Math vs. The Tonight Show

D. Miller M. Searles N. Wolford

February 22, 2016

Abstract

We'll put a more unified abstract here - maybe combine the two from MathFest, and add on as we find more

Everyone wants their 15 minutes of fame. Viral media consists of a vast amount of information exchanged over a short period of time. The Tonight Show starring Jimmy Fallon features a hashtag game where humorous tweets must be selected from a viral hashtag within 24 hours. Out of the hundreds of thousands of tweets he receives, how does Jimmy Fallon choose the handful that he reads on the show? We implement natural language processing and machine learning techinques to quantify the humor of these entries. Specifically we present a set of features that will train a predictive model to choose tweets that could be read on the show. These features can also be applied to similar data sets to evaluate different levels of humor.

1 Introduction

The Tonight Show Starring Jimmy Fallon is a late night talk show that features comedic sketches and games. Every Wednesday afternoon Jimmy Fallon tweets with a particular hashtag that begins the "Hashtag Game" for that week. Over the next 24 hours, until The Tonight Show starts at 10:30 EST on Thursday, people can tweet with the hashtag introduced by Jimmy Fallon. During Thursday's show, the Hashtag Game is played and Jimmy Fallon reads a handful of tweets selected from Twitter.

Using techniques from data analytics, we wanted to better understand the Hashtag Game segment from The Tonight Show and ultimately we decided on two main goals.

The first goal is to understand how Jimmy Fallon chooses the tweets that are read on his show. We want to find out if this process is efficient and is resulting in the desired outcomes for his show. Do they use a scientific method of choosing tweets or is it merely a human that is going in and hand-selecting tweets based on their personal preference?

The second goal is to write an algorithm that more efficiently chooses the optimal tweets to be read on his show. If the way that Jimmy Fallon is choosing his tweets is not the most efficient way, is there a way that we can do it better? Can we automate the process so that someone doesn't have to search through all of twitter to find potential segment material?

2 Problem 1

2.1 Initial Thoughts

Before we began collecting any data, we brainstormed how we thought The Tonight Show chose their weekly tweets for the show.

Considering The Tonight Show is a very popular, and probably well-funded show, we considered the idea that the tonight show already has its own algorithm. Another thought that we had relates to network theory. Many people are interrelated on Twitter. Even Jimmy Fallon and The Tonight Show twitter accounts follow thousands of people. If you have a lot of followers, chances are your tweet will be favorited more often and people will also retweet your tweet more often. The more people that favorite and retweet your original tweet, the more likely it is that Jimmy Fallon will have the opportunity of seeing your tweet.

We were also initially unaware of how many people participate weekly in the Hashtag Game. If only a handful of people participate in the game, then it would be easy for Jimmy Fallon to personally read through all the tweets and choose the best tweets for his show.

2.2 Data Collection

The Tonight Show uses Twitter feed as material for its Hashtag Game segment, so we knew that we would also need to work with that same data from Twitter. Twitter has a rest API and a streaming API. API stands for Application Program Interface and through Twitter's streaming API we were able to collect live Twitter feed only applicable to the hashtag game.

We choose Python programing to write our code that would access Twitter's streaming API. All the documentation to access this is available online, but in order to access it with python you need to download a python library. Tweepy is a python library that users have made accessible via sites like GitHub and other code sharing sites, so we chose Tweepy to access the twitter stream. Once we installed that we looked at Twitter's and Tweepy's documentation to see which commands in Tweepy's library will help us access certain information from Twitter's API. We then had to get access codes from Twitter's developer site by setting up an account with them. By doing so we received a set of tokens and keys that we inputted inside our program that gives us authorization to use Twitter's API.

Each Wednesday we began running our program when Jimmy Fallon announced the Hashtag Game for the week and our program would stop at 11:30pm EST on Thursday when The Tonight Show began it's show. Our program would collect data during this time period and save the information in an Excel Spreadsheet.

In the beginning phases of running our program we were getting access to more information than we needed. Twitter's streaming API gives you access to extraneous information such as when a Twitter account was created, number of tweets a user has favorite, or background color of the user's Twitter profile. We updated our program to eliminate some variables that we thought provided needless information.

2.3 Initial Findings

In order to cut back unnecessary information and create an efficient program, we first had to find out more information about the hashtag game and the types of tweets that are chosen to be read on the show. In order to find out this information we watched clips from the Hashtag Game segment and searched Twitter relevant data.

2.3.1 Watching Clips and Sifting through Twitter

By watching clips we were able to discover the types of tweets that Jimmy Fallon chooses to be read on his show. These tweets tend to be funny, ironic, and self-deprecating. These tweets are also relatable to a wide range of people of gender, age, background, etc. Depending on the particular hashtag for the week, generally these tweets tend to tell a story.

By searching on Twitter we learned that The Tonight Show has four people who run all their social media accounts (Twitter, Instagram, Facebook, etc). We

also learned that even though some tweets are favorited and retweeted numerous times and may appear on the Top tweets list, it does not mean that it will be read on the show. We found many tweets that fit the humor criteria, but were not favorite or retweeted very often, particularly because they did not have a large follower base. Jimmy Fallon and The Tonight Show's Twitter account follow thousands of people, and that also doesn't necessarily mean that if Jimmy Fallon follows you that your tweet will be read on the show.

There are a handful of people who tweet multiple tweets in order to increase their chances of being read on the show. As we have looked up information of people whose tweets have been read on the show, we have noticed that there are a lot of people who have tended to tweet numerous times. These tweets are either the same tweet (although with a slight variation because Twitter won't let you tweet the same tweet twice) or they have been completely different tweets that are still for The Tonight Show Hashtag Game. We have also found out that The Tonight Show rewords some of their chosen tweets for the show. The wording could be changed to clarify the tweet or to make it more humorous.

2.3.2 Our Program

With this information we were able to build a more refined code. The data that we now receive and save to our Excel Spreadsheet is timestamp, time, username, and text. We wanted to focus on the variable of time at first. As we did this we notice a trend with tweets being chosen around the same time. There is a high concentration of tweets being read on the show that were tweeted within the first two hours once the Hashtag Game was announced. We also noticed that sometimes there were smaller clusters of tweets that weren't tweeted within the first two hours, but were chosen together in a similar time period. However, the majority of the time we noticed that the majority of the tweets were chosen in the first two hours that the hashtag game was announced.

We were also able to figure out the number of tweets that are tweeted for the Hashtag Game each week. The number of tweets can range anywhere from 5,000 to 30,000 plus tweets, and the number of tweets depends on the type of hashtag that was announced. Some hashtags get a lot of traffic such as "Misheard Lyrics" where you tweet out a humorous interpretation of song lyrics that you misheard at one time. A hashtag that didn't produce as many results was "Dino Raps". This hashtag was to inspire poetic rap lyrics about dinosaurs. The second one was harder to relate to and therefore had fewer people tweeting with that hashtag.

Based off this information we were able to form a prediction on how Jimmy Fallon chooses his hashtag.

2.4 How the Hashtags are chosen

After our initial researching running our python program, our first thought is that Jimmy Fallon does not choose the hashtags that are read on the show. We predict that the four people that control his social media are probably the ones

that choose the tweets that are read on the show. Or at least choose multiple tweets that are then sorted through by a writer or a producer.

We also predict that these 4 social media specialist simply go into Twitter soon after the weekly hashtag is announced and pick a handful of tweets. When the hashtag game is first announced there are a great number of people that initially tweet. Oftentimes this hashtag can become a trending topic within the first half hour, either in the United States or even the world, and then the number of tweets dwindles until there are only a few tweets every hour right before the show starts on Thursday. However, if they didn't find a good number of funny tweets to be read on the show within the first two hours (maybe the writers or producers did not think that the tweets chosen produced all the right material they were looking for), the social media specialist will go back into Twitter and choose a few more tweets at other time periods. This theory would explain why we see a large number of tweets chosen within the first two hours after the hashtag game is announced, and also explains why we sometimes see clusters of tweets chosen at other time periods.

2.4.1 Naive Bayes

If an individual wanted to find a way to get their tweet read on the show, you would naturally think then that it would be prudent to tweet within the first two hours that the hashtag game is announced. This relates to Naive Bayes. Naive Bayes is the idea that there are different independent probabilistic classifier and when all of these classifiers are applied it increases the probably that something will happen. We can relate this idea to the Hashtag Game.

One independent classifier we can assume is time. If a tweet is tweeted within the first two hours the hashtag game is announced, we have seen that you have a greater probability of your tweet being read on the show.

Another independent classifier we can assume is wording. The Tonight Show is an NBC program. Since NBC is a public television station. Since there are certain profane words you are not allowed to say on television, if a tweet has one of those words, it would naturally not be in the running for being read on the show.

Another possible independent classifier could be length of a tweet. The tweets read on the show generally provoke humor or tell a story. If a tweet is too short, say 20 characters, it might not be long enough to deliver the desired content and therefore not a contender for being read on the show.

If a tweet fulfills all of these independent probabilisitic classifiers, the probably that your tweet will be read on the show would be greater than someone who didn't have all of these variables.

We can explore this further by looking into time. The #MyDumbInjury was a popular hashtag that had 20,341 total tweets within the extent of Hashtag Game competition. There were 9,619 tweets tweeted within the first two hours since the hashtag was announced. 7 tweets were read on the show and 6 of the 7 tweets that were read on the show were tweeted within the first two hours.

The probability given that your tweet was read on the show and it was

tweeting in the correct period of time is extremely low (0.000258). However, the probability that your tweet was read on the show given that it was not tweeted within the first two hours is even lower (0.000042). The odds of your tweet being read on the show, no matter the time that you tweeted, are very slim.

What can we conclude from this?

2.5 Conclusion for Problem 1

Since there is so much data to sort through and such a small amount of time to sort through it, it would be impossible for The Tonight Show to find the best tweets. This then begs the question, is there a better way to do this? Is there a way that using mathematics, we can write an algorithm that will make this an efficient process that chooses the most optimal tweet?