Arnav Sharma

Reddit Sentiment Analysis

6th January 2023

OVERVIEW

Sentiment Analysis is a subset of Machine Learning that focuses on building predictive models that can identify the “emotion” of a text. While being a relatively new area of study, it has potential to be of heavy use in many sectors, most notably marketing.

The focal point of this data science project is to explore what other ways sentiment analysis can be used, and how it can possibly be improved. To do this, comments from the social media site Reddit will get judged based on their sentiment for a variety of purposes.

GOALS

1. **Political**: Find and explain the differences in sentiments between liberals and conservatives.
2. **Social**: Be able to draw conclusions on news articles based on the sentiments of their comments.
3. **Economic**: Use sentiments of traders to potentially predict the stock market.

SPECIFICATIONS

The model will be constructed to predict whether a specific comment’s tone is positive, neutral, or negative. In order to do this, a dataset consisting of tweets along with their respective sentiments will be used. This dataset will have its tweets cleaned and will be shifted to have an equal amount of comments(2363) for each possible sentiment.

Due to the relatively low amounts of data, neutral networks will not be used; instead, a pipeline consisting of vectorization, tf-idf, max abs scaling, and finally a Logistic Regression algorithm will be trained on 75% of the data.

In order to get the Reddit data, the library praw will be used as a Reddit scraper. For Goal #1, the 1000 “hottest” Reddit posts from the subreddits r/Conservative and r/politics(typically liberal) will be used, with each post having a maximum of 10 comments scraped for speed purposes.

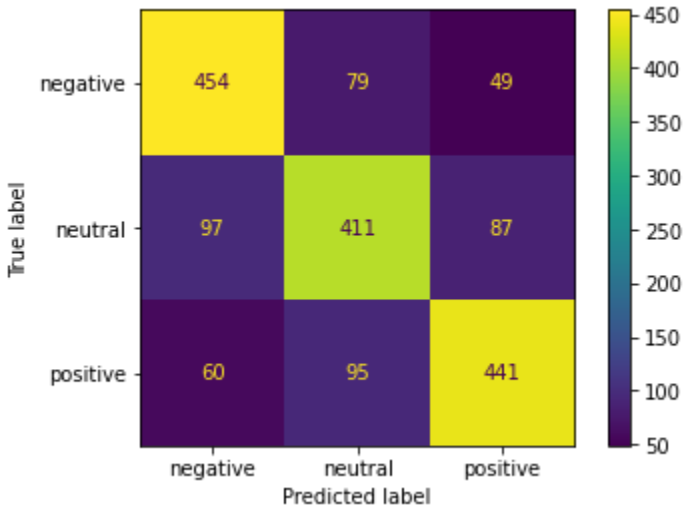
For Goal #2, the 100 most upvoted posts from the subreddit r/news will be scraped for 10 comments per post. The averages of the sentiments for each of the posts will be analyzed.

For Goal #3, r/wallstreetbets, a subreddit dealing with stocks, will be used. In particular, 10 comments per “weekly discussion” post for each week from August 13, 2022 to December 31, 2022 will be scraped. The sentiments of the comments will be analyzed along with the S&P 500 values at the end of each said week.

RESULTS

Sentiment Model

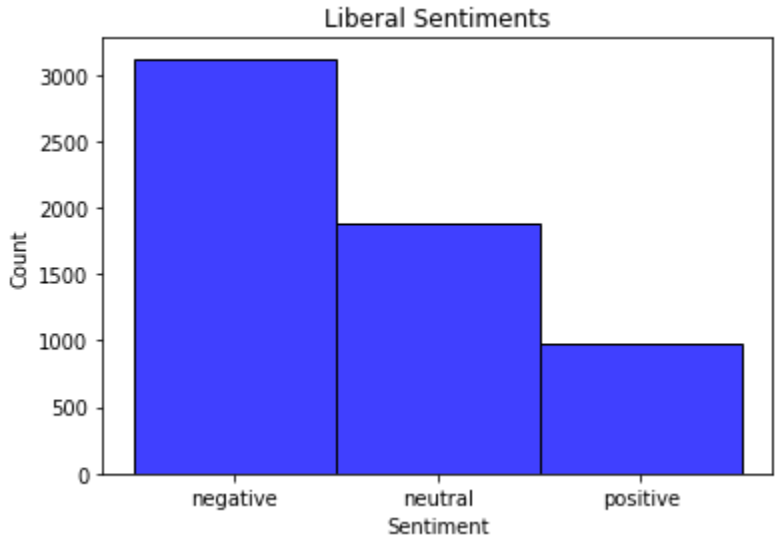
The Logistic Regression model got a 73.7% accuracy on the test data, higher than any other type of algorithm used throughout the predictive process. The confusion matrix is as follows:

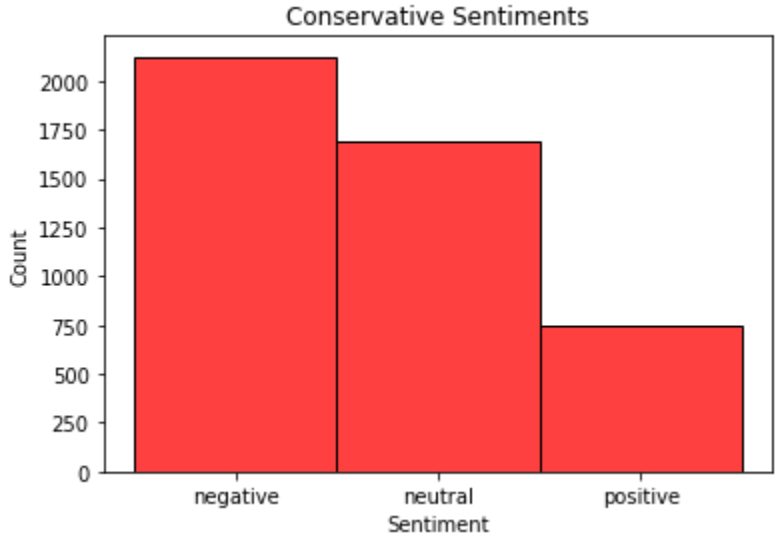


The model is relatively accurate with little differences in accuracy among the different sentiments. As well, among the true negatives, the model rarely predicted positive, and among the true positives, the model rarely predicted negative. Among the true neutrals, there was an even split among the false positive and false negative predictions, showcasing the lack of bias in the model.

Goal #1 - Political

Modeling the distribution of the resulting sentiments of each comment gives the following:

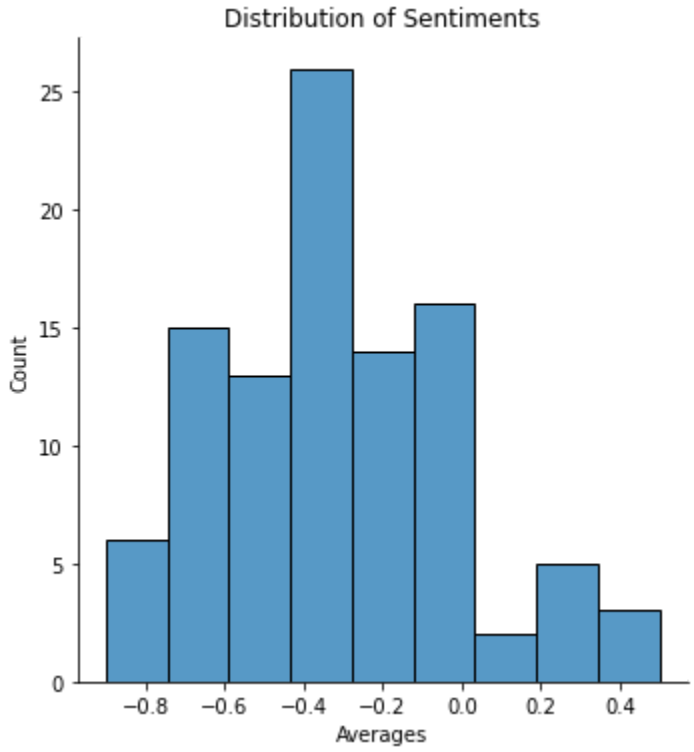




Based on the histograms, it is clear that both liberal and conservative online comments are heavily skewed negatively. However, conservatives generally give slightly more neutral comments than liberals, showcasing how left leaning individuals tend to be more emotional in their comments. This could be due to a variety of factors, such as liberals tending to be younger than conservatives, or liberals tending to discuss less economic issues and more social issues(which tend to bring out more emotion). Despite this, both liberals and conservatives give very few positive comments, indicating the overarching negativity in the political sphere.

Goal #2 - Social

After achieving the average sentiments of each post, the distribution plot of the average is as follows:

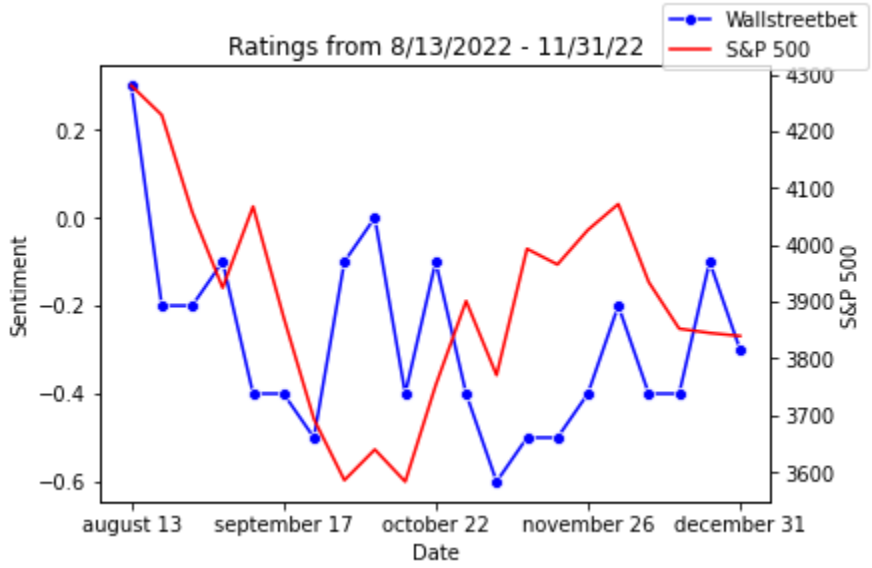


Similar to the liberal and conservatives sentiments, the average tone of each news article skews towards the negative side. This is likely due to the fact that news articles are mostly political, and shows that generally the news makes us less happy.

Looking at the specific news articles along with their ratings, their sentiments are slightly surprising. The most negatively rated post was “Family of 11-year-old boy who died in Texas deep freeze files $100 million suit against power companies.” Despite one guessing that this post may be seen more positively due to people supporting the suit, the sentiment was highly negative due to the comments mainly focusing on the Texas freeze. On the other end, the most positively rated post was “Report: Stan Lee dead at 95 - Story”. Although one might guess that this post may be seen negatively due to the death of Stan Lee, the sentiments were positive because most comments praised Stan Lee.

Goal #3 - Economic

Mapping the plot of the S&P 500 values along with a plot of the average sentiments for the stock market based on r/wallstreetbets gives the following graph:



The trend between the S&P 500 and the tone of r/wallstreetbets is slightly correlated: generally, when the S&P 500 falls, so does the average sentiment of the discussion posts, and generally when the S&P 500 rises, so does the average sentiment.

Despite this, the correlation is far from good enough to be used to predict the stock market: The correlation coefficient between the sentiments of each discussion post and the S&P 500 value the following week is only 19.7%, a very weak correlation.

CONCLUSION

Although analyzing the sentiments of comments provides some insight, this insight is very general and doesn't provide much use. This is due to the complexity of human comments; simply categorizing every comment as positive, neutral, or negative doesn't tell the full story of said comment.

Possible improvements to future sentiment analysis projects include:

- Using sentiment models trained on comments specific to the area being analyzed. For example, building a model from data specifically for r/wallstreetbets comments would give more accurate results
- Constructing a more complex sentiment model that can provide more information, such as the subjects of the comment and whether the subject is being perceived as positive or negative
- Using social media sites other than Reddit, as Reddit tends to be more left-leaning. This would result in less biased results politically.