Udacity Machine Learning Nanodegree Program Capstone Project Proposal

Yifei Tong

***Domain Background**:* The research is derived from the field of finance. Specifically, the research intends to study the effect a financial news article will have on the movement of related stocks. If a financial news article can be used to quickly predict future short-term price movement, the model can potentially generate profit by telling investors to act sooner after a news articles is published.

Problem Statement: The primary problem this research project is trying to solve is a subset of the larger problem described above. Specifically, this project intends uncover the short-term effect a related financial news articles on the stocks of four of the largest technology companies, namely Google, Facebook, Amazon, and Microsoft. The short-term effect in this problem is defined as the one-day percentage change in stock price. The problem then becomes a regression problem where the input is a financial news article and the output is the predicted one-day percentage change in stock price.

Dataset and Inputs: I generated the relevant dataset from a large set of financial news articles found at the Kaggle "US Financial News Articles" data page. From all the json files in the page, I read the "text" sections and the "published" sections, which correspond to the news articles and the published time of the articles. The articles are filtered using keywords "Google", "Facebook", "Amazon" and "Microsoft". Then I used the pandas-datareader package to read stock data from Yahoo Finance and calculate the percentage price change of the stocks on the published date. After filtering out the news on non-trading days, the sizes for the datasets are 4804 articles for Google, 6276 articles for Amazon, 11747 articles for Facebook, and 2392 articles for Microsoft. The average length of all articles in the filtered dataset is about 695.4. The input in this problem will be a preprocessed and vectorized representation of a news article.

Sample text from dataset: a slight gain on friday was not enough to stop stocks from posting a los s this week, weighed down by fears of a possible trade war and white house turmoil.\nthe s&p 50 0 notched a 1.2 percent loss for the week, despite a 0.2 percent gain on friday. the dow jones ind

ustrial average also fell 1.5 percent on the week as shares of boeing dropped 6.8 percent on the tr ade tensions. the dow closed 72.85 points higher on friday at 24,946.51.\nthe nasdaq composite c losed flat at 7,481.99 amid a 1.4 percent decline in google-parent alphabet and a 0.7 percent fall i n amazon shares. the index also fell 1 percent for the week.\ntariffs on steel and aluminum impor ts are expected to come into effect in the coming weeks, after trump signed two declarations last week. while canada and mexico are exempt from the deal, investors worry that countries around t he world including china may strike back.\n"the market is still vulnerable to headlines, particular ly with regard to trade and any retaliation," said quincy krosby, chief market strategist at prudenti al financial. we\'re "waiting for reaction from the european union and reaction from the chinese i n terms retaliatory responses."\nalso on investors\' minds, krosby added, is the two-day federal o pen market committee monetary policy meeting next week, with wall street preparing for new fe deral reserve chair jerome powell to lead bankers in raising rates.\n"it\'s jerome powell\'s first co nference and the market expects a rate hike ... [investors] will also be paying attention to his com ments and the press conference," krosby said. "he\'s extremely fluent in the language of the fed, e xtremely fluent in the thinking of the fed."\nin political news, president donald trump has reporte dly decided to remove national security advisor h.r. mcmaster from the u.s. administration, accor ding to a thursday report by the washington post. the white house has, however, denied that any c hanges are set to emerge within the national security council.\nadding to the political drama, cbs news reported on friday that white house chief of staff john kelly, too, could depart the administr ation as early as today. fears that the chief of staff could be on his way out were kept at bay, how ever, after the wall street journal reported that trump and kelly settled on a temporary "truce."\nk elly, rattled by president trump\'s abrupt firing of secretary of state rex tillerson via twitter earlier in the week, had told colleagues to start looking for new jobs, the journal reported. tillerson\'s di smissal comes a week after gary cohn resigned as the national economic council\'s director.\nbre ndan mcdermid | retuers nyse trader on the floor "i think the market has understood for a while th at this is a chaotic white house," said michael shaoul, chairman and ceo of marketfield asset man agement. he noted that stocks have been trading in a close range recently. "i think it will take mo re economically driven or corporate-driven news for the market to make up its mind."\nthe com merce department said housing starts declined 7 percent in february, a bigger-than-expected fall. building permits, meanwhile, fell 7.7 percent last month.\nelsewhere, consumer sentiment rose to a level not seen since 2004 in march, according to a preliminary reading from the university of

michigan. meanwhile, the labor department said job openings increased to 6.3 million in january, a record.\n"this week investors have been focused on washington," said jeff kravetz, regional inv estment strategist at u.s. bank wealth management. "but the narrative seems to be changing with strong economic numbers."\n"i think investors are going to be focused on economic data" and th e federal reserve next week, kravetz said.\nin corporate news, adobe systems reported better-than -expected quarterly earnings, sending its stock up 3.1 percent.\nmeanwhile, walmart responded t o accusations of issuing misleading e-commerce results, calling the person a "disgruntled former associate." walmart shares rose 1.9 percent.\n—cnbc\'s jacob pramuk contributed to this report.

Solution Statement: To develop solution for this problem, I intend to train regression models using different algorithms on the dataset for each technology company. The resulting solutions will be different regression models to compare the performance on.

Benchmark Model: No benchmark model addressing exactly the same problem as the one proposed here can be found online. However, the model can be benchmarked to actual stock data on a date a news articles is published.

Evaluation Metrics: The evaluation for this problem can be similar to other regression problems. We can use metrics like mean absolute error or root mean squared error to represent the error made by our models.

Project Design:
This section describes the theoretical workflow for solving this problem. The collected dataset is already saved as a pickle file locally.
The first step will be to upload the said pickle file to the AWS S3 bucket.
Then we can create a notebook instance from which we create models on. In the training notebook, we load the pickle file and read the dataset. Afterwards, we can preprocess the news articles by removing stop words and keeping only the stems of words. Then we can vectorize the articles by constructing a large dictionary for the articles. The vectorized articles will then serve as the input for the regression models.

Then we can use Amazon SageMaker's LinearLearner to train a classic linear regression model on the input. Besides LinearLearner, we can also take advantage of Amazon SageMaker's custom model functionality to develop our own LSTM models. We then train both models on the training dataset which contains the vectorized news articles and the percentage price changes. After training the models, we can deploy endpoints that we can send test data to and evaluate the performances of the models with.

Reference:

"US Financial News Articles", Kaggle, https://www.kaggle.com/jeet2016/us-financial-news-articles