Kevin Foo

FIN 6368.001

2/3/2024

**Social Media Sentiment Analysis for Predicting Stock Prices**

*Introduction*

Social media is a valuable source of big data and provides a prime opportunity for many data mining methods. In this project, my program will only focus on sentiment mining from text using a natural language processing (NLP) model. The research questions this project will attempt to answer are the following: How accurately can an NLP model rate a statement? Can sentiment mining data be used to predict stock prices? And lastly, how does sentiment analysis compare with ratings of financial analysts? This project will be programmed in Python and will include a dashboard of various data visualizations to display its results.

*Methods*

There are many packages that exist with pre-loaded NLP models on Python. For this project, I will be using the RoBERTa model on Huggingface, which was trained on 124 million tweets, for sentiment mining. The social media source that will be used in this project is Reddit, which has extensive data on users’ attitude towards companies/stocks.

To extract data from the Reddit API, the program will require the use of the package PRAW (Python Reddit API Wrapper). Using PRAW, the data will be converted into .xlsx format for ease of shareability and will be used in the NLP model. Typically, NLP models will output labels (i.e. positive, neutral, negative) with an associated probability indicating the likelihood that the input matches the label. From there, the program will take the weighted average of the labels and produce a sentiment score from -1 to 1 (-1 indicating negative, 1 indicating positive, and 0 indicating neutral).

Next, the program will regress sentiment scores related to a specific company/stock on the stock’s return and test for significance. The model will use a large sample of stocks (top 100 S&P 500 companies) to avoid any firm-specific biases/trends and to show whether sentiment analysis is applicable for a wide range of stocks or if it is only applicable to firms that meet certain criteria. We can then use the significant regression estimators in historical back-testing to confirm whether sentiment scores are effective indicators for stock price movement. Lastly, by using the Fama-French 3 Factor Model as a benchmark, we can combine it with a sentiment factor to measure the marginal effect of sentiment scores on stock returns.