Stock Market Prediction using NLP

Jyothsna Aitipamula
Sai Deekshith Miyapuram
Vaishnavi Sai Malapati

Department of Data Science University of
New Haven

Abstract:

We used Natural Language Processing models to evaluate data pertaining to the stock market of corresponding time, to predict stock trends. In this project we are using BERT model to analyze the sentiment between tweets and their affect on stock prices for investor-based decision. we used LSTM model which shows good ability to get long-term prediction. In this project our goal is to built a model which increase the accuracy of stock prediction.

I. Introduction:

A stock market is the aggregation of buyers and sellers of shares of their business or ownership. Therefore, prediction of stock price movement plays an important role in reducing risk of loss and increase profits which helps in individual investor's interest and country total economy trend.

Natural Language Processing is a technique used by computer to understand and manipulate natural languages. Which is used to analyze text and let machine derive meaning from the inputs. This human-computer interaction allows us to come up with many different applications to bring humans and machines as one.

In stock market the public sentiment plays a key role in which it can change the trend of stock exchange.so we need to conduct sentiment analysis of data. Sentiment analysis is a natural language processing (NLP) technique used to determine whether data is positive, negative, or neutral. To collect the data, we used social platforms like twitter where we can get the fastest information about the stock exchange but the data is messed up so we analyzed the text data using NLP tools and finally quantified the unstructured data. We did sentiment analysis using the BERT model.

Apart from the features we used to make predictions choosing the best technique plays a major role because for making predictions we need to ability to memory long term. So, we used long short-term memory (LSTM) neural networks, which is more suitable in this case since stock price is a long-time sequence with random and nonstationary characteristics.

II. Related work and technology background:

a) Related work:

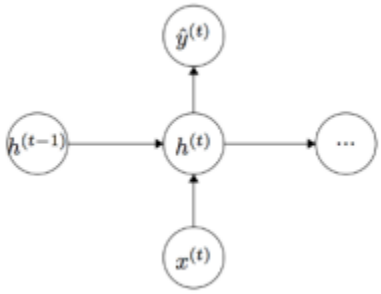
Stock markets can reflect anything Individual stock information. Overall stock market Efficient i.e., neither fundamental analysis nor technical analysis, it could help investors make more money from stock forecasts. Historical stock prices, fundamental analysis and sentiment analysis could help investors predict the stock market better. Many past studies have developed various machine learning techniques to achieve inventory forecasting. Their final results show that the LSTM model, a special case of RNNs belonging to deep learning, outperforms other methods in predicting stock closing prices.

b) NLP (Natural Learning Processing):

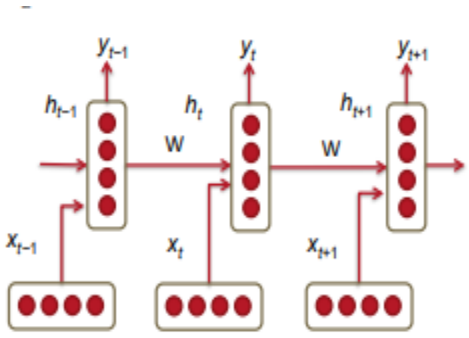
Natural Language Processing (NLP) is a study and technology field that investigates how computers could be used to interpret human language text. As a sub-field of NLP, sentiment analysis helps people to extract emotion factor in given text. Through measuring and analyzing attitudes within the text, it provides valuable insights about the possible relation between stock movement and human sentiment, therefore enable us to make a better prediction of future stock price.

c) RNN (Recurrent Neural Network):

Unlike the conventional translation models, where only a finite window of previous words would be considered for conditioning the language model, Recurrent Neural Networks (RNN) are capable of conditioning the model on all previous words in the corpus.



Input and output of RNN

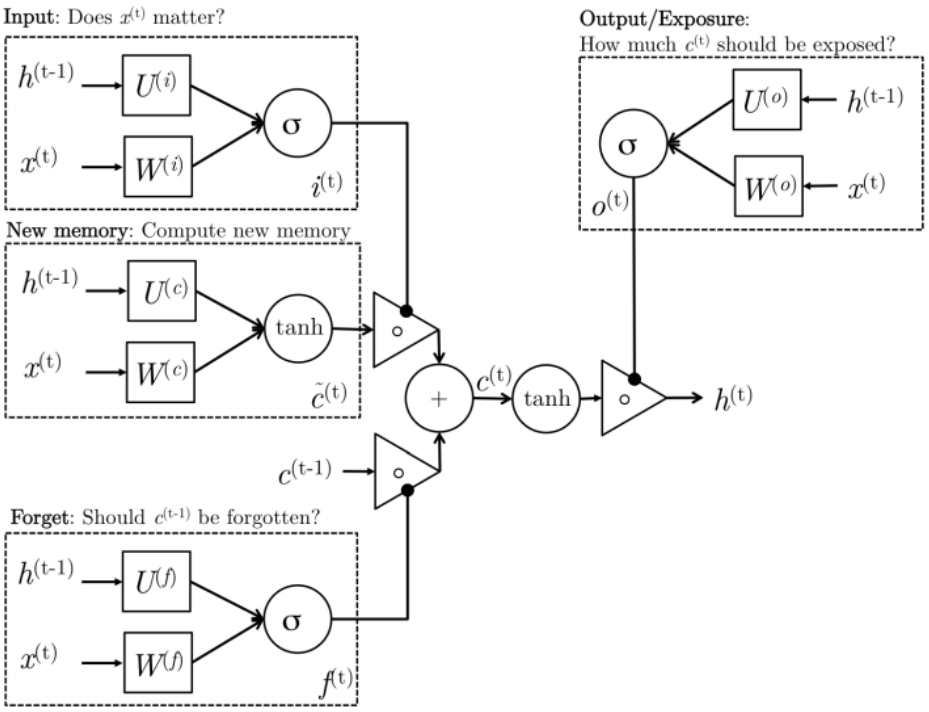


Architecture of RNN

d) LSTM (Long Short-Term Memory):

Compared to traditional neural network, recurrent neural network (RNN) has internal memory, they memorize all information stored in the past and use it to make decisions in further step. Even RNNs work well when dealing with short sequence data, they suffer from two major problems: gradient vanishing and gradient exploding.

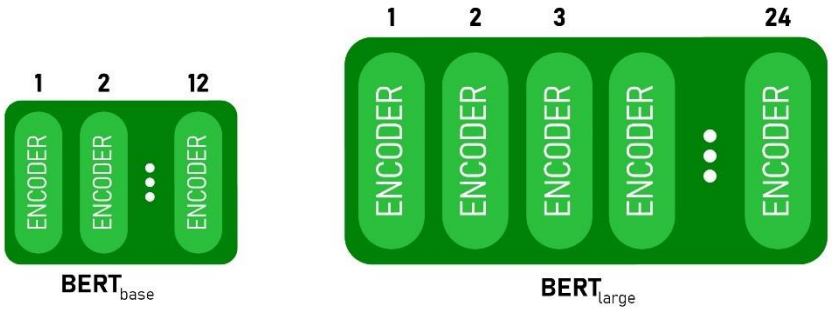
To solve the problem of RNN, Long Short-Term Memory (LSTM) should be considered. LSTM acts as a specific RNN with the improvement on identifying long-term dependency in sequence data. Unlike RNN, LSTMs perform a more complex way in computing the hidden state with replacing the traditional hidden layer neurons by sets of memory cell. Information can be regulated through three gate structures. These gates update information selectively by learning what kind of information in the sequence is important to keep or identifying information that not useful and then throw them away.



LSTM architecture

e) BERT (Bidirectional Encoder Representations from transformers):

BERT is a natural learning processing model. BERT is basically the encoder stack of the Transformer architecture. The Transformer architecture is an encoder/decoder network that uses self-awareness on the encoder side and attention on the decoder side. This model first takes a CLS token as input and then a string of words as input. where CLS is the classification token. Then transfer the input to the layer above. Each layer applies self-recognition and passes the result through a feedforward network to the next encoder. This model outputs the hidden magnitude vector (768 for BERT BASE). In this project we use BERT for sentiment analysis



BERT (base) BERT (large) architecture

III. METHODOLOGY AND PREDICTION MODEL

In this section we explore the models, techniques and approaches used by researchers for stock market prediction: what methods are used, what kind of data is used, what are the implemented preprocessing techniques, while discussing the achieved results of the model.

A. Model Selection

The stock market is unpredictable and affected by numerous factors such as, the effectiveness of Company and public sentiment etc. It is challenging to apply several conventional statistical models to financial time series prediction with high precision due to the characteristics of high volatility and noise. According to several studies, time series forecasting using machine learning techniques like RNN, Random Forest, SVM etc. Same goes for sentiment analysis as there are lot of finetuning models like NLTK's Tweet Tokenizer, Genism, SpaCy but among them BERT works best for our project. The BERT model is one of the first examples of how Transformers were used for Natural Language Processing tasks, such as sentiment analysis (is an evaluation positive or negative) or more generally for text classification.

Therefore, we use a pre-trained BERT model that has been trained on a huge dataset. Using the pre-trained model and try to "tune" it for the current dataset, i.e. transferring the learning, from that huge dataset to our dataset, so that we can "tune" BERT from that point onwards. The LSTM is a unique example of an RNN model with a memory selection mechanism. Using its memory cells, it retains information that is beneficial and discards information that is inappropriate, thereby capturing the structure of data over a long period of time with great precision. As a result, this study believes that LSTM is a solid option for predicting stock price and BERT for sentiment analysis.

B. Data Collection

- In this project we used two main data sets for the prediction of stock market. The first one is the APPLE company stocks which are taken from Kaggle (https://www.kaggle.com/datasets/berkayalan/apple-stock-data-between-2015-and-2022?resource=download\) It contains features like date(Date of the stock price), open price (opening stock price in a related day), Close price (closing stock price in a related day), high price (highest stock price in a related day), low price (lowest stock price in a related day), Adj Close (The closing stock price after adjustments for all applicable splits and dividend distributions in related day) , volume (sales volume of the stocks). This dataset contains Apple's(AAPL) NasdaqGS - NasdaqGs daily stock price information between 2015 and 2022. All data is collected from Yahoo Finance, and they are in USD.The size of the file is 43kB. The csv file can be read using read_csv() method. After loading the dataset it is then preprocessed for cleaning the data, removing the null values ,etc.

- The second dataset is the APPLE company Tweets d ataset. Recent studies on sentiment analysis in stock prediction have found that Twitter can be a useful tool for gauging public attitude. They suggested that data taken from the microblogging platform is effective for studies about marketing trends and social behavior because more individuals are using twitter to share their thoughts due to its ease of accessibility. For this we have real-time tweets from the twitter that are related to APPLE company. We scrapped the real-time data from twitter using snslibrary (https://drive.google.com/drive/u/0/my-drive). After scrapping the final dataset consist of features like DateTime (The date and time the tweet posted), Tweet Id (Id of the tweet), Tweet (news), Username (person posted). The dataset contains around 5k tweets which are later used for sentiment Analysis. In order to quantify the public sentiment towards company stock, this paper collect all news tweets from 2015 to 2022that contain the APPLE company. Here for the project we used libraries like Keras, Tensor Flow, snscrape, BERT Transformers.

C. Sentiment Analysis

Sentiment analysis is a procedure that uses Natural Language Processing to automatically mine attitudes, opinions, perspectives, and emotions from text, audio, tweets, and database sources (NLP). When applying sentiment analysis on social media, the text itself has been the main focus. Sentiment analysis task is very field specific.

Tweets are classified as positive, negative and neutral based on the sentiment present. Out of the total tweets are examined by humans and annotated as 1 for Positive, 0 for Neutral and 2 for Negative emotions. For classification of nonhuman annotated tweets, a machine learning model is trained whose features are extracted from the human annotated tweets. It was discovered that accounts with less than 171 followers on Twitter who tweeted about a company had a bigger influence on the stock's results the next trading day than accounts with more followers. Algorithms apply a systematic approach to sentiment analysis to extract things like polarity, subjects, and opinions from the text. Two approaches that can be used are rule-based language modeling and artificial intelligence-based computation of hidden patterns.

Sentiment analysis is seen as difficult since computational models cannot simply summarize and describe the grammar and structure of language. One of the issues is sarcasm detection and word ambiguity. Depending on the context and the usage of literary devices like irony, words can have numerous meanings. Additionally, social media text is condensed and uses emojis, abbreviations, and capital letters to accentuate meaning and emotions. Harmonized sentiment analysis is made more difficult by the fact that social media messages often communicate emotions in ways that are different from ordinary text.

	tweet	polarity	subjectivity	sentiment_score	sentiment
4433	HP's new omen 15 gaming laptop promises big pe...	0.094545	0.381818	0.094545	Positive
1899	Remember the slogan for Apple company when the...	0.000000	0.600000	0.000000	Positive
2878	The worst thing about this apple company is th...	-1.000000	1.000000	-1.000000	Negative
3793	Apple company ko hua 1.5 lakh ka nuksaan 😢\nKy...	-0.500000	0.500000	-0.500000	Negative
4595	@appleinsider Apple company and there supplier...	0.000000	0.000000	0.000000	Positive
98	References 2 Apple Company\n\nReferencias 2 Em...	0.000000	0.000000	0.000000	Positive
889	How Did Apple Change the Future of the World? ...	0.000000	0.125000	0.000000	Positive
4765	Apple company's services are the worst in Indi...	-0.133333	0.866667	-0.133333	Negative
4115	Apple company ko hua 1 lakh ka nuksaan😕\n ...	-0.500000	0.500000	-0.500000	Negative
2310	@Crassus_K @MigunaMiguna @JoshuaBaraka11 @Appl...	0.350000	0.650000	0.350000	Positive

Output of Sentiment Analysis

The findings are then compiled by date to measure the sentiment for each trading day. This study uses the mean sentiment score for days with multiple article recordings.

D. Data processing and model construction

1. Scaling

Before building the LSTM model, it is important to combine sentiment scores and stock historical data together and transform them into same scale range. Table.3 below shows the data frame of combine data, the last two columns are sentiment score got from article sentiment

analysis, rest are stock related indicators which include Date, Open, High, Low, Close, Adj Close and Volume.

	Date	Open	High	Low	Close	Adj Close	Volume	polarity	sentiment_label
0	2015-08-31	28.007500	28.632500	28.000000	28.190001	25.740059	224917200.0	0.18000	1
1	2015-09-01	27.537500	27.969999	26.840000	26.930000	24.589560	307383600.0	-0.06875	0
2	2015-09-02	27.557501	28.084999	27.282499	28.084999	25.644180	247555200.0	0.00000	1
3	2015-09-03	28.122499	28.195000	27.510000	27.592501	25.194494	212935600.0	0.00000	1
4	2015-09-04	27.242500	27.612499	27.127501	27.317499	24.943388	199985200.0	0.00000	1

It can be seen that the magnitude of stock price is different, and the data range of sentiment score is different from both stock price and stock volume. Using data in this format will negatively influence the prediction result, thus normalization is important to eliminate the magnitude difference between data. Normalization is necessary to remove the magnitude difference between the data since using the data in this format will have a detrimental impact on the prediction outcome.

In this project, we chose MinMax scalar to normalize the data.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

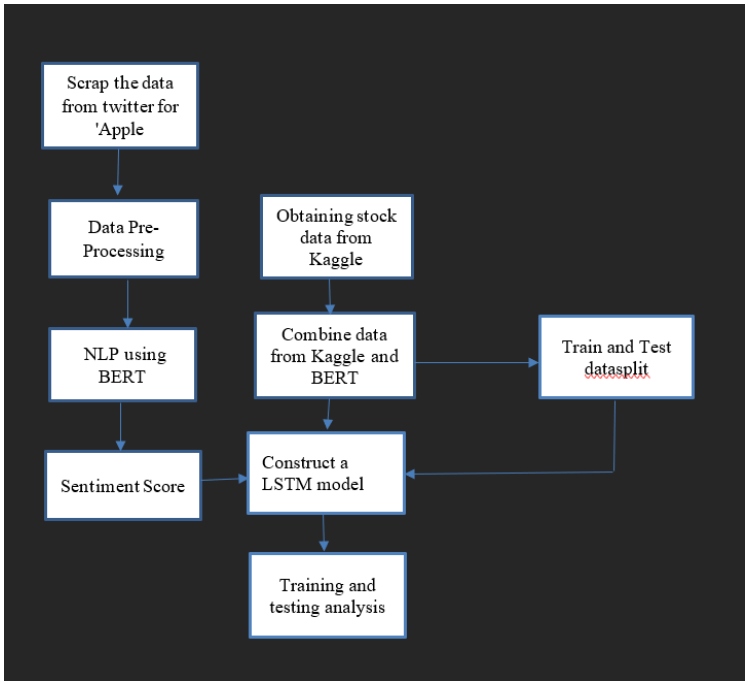
2. Dataset Split and parameters setting

All data collected have range from 2015.1.1 to 2020.8.13, this paper splits training and testing dataset follow the proportion 80:20. Thus, our final data used for training and testing are from 2015.1.1 to 2019.6.29 and 2019.6.30 to 2020.8.13, respectively. Within the training dataset, 20% are used for model validation. We have used the data with tweets and without tweets to actually compare the effect on stock prices on lstm model so the results can be analyzed for good understanding.

During the model fitting, this paper tried different values for various parameters of the LSTM model in order to get better prediction accuracy. Epoch measures how many times the entire dataset pass through the neural network, improper epoch number would cause underfitting or overfitting. Due to the large size of dataset, it is impossible to pass whole dataset through the neural net at once, thus, dividing dataset into several sets (batch) is necessary. Batch size is the number of training examples in each single batch. After tuning parameters, the final LSTM model has a two-layer network structure with 128 hidden nodes per layer. The batch size is 128 and the number of epochs is 10.

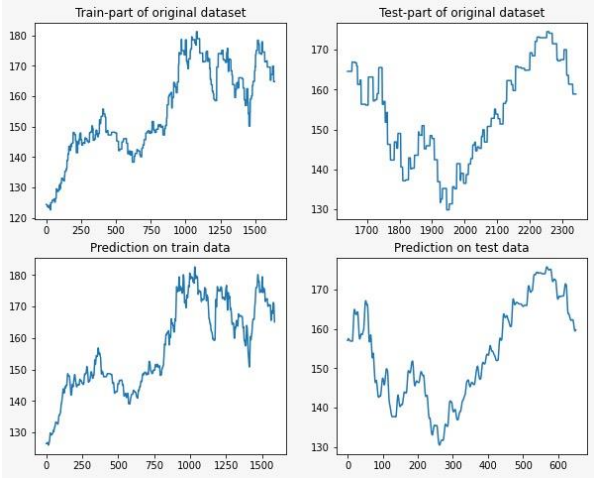
3. Model Construction

In this paper we have chosen Apple Company stocks as we are targeting one specific company tweet sentiments and how stock prices are varying respectively. BERT sentiment scores and features from the stock data are merged and used as input to LSTM model.



IV. Model Training Analysis

The learning shows the learning effect of the model over time, on the train and the test data set, a learning curve can be used to check if the model is underfitting, overfitting or well-fitting. BERT model was fine tuned to perform sentiment analysis of tweets. Below figures shows the plots for the training part and prediction on train dataset. The plots give the predicted adjacent closing stock prices for the apple company based on sentiment analysis.



V. Prediction Result Analysis

The performance of the LSTM stock price prediction model could be evaluated based on the mean square error (MSE). MSE represents the average of differences between actual and predicted values in dataset. Smaller MSE indicates that predicted values are closer to true values, model with smaller MSE provides a better prediction accuracy. Adam optimizer is used in our project to reduce the loss and MSE values. For results we have used graph plots to show the difference between using sentiment and without sentiment. This clearly explains the how sentiment can play vital role in stock prices on daily trading.

MSE with sentiment	MSE without sentiment
0.034	153.18

VI. Conclusion

This project only touches the surface of how the latest natural language processing techniques and various models are used to extract meaningful information from twitter tweets to perform sentiment analysis for APPLE company stocks. The BERT model used in the project was able to successfully achieve accuracy of 67% though initially we aimed for 62% accuracy. The LSTM model graphs clearly indicate that sentiment can be used effectively by investors before making decisions to invest in a particular stock. During the project, there were a couple items like comparing models to check which fits perfectly for this project could have been done, but due to time constraints we limited our project to BERT and LSTM model.

Future research may include feature selection for stock technical indicators and the incorporation of very significant characteristics, which could further enhance model quality. To make the sentiment analysis system more complete, sentiments from other news websites might also be helpful. Additionally, considering additional stocks from various industries may be beneficial in identifying intriguing company distinctions for the impact of sentiment analysis on stock prediction.

VII. References

[1] 2021 Asian Conference on Innovation in Technology (ASIANCON) Pune, India. Aug 28-29, 2021 Stock Market Trend Analysis on Indian Financial News Headlines with Natural Language Processing . Anshul Saxena , Dr Vandana Vijay Bhagat , Amrita Tamang.

[2] Nils Reimers. Sentence transformers: Multilingual sentence embeddings using bert / roberta / xlm-roberta & co. with pytorch.

[3] L. Bing, K. C. C. Chan and C. Ou, "Public Sentiment Analysis in Twitter Data for Prediction of a Company's Stock Price Movements," 2014 IEEE 11th International Conference on e-Business Engineering, Guangzhou, 2014, pp. 232-239. doi: 10.1109/ICEBE.2014.47

[4] W. Bouachir, A. Torabi, G. A. Bilodeau and P. Blais, "A bag of words approach for semantic segmentation of monitored scenes," 2016 International Symposium on Signal, Image, Video and Communications (ISIVC), Tunis, 2016, pp. 88-93. doi: 10.1109/ISIVC.2016.7893967

[5] NLP for Stock Market Prediction with Reddit Data Stanford CS224N Custom Project Muxi Xu
[6] V. U. Thompson, C. Panchev and M. Oakes, "Performance evaluation of similarity measures on similar and dissimilar text retrieval," 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, 2015, pp. 577-584
[7] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1-8, 2011.
[8] Z. Jiang, P. Chen and X. Pan, "Announcement Based Stock Prediction," 2016 International Symposium on Computer, Consumer and Control (IS3C), Xi'an, 2016, pp. 428-431. doi: 10.1109/IS3C.2016.114 Andrew Han. Financial news in predicting investment themes. 2020.
[9] Marcus Jun Rong Foo, Chi Seng Pun. 2022 4th International Conference on Natural Language Processing (ICNLP) Stock Movement Prediction with Social Sentiments and Interactional Data: Integrating NLP and Bayesian Frameworks
[10] Yuqiao Guo. 2020 2nd International Conference on Economic Management and Model Engineering (ICEMME). Stock Price Prediction Based on LSTM Neural Network: the Effectiveness of News Sentiment Analysis