Scope of Work for Data Science Final Project

Prepared by Group #5
Karan Bhandarkar, karanbhandarkar@gmail.com
Vivek Mishra, iblpvivek@icloud.com

Project Statement and Background

There are an average of 6,000 tweets produced on Twitter per second. Twitter posts are mostly public and can be easily collected using Twitter's developer platform API. Also, frequent use of hashtags makes it more interesting to draw conclusions.

The role of so-called social media "bots" automated accounts capable of posting content or interacting with other users with no direct human involvement has been the subject of much scrutiny and attention in recent years [1]. These accounts can play a valuable part in the social media ecosystem by answering questions about a variety of topics in real time or providing automated updates about news stories or events. At the same time, they can also be used to attempt to alter perceptions of political discourse on social media, spread misinformation, or manipulate online rating and review systems [2]. As social media has attained an increasingly prominent position in the overall news and information environment, bots have been swept up in the broader debate over Americans changing news habits, the tenor of online discourse and the prevalence of fake news online [3].

In order to detect bots, classification models and/or natural language processing techniques such as topic modeling and sentiment analysis can be incorporated. This project will involve feature engineering and will provide a real-world data collection experience.

Goal: In this project, the goal is to detect Twitter bots using tweets data from the Twitter developer API by utilizing machine learning techniques.

Data Resources

We will collect your own data for this project. We were provided a basic Python script, tweepy_script.ipynb, that utilizes the tweepy library [4] to access the Twitter API. The provided tweets_sample.json file demonstrates a sample of what you will collect and the attributes of a tweet made available by the Twitter API [5].

High-level project goals

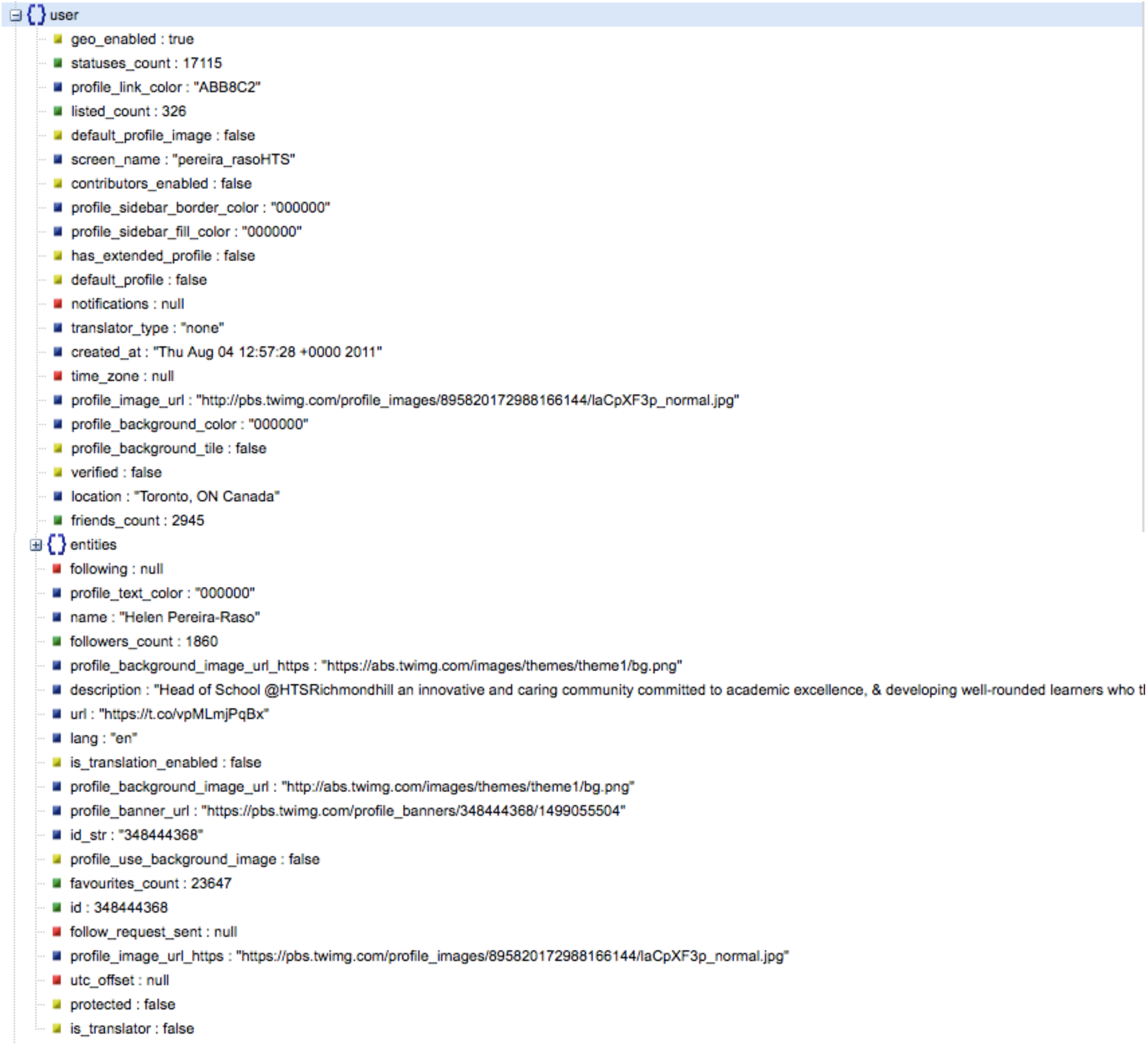
1. The first step is to create our own dataset. We will mine the data for the project using the Twitter API and utilize feature engineering and pre-processing techniques to prepare the data for analysis.

2. Create several models to determine characteristics of different types of twitter users. Create at least one model that uses natural language processing techniques, such as topic modeling [6], and at least one model that uses a classification algorithm. We may decide to have models that use both. We will provide evidence of success at detecting bots when compared to human users or explain why it wasn't possible.

3. Perform a comparison of your models. This will include an error analysis and an evaluation of the predictive quality of your models.

Preliminary EDA

1. Initial EDA will be for selection of variables. Looking at the JSON response from the twitter API, we could use features directly from the user entity, like:



```
⊟ { } user
    geo_enabled : true
    statuses_count : 17115
    profile_link_color : "ABB8C2"
    listed_count : 326
    default_profile_image : false
    screen_name : "pereira_rasoHTS"
    contributors_enabled : false
    profile_sidebar_border_color : "000000"
    profile_sidebar_fill_color : "000000"
    has_extended_profile : false
    default_profile : false
    notifications : null
    translator_type : "none"
    created_at : "Thu Aug 04 12:57:28 +0000 2011"
    time_zone : null
    profile_image_url : "http://pbs.twimg.com/profile_images/895820172988166144/laCpXF3p_normal.jpg"
    profile_background_color : "000000"
    profile_background_tile : false
    verified : false
    location : "Toronto, ON Canada"
    friends_count : 2945
  ⊞ { } entities
    following : null
    profile_text_color : "000000"
    name : "Helen Pereira-Raso"
    followers_count : 1860
    profile_background_image_url_https : "https://abs.twimg.com/images/themes/theme1/bg.png"
    description : "Head of School @HTSRichmondhill an innovative and caring community committed to academic excellence, & developing well-rounded learners who tl
    url : "https://t.co/vpMLmjPqBx"
    lang : "en"
    is_translation_enabled : false
    profile_background_image_url : "http://abs.twimg.com/images/themes/theme1/bg.png"
    profile_banner_url : "https://pbs.twimg.com/profile_banners/348444368/1499055504"
    id_str : "348444368"
    profile_use_background_image : false
    favourites_count : 23647
    id : 348444368
    follow_request_sent : null
    profile_image_url_https : "https://pbs.twimg.com/profile_images/895820172988166144/laCpXF3p_normal.jpg"
    utc_offset : null
    protected : false
    is_translator : false
```

i. Does the name, screen_name or description contain the word 'bot'?
ii. Is the geo_enabled true or false?
iii. Is the location value 'null'?
iv. Is it a Twitter verified account?
v. Is the followers_count extremely low?

2. The next stage of EDA could be to see if there are any trends in the issues covered by the tweets. Do they all belong to the same category?

3. Further EDA could include checking if there is any trend in the grammar of sentences tweeted by bots.

References

1. Stefan Wojcik, “Bots in the Twittersphere”: http://www.pewinternet.org/2018/04/09/bots-in- the-twittersphere/

2. Chris Baraniuk, “How Twitter Bots Help Fuel Political Feuds”: https://www.scientificamerican.com/article/how-twitter-bots-help-fuel-political-feuds/

3. Chengcheng Shao et al., “The spread of low credibility content by social bots”: https://arxiv.org/pdf/1707.07592.pdf

4. The tweepy Python library: http://www.tweepy.org

5. Twitter’s developer resources to learn about the API: https://developer.twitter.com

6. Asbjan Ottesen Steinskog et al., ”Twitter Topic Modeling by Tweet Aggregation": http://www.aclweb.org/anthology/W17-0210