Utilizing Big Data in the Financial Industry: A Study on Forecasting Stock Values and Analyzing Market Data

Siddharth Nilakhe
snilakhe@stevens.edu
BIA 678: Big Data Technologies
Prof. Mohammad Nikouei, PhD

Introduction

A recent survey by Gallup tells us that about 150 million people in the United States have invested in the stock market, accounting for 58% of adults in America. The average return from the stock market has been 10% per year for nearly the last century, as measured by the S&P 500 index. Every young adult in today's world is given the same advice to start investing at an early age. But identifying stocks that will provide you with a good return is not easy. Each and every investor must be able to forecast stock prices in order to make wise buy or sell decisions. The major issue for investors is making market predictions. Market predictions can be made using information from sources like news and social media. In this paper, we will research how big data helps in forecasting stock values and also find out what role big data play in the financial industry at large.

Recently, the availability of big data has piqued the curiosity of academics studying finance. These datasets have garnered a lot of interest because they have the potential to add fresh data to the field of investment management. While massive datasets certainly have a lot of potentials, there are several limitations to their application that should be made clear if we are to significantly advance our understanding of equity markets and hone in on trustworthy signals that can increase the effectiveness of money managers. Unlike earlier systems like relational databases, big data comprises vast amounts of structured and unstructured information that can be handled by computer algorithms. We will find out ways in which investment institutions are using big data to identify patterns and trends in stock values, and what methods are they using to analyze this cosmic amount of data.

Big data analytics assist in managing massive volumes of data produced by various businesses and can be used to forecast trends. We will also try and discover how big data technologies are used for back-office data management and analytics,

front-office algorithm trading, and analyzing high-volume structured data and social news to reveal hidden market opportunities. Due to the advanced big data technologies used in larger financial institutions, we will also try and find out whether there are any repercussions that the smaller financial institutions are facing due to this.

Finance was a small-data discipline 20 years ago. The main cause is a lack of data. Just Open, High, Low, Close (OHLC) four prices per instrument per day were offered by the majority of exchanges. Even for the largest market markers, further intraday data was not preserved past what was necessary by the regulations. For instance, until 13 years ago, commodity trading floors only retained intraday data for the previous 21 days. Nowadays, the influx of data has drastically altered the financial sector, impacting not only risk management and portfolio analysis but also retail banking and credit scoring. Capital firms have been looking for ways to make Big Data more manageable and to condense massive amounts of information into usable insights in order to maintain their competitive edges in the market. This is due to the ever-increasing volume, velocity, and variety (3V's) of financial data. This transformation towards Big Data has brought a lot of attention to financial institutions. We will try and find out how Machine Learning algorithms and Deep Learning algorithms approaches can be used to predict market trends.

Related Work

As investors want to estimate market values correctly, forecasting is difficult, and numerous models have been put forth. Each researcher makes an effort to forecast the market correctly. Many methods involving both organized and unstructured data are used to predict the stock market's future. Some argue that investors consider historical data and customer views to forecast stock values. While I found out that some even use Twitter’s data to predict share price movements, there should be necessary data pre-processing steps taken before prediction as the data in Twitter will be unstructured. Steps such as removing stop-words should be used before applying the model to predict the share price.

I found a report with 50% accuracy in predicting stock price movements using Twitter data. The report mentioned that the accuracy of results was affected as the stock market was closed during the weekend. Several other research papers suggested that using news and other external factors alongside the chart data will help us attain good accuracy. Chart data here is the historical data of that particular stock. Several studies use various deep-learning models such as ANN, RNN, and LSTM RNN. Many studies have been made on machine learning models, including models such as SVM, decision trees, linear regression, and many more.

Another famous study Tetlock is a well-known analysis of how big data is used in

Finance. This study examines whether the Wall Street Journal's well-known "Abreast of the Market" column's textual content accurately forecasts stock returns. Another example of big data analysis considers product quality ratings by Amazon customers to predict stock prices. Above, you can clearly see how many different approaches people have brought to use big data in predicting stock prices.

This paper will see how different models are applied to historical data, Twitter, and news related to different companies. We will also look at different pre-processing steps that are performed before the data is passed on to the model for stock price predictions.

The approach of Big Data toward Stock Price Predictions

Dataset

Initially, we must have accurate data to feed into our models. We will require two types of datasets - which are historical datasets and the other one is data which is collected from news, blogs, and Twitter. We can get the historical data from Yahoo Finance or Tiingo using API. I have used Tiingo to get data, which is quite simple. Please take a look at the figure below to have an idea about how the data looks when we import it.

	symbol	date	close	high	low	open	volume	adjClose	adjHigh	adjLow	adjOpen	adjVolume	divCash	splitFactor
0	TSLA	2018-01-29 00:00:00+00:00	349.53	350.85	338.28	339.85	4722025	23.302000	23.390000	22.552000	22.656667	70830375	0.0	1.0
1	TSLA	2018-01-30 00:00:00+00:00	345.82	348.27	342.17	345.14	4696465	23.054667	23.218000	22.811333	23.009333	70446975	0.0	1.0
2	TSLA	2018-01-31 00:00:00+00:00	354.31	356.19	345.19	347.51	6076998	23.620667	23.746000	23.012667	23.167333	91154970	0.0	1.0
3	TSLA	2018-02-01 00:00:00+00:00	349.25	359.66	348.63	351.00	4163194	23.283333	23.977333	23.242000	23.400000	62447910	0.0	1.0
4	TSLA	2018-02-02 00:00:00+00:00	343.75	351.95	340.51	348.44	3674909	22.916667	23.463333	22.700667	23.229333	55123635	0.0	1.0

Figure 1 - Historical structured dataset sample

The above sample shows the historical data for the TESLA stock.The data for sentiment analysis can be collected from various platforms such as Twitter, News websites, Kaggle, and other sources.

Data Preprocessing Steps

There may be some missing values in the historical data. To overcome this issue we can either remove the data which has missing/ null values or fill them in using

snilakhe@stevens.edu

various mathematical techniques. Removing data can result in a drop in our accuracy; hence we can perform mathematical techniques such as taking the average of stock prices beside the missing one. For example

X = Previous day + Next Day / 2

We can perform various operations like this to eliminate our missing values problem. We use a few text-processing techniques on unstructured datasets, including news and social media. Unstructured datasets cannot be used in machine learning models, therefore spam is removed using text processing and tokenization techniques. To process datasets, Spark and Python were employed in Databricks. A Spark context can be used to read a dataset. Resilient distributed datasets (RDDs), a Spark structure that enables clustering of data, are created from datasets. Additionally, we eliminate unnecessary columns and convert datasets with non-numeric values to numeric values.

Please have a look at the below figure to check what are approach will be toward sentiment analysis for forecasting the stock market.

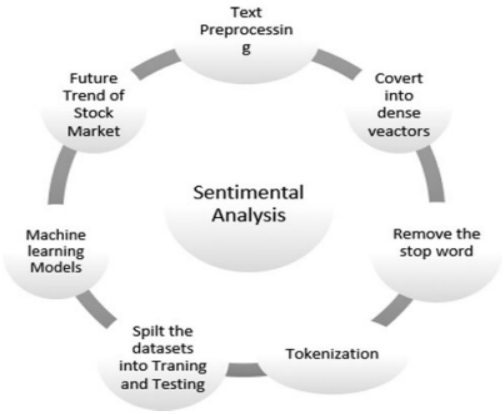


Figure 2 - Sentimental Analysis Model for Forecasting the Market

We eliminate extraneous colons and stop words that could skew the model's predictions. Using the NLP toolbox, each word from the dictionary can be compared and any matches can be eliminated. The analysis used the term frequency (TF) and inverse document frequency (IDF) methods. It is a feature function that shows the relationship between words and determines their weight. We can divide the information frame into training and testing after using all of the functions in the data frame. Using Python and Spark MLlib, the machine learning library is applied. The machine model is then described.

Machine Learning Techniques

Linear Regression

LR is used to establish if there is a relationship between two variables. More specifically, used to establish if there is a statistically significant relationship between the two. LR comes in very handy when we have to forecast new values based on historical data.

In LR, a similar idea can be applied to determine the precise value for Spark. The supervised machine is necessary for the LR model. It projects what the stock price will be. The model sets values as targets based on independent or dependent changing values. The LR model makes predictions about prices based on independent values. This model can be used to predict future stock price values using datasets from various companies.

Please take a look at the figure below, which tells us the forecasting results that we get after using LR to forecast the AAPL (Apple Inc.) stock.

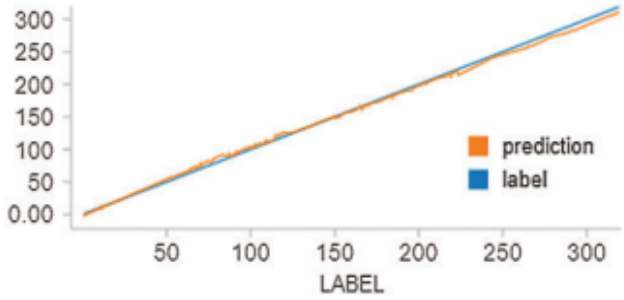


Figure 3: Forecasting Results

Decision Tree Algorithm

A decision tree is a non-parametric supervised learning approach that can be used for both regression and classification applications. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes. Investors may comprehend the decision-making process clearly with the help of decision trees by viewing how the algorithm arrived at a specific prediction. This makes it simpler for investors to decide wisely based on the forecasts of the algorithm. Decision trees are appropriate for a variety of stock market data types since they can handle both numerical and categorical data. They are more resistant to problems with data quality since they can handle missing data.

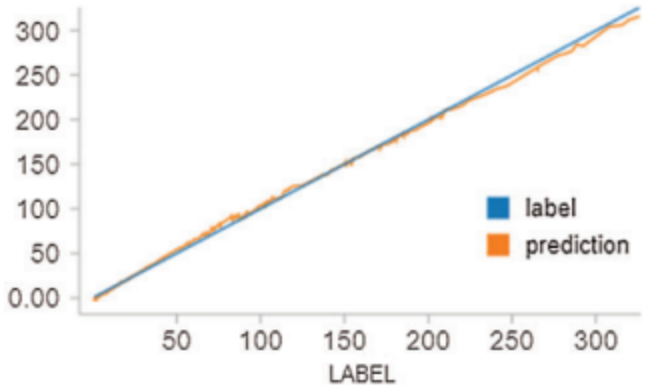


Figure 4: Forecasting Results using Decision Tree

Overall, the decision tree algorithm is a powerful tool for predicting stock prices, providing investors with a clear and interpretable understanding of the decision-making process while also being flexible and scalable enough to handle large amounts of data.

For Sentimental Analysis, we can use Naive Bayes and Logistic Regression. Before using this model, we do text classification and tokenization. Using this approach, we examine text data taken from the Yahoo! Finance database to forecast changes in stock values based on investor feedback. Compared to the results of the logistic-R model, the results of the NB model are more accurate.

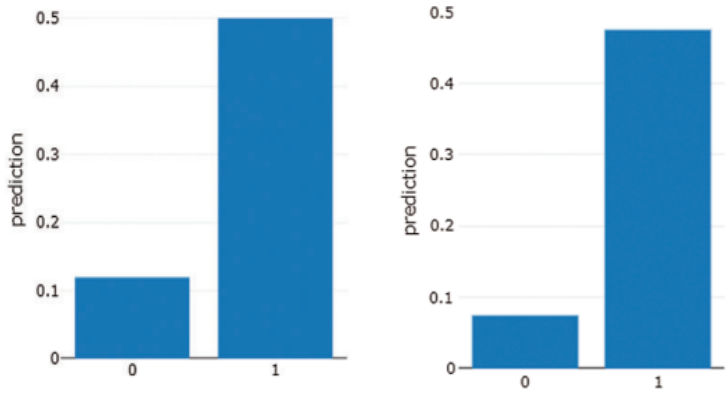


Figure 5: Results from Naive Bayes and Logistic Regression Model

Performance Evaluation of Machine Learning Models

I could find various comparisons of the models we discussed in the above sections. Finding the best model that gives us the desired accuracy is very important. Here are some figures that tell us more about the performance comparison.

The below picture shows us the price predicted by each of the different machine learning models.

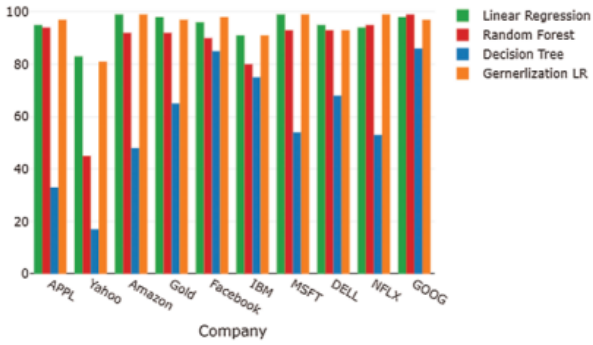


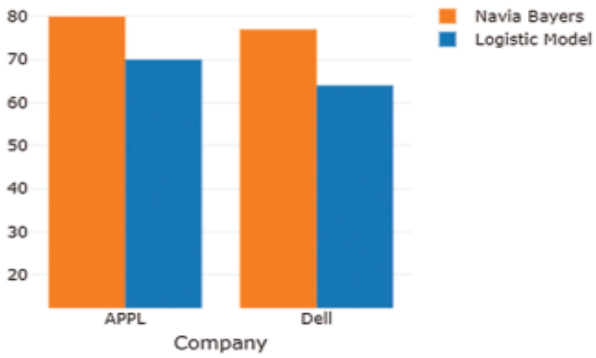
Figure 6: Price Predicted for 10 Companies

Please refer to the below image for the performance evaluation of each model by using the R2 and Root Mean Squared Error - RMSE metric. This will help us to identify which model has the highest accuracy.

Model	R2	RMSE	Accuracy (%)
Linear Regression	0.098	3.143	95
Decision Tree	0.63	5.224	37
Random Forest	0.995	6.677	89
Generalized Linear Regression	0.998	3.114	**97**

Figure 7: Performance Comparison

Now, let us have a look at the sentimental model results.



snilakhe@stevens.edu

Model	APPL (%)	DELL (%)
Naive Bayes	80	79
Logistic Regression	70	77

Figure 8: Sentimental Model Results

After researching whether we should use sentimental analysis alongside historical data, I found a figure that clearly explains to us that in some cases sentimental analysis does help us get higher accuracy. In conclusion, sentiment analysis is a crucial tool for forecasting stock prices since it may shed light on how news and social media affect stock prices, detect new trends, and increase the precision of forecasting models.
Here is a figure which depicts the same.

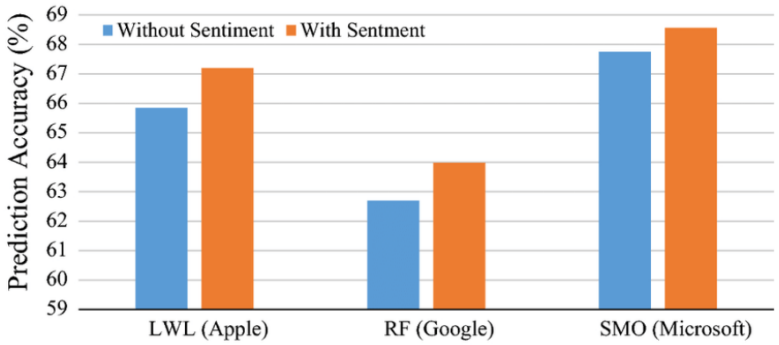


Figure 9: With Sentiment vs Without Sentiment

Deep Learning Approach

After going through various research papers, I found out that most of them suggested the deep learning approach in future work and suggestions. So I decided to try out the deep learning approach myself and see the outcome.
My goal was to learn more about stock market prediction and forecasting, with a focus on one company's stock price and an attempt to apply stacked LSTM. It's crucial not to apply this paradigm to personal investing, though.
Here, I used the Tiingo tool to collect all the valuable data for the desired stock. Use "import pandas datareader as pdr" to import the required tools. By the "get data tiingo" function we can get our data for the stock by the API key.
The next step is to apply min-max scaling, which transforms numbers between 0 and 1. You may accomplish this by using the commands "import numpy" and "from sklearn.preprocessing import MinMaxScaler".

Following the completion of these stages, the data frame will be transformed into an array of values that can be utilized for forecasting and prediction. Now, as we know Time series data is always dependent on previous data. Here, I will keep my training size 65% of the data. The rest will be the test size. In this way, we will separate our training data and testing data. The first 65% of my data will be the training data. In forecasting, we will have to consider timesteps i.e to predict a stock price on a particular day how many previous values(days) should be considered. In our approach here, I have considered timesteps as 100. As we will predict a series of values, the output from our first prediction will be included in our input in the 100th place for the second prediction. In this way, we keep going till our last prediction. Now we create our stacked LSTM model. Input and output dimensions, as well as the number of layers and neurons per layer, can all be changed when configuring stacked LSTMs. Its adaptability enables the model architecture to be tailored to the particular needs of stock price prediction jobs. Please have a look at the summary of our model which we created below.

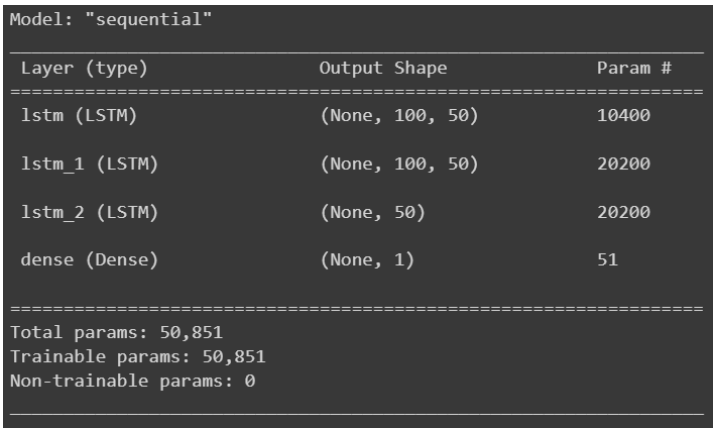


Figure 10: Stacked LSTM Model Summary

Now we fit our model and keep the epochs as 100 and batch size as 64. Let us have a look at the output plot which we got from our model. In the figure, the blue line is the original stock price line that we have considered, the orange line is our training data and the green line is one that we have predicted.
A helpful visualization tool that enables us to contrast the actual stock price data with the model's predictions is the plot we were able to acquire from our model. We can assess how effectively our model has picked up on patterns and trends in the data by comparing the orange line (training data) with the blue line (original stock price data). The stock values that were anticipated based on the training of our model are shown by the green line. The degree to which our model has generalized to new data can be determined by contrasting the green line with the blue and orange lines. The green line

should ideally closely resemble the blue line, showing that our model has correctly identified the patterns and trends in the data.

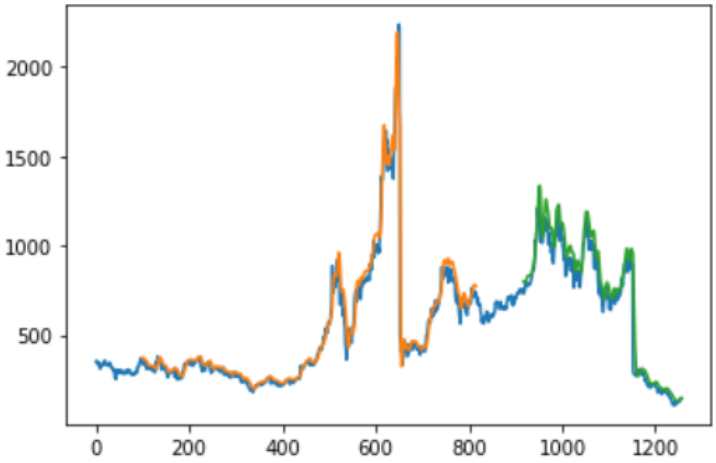


Figure 11: Model prediction output

Overall, the output plot produced by our model allows us to visually assess the precision of our predictions and pinpoint any potential areas for model improvement.

By keeping track of data from earlier time steps and using it to enhance predictions for the future, stacked LSTMs are beneficial for understanding long-term dependencies in stock price time series data. In addition to predicting stock prices, one can also predict trends for a given time frame, such as the upcoming 30 days. The stock price during the last 100 days is represented by the blue line on the output plot from our model, and the anticipated trend for the following 30 days is shown by the orange line. The output plot makes it possible to see the expected trend, which gives investors important information when making stock-related decisions.

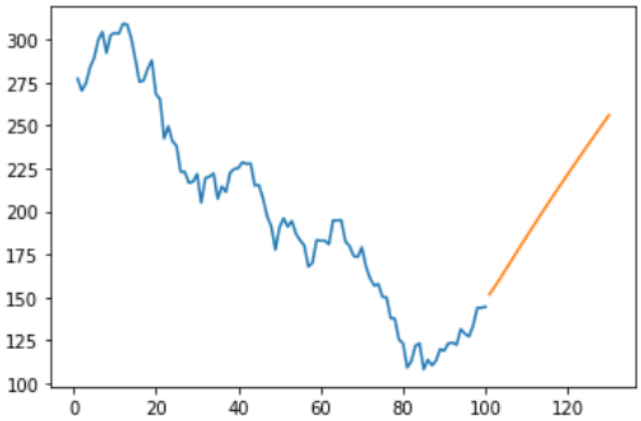


Figure 12: Next 30 days prediction

Conclusion:

To forecast stock price changes using the Spark big data platform, we saw how a number of machine learning models perform. We were able to understand how the use of big data plays a drastic role in the prediction and forecasting of stock prices. We also compared various machine learning models. We also went through sentiment analysis for the prediction of the market. We saw whether combining sentiment analysis along with chart-based analysis makes a difference. After looking at various suggestions to use LSTM for predicting the market, we were able to deploy our own LSTM model and see how the approach works deeply. We were also able to achieve good accuracy after using the deep learning approach. From our study, we could identify and learn the approach of big data towards the finance industry and the stock market in particular.

References:

[1] Juliane Begenau, Maryam Farboodi, Laura Veldkamp, *Big data in finance and the growth of large firms*
[2] Mazhar Javed Awan, Mohd Shafry Mohd Rahim, Haitham Nobanee, Ashna Munawar, Awais Yasin and Azlan Mohd Zain, *Social Media, and Stock Market Prediction: A Big Data Approach*
[3] Bin Fang and Peng Zhang, Chapter 11: *Big Data in Finance*
[4] *Hansol Lee, Eunkyung Kweon, Minkyun Kim, Sangmi Chai, Does Implementation of Big Data Analytics Improve Firms' Market Value? Investors' Reaction in Stock Market*

snilakhe@stevens.edu