Jonathan Wiseman

February 17, 2020

CS495

Milestone 1 Report

**Original Goals:**

1. Dataset collection
2. Dataset cleaning
3. Selection of sentence embedding algorithm

**State of the Project:**

The fundamental goal of the project remains unchanged: produce a sentiment analyzer trained on Reddit comments for gauging reactions to NFL draft picks. The overall structure of the project is shown in figure 1:

**Figure 1: New Project Flow**

Once all comments have been scraped and separated into their respective threads, they will be consolidated into one dataset file (tentatively called comments.json). The script *label\_comments.py* allows a user to label comments. After comment labeling, *clean\_comments.py* will remove URLs, links to other subreddits and users, emojis, parentheses, and brackets. After labeling has been complete, the project will focus on model creation.

**Progress**

1. Dataset collection

The script *scrape\_comments.py* accomplishes this objective. Using PRAW, *scrape\_comments.py* collects all top- and second-level comments for a given player’s draft thread. If the player is not found in the thread, then the program raises a PlayerError. If the player is found, then all top- and second-level comments are scraped. There is no cleaning performed in this step, so comments consisting of URL links, auto-moderated comments, and deleted comments are all present; instead, the task of cleaning is left to INSERT SCRIPT NAME.

Running *scrape\_comments.py* for the target 12 players yields the following results:

* Baker Mayfield: 1381 comments
* Mitchell Trubisky: 1945 comments
* Daniel Jones: 3165 comments
* Kyler Murray: 1571 comments
* Lamar Jackson: 889 comments
* Dwayne Haskins: 948 comments
* DeShaun Watson: 910 comments
* Sam Darnold: 637 comments
* DeShone Kizer: 589 comments
* Josh Rosen: 673 comments
* Josh Allen: 766 comments
* Patrick Mahomes: 959 comments

This yields a total of 14434 comments that will have to be labeled. For each player, a .json file with the following fields was created: comment\_id, post\_id, comment. The comment\_id is the unique comment identifier, the post\_id is the parent post (i.e. the draft reaction thread), and the comment is the text data.

1. Dataset cleaning
2. Selection of sentence embedding algorithm
3. Other