|  |
| --- |
| Sawyer Jacobson |
| Aran Miller |
| Sameer Patel |

DS Tools 1 – Fall 2020

|  |  |
| --- | --- |
| Political Sentiment Analysis via Webscraping and NLP  University of Denver - Ritchie School of Engineering | Abstract  An exploration of overall political sentiment using Python webscraping and natural language processing (NLP) functionalities. Data was collected from Reddit and Twitter and subsequently cleaned and analyzed using existing Python data science libraries and packages.  Co-Contributors |

# Data Set and Motivation

For this project the team decided to look into general political sentiment based on analysis of user-submitted content to two of the most popular websites in the Unites States – Reddit and Twitter. In order to do so, the data would need to be analyzed through techniques common to the Natural Language Processing (NLP) field of data science.

The data collection, cleaning, and analysis process was conducted as follows:

1. Collection of “buzz words” from two popular politically aligned subreddits - /r/Liberal and /r/Conservative
   1. Store 1,000 headlines from each subreddit using the Python praw library
      1. Headlines are first filtered to include the top (most upvoted) user-submitted posts over the past year
   2. Removing stop words using the Python NLTK package
   3. Identifying the top 5 unique “buzz words” for each political ideology based on usage frequency
2. Webscraping Tweet data by searching Twitter for the words identified in Step 1
   1. Extracting Tweet text and removing certain types of irrelevant information (eg. emojis, special characters, extra blank spaces)
3. Training a Sentiment Analysis model using the cleaned data
   1. Employing (Python library or web application) and manually tagging a set of x tweets in order to train the model to make future predictions
4. Graphically visualizing overall sentiment of the “buzz words” identified based on model performance

The metadata important to conducting this analysis includes the following:

* Reddit
  + Post title
  + Number of upvotes
  + Number of comments
  + Creation date
* Twitter
  + Tweet text
  + Creation date
  + Retweet count

# Research Question

Can we make a prediction about the November 2020 election based on sentiment analysis of Twitter data containing “buzz words” specific to both liberal and conservative messaging boards/platforms? As briefly aforementioned in the Data Set and Motivation section, the input data necessary to answer this question comes in the form of Tweet text, and the output is a visualizing of overall sentiment of the buzzwords identified from the subreddit webscraping.

# Literature Review

NLP is an extremely complex field of data science, with a multitude of methodologies and models currently existing to perform specific types of analyses. In trying to stick to our goal of sentiment analysis, the team gleaned techniques from well-established data science entities such as Monkey Learn and Towards Data Science in order to guide us in this analysis. The specific model chosen to conduct our analysis was …

# Quality of Cleaning

## Data Cleaning and Type Conversion

All forays into NLP begin with cleaning of the text data; including processes such as tokenization, stemming/lemmatizing, removal of stop words/accessory data (eg. excess blanks, emojis, etc.), and others. For this specific project, the praw library (<https://praw.readthedocs.io/en/latest/>) was employed in combination with the nltk library to extract meaningful “buzz words” from Reddit headlines. Code chunks demonstrating this process are presented below:

## Missing Values

Missing data (NaNs/nulls) was not an issue for this analysis, as data was scraped directly from the top 1,000 headlines (sorted by the most upvoted content over the year in both the /r/Liberal and /r/Conservative subreddits), and post titles cannot be blank.

## Attribute Creation, Summary Statistics, and Interpretation

The words returned from the list of 1,000 headlines (per subreddit) were combined into a total list of words, stop words were removed (using the Python nltk library), and a dictionary of counts was created showing the most frequently-used words from /r/Liberal and /r/Conservative. From there, 10 unique “buzz words” were identified (5 for each subreddit) and subsequently used to scrape Twitter data. Code chunks demonstrating this process are presented below:

# Visualization

## Data Visualizations

A graph of overall sentiment relative to the “buzz words” identified is presented below.

## Description/Interpretation of Visualizations

From the above graph, the team arrived at the following conclusions regarding overall sentiment of the specified terms:

## Connection to Understanding Data Distribution

The specific line in the rubric is “What data visualization helped you understand about data distribution.”

My guess is he wants us to classify our data to a specific stats distribution? Would it even be binomial? Only two options, but no relation to probabilities with sentiment analysis at all, so not sure how we can classify our output if a probability density function can’t possibly exist for it.

## Outliers/Other Issues

I would need to see our completed analysis in order to complete this section. I get the feeling some issues we’re definitely going to have would be dealing with things like emojis/blanks/etc. However, if we opt to use the Monkey Learn app, those issues (and all other issues) get taken care of automatically.

# Conclusion

Based on the analysis and visualizations presented in the preceding sections, our group predicts …

Works Cited

Gupta, Shashank. “Sentiment Analysis: Concept, Analysis and Applications.” *Medium*, Towards Data Science, 19 Jan. 2018, towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

Pascual, Federico, et al. “Twitter Sentiment Analysis with Machine Learning.” *MonkeyLearn Blog*, 4 Aug. 2020, monkeylearn.com/blog/sentiment-analysis-of-twitter/.

Tanner, Gilbert. “Scraping Reddit Data.” *Medium*, Towards Data Science, 12 Feb. 2019, towardsdatascience.com/scraping-reddit-data-1c0af3040768.