Stock Index Prediction using Sentiment Analysis of Financial Tweets

Abhik S. Basu
abhik20165@iiitd.ac.in
Ishaan Marwah
ishaan20068@iiitd.ac.in
Sohum Sikdar
sohum20339@iiitd.ac.in

Abstract

As young engineers about to go into the workforce with a majority of our compensation being in stocks of established companies, we want to build a model that when fed the business news of a day it predicts the market sentiment and finally gives us a binary result of whether the index of a market (BSE/NSE for example) will have an uptick or a downtick. The link for the GitHub repository of the project is : ML-Endterm-Project

1. Introduction

The problem statement is to predict whether a particular index of market goes up or down at the end of the day by analyzing the data of business related news and commentary. This means that we have been given a particular set of stocks along with their opening and closing prices for a particular day and some number of comments. Given the comments, we have to predict correctly using the comments whether the price of the stock rises for the particular day or goes down. Rising of a stock on a particular day means that the closing price is greater than the opening price and going down of a stock means than opening price is greater than the closing price.

Input (Dataset): Set of n possible instances of data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
where $x_i = (open_i,\ close_i,\ high_i,\ low_i,\ volume_i,\ comments_i,\ average\_price_i,\ sentiment_i,\ sentiment_{i-1},\ diffprice)$ and $open_i$ is the opening price of the stock for the i^{th} day, $close_i$ is the closing price of the stock for the i^{th} day, $high_i$ is the highest price of the stock for the i^{th} day, low_i is the lowest price of the stock for the i^{th} day, $volume_i$ is the volume on a particular day, $comment_i$ is a set of comments related to the market of that stock on that particular day, $average\_price_i$ is the average price of the stock on that day, $sentiment_i$ is the sentiment of that day, $sentiment_{i-1}$ is the sentiment of the previous day, $diffprice$ is the difference in the average prices of the day and the previous day and y_i =actual sentiment of stock index.

Set of hypothesis functions: Let X be the set of all x_i and Y be the set of all y_i. Then, we have a set of functions from X to Y say $H = \{h | h : X \to Y\}$.

Loss function: $L(f(x_i), y_i) = (f(x_i) - y_i)^2$. This function basically tells us the error in the predicted value and the actual value.

Output: return back $h^* = argmin_{f(x) \in H} \sum_{i=1}^{n} L(f(x_i), y_i)$.

So, our basic goal is to minimize the loss function.

2. Literature Survey

- **Twitter mood predicts the stock market :** They found a positive correlation between the sentiment of the tweets and DJIA values, and how simple text processing techniques can be used to track the sentiment of tweets, it also said how socio-cultural events also cause a lot of mood shifts which may or may not be predicted correctly by a machine learning model.
- **Stock Prediction Using Twitter Sentiment Analysis :** They built upon the paper cited above, to neural networks and concluded that they perform quite well in terms of predicting DJIA values when trained on a feature set of DJIA values.
- **Twitter Sentiment Analysis and Influence on Stock Performance Using Transfer Entropy and EGARCH Methods :** We use the idea to determine the rate of change of the stock by subtracting the highest values of two consecutive days.

- **Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts :** This paper lays down a framework for the sentiment of a financial/business related text from the point of view of a retail investor. This framework can be used for benchmarking and training our model. A dataset also exists which has implemented this framework and used it for sentimental analysis of financial news (https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news)

- **A Survey on Stock Market Prediction Using Machine Learning Techniques:** This study analyses different techniques for stock market prediction and accordingly tells the drawbacks and advantages of them. This can be used for helping us in making smarter decisions while creating our own stock market analysis model.

- **Twitter Sentiment Analysis Based on Ordinal Regression :** This study aims to perform a detailed sentiment analysis of tweets based on ordinal regression using machine learning techniques. The proposed approach consists of first pre-processing tweets and using a feature extraction method that creates an efficient feature. Then, under several classes, these features scoring and balancing

- **A web scraping framework for stock price modelling using deep learning methods:** This resource helped us in understanding how we scrape data from websites which was very beneficial for our progress till now.

3. Dataset

For generation of data, we did the following:

Tweets on Company: The original dataset has tweets from all other major tech companies from which we only focus on Apple in this study. This dataset contains tweets, retweets, epoch timestamps, comments, likes and the name of the user who has made the tweet. (source: Kaggle)

Historical Stock Data: Contains the open, close, high, low, volume traded for the day.

Final Dataset: for each day, contains the twitter sentiment (Discussed Later), twitter sentiment for previous day, open of current day, close of previous day, high of previous day, low of previous day, and volume of previous day, and we try to predict the close of current day.

We did the data preprocessing to clean the data in order to use it for analysis. We did the following steps for preprocessing the company tweet data:

1. Conversion from epoch timestamps to human readable date formats.

2. Converting the format of date from the format obtained in dataset to the format which will be used by us (YYYY-MM-DD).

3. Creation of an engagement parameter which is sum of likes,comments and retweets.

4. Taking top 20 tweets based on engagement.

5. Cleaning of tweet text to allow sentiment analysis and then calculate polarity of tweets.

6. Clustering by dates from 2015 to 2016 and taking weighted average of polarity and engagement.

The final dataset that we used can be found on the following link:
dataset_project

4. Methodology

We did the following steps for our analysis:

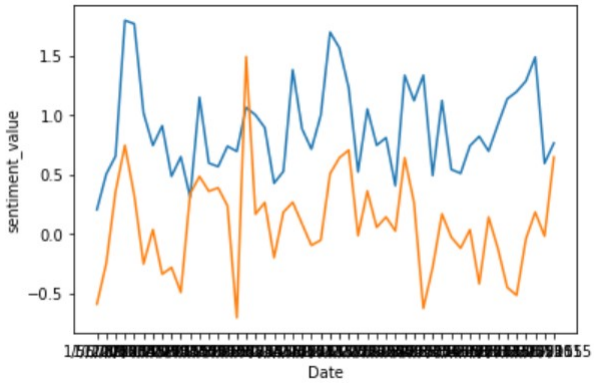
1. Scraped the tweets of Apple from various sources and cleaned them to get the final data set. While cleaning we took care of the NaN values which could have created a problem if we used the data along with them

2. We ran a sentiment analysis on all the tweets in the dataset to get the polarity of all the tweets that were available with us. Sentiment analysis basically gives a number that quantifies the impact of that tweet on people. If sentiment is a positive value then it means that the sentiment is positive and good otherwise if it is negative then we sa that the sentiment is bad. Here, we Used TextBlob as the model for each tweet in setiment analysis.

3. For predicting nature of growth (up or down) for a day we needed the overall sentiment of the day instead of the sentiment of all the tweets, So, we defined the average of all sentiments of a day is the sentiment of a day. This value correctly represents the expected value of sentiment of a particular day.

4. We plotted the sentiment, change in stock price vs day to check if there is a relation between the two that can be learnt. If there is some amount of correlation between these two values then we can easily use this fact to our advantage. This study becomes pointless in case there is no correlation between the two, however as seen in the plot, where the blue line represents the

change in stock value from previous day, and the orange line plot represents the current sentiment. Can be seen how higher sentiment value increases the stock price. The correlation between the two was coming out to be 0.63 which indicates they are positively correlated. This implies that the sentiment of the tweets obtained clearly influences the stock prizes.



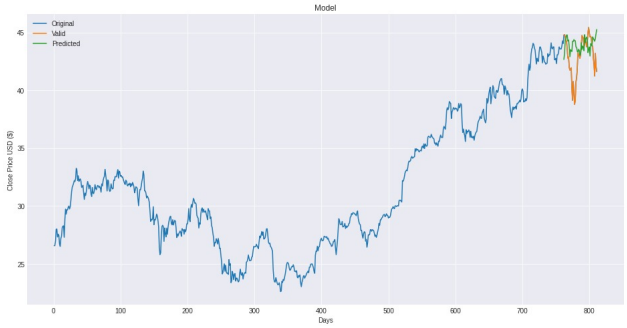
5. We needed to either solve this as a classification problem or a regression problem, use different models, Linear Regression, Decision Trees, Random Forests, Support Vector Machines, and finally LSTM Neural Networks.

5. Results and Analysis

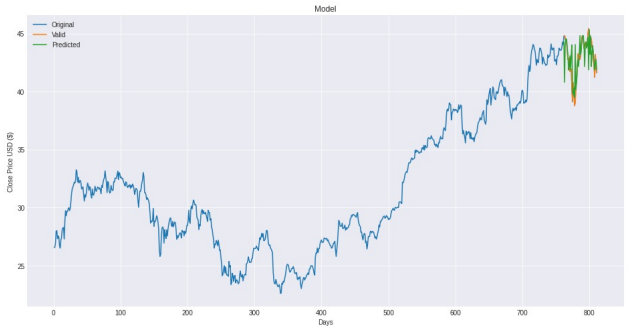
We observe the results as follows:

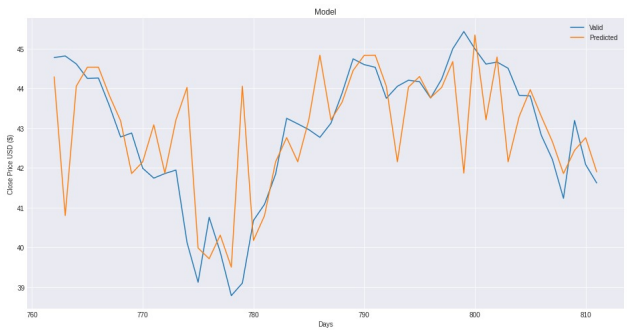
1. **For Support Vector Machine(SVM):** We used SVM to classify the data and we obtained an accuracy of 0.4723926380368098 which is not a good enough accuracy. The plot was as follows:

2. **For Linear Regression:** We run linear regression to predict the value of the stock for next fifty days. The graph was as follows and we can see that the graph for the predicted value is very different from the graph for actual value. The root mean squared error in this case is 2.1414 , so we conclude that this is not a good model and we have to do better than this.

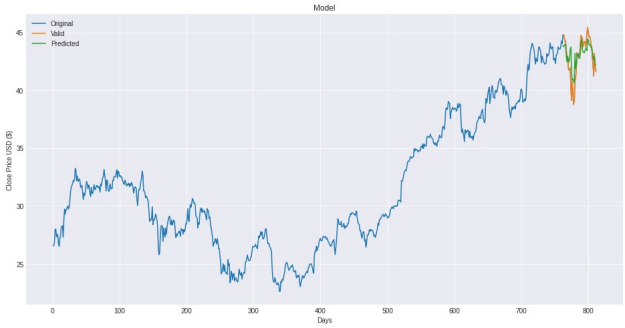


3. **For Decision trees:** Then, we used Decision Trees for the problem and we got an RMSE value of 1.2647. This is better as compared to linear regression on the data as shown by the figures.

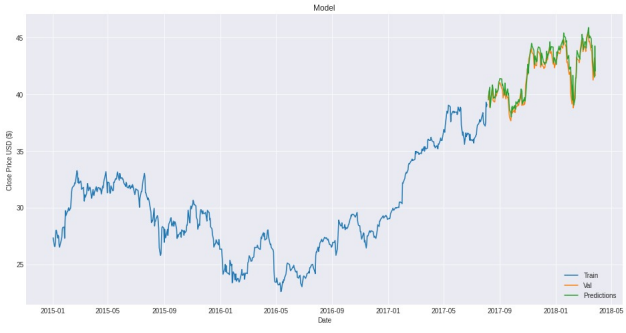




4. **For Random Forests:** Then, we used Random forest for the problem and we got an RMSE value of 1.0505. So, the RMSE value goes down on switching to random forest. This is better as compared to all the models till now on the data as shown by the figures.



5. **For LSTM(Long Short Term Memory):** We finally used the LSTM model(long short term memory), which greatly improved the results as shown in the figures. The RMSE value for this model was 0.3650404639802157 which is very small as compared to all the models till now on the data as shown by the figures.



6. Conclusion

From the models presented, we get two things, Firstly this problem is better scene as a regression problem than a classification problem as seen from the errors. Also the same was discussed in the research papers too. Another interesting thing is that yes there is a correlation between the stock price and the sentiment but it is delayed by a few days, and hence LSTMs work so well giving a RMSE 0.36, which was the best model out of all covered and Random Forests having 71% accuracy. Lastly, stock prediction is a very complex problem, a lot of unrelated events influence stock prices which are very hard to factor in models, even the best models in literature review, if solved using a classification approach, had a 75% accuracy, and thus this study comes very close to them.

References

[1] Arpit Goel Anshul Mittal. Stock prediction using twitter sentiment analysis.

[2] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[3] SHIHAB ELBAGIR and JING YANG. Twitter sentiment analysis based on ordinal regression.

[4] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.

[5] Marta. The saddest day on twitter: Sentiment analysis engagement trends in company's tweets.

[6] Román A. Mendoza-Urdiales, José Antonio Núñez-Mora, Roberto J. Santillán-Salgado, and Humberto Valencia-Herrera. Twitter sentiment analysis and influence on stock performance using transfer entropy and egarch methods. *Entropy*, 24(7), 2022.

[7] Dr Polamuri, Kudipudi Srinivas, and A. Mohan. *A Survey on Stock Market Prediction Using Machine Learning Techniques*, pages 923–931. 05 2020.

[8] vivek rathi. Sentiment analysis of financial tweets.

[9] Mustafa Dogan ÖMER METIN. Tweets about the top companies from 2015 to 2020.

[2] [1] [6] [4] [7] [5] [8] [3] [9]