward look forward everi forward everi everi week podcast everi learn much absolut love best friend podcast alway great content first episod found podcast listen sinc well done love hear

In Table 4 above, we clustered the reviews through topic modeling and included a snippet of what some of the topics that show up the most represent. We can see that in topic 0, users who included a review mentioned “love podcast”, “absolutely love”, and “really love” a lot, thus these users can be grouped into a “love” topic. Furthermore, in topic 1, the word “listen” occurs often and in topic 2, the words “look forward” and “every week” appear frequently as well, thus these topics can be grouped into “listeners” and

"optimists", respectively. Through this snippet, we can see that users are generally very optimistic and positive in their reviews.

Table 5

	tfidf
love podcast	3.699773
listen podcast	3.925193
look forward	4.324125
feel like	4.398882
podcast listen	4.732202
true crime	4.738565
love listen	4.760473

Table 5 represents the TF-IDF score. The TF-IDF score represents which words are showing up the most through our data. A lower TF-IDF score means the words have a higher frequency in the dataset.

We decided to analyze bigrams and trigrams, which are 2 word and 3 word phrases, and sorted these n-grams based on their lowest TF-IDF score. We can see that phrases such as "love podcast", "look forward", and "love listen" appear the most, reinforcing our findings from topic modeling results in Table 4 above and that our dataset is heavily skewed towards positive, 5-star ratings.

IV. Sentiment Analysis

Sentiment Analysis can help us understand the sentiment and gather insightful information in text. We used TextBlob under NLTK to perform sentiment analysis. TextBlob returns a polarity and subjectivity score in a text. Polarity lies between -1 and 1, inclusive, where -1 has a negative sentiment and +1 has a positive sentiment. Subjectivity lies between 0 and 1, inclusive, where subjectivity measures the amount of personal opinion and factual information in the text. The higher the subjective sentiment score means the text contains more personal opinion rather than factual information. We opted to use TextBlob's default sentiment analysis implementation called PatternAnalyzer. In creating our sentiment analysis function, if the polarity score returned a positive number, then we categorized the profile as "Positive", if the polarity score returned a negative number, then we categorized the profile as "Negative", and if the polarity score equalled to zero, then we categorized the profile as "Neutral". Looking at the overall results, Figure 8B shows 153,979 reviews were positive, 16,304 reviews were negative, and 14,707 reviews were neutral. We decided not to include subjectivity within our analysis as reviews are already a users personal opinion and would not provide much insight for the purposes of this project.

A. General Sentiment Analysis

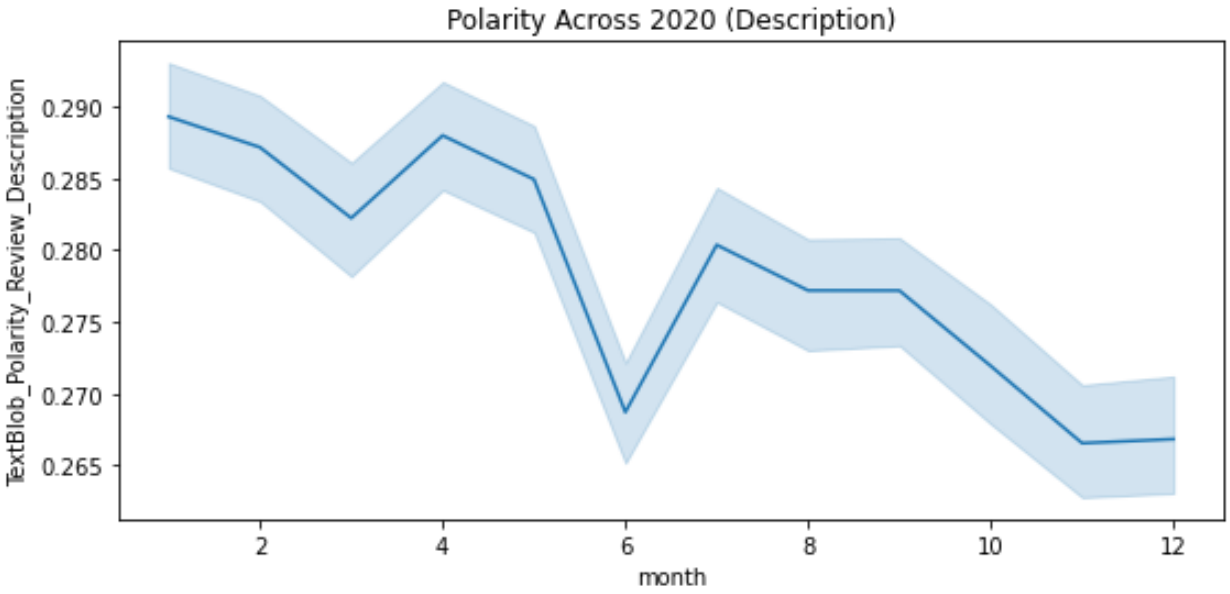

Figure 7

The line plot above shows sentiment through 2020 and is generally positive but decreasing throughout the year, with a large dip in polarity in June, which will be further analyzed in section IV.B.

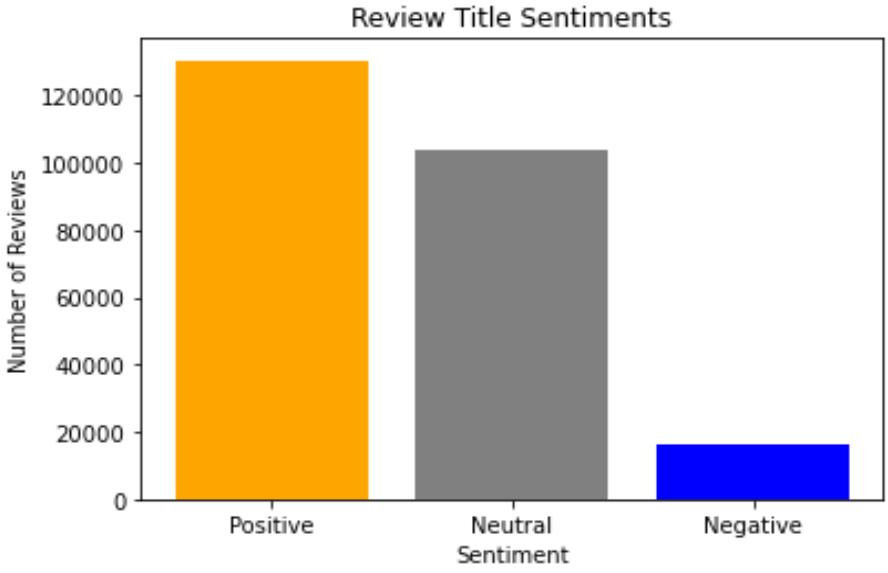

Figure 8A

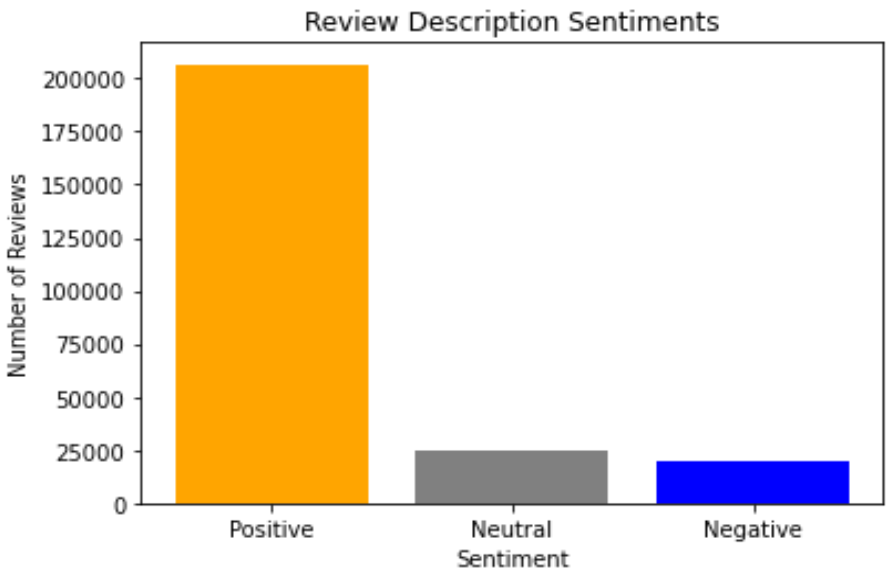

Figure 8B

Figures 8A and 8B represent the sentiment for review titles and descriptions respectively. Notice that review titles are more neutral whereas review description has a higher proportion of positive sentiments compared to negative and neutral. Review titles are usually to the point and have a much smaller character limit compared to description so people cannot express their full sentiment properly in the title alone. However, the description provides more detail about the sentiment of the review. This reinforces the finding that titles are shown as more neutral, while descriptions tend to have more positive or negative sentiment added. Therefore, looking at the title alone may not be useful in determining how well a podcast is received by the users.

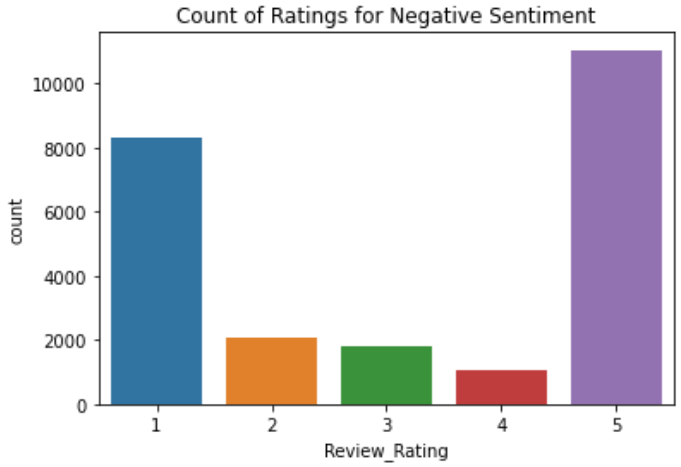

Figure 9A

Huh?	!!!!!!!!!!!!!!!!!!!! What the heck is this ...
Repetitive	Worst podcast i have listened to.

Example 9A

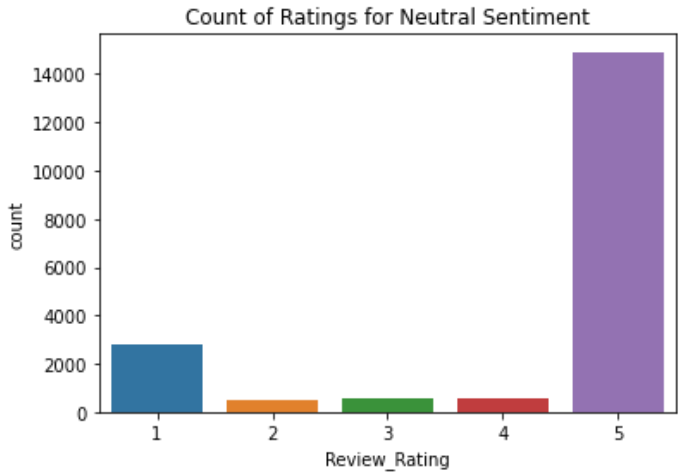

Figure 9B

Review_Title	Review_Description
Gabe	I used to eat my bacon now I save it and liste...
5 stars	Quality Blink discussion. But please, do the l...

Example 9B

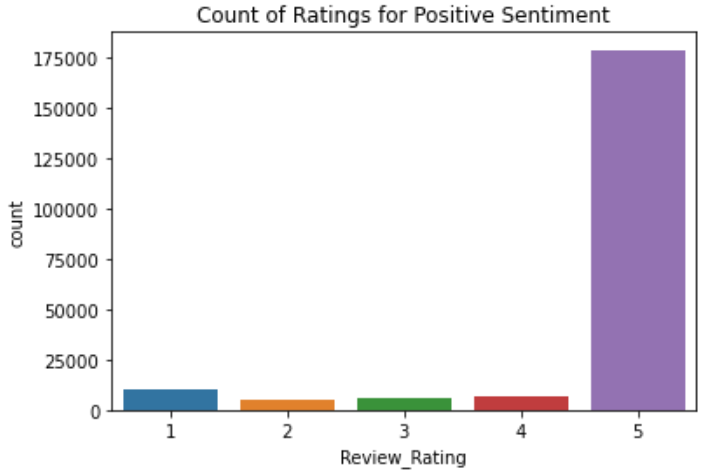

Figure 9C

Review_Title	Review_Description
So good	Keep up the great work!! This podcast always h...
Addicting!	Best podcast! Got my mom hooked on it too!

Example 9C

In Figures 9A-9C and examples 9A-9C, we plotted the count of ratings against each sentiment. We can see that the count of reviews with a Rating of 1 has a higher proportion compared to the other figures in Negative sentiment. Furthermore, the count of reviews with a Rating of 5 increases as the sentiment becomes better from a count of a little more than 10,000 for negative sentiment, approximately 15,000 for neutral sentiment, and a high of approximately 175,000 for positive sentiment. For each figure, we showed a snippet of the corresponding review descriptions. For example, we see that the word "worst" shows up for negative sentiment and the word "best" shows up for positive sentiment.

B. Sentiment Analysis by Historic Trends

a. General statistics of rating trend in 2020.

i. COVID-19, Black Lives Matter, and the Election

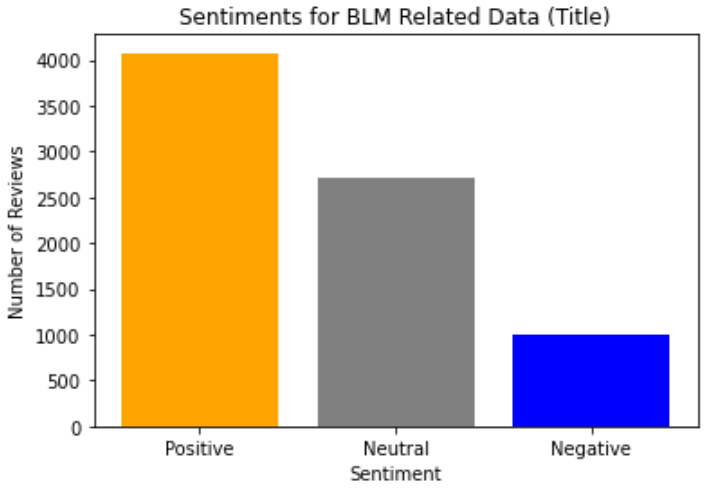

Figure 10A

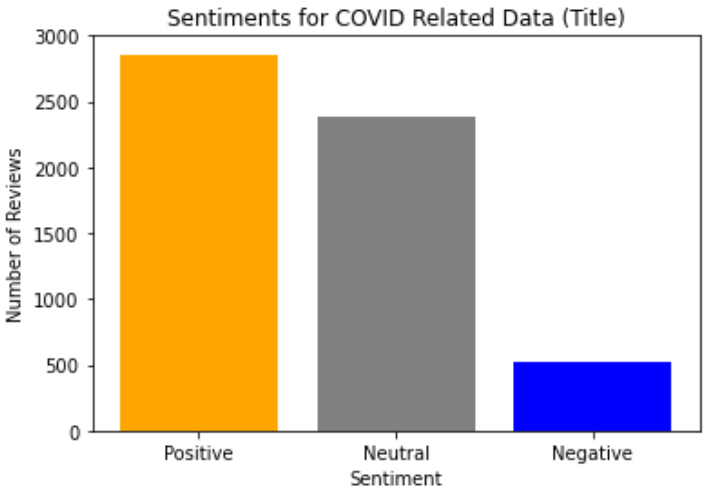

Figure 10B

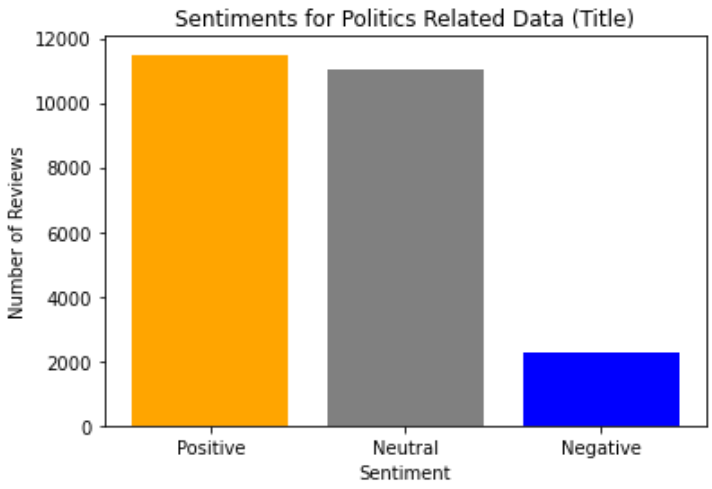

Figure 10C

Before conducting causal impact analysis, we plotted sentiment frequency for review titles of reviews related to historical events of COVID-19, Black Lives matter protests, and the election. Notice in Figure 10A that review titles related to the Black Lives Matter protests have a higher proportion of negative sentiment than those in Figure 10B for COVID-19 and, Figure 10C the election.

ii. Further analysis for BLM protests

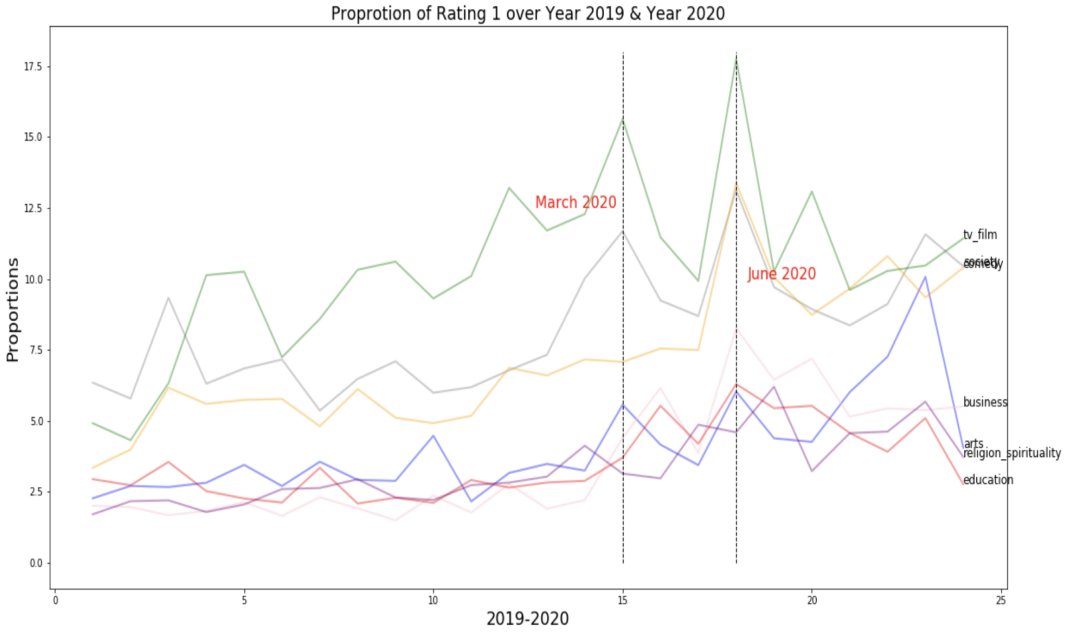

Figure 11A

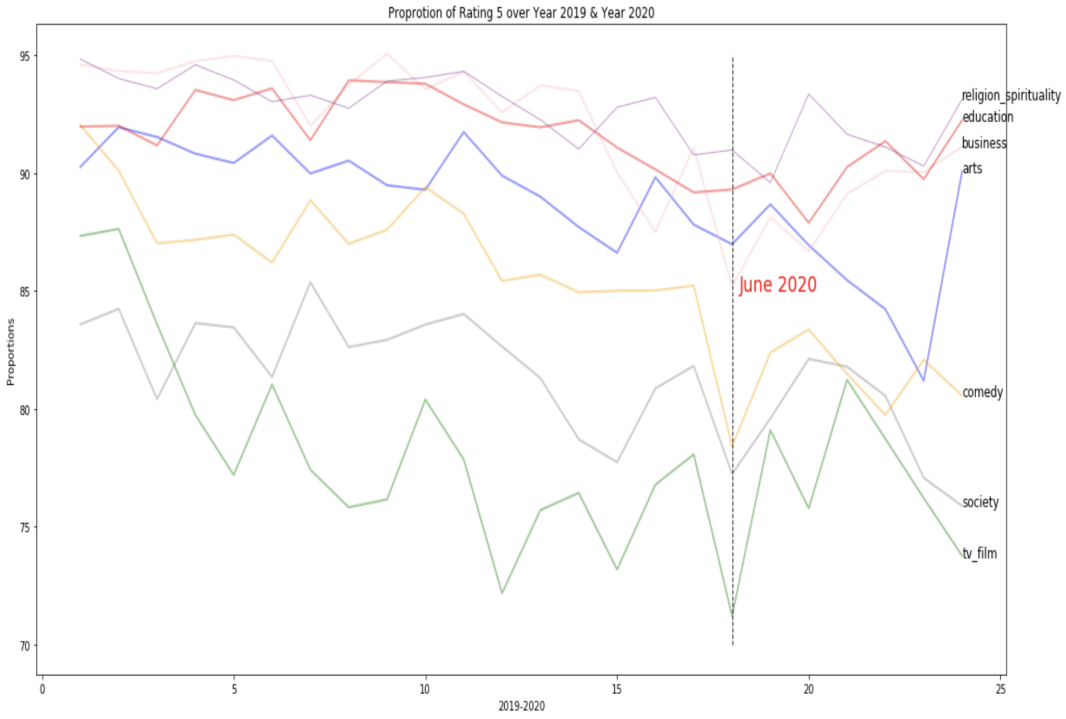

Figure 11B

Next, we graphed the trends of proportion of rating 1 and proportion of rating 5 respectively of top 7 categories having the most reviews in 2020. Figure 11A shows that there are two spikes for the proportion of rating 1 that occur in the months of March and June. In these two months, two events were taking place, first, when the Covid 19 pandemic hit America, and second, when Black Lives Matter protests spread across the country, respectively.

b. Causal Impact Analysis

Based on the observations displayed in part a above, we wanted to identify if historical events, such as Black Lives Matter protests, have a causal effect on podcast ratings.

We decided to run a causal analysis on the effects of the Black Lives Matter movement on podcast ratings. To do this analysis we used Python's causal impact package . We use the Google trends identified keyword "Black Lives Matter " as the treatment to fit into the causal impact model. We learned that in 2020 the Black Lives Matter movement does have a significant positive effect on the proportion of rating 1. Additionally, the protests also have a significant negative impact on the proportion of rating 5.

```
The probability of obtaining this effect by chance is very small
(Bayesian one-sided tail-area probability p = 0.0).
This means the causal effect can be considered statistically
significant.
```

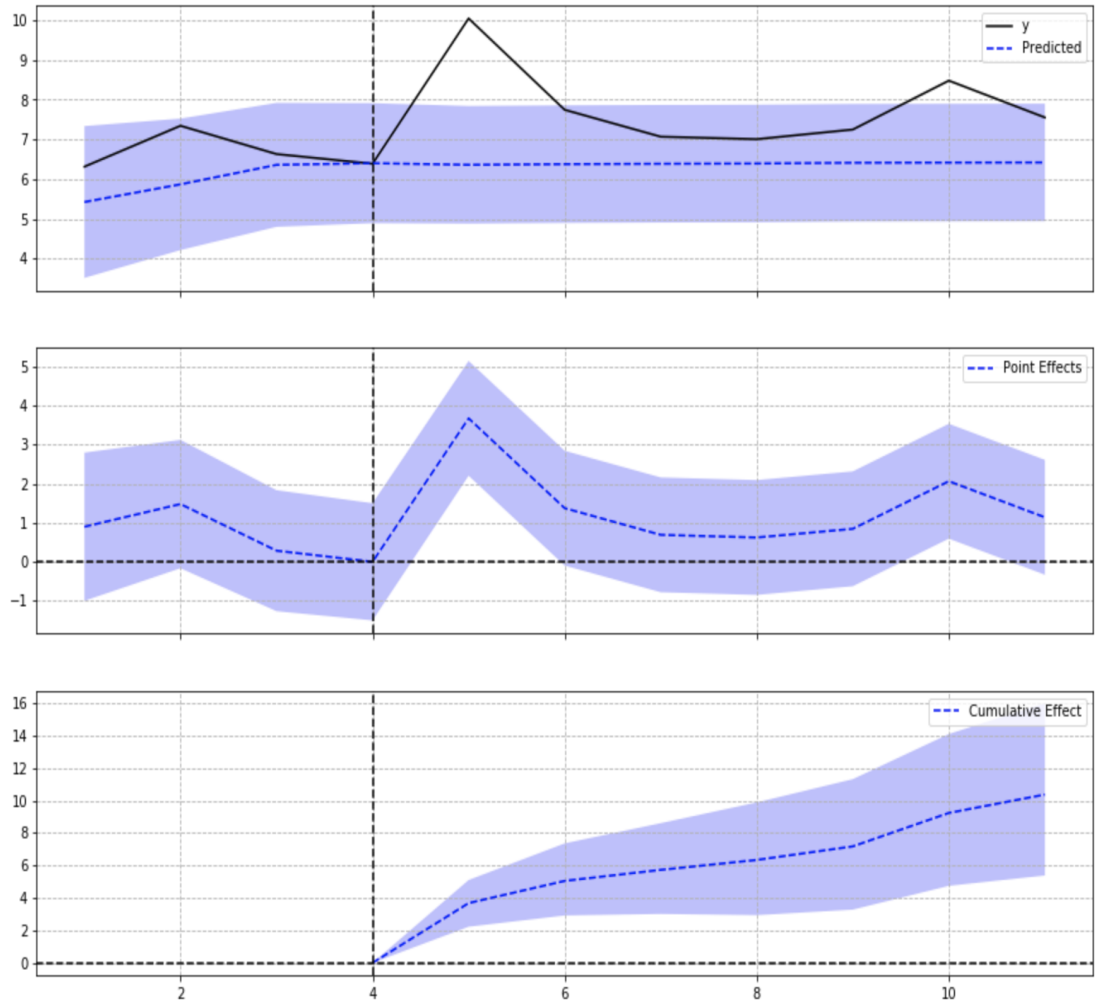

Figure 12A: Results of causal analysis for rating 1

Figure 12A shows the results for fitting the model on proportion of rating 1. The treatment that we are using is the Black Lives Matter protest. The blue dotted line represents the predicted trend of proportion of rating 1 without treatment, whereas the black solid line represents the actual trend of proportion of rating 1 with the treatment. The results determine a p-value equal to zero which indicates that the treatment has a significant impact. The difference between the predicted trend and the actual trend is significant.

```
The probability of obtaining this effect by chance is very small
(Bayesian one-sided tail-area probability p = 0.0).
This means the causal effect can be considered statistically
significant.
```

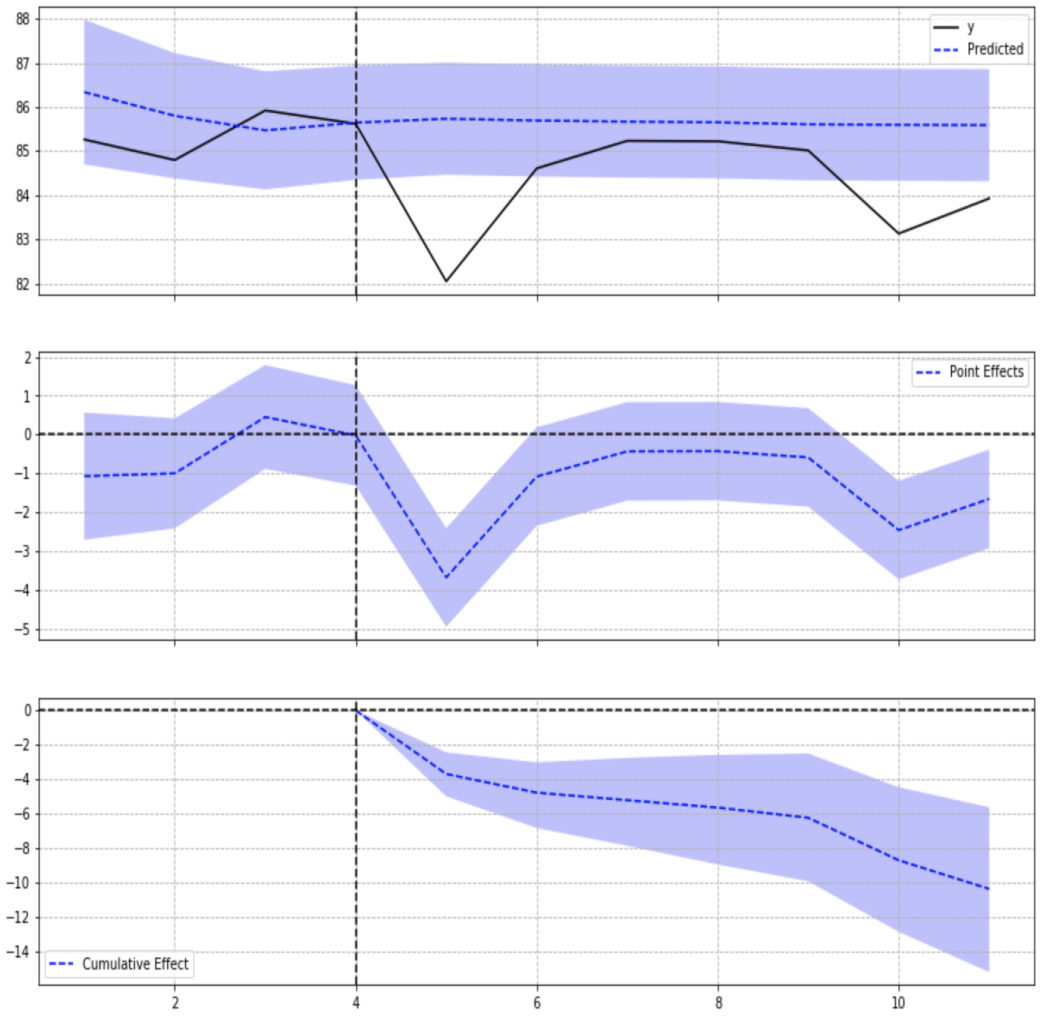

Figure 12B: Results of causal analysis for rating 5

Figure 12B shows below for fitting the model on proportion of rating 5. The treatment that we are using is the Black Lives Matter protest. The blue dotted line represents the predicted tread of proportion of rating 1 without treatment, whereas the black solid line represents the actual trend of proportion of rating 1 with the treatment. The results determine a p-value equal to zero which indicates that the treatment has a significant impact. The difference between the predicted trend and the actual trend is significant.

Rating: 1
Word Cloud for arts-fashion-beauty

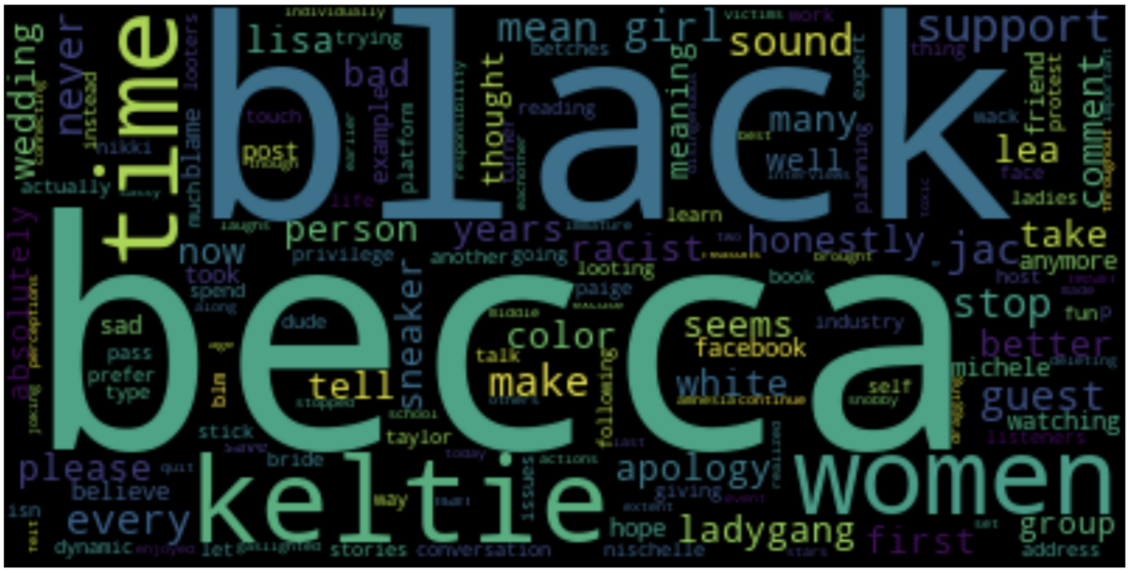

Figure 13A

Rating: 1
Word Cloud for tv-film

Figure 13B

Rating: 1
Word Cloud for comedy

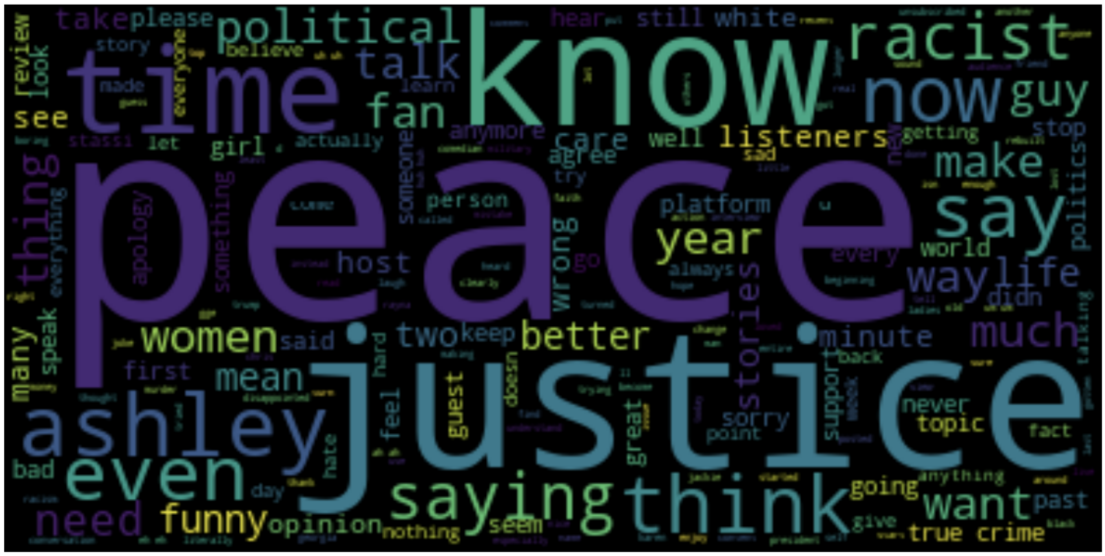

Figure 13C

Rating: 1
Word Cloud for society-culture

Figure 13D

Rating: 1
Word Cloud for arts-performing-arts

Figure 13E

To further validate the effect, we also created word clouds for the reviews description and review titles of those podcasts having the highest proportion of rating 1 in 2020. Figures 13A through 13E show these word clouds. The most frequently occurring words as shown are highly relevant to the protests and explain the rise of proportion of rating 1, which is consistent with the outcome obtained from causal impact analysis. We can see that words such as “black”, “racist”, “peace”, “white”, and “justice” appear the most.

C. Sentiment Analysis by Category

There were 26 categories in the dataset, so we generalized them into 10 main categories: Arts, Society-Culture, Business, Technology, TV-Film, Comedy, Kids-Family, Religion-Spirituality, Education, Music.

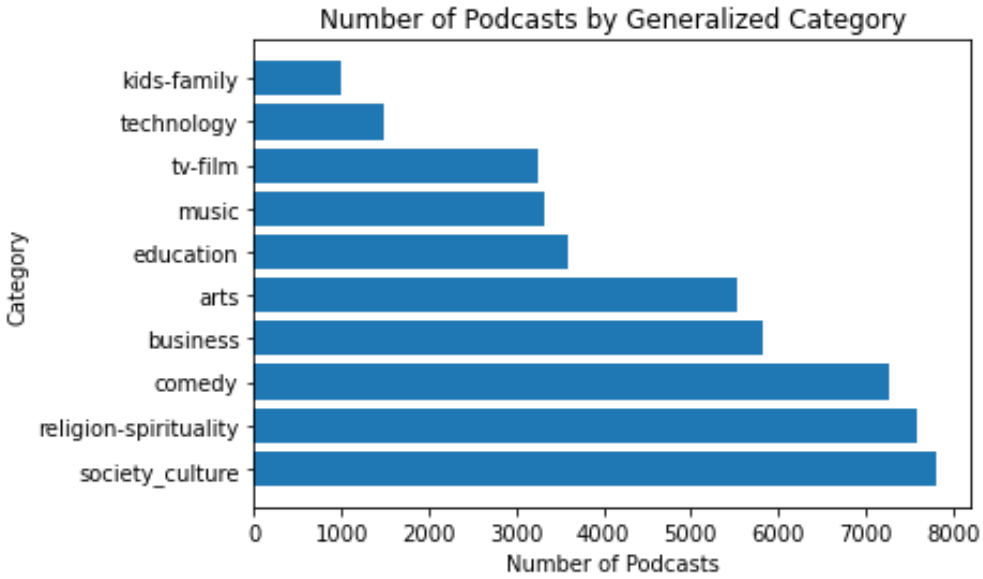

Figure 14A

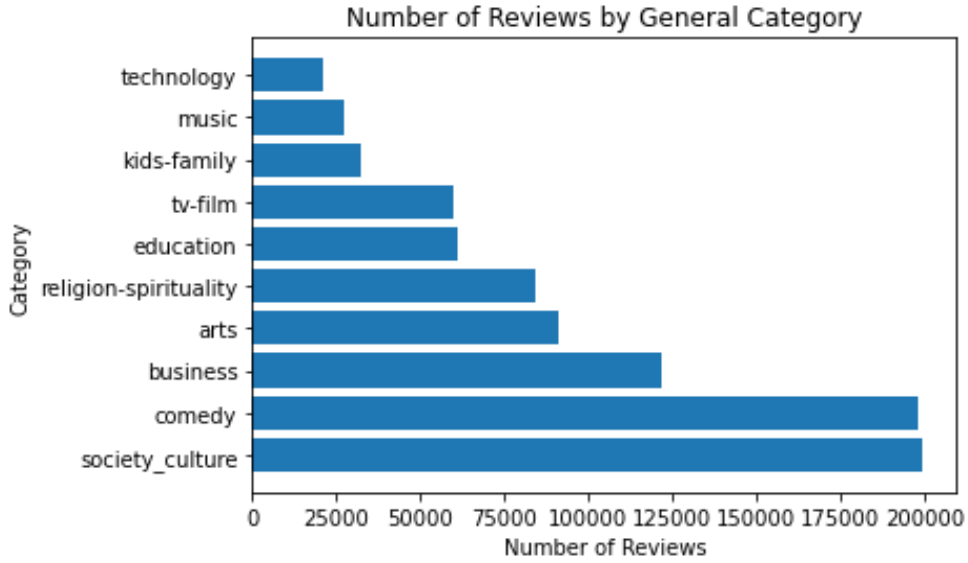

Figure 14B

Figure 14A and 14B shows us that the most reviews and number, respectively, of podcasts were for the category Society-Culture. Comedy came in a close second for number of reviews and third for number of podcasts.

In the previous section, we discussed the causal impact of historic events such as COVID-19, BLM movement, and the election. What is interesting to note is that in the year 2020 the two categories that had the most reviews during the period of the Black Lives Matter Protests in June were TV-Film and Society-Culture. During this time, people were expressing their positive and negative opinions of different podcasts hosts through their reviews.

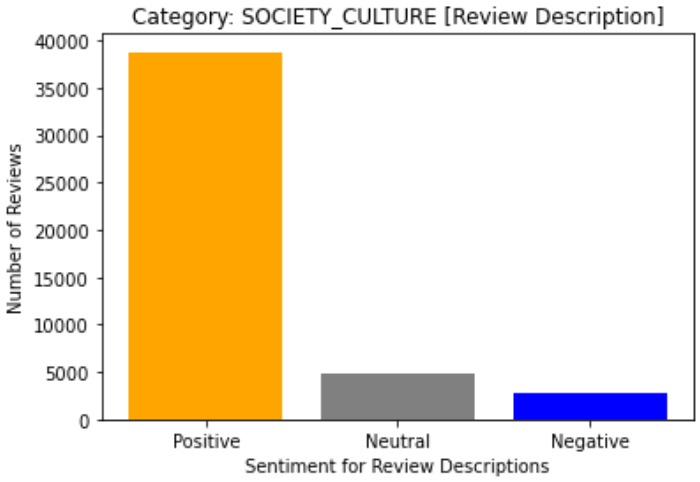

Figure 15A

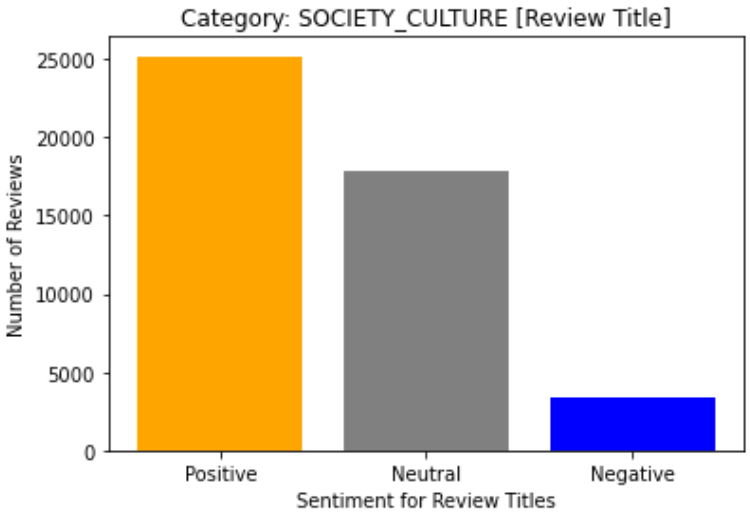

Figure 15B

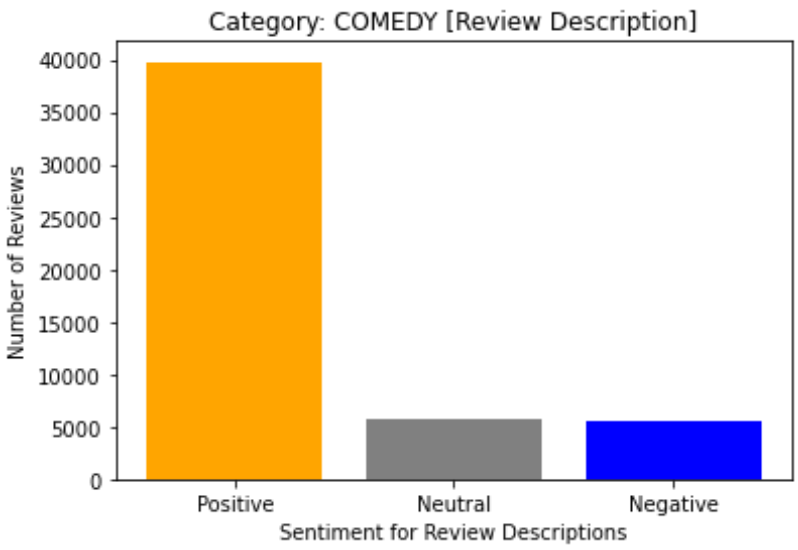

Figure 16A

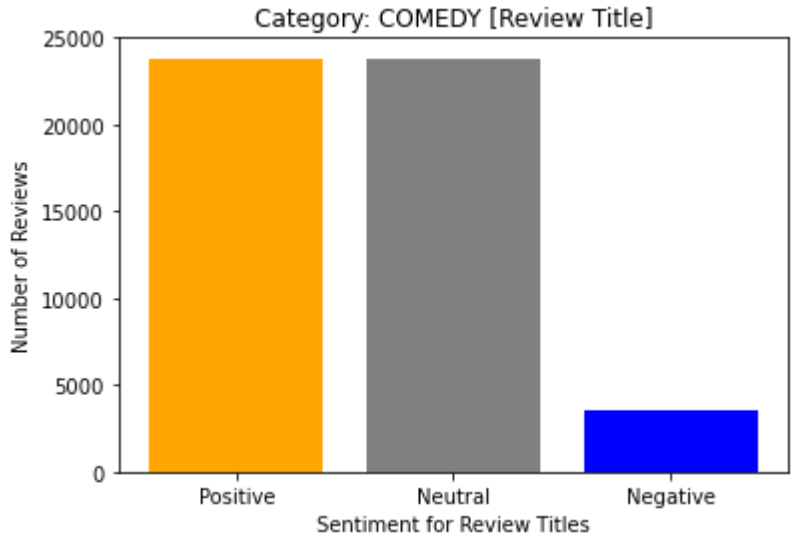

Figure 16B

We have highlighted review description and title sentiment for the Society and Culture category above in Figures 15A and 15B which has the most reviews. These graphs are representative of a similar trend noticed in all the other categories as well. It seems that most reviews were positive. However, what is interesting is that for review titles, there was a greater percentage of neutral review titles, and in the case of comedy category, it was almost equal to that of the positive reviews which can be seen in Figures 16A and 16B.

Noticing this, we wanted to explore the sentiment of the combination of review titles and descriptions.

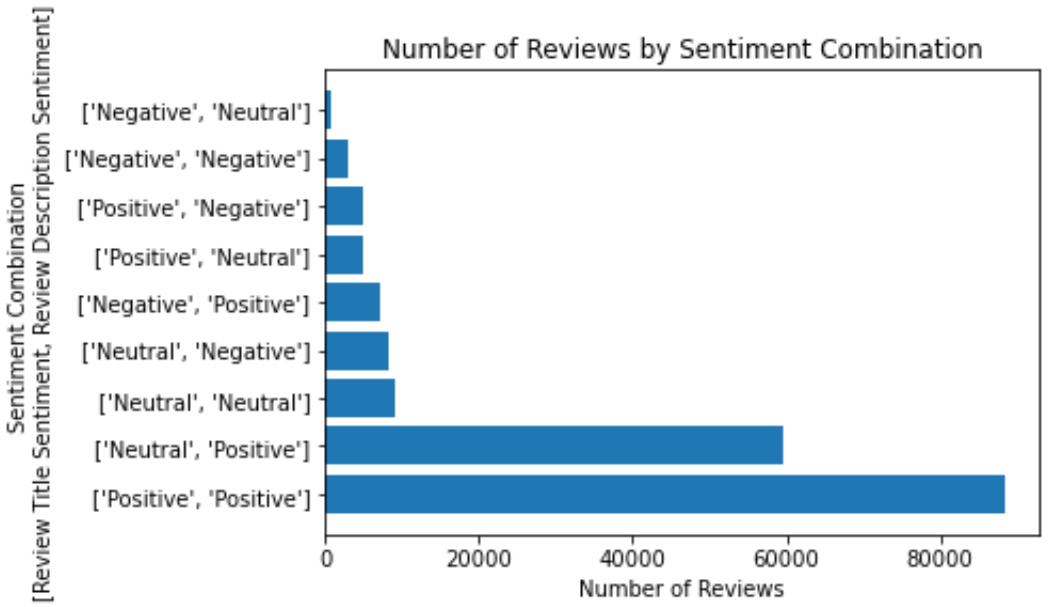

Figure 17

Notice that Figure 17 most reviews seem to have a combination of the description and title both being positive, with a neutral title and a positive description as the second most occurrence. In the case of negative reviews, most reviewers seem to more likely put a neutral title with a negative review as opposed to putting a negative title and review. This is likely due to the fact that there is a lower character limit for titles and people tend to give more details in their actual review descriptions, as opposed to their titles.

V. Future Considerations & Improvements

In the future, we hope to perform analysis on peer influence. Peer influences can encourage users to review a podcast with a high or low rating. We hope to perform social network analysis with listener data to understand what peer influences might be apparent.

Additionally, having information on demographics of users and podcasters may also provide some helpful insights into what may make a podcast more successful or what demographics to target to obtain more listens or higher ratings.

We also would like to perform more analysis of the full dataset to see if there is any seasonality in the sentiment or ratings of the data and make conclusive statements. Also, based on our analysis we found that historical events do seem to have a causal impact on 1 and 5 star ratings. We would like to take that further in creating a model to possibly prepare podcasters on how to work with and around historical events topics in their podcasts.

Furthermore, an important addition is to find data related to the number of listens and subscribers of a podcast to relate popularity of a podcast at time and their average rating at the time.

VI. Recommendations & Conclusions

From our analysis, we have some recommendations for podcasters, sponsors, as well as the Apple Podcasts Platform:

1. **Exploratory Data Analysis**
   First, we are assuming that most people review podcasts directly after listening to them. That being said, through exploratory data analysis we noted the best time and day for podcast releases and promotions to receive the most feedback and coverage. The time of the week best suited for this goal includes early weekdays of Monday through Wednesday, and time wise to be in the early morning hours of 8 am to 11 am. The most ideal time we deduced was to be Tuesday at 10 am[1]. We recommend that both sponsors and podcasters try to promote/release their podcast and brands during these peak hours of engagement as it has shown to create more traffic and brand engagement in terms of feedback and reviews, of which mostly has been positive.

2. **Causal Analysis**
   Through causal analysis, we found that historic and current events have significant impacts on the proportion of 1-Star and 5-Star ratings. These ratings are really important for podcasters as they are representative of how listeners receive their podcast. We recommend that podcasters be more cognizant of what events are happening around in the world and what they say associated with those events as it can have extreme effects on ratings and reviews.

   We recommend doing research properly on the events going on and understanding it before putting the opinion out to be scrutinized by the public, as it can have both extreme positive and negative impacts on the state of the podcast itself.

3. **Sentiment Analysis**
   From the sentiment analysis, we found that the polarity score decreased over the year, i.e. the reviews became more and more negative. Podcasters should be aware of sentiment at different times of the year. They can plan out more efficiently if they are aware of how the general public of podcast listeners are

[1] Check Figure 4 and 5 for visual representation of this finding.

feeling. The time of year can affect a podcast's ratings, as users seem to be more critical at certain points of the year.

4. **Apple Podcasts Implementation**
   Apple Podcasts can leverage that at the moment, they are the only platform that allows users to provide ratings and reviews to podcasts. Through our analysis, we would recommend Apple Podcasts to create a platform that connects sponsors and podcasts. Through reviews, they are offering the indicator of how well a podcast is received by an audience. This could possibly become a one-stop platform for future podcasters by allowing sponsors and podcasters to work together.

Podcasting is a way to get information and opinions out into the public. It is important to get feedback for the content that podcasters are putting out there. Reviews stand as a motivation for podcast creators to keep putting out relevant content for their listeners. With our analysis, we hope to help podcast creators understand what they can do to obtain more feedback and better reception to maximize the benefits from producing their podcast.