Stock Price Prediction using Sentiment Analyses of Tweets

Anand, Mayank
my321532@dal.ca

Bhupathiraju, Akhilesh Varma
ak445438@dal.ca

November 2021

1 Problem Statement

Researchers have been predicting Stock Market using machine learning like ARIMA by taking the past data for a while now. In addition to the past data, they have also included other features, as it is well known that there are various other factors that affect the stock price. Moreover, with the recent advancements in deep learning space, neural networks specially RNNs and GANs have been proven to perform better than previous ones. In our research, we will be trying to discuss about expanding one such idea of using GAN as implemented by Priyank Sonkiya [5] and looking at particularly how the sentiments of tweets related to the stock company would be effective in predicting the stock price. Furthermore, we will be using additional features instead of just looking at the historical data for example 7-day moving average, 21-day moving avg, Exponential moving average, Bollinger bands as implemented by Priyank Sonkiya [4] and as Anshul Mittal [5] in his research used POMS score along with No of tweets containing the stock name which showed how tweets sentiments can affect stock prices. And to get the sentiment scores we will be doing the sentiment analysis of tweets from twitter accounts of CEO/CTO, company news, public news. To get the sentiment scores we will be using State-of-the-Art models like BERT which is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. As for the actual prediction of future stock prices, we will be exploring various models from machine learning (ARIMAX) to deep learning (LSTM and GAN) for testing out the results and compare them. Our main model would be GAN where we would pass in various technical parameters along with the previous day closing prices of stock, and the sentiment scores. The data for our purpose would include stock market data of SP 50 stock companies extracted from Yahoo Finance website. As for the tweets, we will be extracting them from twitter from different sources like Company news account, CEO/CTO official account, Stockwits account. Stock market being a highly volatile market, there are various things which can trigger a sudden rise or fall of the price, one of them being the real time news related to that company. For instance, a highly optimistic tweet from the CEO of the company could grab the eyeballs of many investors around the world leading to spike in the price. This is why looking at the sentiments of the tweets can help us understand the market change and may be useful in the prediction the price of company's stock.

2 Potential Approaches

2.1 Data Collection

We are using Twitter API for python to get the tweets from twitter based on keyword search. And also, tweets from user specific account. One challenge would be that since we will be having many tweets from different accounts, we will be needing a single sentiment value for a day as the data for the stocks present is day wise. For this, we will be considering having a normal average for all the sentiment scores of tweets within the day and a weighted average for the sentiments of tweets between different accounts. i.e. a tweet

from the CEO of the company would be given more importance than a tweet from some random source on twitter.

2.2 Data Prepossessing

As our one of our main source of data will be from twitter and we are expecting to have a lot of noise in it. To remove that noise we will be using NLP techniques for example one of the basic is removing links from tweets using REGEX. Another approach that we are planning to use is to remove stop words such as is, the, etc. followed by stemming/lemmatization from tweets before pushing it for sentiment analysis to VADER [3]. We will not be removing stop words in our other approach where we will using BERT as it gives sentiments based on pragmatics or whole sentences which might be beneficial for predicting accurate sentiments.

2.3 Sentiment Analysis

For predicting the sentiments of our tweets, we will be exploring two approaches. One would be VADER (Valence Aware Dictionary for Sentiment Reasoning) which is rule-based (lexicon based) analysis tool that helps us get sentiment score which can be either positive, negative, neutral or compound based on the twitter sentence that we pass. Another one would be a deep learning transformer based model called BERT by Google [1] that has been trained on huge corpus of data. In addition to BERT there is finBERT [6] which is a fine tuned version of BERT trained on financial data. We would be comparing the results of these two. The sentiment scores is passed to the generator model of the GAN as a latent vector similar to the one done in the paper [5]

2.4 Potential Models

2.4.1 ARIMAX

ARIMAX stands for Auto-regressive integrated moving average and X stands for exogenous variable which is nothing but the additional variables that can included for better prediction. In the ARIMAX model we will be passing in the historical data of the closing price, along with other key parameters that we believe will add value to our model prediction. This model will act as a metric to rate our deep learning models. As from previous researches it has been proven that deep learning models are much efficient in time series data.

2.4.2 Long Short Term Memory (LSTM)

LSTM models were created as a solution of short term memory because of vanishing gradient problem of Recurring Neural Networks(RNN) and perform extremely well with time series data for generating sequences. Internally LSTM have three gates Forget, Input and Output. Forget gate decides what information should be throw away or kept. Information from previous hidden state and information from current input is passed through the sigmoid function which give value between 0 to 1.Value closer to 0 will be thrown away.Input gate is used to update cell state and the Output gate decides which information will be propagated to next cell.

2.4.3 Gated Recurrent Units (GRU)

GRU are pretty similar to LSTM only difference is we have only two gates in GRU as compared to three in LSTM. GRU's have fewer Tensor flow operation; therefore, they are little speedier to train then the LSTM's. And it is observed that GRU performs well on small datasets.

2.4.4 GAN

Generative Adversarial Network is deep learning model designed by Ian Goodfellow in 2014 [2] . GAN contains two neural networks called Discriminator and Generator which compete with each other in a minmax game. Zhang et al [7] has taken MLP (multi layer perceptron) as the discriminator and LSTM as the generator for the forecasting the future closing prices. Priyank Sonkiya [5] on the other hand has taken 1D CNN as the discriminator and GRU as the generator. Both the researchers have shown some exciting results. Influenced by their work we will be stacking LSTM and GRU as the generator and for the discriminator we will be trying both MLP and 1D CNN and compare the results. Our model will be different from the other two in such a way that as we are stacking the different RNN models and using additional features such as POMS, weighted sentiment scores. Stock data with additional variables are passed to the discriminator while in the generator the sentiment scores of the tweets will be sent as a latent vector.

2.5 Project Timeline

November 6 – December 6			
Week 1	Week 2	Week 3	Week 4
Extracting stock Data, tweets data from twitter, cleaning and pre-processing of text	Feature generation, sentiment scores using VADER and BERT	Modeling ARIMAX and LSTM	Modeling GAN, comparing results and project report.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[3] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[4] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[5] Priyank Sonkiya, Vikas Bajpai, and Anukriti Bansal. Stock price prediction using bert and gan. *arXiv preprint arXiv:2107.09055*, 2021.

[6] Yi Yang, Mark Christopher Siy UY, and Allen Huang. Finbert: A pretrained language model for financial communications, 2020.

[7] Kang Zhang, Guoqiang Zhong, Junyu Dong, Shengke Wang, and Yong Wang. Stock market prediction based on generative adversarial network. *Procedia computer science*, 147:400–406, 2019.