HoKwong Andrew Fu
3/25/2020
Problem Solving with Python
Professor Zhi

Reddit Sentiment Analysis Write Up

I used NLTK VADER to do a sentiment analysis on the headlines of the /r/Coronavirus thread in hopes to use it as a sample to gauge the current American consumer confidence on the economy as a factor to determine if index ETFs like the VTI will go up or down in the short term future.

Firstly, I used a without-limit-for-loop to iterate through the headlines within the subreddit. Since Reddit caps it at around 1000, the script ended up giving around 950, give or take 5 or 6. Then I use VADER to determine the headline's sentiment and append the results into a list. Afterward, I found the most common words for the positive and negative headlines to gauge the content of the positivity and negativity. I used bar charts and line graphs to visualize the gathered data.

One design decision where I had to choose between multiple alternatives is the choice to define the function on line 95, process_headlines, rather than just writing the code out itself. By writing a function, I can call it later, as in line 111, and line 148, even though it is just in two other instances.

From the results of the script, I found that the out of the 950 headlines, most were neutral, at 49%, then negative at 30% and then positive at 20%. (See Appendix A for a bar chart) To reach a score that decides whether a headline is positive, neutral, or negative, I had to choose where the threshold is. For example, is the score greater or less than .2 is positive or negative. The higher the value, the greater the confidence and the smaller the margin of error in the results. However, the sample size also decreases as confidence goes up. To test different margins of error, I used ± .1 and increased it by .05 until I reached .3. Overall, even with different margins of error, negative is higher than positive but an average of 10%, while people are predominantly

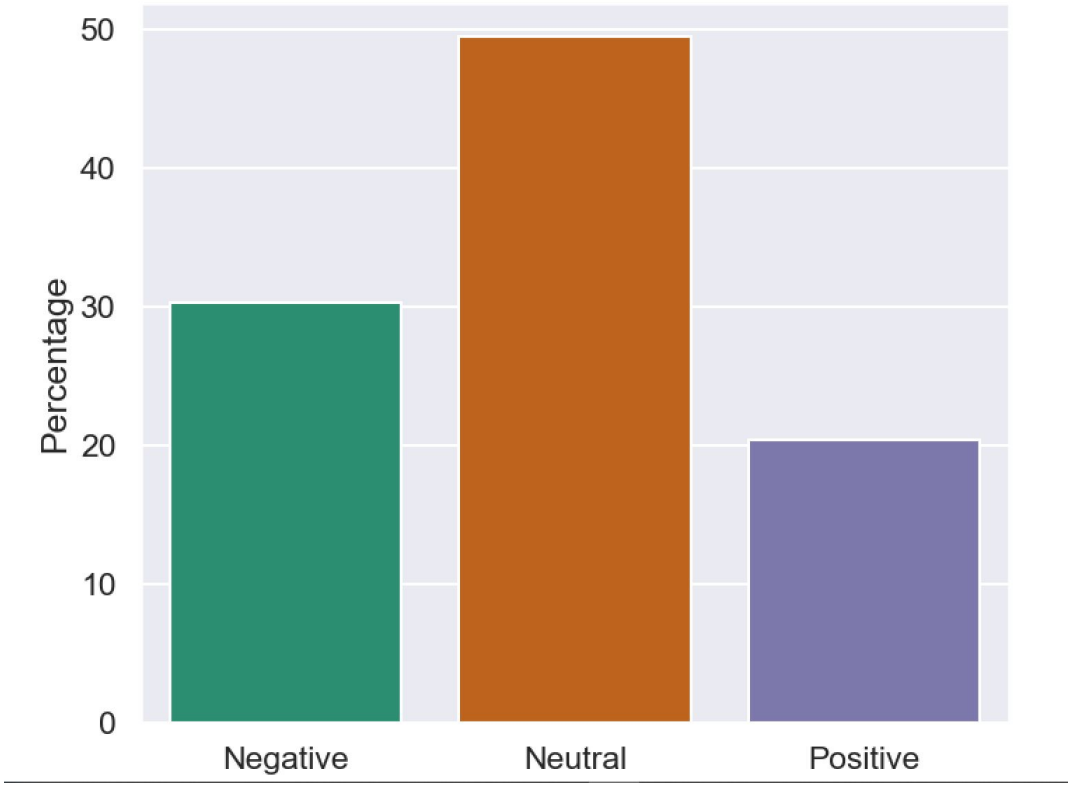
neutral. It's worth noting that as the margin of error decreases, so does the difference between the positive and the negative categories. (See Appendix B for a table of the results).

Although further analysis is required, a generalization can be concluded from the data: people are mostly uncertain about the virus while there are more people who thinks negatively of it than positively on 3/25. If this data does have a positive relationship to consumer confidence on the American economy, then it means the market should have been down today rather then up. If the data is reliable, it means that the market reacted irrationally today and should soon reflect the true state of consumer confidence in the economy, returning to a lower equity valuation of the index companies. If the market goes down tomorrow, 3/26, then perhaps the data has validity. If not, further exploration is required to find the relationship between the two.

I think getting data that reflected my own belief about the market, against consensus, was encouraging. I also thought that building a script like this would be much harder than it actually was. I think if I had more time, I could have ran regressions against actual market data to look for an indication of a positive relationship between the collected sentiment and the index ETFs. I am going to refine my abilities in data analysis as I thought this was particularly interesting and it has real-world applications. I actually exited all of my positions today, partly because of this data, in the early afternoon, which turned out to be an excellent decision as the index ETF, VTI, fell sharply after that.

I wanted to work in a team but that didn't end up happening. Though, now looking back on it, I think I learned more as a result and had a higher drive to do the project because I was working alone.

Appendix A



Appendix B

Numbers						Average
0	417	444	470	508	546	477
-1	315	300	286	265	240	281.2
1	219	205	194	176	165	191.8
Total	951	949	950	949	951	950
Percentages						
0	44%	47%	49%	54%	57%	50%
-1	33%	32%	30%	28%	25%	30%
1	23%	22%	20%	19%	17%	20%
Total	100%	100%	100%	100%	100%	1
	± .1	± .15	± .2	± .25	± .3	