|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | ID | Section Contributed | Section Edited | Other Contributions |
| Thompson, Josh | 301421008 | Written component |  |  |
| Ty, David | 301334955 | Coding component |  |  |

**Assignment 2 – Group 9**

**Description of Program**

The goal of our program was to conduct a sentiment analysis on reddit comments from various posts. We chose to look at comments that were sorted into the controversial category underneath the reddit posts. Controversial comments on reddit are essentially comments that are downvoted and hidden from most users. Our belief was that these comments were downvoted because they contained negative sentiment. Comments were gathered by scraping various posts, while sentiment analysis was conducted through a reimplementation of vader.

As noted, the input was gathered through scraping reddit posts. These overarching topic of these reddit posts was world news. Specifically, we chose from 5 posts from r/news with topics on a Harriet Tubman monument, Saudi Oil, Charges dropped in a road rage incident, Silicon Bank, and Norfolk. From these posts our scraper gathered comments that had been sorted into the controversial category, with all but 1 of these posts resulting in 10 comments each. The scrape on the reddit post regarding Norfolk had only had 1 controversial comment at the time of scraping. Likewise, the scrape from the Harriet Tubman monument post contained 1 deleted comment, for which sentiment would go onto to rate as neutral. Resulting in 41 “controversial” comments for our program to conduct sentiment analysis on.

By reimplementing vader, we were given the categories of positive, neutral, or negative. Each individual comment was given a sentiment for each of the 3 categories, resulting in a summary of either being positive, neutral, or negative. At the same time each of the 5 topic categories received a percentage for each category.

The 10 comments from the reddit post on Saudi oil giants received a sentiment of 20% positive, 70% negative, and 10% neutral, with the average sentiment being -0.32724. The comments on the Harriet Tubman monument story received 50% positive, 20% negative, and 30% neutral, with an average sentiment of 0.22934. It should be noted that one of the comments from this story was deleted which increased the neutral percentage. The comments from the story on charges dropped in a road rage incident were 10% positive and 90% negative, with an average sentiment of -0.47107. The comments from a story about Silicon Valley bank were 40% positive, 50% negative, and 10% neutral, with an average sentiment of -0.20811. The single comment on the story about Norfolk was positive and had a sentiment of 0.8999. Of the total 41 comments, 56.09% or 23 of them were negative, while 31.7% or 13 were positive, with the remaining 12% or 5 comments being neutral. This output lends support to our original hypothesis that controversial comments would be more negative in nature.

**Code**

Our code that we used can be accessed here: <https://github.com/Sadieka/LING250Group9Assign2>

**Sources**

We used information about vader from to help set up and reimplement our sentiment analysis in python (Hutto & Gilbert, 2014; Hutto, 2018). For our scraping we used information about reddit scraping from Boe (2021; 2023a; 2023b; 2023c), we also required information about where controversial comments are stored in Reddit’s API, information for this was gathered from Wotanni (2018). To prevent errors when no controversial comments could be detected we used information from an online forum (Jscriptor, 2022).

**Reference Notes**

cjhutto : repository for vader, installation and use instructions

*Comment Extraction and Parsing* : how to gather the comments of a submission

Hutto, C.J & Gilbert, E.E : paper on vader, asked to cite whenever using vader

*praw-dev/praw* : repository for praw, installation instructions

*python - DeprecationWarning: The default dtype for…* : prevents error when passing an empty series, ie when no controversial comments exist

*Quick Start — PRAW 7.7.0* : how to iterate through submissions on a subreddit

*Quick Start — PRAW 7.7.1.dev0* : how to find the attributes of a comment, to find the “controversiality” variable used in the program

wotanii : how reddit marks controversial comments in the api, ie “controversiality”

**References**

Hutto, C.J. (2018, December 19). *cjhutto/vaderSentiment.* GitHub. https://github.com/cjhutto/vaderSentiment

Boe, B. (2023a) *Comment Extraction and Parsing — PRAW 7.7.1.dev0 documentation.* (n.d.). Praw.readthedocs.io. Retrieved March 15, 2023, from https://praw.readthedocs.io/en/latest/tutorials/comments.html#

Hutto, C.J. & Gilbert, E.E. (2014). *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.* Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Boe, B (2021) *praw-dev/praw.* (2021, March 28). GitHub. https://github.com/praw-dev/praw

Jscriptor (2022) *python - DeprecationWarning: The default dtype for empty Series will be “object” instead of “float64” in a future version warning.* (n.d.). Stack Overflow. <https://stackoverflow.com/questions/62158734/deprecationwarning-the-default-dtype-for-empty-series-will-be-object-instead>

Boe. B. (2023b) *Quick Start — PRAW 7.7.0 documentation.* (n.d.). Praw.readthedocs.io. Retrieved March 15, 2023, from https://praw.readthedocs.io/en/stable/getting\_started/quick\_start.html#submission-iteration

Boe, B. (2023c) *Quick Start — PRAW 7.7.1.dev0 documentation.* (n.d.).Praw.readthedocs.io. Retrieved March 15, 2023, from https://praw.readthedocs.io/en/latest/getting\_started/quick\_start.html#determine-available-attributes-of-an-object

wotanii. (2018, April 5). *decide if a comment is controversial via praw.* https://old.reddit.com/r/redditdev/comments/89zzd2/decide\_if\_a\_comment\_is\_controversial\_via\_praw/