Project Proposal
MPCS 53113: Natural Language Processing

Shobhit Verma, Ashish Verma
shobhitv@uchicago.edu, ashishv@uchicago.edu

June 27, 2022

1 Title and Team Members

Our project is '`Prediction of stock prices via sentiment analysis`'. The team members are - `Shobhit Verma (shobhitv@uchicago.edu)` and `Ashish Verma (ashishv@uchicago.edu)`.

2 Abstract

Our project aims at tackling the age old problem of predicting stock prices. Our assumption - which is widely accepted - is that the share price is hugely correlated to 'public sentiment'. In this project, we first aim at quantifying public sentiment by sentiment analysis of related tweets, and then use that to predict the change in stock price. The project is composed of three parts:

1. Curate a dataset of tweets that capture the sentiment of the stock market and stock prices.
2. Building a classification model for sentiment analysis of tweets - given a tweet, the model classifies the tweet into one of the following 5 classes: extremely negative, negative, neutral, positive, extremely positive.
3. Building a regression model that maps the sentiment (or a group of sentiments of all the tweets made in, say, a day) to the change (delta) in stock price.

The classification task allows us to choose from many models - RNNs/LSTMs, transformers, etc. and similarly, in regression, we will experiment with linear/logistic regression, neural networks, SVMs, etc. We build our own dataset from scratch - tweets are fetched from the Twitter API and the stock prices are fetched from yahoo finance API. We use pre-existing datasets to train our sentiment classifier, and then use that to classify our tweets into sentiments. Once we have the aggregated 'sentiment' and the corresponding fluctuation of

stock price in a day, we use this to train our regression model. Tweets and stock prices will be fetched for only a single company at first to reduce the complexity of our task.

3 Related work and literature survey

According to [1], a stock market, equity market, or share market is the aggregation of buyers and sellers of stocks (also called shares), which represent ownership claims on businesses. From [2], the share price is determined by 'demand and supply' of the shares in the market. Typically, shares with higher demand have higher prices and shares with higher supply have lower prices. In this section, we first give a brief overview of existing works in sentiment analysis and then we discuss previous attempts in utilizing sentiment analysis for prediction of stock prices.

[3] gives a perfect survey of existing sentiment analysis algorithms and applications. It organizes these algorithms into two broad categories - the machine learning approach and the lexicon based approach. We adopt the former category in this project. Although not used in this project, it is worth describing lexicon based approaches - these rely on a sentiment lexicon, a collection of known and precompiled sentiment terms. Further division is into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. [3] surveys and compares the following ML approaches - Probabilistic classifiers (Naive Bayes classifier, Bayesian networks, Maximum Entropy classifier), Linear classifiers (Support Vector Machines, Neural Networks), Decision Tree classifiers and rule based classifiers. They, however, do not mention models used commonly in NLP tasks - RNNs/LSTMs, GRUs and transformers.

As stated by [4], the main problems of the supervised methods mentioned above are that they need a large amount of training data and are usually slow, and while the lexicon based approaches are much faster, they usually fail at hitting a practical accuracy target. [4] shows that an attention based CNN-RNN model works better than the aforementioned supervised methods. They however, do not mention usage of transformers.

[5] presents another similar related work which attemps to solve the same problem of stock price prediction using Tweets. It uses pre-trained Glove and Sentiment-Specific Word Embeddings and NLTK tokenizer, and then predicts a positive or negative sentiment for a particular tweet using a softmax classifier. Using this positive or negative sentiment, a prediction is made if the stock price will increase, or decrease. However no absolute value of the delta is predicted. As a direct improvement, in this project we'll attempt to predict the actual delta in the stock price by using a multi-class sentiment from a given set of Tweets rather than just positive or negative, and then predict the magnitude using that.

4 Plan of action

The action items are broadly categorized into 3 steps, mentioned below.

- Curate dataset specific to comapanies - using twitter API fetch tweets, other APIs from financial news channels and using yahoo finance api fetch stock price. Specifically, we are interested at the opening and closing price of the stock given a day. We then attempt to correlate the sentiment of tweets for a day to the stock price fluctuation.
- For the purpose of sentiment analysis, we use a classifier to quantify the notion of sentiment into five categories. We use the following models - Feed Forward Neural Networks, CNNs, RNNs/LSTMs, Transformers. We choose the model with best performance. We will use word2vec dataset for getting accurate word embeddings.
- Once we have confidence on the quantified value of the sentiment by day, we will correlate this to the fluctuation of the stock price on the same day using a regression model. Here, some exploration will be required - most of the previous work in this domain focuses on correlation of a sentiment of a day to the stock price fluctuation on the same day, however, we feel that the stock price may depend on the overall sentiment of a certain number of days instead of one. For example, a highly negative sentiment (say, outbreak of war) will continue to effect stock prices for more than a day.

5 Evaluation criteria

- Train till day k, predict stock price at end of $k+1$, compare with ground truth. We'll build a training set comprising Tweets and corresponding fluctuation in stock price till day k, and then predict the stock price delta for the next day and compare it with the ground truth for next day.
- Train till day T - k, predict from T - k to T, compare with actual/ground truth. We'll only train on a dataset which contains Tweets till day $T-k$, and then predict the consecutive fluctuations in the stock prices for the next k days and compare with ground truth.
- We'll compare our model's accuracy with the existing work in this domain.

6 Division of work

- Dataset curation. The dataset comprises of 2 sets of data - the Tweets (input X), and the corresponding fluctuation in stock price for the next day (label y). We'll divide this curation of X and y, corresponding to each organization.

- Study the existing work in this domain by dividing the work/research-papers.
- We'll experiment with different model designs to land at the best possible network. We'll individually experiment with different network configurations after discussing.

7 References

- [1] https://en.wikipedia.org/wiki/Stock_market
- [2] https://investopedia.com/ask/answers/how-companys-stock-price-and-market-cap-determined
- [3] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams engineering journal 5, no. 4 (2014): 1093-1113.
- [4] Basiri, Mohammad Ehsan, Shahla Nemati, Moloud Abdar, Erik Cambria, and U. Rajendra Acharya. "ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis." Future Generation Computer Systems 115 (2021): 279-294.configurations after discussing.
- [5] Michael Jermann "Predicting Stock Movement through Executive Tweets". https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2743946.pdf