Jonathan Wiseman

March 16, 2020

CS 495

Milestone 2 Report

**Goals:**

1. Complete Dataset Labeling
2. Move into model creation

**Progress:** <https://github.com/jonwiseman/DraftSense>

The fundamental project goals remain unchanged. There is one additional file that has been added: “Exploratory Data Analysis”.ipynb, located under the /docs folder. This file contains EDA on the labeled dataset. Additionally, there have been several edits to *label\_comments.plabel\_comments.py*. There were 3 big code bugs that were solved to reach Milestone 2:

1. Mismatched indexes: this problem was caused by naming the DataFrame’s field “rating” rather than “label”
2. Script exiting after loading file: after >50% of the dataset was labeled, the current index exceeded the number of remaining comments; exit conditions were changed
3. Script failing to save progress: again, this was related to >50% of the dataset being labeled

The file “comments.json,” under /Data/Dataset contains the labeled dataset. See the EDA notebook for a quick exploration!

**Lessons Learned:**

Labeling a dataset is tedious, time consuming, and hard. While it may be no problem to label a batch of 1000 comments (it takes ~30 minutes), the work becomes monotonous and tedious. There’s only so many times you can see the comment “lol” before you get extremely bored. Some comments are also hard to label: is a comment criticizing the trade used to get a quarterback speaking negatively of the quarterback? Is a comment that contains positives and negatives about a player a positive or negative reaction? Why are there so many off-topic comments?

The majority of the dataset (~50%) turned out to be off-topic or irrelevant comments. While these may be useful in some context for others, they serve no real purpose for this project. There are some key differences between positive and negative comments which should make sentiment analysis robust, but the dataset’s final size is a bit disappointing: 1616 and 2652 positive and negative comments, respectively.

Finally, there are some more changes that will have to be done on the negative comments to bring them to a usable state: many are disambiguations of acronyms (like “lol” vs “loooool”, “haha” vs “hahahahahaha,” “lmao” vs “lmfao” vs “lmfaooooo,” etc.) Some also have weird spacing (like B U S T), which is the result of a meme. I will train some initial models with the dataset as is, and then move into more text cleanup if needed.

**Deliverables:**

1. comments.json: labeled dataset
2. label\_comments.py: final labeling script
3. “Exploratory Data Analysis”.ipynb: EDA notebook