Design Phase

Team: Mount Harvard
Members: Jenna Beutler, Seth Grossman, Hua Miao, Samuel Statton, Nhi Vo

INTRODUCTION:

Purpose: To investigate what aspects of a tweet drives higher engagement in tweets related to the Coronavirus pandemic.

Scope: We will be performing predictive analysis on a dataset of over 8 million tweets in english, between March 4 and March 28.

Definitions/ things to know:

- Engagement: number of likes a tweet receives

DESIGN DATA

Data introduction: We decided to look into data relating to the current global pandemic and social media. The dataset we chose includes 500,000 original tweets (not retweets) scraped from Twitter each day between March 4 and March 28 of 2020, all of which include one of the following hashtags: #coronavirus, #coronavirusoutbreak, #coronavirusPandemic, #covid19, #covid_19. In tweets after March 17, the data also scraped tweets that included hashtags #epitwitter and #ihavecorona. The original dataset (after merging each day) includes over 8 million rows, and 22 columns, describing different aspects of each individual tweet, like username and tweet ids. This data is not completely comprehensive due to the large volume of tweets being sent each day.

Our dataset can be found here:
https://www.kaggle.com/smid80/coronavirus-covid19-tweets#2020-03-00%20Coronavirus%20Tweets%20(pre%202020-03-12).CSV

Data dictionary: the following list includes all original 22 columns in the data.

status_id: The ID of the actual Tweet.
user_id: The ID of the user account that Tweeted.
created_at: The date and time of the Tweet.
screen_name: The screen name of the account that Tweeted.
text: The text of the Tweet.
source: The type of app used.
reply_to_status_id: The ID of the Tweet to which this was a reply.
reply_to_user_id: The ID of the user to whom this Tweet was a reply.
reply_to_screen_name: The screen name of the user to whom this Tweet was a reply.
is_quote: Whether this Tweet is a quote of another Tweet.
is_retweet: Whether this Tweet is a retweet.
favourites_count: The number of favourites this Tweet has received.

retweet_count: The number of times this Tweet has been retweeted.
country_code: The country code of the account that Tweeted.
place_full_name: The name of the place of the account that Tweeted.
place_type: A description of the type of place corresponding with place_full_name.
followers_count: The number of followers of the account that Tweeted.
friends_count: The number of friends of the account that Tweeted.
account_lang: The language of the account that Tweeted.
account_created_at: The date and time that the account that Tweeted was created.
verified: Whether the account that Tweeted is verified.
lang: The language of the Tweet.

Operations performed:

Dropped the following columns in order to simplify the dataset to variables that we need for analysis:

- Status_id
- user_id
- reply_to_status_id
- reply_to_user_id
- reply_to_screen_name
- is_quote
- is_retweet
- place_type
- friends_count
- account_lang
- account_created_at
- country_code
- place_full_name

Subsetted our tweets by English only.
Converted created_at column to datetime, renamed to ‘date.’
One-hot encoded verified column

Machine Learning:

Running linear regression on the data to predict the number of favorites a tweet receives based on retweets, followers, verified status, and text features. On the small data set, the training model has a 0.80 r-squared value.

ARCHITECTURE:

Tools used:

- myBinder: for collaboration on code, and connecting to our Github repo.
- Pyspark: for data operations/analysis.
- AWS: for data processing and storage
- Github: for sharing, organizing code, project documents, etc.

SOLUTION BREAKDOWN

Deployment strategy

Currently, our code is still in the debugging process on a sample of ten percent of the data. As soon as we iron out the details, we plan on deploying our Python 3 notebook in an EMR cluster through AWS on a full dataset.

Other things to note