Stock Market Price Prediction using LSTM and Sentiment Analysis of News Articles

Robert Li

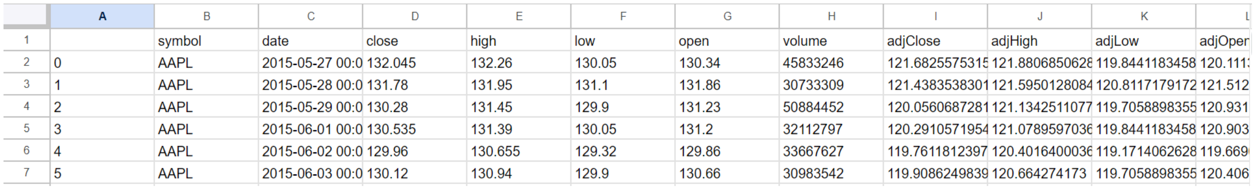
Tom Lewis

Introduction

In the dynamic landscape of financial markets, the ability to accurately predict stock prices has long been a pursuit of great interest and significance. The intertwining factors of market behavior, economic indicators, and global events create a complex ecosystem that challenges traditional forecasting methods. As technology advances, the integration of artificial intelligence and machine learning techniques has opened new avenues for predicting stock market movements with enhanced precision. This scientific report explores the topic of stock market price prediction by using the power of two innovative technologies: Long Short-Term Memory (LSTM) neural networks and Sentiment Analysis of news articles. LSTM, a type of recurrent neural network (RNN), is well-suited for capturing intricate patterns and dependencies in time-series data, making it a promising candidate for predicting stock prices that exhibit temporal dependencies. Concurrently, Sentiment Analysis, a subfield of Natural Language Processing (NLP), offers the capability to gauge public sentiment by analyzing news articles, thereby providing valuable insights into market sentiment and potential price movements. The amalgamation of LSTM and Sentiment Analysis not only aims to enhance the accuracy of stock market predictions but also to uncover the underlying connections between market sentiment and price fluctuations. By exploring this symbiotic relationship, we endeavor to contribute to the evolving landscape of financial forecasting and provide a nuanced understanding of the interplay between market dynamics and information dissemination. Throughout this report, we will navigate the theoretical foundations of LSTM networks and Sentiment Analysis, elucidate their integration in the context of stock market prediction, and present empirical findings derived from real-world data. This research not only holds implications for traders and investors seeking informed decision-making tools but also contributes to the broader discourse on the intersection of artificial intelligence and financial markets. As we embark on this exploration, we anticipate shedding light on the intricate mechanisms that govern stock price movements, paving the way for more robust and effective forecasting methodologies.

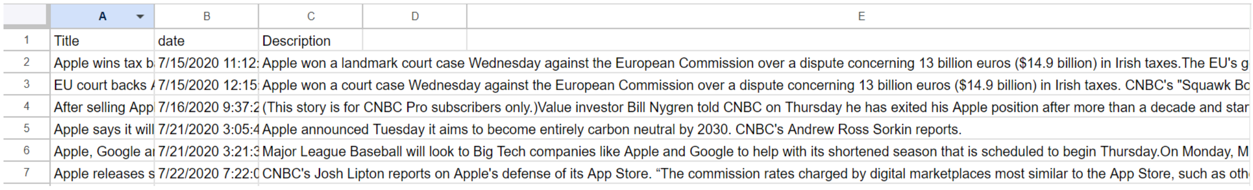
METHODOLOGY AND DATA

The first dataset consist of 1257 entries of Apple stock market prices from 2015 to 2022, the dataset includes data about stock's closing prices, stock volumes, lowest price or highest price that stock reached that day and ect. We assume that stock's closing prices, stock volumes are the most important parameters as for example information about the lowest price or highest price that dataset has doesn't give much information as for example price in the end of the day when news about stock already affected price and etc.. Sentiment score is an important factor we decided to use to predict price is the “mood” of the market about the company we want to predict the stock price. To do so we make a sentiment analysis of the news articles regarding mentioning it on the CNBC website. We used a free api to scrap all articles from 2015 to 2022 about Apple company, we took the information about the day on which article was published, url, article title and description of the article. In the end we were left with 15001 publishings.



	A	B	C	D	E	F	G	H	I	J	K	L
1		symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOper
2	0	AAPL	2015-05-27 00:0	132.045	132.26	130.05	130.34	45833246	121.6825575315	121.8806850628	119.8441183458	120.111
3	1	AAPL	2015-05-28 00:0	131.78	131.95	131.1	131.86	30733309	121.4383538301	121.5950128084	120.8117179172	121.512
4	2	AAPL	2015-05-29 00:0	130.28	131.45	129.9	131.23	50884452	120.0560687281	121.1342511077	119.7058898355	120.931
5	3	AAPL	2015-06-01 00:0	130.535	131.39	130.05	131.2	32112797	120.2910571954	121.0789597036	119.8441183458	120.903
6	4	AAPL	2015-06-02 00:0	129.96	130.655	129.32	129.86	33667627	119.7611812397	120.4016400036	119.1714062628	119.669
7	5	AAPL	2015-06-03 00:0	130.12	130.94	129.9	130.66	30983542	119.9086249839	120.664274173	119.7058898355	120.406

1. Dataset of the Apple’s stock data



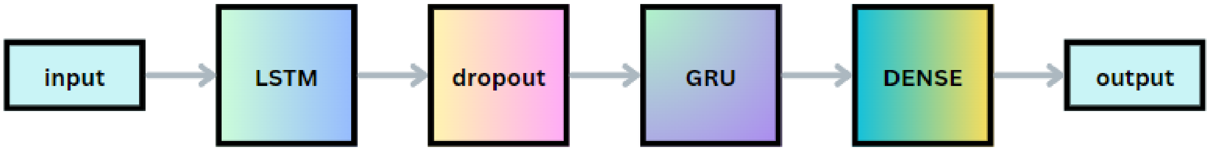
	A	B	C	D	E
1	Title	date	Description		
2	Apple wins tax b	7/15/2020 11:12:	Apple won a landmark court case Wednesday against the European Commission over a dispute concerning 13 billion euros ($14.9 billion) in Irish taxes.The EU's g		
3	EU court backs A	7/15/2020 12:15:	Apple won a court case Wednesday against the European Commission over a dispute concerning 13 billion euros ($14.9 billion) in Irish taxes. CNBC's "Squawk Bo		
4	After selling App	7/16/2020 9:37:2	(This story is for CNBC Pro subscribers only.)Value investor Bill Nygren told CNBC on Thursday he has exited his Apple position after more than a decade and star		
5	Apple says it will	7/21/2020 3:05:4	Apple announced Tuesday it aims to become entirely carbon neutral by 2030. CNBC's Andrew Ross Sorkin reports.		
6	Apple, Google ai	7/21/2020 3:21:3	Major League Baseball will look to Big Tech companies like Apple and Google to help with its shortened season that is scheduled to begin Thursday.On Monday, M		
7	Apple releases s	7/22/2020 7:22:0	CNBC's Josh Lipton reports on Apple's defense of its App Store. “The commission rates charged by digital marketplaces most similar to the App Store, such as oth		

2. Dataset of the news articles published on the CNBC site regarding Apple company

After getting these two datasets, we merged them and obtaining dataset with 6187 entries, we got such number since news are not coming out everyday and in some days on the other hand, many news could be published.

Next we had to make sentiment analysis of article titles this later will be added to the stock price dataset and considered by LSTM in price prediction. The logic of this is that the ‘mood’ of the market about the given company is a good heuristic for the prediction. Making a good sentiment analysis is a hard task, requiring a lot of data for this reason we use pre trained model “BERT” - its Bidirectional Encoder Representations from Transformers is a language model based on the transformer architecture from Google. It also includes a pre made word embeddings. Next we took the article names and used sentiment_score method to get the score that ranges from -1 to 1, where number closer to -1 means that article title is most likely negative, number closer to 0 means that title is neutral and if number is leaning towards +1 then article title is most likely positive. After obtaining sentiment score for each article title we can pass it later to the model that will consider this factor along with the other parameters to predict stock price for the next day.

For the model we decided to use following structure: LSTM layer with 50 units, dropout layer, GRU (**Gated recurrent units)** with 50 units and then dense layer.



3. Structure of the model

For training we decided to use a batch size of 32 and 900 epochs as learning with more epochs showed worse results. Dataset of size of 6188 was split as follows: 70% training data and 30% validation.

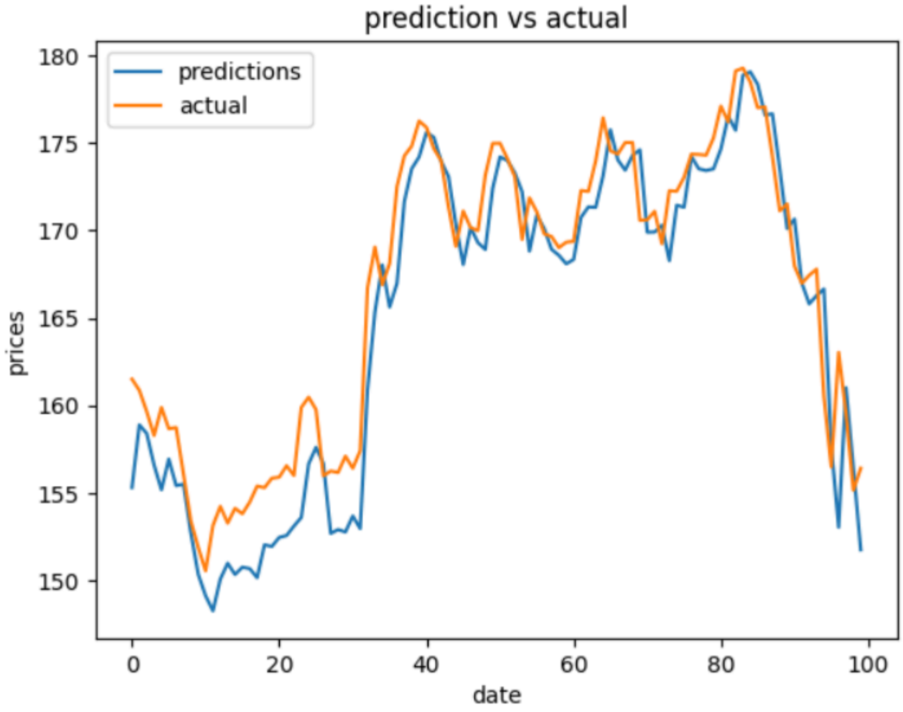
Naive LSTM

We have also built a naive LSTM that takes input the closing price of the last 60 days and outputs the closing price of the next day. The LSTM structure was the same as the one we saw before. In this method, we need a lot more data than the LSTM with sentiment analysis. The LSTM can use data from a single day and predict the next independent of the context of the outside world that has a great impact on the market. The advantage of this model is that it can predict data for several days ahead as opposed to only the next day as it doesn't need news articles and volume that is impossible to predict. To achieve multiple-day predictions we add to the input the predictions. This model was trained on multiple different companies (not only apple) the companies used were google, tesla and netflix and therefore it is not bound to a single company. The reasons for this training was to increase the training data size to avoid overfitting and multi-company predictions.

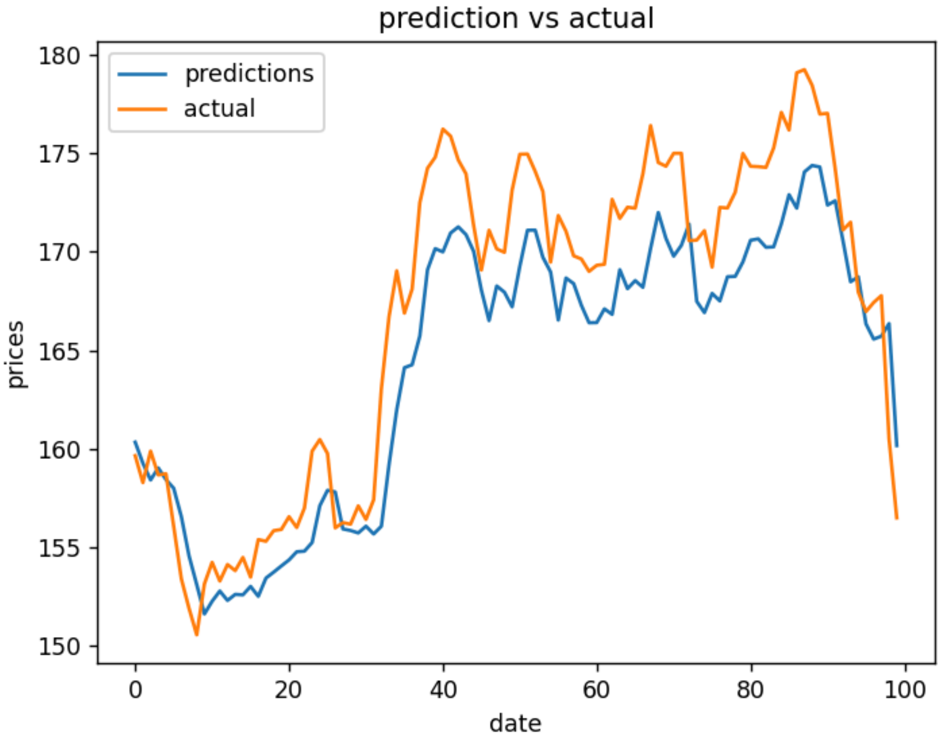
Naive LSTM VS LSTM with sentiment analysis

After training both models we were able to compare them on the same sample. The LSTM with sentiment analysis predicted the values better than the naive LSTM as the following graphs demonstrate:

LSTM with sentiment analysis:



naive LSTM:



This can also be seen after a small simulation when "investing" 10000$ over 100 days. This is done by predicting the next day and "buying" the stock only if the model thinks the stock will go up ("selling" it at the end of the day and repeating it every day").
The naive LSTM after 100 days gained -90$ The LSTM with sentiment analysis gained 850$
If the LSTM would predict accurately every day it would have gained 5700$

Conclusion

In conclusion, we can deduce that the analysis of the news is a strong help in predicting the stock price but maybe we could add other elements to help the model get closer to the maximum gain. Eventhough usage of sentiment analysis of the news articles didnt show an explicit boost on model’s result, we believe that this still could be an important parameter that will be used in stock market analysis.