Stock Market Prediction

Mihir Gajjar, Veekesh Dhununjoy, Tommy Betz, Gaurav Prachchhak
Simon Fraser University

SFU SIMON FRASER UNIVERSITY

Motivation

- To help a novice investor make better informed decisions for short-term investments.
- To explore the effects of news and twitter on future stock prices.
- To determine how accurately one can predict the future stock prices with incorporating news and twitter data along with the traditional stock data.
- To get a better understanding of stock market domain for personal and professional point of view.

Introduction

We used multiple datasets (social media, news and financial data) to explore how each dataset contribute to the overall closing price prediction for a particular company.
To achieve that:

- We preprocessed our data to retain the useful information needed.
- Features like Volume Weighted Average Price (VWAP) and sentiments were generated from the preprocessed information.
- The extracted features were given as an input to various Machine Learning Models (LSTM, BLSTM) with different configurations to get the best possible model.
- We predict the closing stock price for the next day using the features for the past days using the trained model.

Model Architecture / Data Science Pipeline

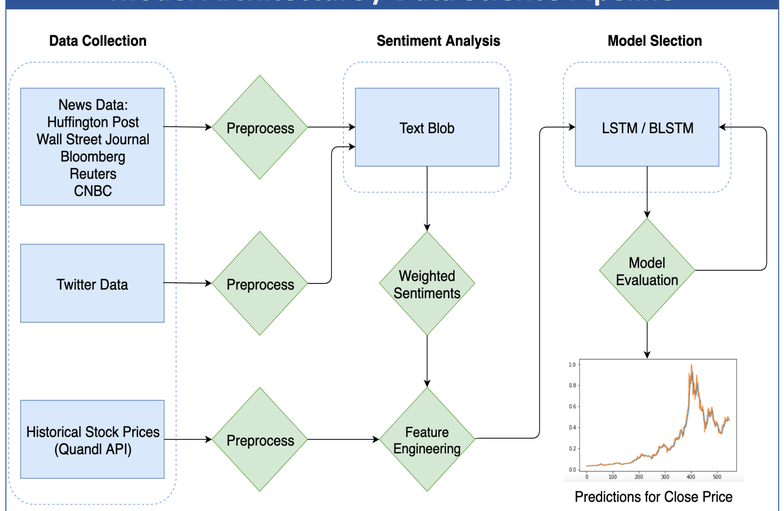

Figure 1. Data Science Pipeline

Results

We tried several approaches to get the best possible close price predictions using different:

- Data Sources
- Tuning Parameters

The results are as follows:

- We got the best result when we used a combination of Stock and News Data.
- Twitter Data did not provide better stock predictions.
- Including Twitter data with Stock and News data combined had a negative effect on RMSE.
- Learning Rate of 0.002 was found to be most effective.
- A lookback of 15 was found to be the best given our data sources.

Table 1. Comparison of Source and Parameter Tuning results

Case	Stock Data	Twitter Data	News Data	Look Back	Learning Rate	Epochs	Min RMSE
1	✓			12	0.002	100	**5.64**
2	✓	✓		15	0.002	100	**5.78**
3	✓		✓	15	0.002	100	**4.72**
4	✓	✓	✓	12	0.002	100	**6.25**

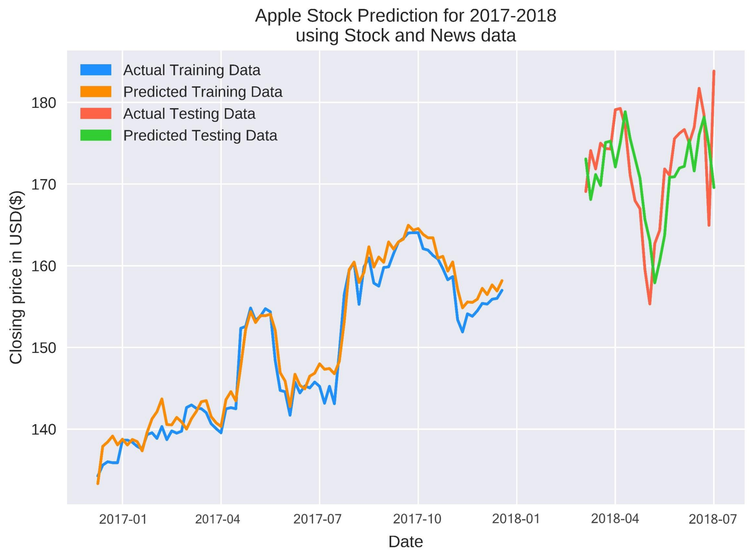

Figure 2. Apple Close Price Prediction for 2017-2018 Using Stock and News data

Datasets and Tools/Technologies Used

Datasets Used:

- Internet Archive Twitter Stream Data.
- Historical Stock Prices Data (Quandl API).
- Financial News Data: Huffington Post, Wall Street Journal, Bloomberg, Reuters and CNBC.

Tools/Technologies Used:
Google Cloud Dataproc (Cloud-native Apache Hadoop and Apache Spark), Apache Spark, Pandas, NumPy, TextBlob (Simplified Text Processing), Keras, Google Colaboratory (Colab), Tableau

Learnings

- Processing massive datasets (3 TB) in a distributed manner in Google Dataproc cluster.
- Extracting features and aggregating large datasets using various data wrangling techniques.
- Creating features that are important for predictions for a model in the stock market domain.
- Performing feature engineering to increase the value of a feature by combining multiple features.
- NLP techniques to extract features like sentiments from text data.
- Creating LSTM and BLSTM models over time series data for future price prediction using Keras.
- How to decide which feature has the most impact on the prediction.
- With the help Instructors and TAs we were able to learn more about the stock market domain and to know how data of this domain is dependent on various features like VWAP and sentiments.

Conclusion

- We achieved the best performance using a combination of Stock and News datasets.
- We were able to predict closing prices closely to the actual values by using weighted average sentiment scores rather than just a simple average sentiment.
- Due to the vast amount of tweets not being directly related to the performance of a company, it is difficult to use these data for stock predictions.
- Relying on time series data to make predictions, it is essential to have large and consistent datasets available.

Future Work

- Gather more News Data relevant to the specific companies.
- Understand more features specific to the stock market field from financial experts and incorporate them while training our model.
- Use different statistical techniques to generate additional useful features.
- Perform more advanced feature engineering and experiment with the different features.
- Use more enhanced sentiment analysis techniques which can identify parts of different articles that are directly related to the concerning company to get more accurate sentiment analysis.
- Predict the stock prices for other major corporations on the new York stock exchange (NYSE).
- Experiment with different ML models with different architectures and features.
- Develop an application that would help anyone to invest in the stock market with more customized and interactive UI.