Predicting the S&P with LSTMs and GloVe (unsupervised learning)

Winfield Chen, Khizr Ali Pardhan, David Pham

Introduction

The simultaneous rise in social media and discount brokerages has spurred the proliferation of online communities dedicated to investing and trading on the stock market. The most popular of these communities is the subreddit r/wallstreetbets (WSB). At over 1 million active subscribers, it is the fourth most popular subreddit at this time. Bloomberg Businessweek recently featured the subreddit in an article describing the message board's ability to reshape the options market and spark stock rallies. With such a large following and arguable influence, the question arises as to whether the community's sentiment can be successfully used as a predictor of stock performance.

Data

Stock Data:
Wharton Research Data Services (WRDS), a division of the Wharton School of Business at the University of Pennsylvania, is a highly-regarded source in academic research and has a dedicated quality control analyst at the NYSE. Once granted access to their database, we created a hourly stock price dataset from their consolidated Millisecond Trade and Quote (TAQ) dataset. Our dataset consists of hourly stock prices for Apple, Amazon, Boeing, SPY (S&P), and Tesla from Jan-01-2017 to Nov-31-2019. The selection criteria was for stocks with the most mentions for 2019 with Amazon and SPY far surpassing the other 3.

r/wallstreetbets data from reddit:
Pushshift is the largest publicly-available Reddit dataset containing both comments and submissions. We downloaded the Pushshift Comments dataset for r/wallstreetbets from Jan-01-2017 to Nov-31-2019.

GloVe - Global Vectors for Word Representation:
GloVe is an unsupervised learning algorithm for obtaining word vector representations from a word corpus. We chose to use Magnitude's pre trained word vector datasets, GloVe Twitter (2 billion tweets containing 27 billion tokens, with a vocabulary size of 1.2 million) for 25-dimensional word embeddings and GloVe Common Crawl (web archive containing 840 billion tokens, with a vocabulary size of 2.2 million) for 300-dimensional word embeddings.
Link : https://nlp.stanford.edu/projects/glove/

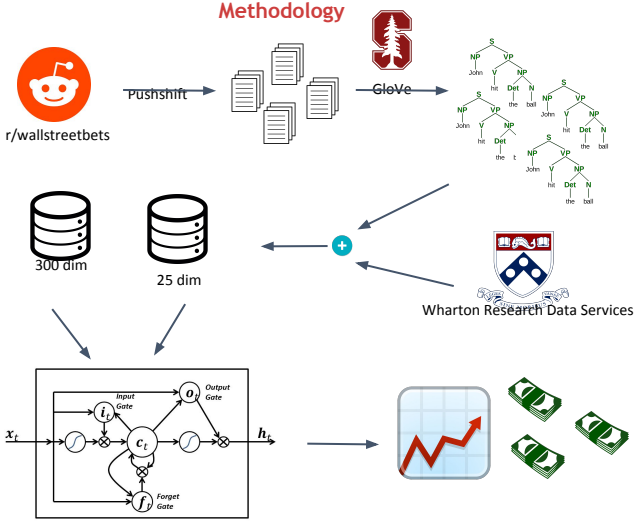
Methods

We embedded each comment by averaging the word vectors of each word in the comment to form a document vector. This generated 2 embeddings, a 25-dimensional embedding and a 300-dimensional embedding using the respective GloVe datasets.

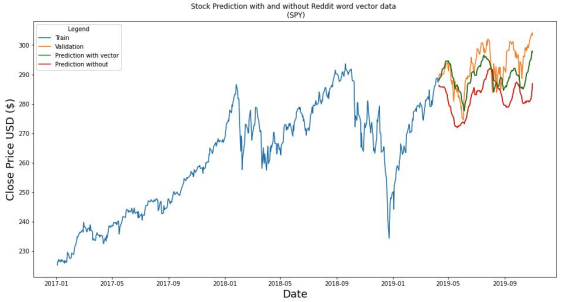
This CPU-bound task alone takes several days on a single machine for one successful run. To speed up this computation, we parallelized the task with ~60 "idle" Computing Science Instructional Laboratory (CSIL) machines each with a 12 core CPU. In addition to the standard document vectors, we appended score, gildings, and word count of each post as additional features.

The word vectors were then aggregated over hours and days separately creating 2 final datasets GloVe datasets, which were then joined with our stock ticker prices.

Methodology

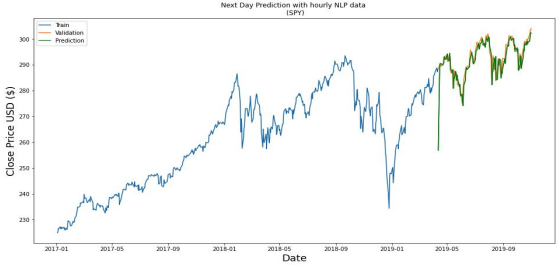


LSTM:
With our fresh new dataset(s), we experimented with several recurrent neural networks to leverage the sequential nature of our dataset, ultimately landing on using stacked long short term models with default activation functions. Using the ReLU and Tanh functions to activate Dense layers were found to undesirable. Stateful LSTMs where taken into consideration, however, we determined them to not be worth the additional cost.



Results

From our experimentation, we found that time series forecasting improves with increasing data granularity. This coincides with our initial expectations because averaging our dataset results in both information and variance loss. As you see from our two graphs, adding both daily or hourly reddit NLP data does yield an increase in prediction accuracy. However, there is a clear advantage to using hourly data. In terms of baselining, we found our LSTM to be competitive against some non neural network based models we evaluated.



Conclusion

For stocks that were highly mentioned in WSB, more data was better. It was not shown in the poster, but for stocks TSLA, BA, and AAPL, there was significantly less improvement in the NLP runs. This was not particularly surprising as their mentions in WSB were not as high. This suggests that perhaps WSB does have some small effect on the market or at least the community responds quickly and appropriately to market news.

Future Work

LSTM Architecture and Hyperparameter Optimization
Due to computational constraints and the size of the dataset, the permutations of LSTM architecture and hyperparameter setting explored was fairly shallow. Future iterations may benefit from Convolutional LSTM layers and optimization hyperparameter tuning.

Dataset
Additional features such as dividends, volume, business sector, and technical indicators, namely the relative strength index, could all be potentially beneficial. However, the trade-off here would be computational cost in both LSTM training and GloVe generation.

Averaging the constituent word vectors of a document to form a document vector is a primitive approach which loses information about word order and importance. Future work should explore the use of a trainable neural network document vector model such as Doc2Vec which extends word vectors to documents in a more sophisticated fashion.

Github.com/alik604/CMPT-419