Predicting fluctuations in stock price using sentiment analysis on the twitter posts.

Author: Shwetha V C(shwva184)

Course code: 732A92

Date: 2022-03-18

ABSTRACT

Factors such as social media posts and news articles influence the supply and demand of corporate stocks. This project aims to find the relationship between the sentiment of general public and the fluctuation in stock trend. It also aims to exploit this relationship to predict future stock fluctuations. Sentiment analysis is performed on the tweets related to Electronic Arts Inc. (EA) stocks. Sentiment score along with the lagged values of the stock variations are used as input to different machine learning models such as Linear model, decision tree model, random forest model, XGBoost model and support vector regressor model to learn the underlying relationship. Most accurate result is obtained with XGBoost model.

1. Introduction

Stocks represent the shares of ownership of corporations. The stock prices go tend to fluctuate based on the supply and demand. Prediction of these fluctuations can be a complex task for laymen who are not experienced in stock trading and related finance. Expert knowledge about multitude of aspects such as market trends, exogeneous events, technical factors etc are not available to everyone.(Galton 1907) utilizes the concept of wisdom of crowds, i.e. the collective knowledge of multiple non-experts to guess the weight of an ox more accurately than the opinion of experts. This project tries to apply this concept to predict the ups and downs of stock price.

Social media is now not just for connecting people, it does more than that. The news around the world has greater reach with social media. It is safe to say that in most households, newspapers and magazines are being replaced by social media platforms like Facebook and twitter. This project attempts to capture and use the sentiment of the crowds by following posts (tweets) of a social media platform twitter, related to a particular stock and to study the fluctuations of stock prices.

2. Related work

(Teti, Dallocchio et al. 2019) performed a study on making use of social media as a tool for investing the relationship between the sentiment of the society and stock prices. The result of this study highlights that sentiment of tweets related to the stock has an effect on price change in the stock. It also implies that it is a good idea for investors to look into the social media content related to the stock for decision making regarding future selling or buying of the stock.

(Bollen, Mao et al. 2011) also embeds the sentiment analysis of the tweets related to a stock to see its effect on the stock price change. The result of this study emphasizes that public mood states or sentiment play a very important role in human decision making and influences stock market prices. Similar studies have been performed in (Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. 2016, October) and (Rao, T., & Srivastava, S. 2012) where sentiment analysis of twitter posts has been shown to be a good feature to predict future stock prices.

3. Dataset:

This project uses two datasets. First is for sentiment analysis of twitter posts, it is a collection of all tweets using hash tags related to Electronic arts twitter dataset-$EA from (followthehashtag). Data is cleaned to retain only the tweet content, date and followers of the account that created the post. This contains the data for dates 2016-03-27 to 2016-06-15.

Tweet Id	Date	Hour	User Name	Tweet content	Followers
714247266752593921	2016-03-28	00:26	Taylor	$EA there's a reason it's stalling right here....	1755.0
714219957093924864	2016-03-27	22:38	Ca$h Ave. Wavy J	RT @TxUndergroundRa: @OriginalVaughn - Ave. $e...	489.0
714218795447877632	2016-03-27	22:33	O.V.	RT @TxUndergroundRa: @OriginalVaughn - Ave. $e...	1024.0
714217112043171840	2016-03-27	22:26	FinSentS NASDAQ	$EA:US Oculus' Virtual Reality Headset To Laun...	2746.0
714211797738381312	2016-03-27	22:05	ProVesting	$EA:\n\nElectronic Arts (EA) Short Interest Di...	737.0

Figure 1. A chunk of data that is used to perform sentiment analysis.

Second is the historical financial data of Electronic Arts Inc. (EA) downloaded from yahoo finance(finance) for the dates 2016-03-27 to 2016-06-15. This data contains opening price, highest, lowest price and adjusted closure. A derived column “change” is added which calculated the difference between adjusted price of the stock and its adjusted price on the previous day. Positive change represents increase in the stock price and negative change represents decrease in the stock price.

Date	Open	High	Low	Adj Close	change
2016-03-29	0.180809	0.252712	0.213282	0.271428	0.090260
2016-03-30	0.325718	0.302489	0.330779	0.296104	0.024676
2016-03-31	0.299608	0.298660	0.293742	0.276624	-0.019481
2016-04-01	0.236292	0.236120	0.257343	0.264286	-0.012337
2016-04-04	0.278068	0.259732	0.273308	0.252597	-0.011689
2016-04-05	0.219321	0.225909	0.252235	0.233766	-0.018832
2016-04-06	0.256527	0.232929	0.265006	0.258441	0.024675

Figure 2. EA stock data downloaded from yahoo finance

4. Experiment:

4.1 Sentiment analysis:

As a first step, sentiment analysis was conducted on the twitter content dataset. Sentiment scores were calculated using vadersentiment package. VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that is specifically designed to study, and the sentiments expressed in social media (Hutto, C.J. & Gilbert, E.E. 2014). As a part of the result, the twitter posts receive a compound sentiment score in the range -1 to 1. Any value between -0.05 and 0.05 are neutral, above 0.05 are assigned a positive score and below -0.05 assigned a negative sentiment score.

Sentiment analysis was done for each twitter post and the compound sentiment score was calculated. This compound sentiment score was multiplied by the number of followers the account has as weights, to consider the impact of the twitter post. As a final step the average of weighted sentiment scores were taken for each day to get a picture of consolidated sentiment of crowd for a day. For further calculations and merging with stock historical data this weighted sentiment scores are scaled from 0 to 1.

Date	Tweet Id	Followers	compound	weighted_sentiment	scaled_weighted_sentiment
2016-03-27	7.142169e+17	1249.000000	0.114700	84.533900	0.208537
2016-03-28	7.144333e+17	3731.804878	0.074698	723.898478	0.232031
2016-03-29	7.147843e+17	4476.458333	0.101644	224.874328	0.213694
2016-03-30	7.151729e+17	1897.916667	0.209617	206.110193	0.213004
2016-03-31	7.155232e+17	4840.074074	0.082563	114.398100	0.209634

Figure 3. Scaled weighted sentiment is constructed as a variable derived from calculated compound sentiment score.

4.2 Analysing sentiment scores and adjusted closure of EA stock.

Variable 'change' in the second dataset (EA stock market data downloaded from yahoo finance) represents the daily change in the 'adjusted closure price' of the stock. Positive 'change' implies the stock price is raising and negative 'change' implies the fall of stock price. In this experiment, it can be observed that raise and fall in the twitter sentiment score resulted in raise and fall in the stock price change in the immediate future.

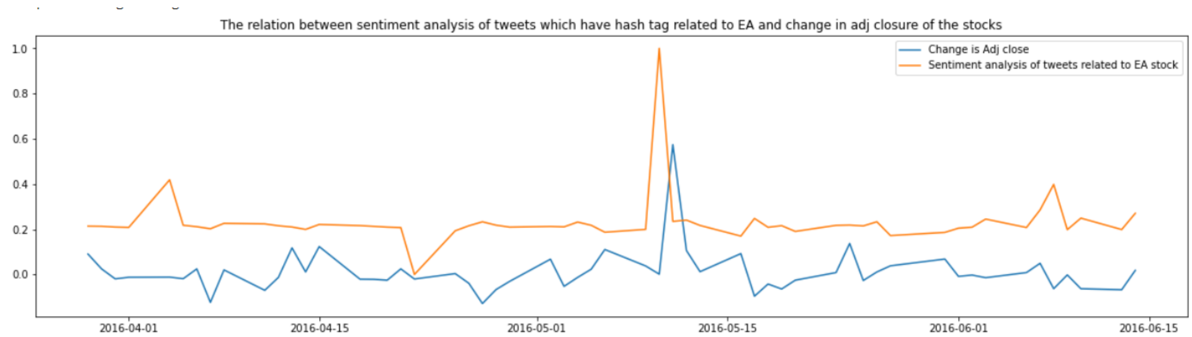

Figure 4. The above plot showcases that the major drops and falls in the sentiment score of related twitter posts are followed by similar raises and drops in stock prices in immediate future.

As we can clearly observe the influence of sentiment of public on the stock prices, this project aims to predict the change is EA stock prices by just learning the sentiment trend of the twitter posts and not taking any aid from experts on the market trends.

4.3 Models

Feature set: Weighted sentiment scores normalised to range 0 to 1, date, time related features such as month, week and day of week and change is the stock price observed on the previous day (lag 1 of change) are used as feature set for all models.

Date	scaled_weighted_sentiment	change_lag 1	month	week	day_of_week
2016-03-30	0.213004	0.090260	3	13	2
2016-03-31	0.209634	0.024676	3	13	3
2016-04-01	0.208101	-0.019481	4	13	4
2016-04-04	0.418465	-0.012337	4	14	0
2016-04-05	0.217599	-0.011689	4	14	1

Figure 5. Feature set.

Target: The change in the adjusted closure stock price.

70% of the data is used for training and 30% of the data is used as test data. As it is a very small dataset, validation data is not used for any hyperparameter search. Linear model, decision tree model, random forest model, XGBoost model and support vector regressor model were applied for the training data. Mean square error was calculated as the loss function. The predictions of this models will be compared and evaluated in the next section.

4.4 Evaluations of models.

Model evaluation has been carried out with two approaches in this experiment. First one is by comparing the MSE scores to evaluate which model performed better. The second approach is to visually analyse which model was able to capture the ups and downs in change in the stock price.

5. Results.

Mean square error was recorded for predicted values of all models. Support vector regressor model provided the least mean square error. Figure 6 shows the MSE scores calculated for all models.

Model	MSE
Linear	0.007264646070439742
Desicion tree	0.007529446099186769
Random forest	0.004589472915330793
SVM	0.002916375655794042 5
XGBoost	0.005918947994361408

Figure 6. MSE score of all model predictions.

But since the main aim of this experiment was to see if the model can predict the raises and falls of stock price, it seems to be more important to visually see if the model was able to predict the raise when the stock price raised and fall when the stock price dropped.

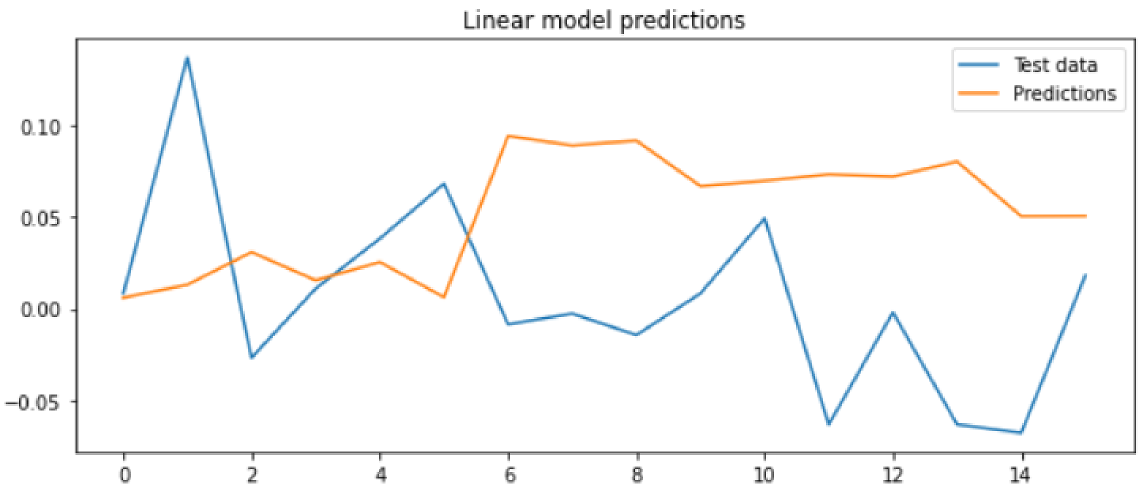

Figure 7. Linear model prediction.

Figure 7 visualizes the predictions from linear model and original test data. Linear model is not very good at learning the change is stock price for the used dataset.

Figure 8. Decision tree prediction.

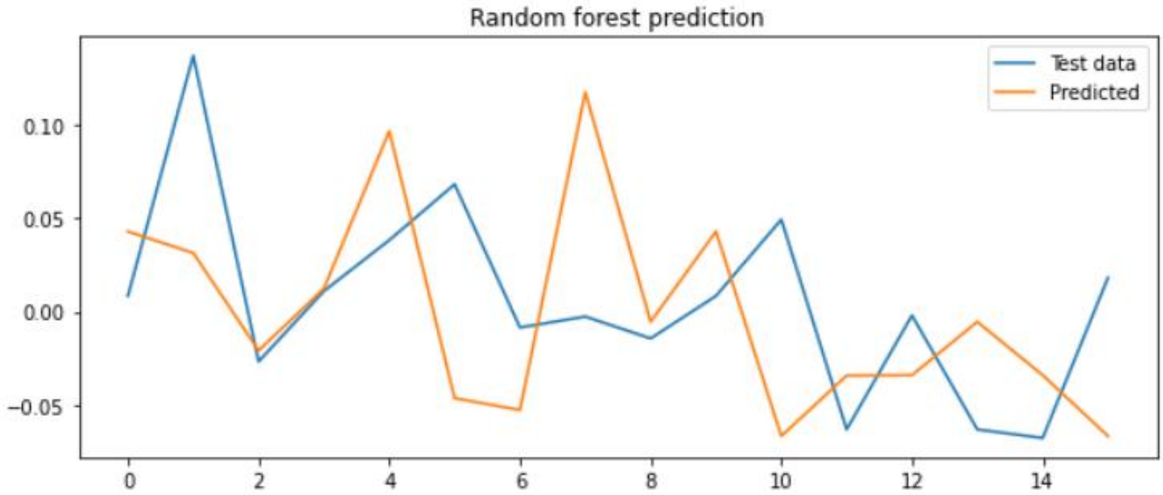

Figure 9. Random forest prediction.

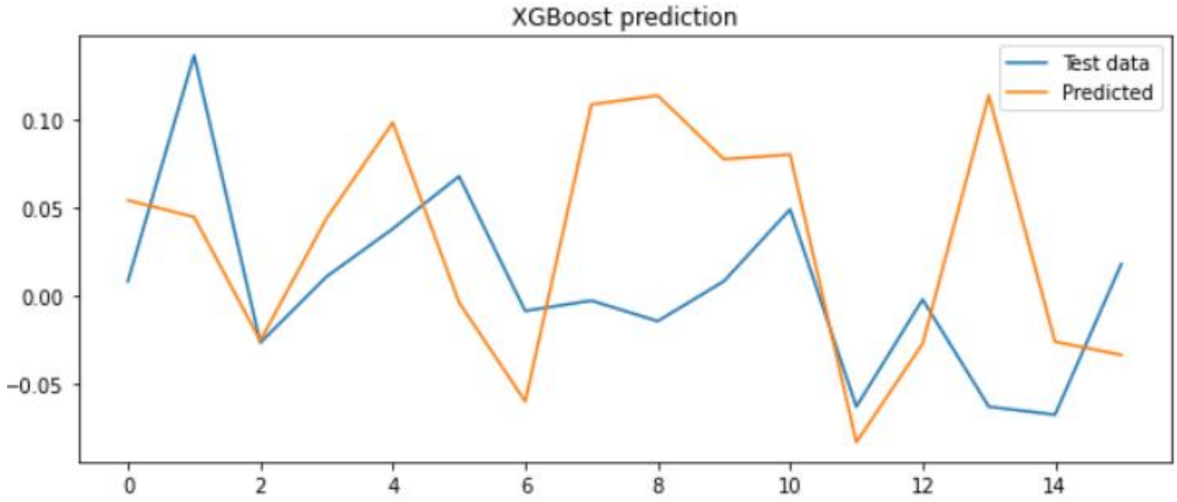

Figure 10. XGBoost prediction.

Figure 8, figure 9 and figure 10 visualizes the predictions from decision tree model, random forest model and XGBoost model respectively. These tree-based models perform better than linear model, it may not be able to capture the exact change, but it still seems to predict if the change is going up or down quite well.

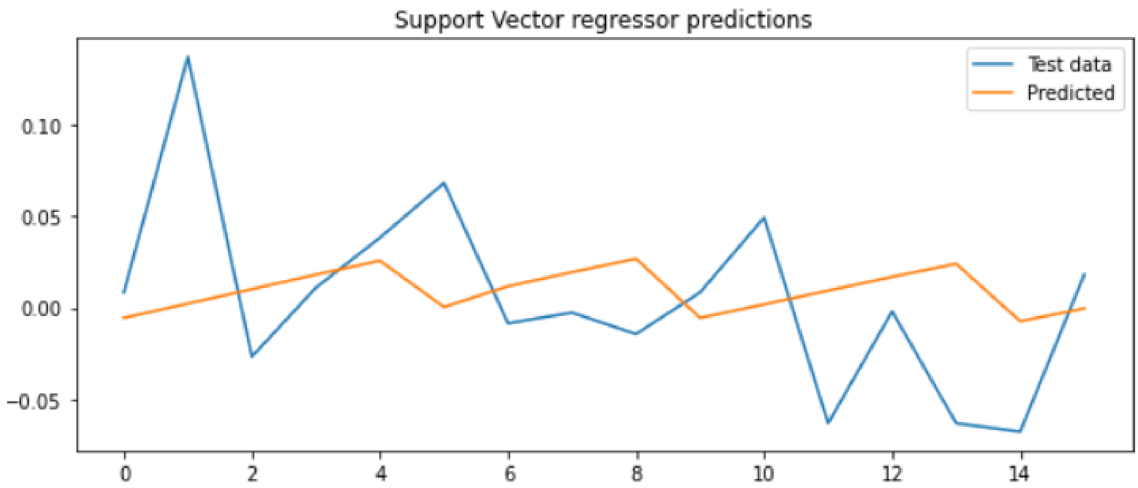

Figure 11. Support vector regressor prediction.

Figure 11 visualizes the predictions from support vector regressor. This model predictions look better than linear model but not as good as the tree-based models.

Even though lowest MSE was observed for the SVM model, it can be inferred from Figure 11 that XGBoost model can predict the ups and downs more accurately than the SVM model.

6. Discussions and conclusion.

The results show that the sentiment scores of twitter posts relevant to the stock clearly influences the immediate fluctuations of corporate stock prices. It would be beneficial for investors to consider the sentiment scores and trends of the stocks in social media for decision making. Adding sentiment score as an additional feature with expert knowledge may lead to better predictions. However even with just the sentiment score from twitter posts related to the stock, the model can predict most of the rises and drops in stock prices.

As further work, combining the sentiment scores from different social media platforms and public news would most likely improve the prediction accuracy. Rises and drops in stock price not always depend only on the sentiment of the public. There are multitude of factors that influence the movement of the stock trend. The sentiment analysis alone is never sufficient to predict the stock trend fluctuations, but addition of this as a feature with other expert knowledge insights will improve the model performance in terms of accuracy. This experiment was performed on a very small dataset of 55 days, larger dataset would probably be able to provide more insights about the relation between sentiment score of twitter posts and stock price fluctuations to the model. Larger dataset would also allow the designer to have validation data to perform hyper parameter tuning, thus making the model more efficient.

7. References

Bollen, J., et al. (2011). "Twitter mood predicts the stock market." Journal of Computational Science **2**(1): 1-8.

finance, Y. "Electronic Arts Inc. (EA),NasdaqGS - NasdaqGS Real Time Price. Currency in USD." from https://finance.yahoo.com/quote/EA/history?p=EA&guccounter=1.

followthehashtag, N. T. from https://data.world/kike/nasdaq-100-tweets?msclkid=ae2d50e6a6b811ec8e24f4e45fd376f5.

Galton, F. (1907). "Vox populi (the wisdom of crowds)." Nature **75**.

Teti, E., et al. (2019). "The relationship between twitter and stock prices. Evidence from the US technology industry." Technological Forecasting and Social Change **149**.

Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using twitter sentiment analysis.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.