Milestone #2

Project Team #6

July 13, 2018

PART A: PROJECT STATEMENT

- We intend to answer the question whether the author of tweet is a human being or a computer bot. We will create a function that takes in a tweet and returns a 1-10 score showing the likelihood the author of the tweet is a human being or a bot.
- The score will be developed using input from both a classification algorithm and an NLP based algorithm.
    - The features of the classification algorithm are discussed below.
    - As for the NLP algorithm, we will focus exclusively on consistency of "messaging" in the user's tweet history. That is, we will attempt to identify whether content of the tweets show a high degree of variability or single-topic focusness. Variety of content will improve the score for human authorship whereas single-topic focusness will improve the score for bot authorship.
- We will show our scoring results and error rates for known real people (i.e., verified users on twitter). We will also show our scoring results and error rates for known bots that we can hopefully find from a reliable source online.
-

PART B: PRELIMINARY EDA

- For our classification algorithm, our initial belief is we will collect and explore the following features of a tweet's user:
    - Related to user profile and followers/following:
        - Date of creation
        - Number of followers
        - Number following
        - Distribution of number of followers as a function of time since creation
        - Distribution of number following as a function of time since creation
        - Count/ratio of follower/followings that are also in ‘known bots” database
    - Related to user's latest month's tweets and likes/liking:
        - Distribution of # of tweets by day
        - Distribution of # of tweets by time of day
        - Distribution of # of likes received by day
        - Distribution of # of likes received by time of day
        - Distribution of # of likes given by day
        - Distribution of # of likes given by time of of day
        - Distribution of measure of text quantity in each tweet
        - Distribution of measure of links quantity in each tweet
        - Distribution of ratio of test/links quantities in each tweet
- Our EDA will consist of the following:

- We would like to collect the above for known real people (verified by twitter) and known bots (from online source)
- Our EDA for the classification algorithm will consist of visualizing the above features for both types of authors--which we will compare visually to see if a pattern is apparent (e.g., real users tweet more erratically over time whereas bots tweet with more consistent timing, etc.
- Our EDA for NLP algorithm will be less visual and just involve starting to use the NLP tools against both types of user data.