Social Media Users' Notions of Popularity: Capstone Milestone Report

Stephanie Sears

I. Introduction

This Capstone Project is a text mining and sentiment analysis study. It focuses on the R&B music of Patti LaBelle, Whitney Houston and Mariah Carey between the period of 1980 and 1995. By analyzing social media comments, it seeks to discover descriptive themes pertinent to popular African American Women R&B singers of that time period. Given the salience of the idea of the "diva," this project seeks to uncover how audiences identify with the top divas of R&B and the ways in which their interactivity with media constructs notions of "the popular." It uses data science to understand the audience's interactive behavior that correlates to popularity and aims to build a self-consistent explanation or model that might interpret that behavior.

Problem

With the advent of social media metrics such as "likes," "follows," "pins," "shares," "comments," and "views," the everyday connoisseur of culture enjoys a greater stake in the social construction of value. Indeed, cultivating fandom, an effect of the ease in which audiences now express "belonging" and "non-belonging" through mediated platforms, is currently a cultural practice. Its prevalence has transformed how people interact and intersect with popular culture. As it is now possible to become a career "influencer" utilizing social media platforms, it is of profound interest to researchers to examine the ways in which cultural consciousness defines what's valuable and how such definitions gain considerable social force.

That an influential social force is garnering strength through mediated platforms is no new phenomenon. But the growing importance of popular culture as a context of public persuasion is compelling. In such a context we witness the transformation of popular expression into impactful commentary that can either launch a public figure into stardom or topple an embattled icon, ensuring a purposively shameful fall from grace. Once relegated a marginal influence, popular culture has became an integral force in the socialization of collective consciousness. As a result, the everyday person participates in the curating of the collective self-image. A vote cast (as a "like", a "share", or a "follow," etc) represents the collective self-image of society and their awareness of the current (historical and political) moment in time.

Religious scholar, Robin Sylvan (2002) explores the ways in which popular culture has morphed into a religious phenomenon, particularly as experienced through popular music. He writes,

> "Observers of culture and scholars of religion have said many things about the slow decline of religion and the death of God in Western civilization….[R]eligion and God are not dead, but very much alive and well and dancing to the beat of popular music; the religious impulse has simply migrated to another sector of the culture, a sector in which religious sensibilities have flourished and made an enormous impact on a large portion of the population….Yet, because conventional

> wisdom has taught us to regard popular music as trivial forms of secular entertainment, these religious dimensions remain hidden from view, marginalized and misunderstood" (p. 3).

Sylvan seeks to make explicit the ways in which popular contexts reveal hidden examples of contemporary religiosity. As a result, he raises the value of the notion of "the popular." As contemporary folks elect a "popular," they raise new deity forms as well (via the collective self-image). In addition, they reconstitute notions of "the sacred" and "ritual" (by casting a vote for what's deemed MOST valued).

Data science offers an exciting possibility to test the relationship between social media user activity and socio-cultural definitions about value. Given Sylvan's claim that popular culture co-creates new forms of divinity, the construct of popularity is in need of examination. Without understanding the ways in which social media users create data, these new forms of value remain hidden. Equally true, without proper attention to this phenomenon, it is likely that unintended psycho-spiritual consequences remained unexamined as well.

II. Data

Two datasets are used for the purpose of this project. Each dataset was created by scraping YouTube channels, using the R package called Tuber. Descriptions of the dataset are as follows:

1. Comments Data

The comments dataset, saved to a data frame entitled all_comments, contains 7901 observations representing user commentary on the videos of Patti LaBelle, Whitney Houston, and Mariah Carey. The key variables of this dataset are the commentary text and the video identification numbers associated with the video from which the comments are drawn.

Code to scrape comments from YouTube channels and then to create all_comments data frame:

```
comments1 <-
  lapply(as.character(videos1$video_id),
         function(x){
           get_comment_threads(c(video_id = x), max_results = 100)
         }
  )

comments2 <-
  lapply(as.character(videos2$video_id),
         function(x){
           get_comment_threads(c(video_id = x), max_results = 100)
         }
  )

comments3 <-
  lapply(as.character(videos3$video_id),
         function(x){
          get_comment_threads(c(video_id = x), max_results = 100)
        }
  )

  all_comments <- data.frame(comment = character(),
                             video_id = character(),
                             stringsAsFactors = FALSE)

j <- 1
for (i in 1:length(comments)) {
all_comments[j:(j + length(as.character(comments[[i]][["textDisplay"]])) - 1), "comment"] <-
  as.character(comments[[i]][["textDisplay"]])
 all_comments[j:(j + length(as.character(comments[[i]][["textDisplay"]])) - 1), "video_id"] <-
  as.character(comments[[i]][[“videoId"]])
 j <- j + length(as.character(comments[[i]][["textDisplay"]]))
}
```

2. Video Statistics Data

The video statistics data, saved to a data frame entitled videostats, contains 111 observations of 7 variables. These variables are date, title of video, view count, like count, dislike count, comment count and a created metric to measure popularity. In order to devise a popularity metric, this project uses the simple equation of views + likes + comments - dislikes. The videostats numerical data was normalized in order to have a consistent standard of comparison across the data.

Code to create the videostats data frame:

```
videostats1 <-
  lapply(as.character(videos1$video_id),
       function(x){
         get_stats(video_id = x)
       }
  )
videostats1 <- do.call(rbind.data.frame, videostats1)
videostats2 <-
  lapply(as.character(videos2$video_id),
       function(x) {
         get_stats(video_id = x)
       }
  )
videostats2 <- do.call(rbind.data.frame, videostats2)
videostats3 <-
  lapply(as.character(videos3$video_id),
       function(x) {
         get_stats(video_id = x)
       }
  )
videostats3 <- do.call(rbind.data.frame, videostats3)
videostats <-
  rbind(videostats1,
            videostats2,
            videostats3)

        videostats$title <-
          c(as.character(videos1$title), as.character(videos2$title), as.character(videos3$title))
        videostats$date <-
          c(as.character(videos1$publishedAt), as.character(videos2$publishedAt),
as.character(videos3$publishedAt))
```

Data Cleaning and Wrangling:

To prepare the data for analysis, some cleaning was in order. I used the text mining package, TM, and the package textclean to assist with helpful data cleaning functions for text analytics projects. Dates were converted to actual dates, factor data was converted back into numeric data and data was normalized.

Code for this step:

```
videostats <-
  videostats %>%
  select(date, title, viewCount, likeCount, dislikeCount, commentCount) %>%
  as.tibble() %>%
  mutate(date = as.Date(substr(date, 1, 10))) %>%
  mutate(viewCount = as.numeric(as.character(viewCount)),
         likeCount = as.numeric(as.character(likeCount)),
         dislikeCount = as.numeric(as.character(dislikeCount)),
         commentCount = as.numeric(as.character(commentCount))) %>%
  mutate(viewCount = viewCount / mean(viewCount, na.rm = TRUE)) %>%
  mutate(likeCount = likeCount / mean(likeCount, na.rm = TRUE)) %>%
  mutate(dislikeCount = dislikeCount / mean(dislikeCount, na.rm = TRUE)) %>%
  mutate(commentCount = commentCount / mean(commentCount, na.rm = TRUE)) %>%
  mutate(pop_score = viewCount + likeCount + commentCount - dislikeCount)
```

Limitations of this data

The Tuber package is great for pulling data from YouTube. However, in its current form, my dataset includes all the videos posted on a particular channel. The code I employed to create the dataset casted a wide net, including some video content that is not relevant to my specific research question. While I believe the data is rich, it is not perfect. Filtering the videos on the channel id proved to be beyond the scope of this current research project. Because of the extra video content, the level of specificity is limited in this project.

III.Preliminary Exploration of the Data

Preliminary explorations of the data reveals linear relationships between the number of views and likes, dislikes, and commentary. An examination of the relationship between commenting activity and popularity suggests that the more popular an artist is, the more likely it is to have commenting on their videos. Could it be that the active commenting on the videos actually contributes to the popularity? In other words, are active fans constructing an ever-increasing notion of what's popular by way of their ongoing activity? Further exploration of the data will seek to delve further into this possibility.

IV. Approach to Overall Project

To complete the analysis of this dataset, I will continue to explore the numerical data to understand its overall distribution. I will be looking for observations of popularity with low scores as well as with high ones. I will build a word cloud for each subset to see if there are language themes associated with the distribution. These themes will be explored to complete the sentiment analysis and to determine if there is consistency between the numeric data and language themes.