

Predicting fluctuations in stock price using sentiment analysis on tweets.

Author: Shwetha V C (shwva184)

Course code: 732A92

Date: 2022-08-25

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 trends. 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 decision tree model, random forest model, XGBoost model and support vector regressor model to learn the underlying relationship and compare it with the baseline of decision tree model without sentiment score as one of its features. All the models had lower error in terms of MSE, MAE and RMSE when compared to baseline model. Most accurate result is obtained with XGBoost model.

1. Introduction

Stocks represent the shares of ownership of corporations. The stock prices tend to fluctuate based on the supply and demand of the stocks. Prediction of these fluctuations can be a complex task for laymen who are not experienced in stock trading and related finance. There are various studies performed in the past to determine the different factors affecting the stock prices, but there is a lack of a generic list of such factors that can be used for all stocks. This is mainly because the factors affecting the stock prices is usually unique for each stock, although there may be some similarities if the stock belongs to same sector or caters to similar customers. Expert knowledge about multitude of aspects such as market trends, exogenous events, technical factors etc relevant to a particular stock is not always 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 by gaining the collective knowledge and emotions of the crowd through their twitter posts.

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. The notion behind using this as one of the factor or features that determines the future price change of a particular stock is that the sentiment of the crowd can be considered to represent the effect of various external events on the emotion of the customers. Sentiment of the crowds can be collected from stock related newspaper or magazine articles, social media posts such as twitter, facebook etc. In this project we will be focusing on stock related twitter posts. This project tries to examine if combining sentiment analysis of the crowd along with the historical stock prices will result in the better prediction of stock price fluctuations when compared to just using the historical data.

2. Related work

Teti, Dallochio 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 et al. (2016) 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. The data is only available for the period starting from 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. Data for this particular period is chosen as the dataset containing the tweet content has the data for only this period and these both datasets are merged for the experiments conducted in this project. This data contains opening price, highest, lowest price and adjusted closure. All these columns have been scaled from 0 to 1 before using for the experiments of this project. 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 also performed by using sentiment analysis functionality of textblob library in python (Text blob, 2020). Sentiment analysis in textblob library is not as equipped as VADER to deal with social media jargons and emojis (Amy, 2022). Hence this required some data pre-processing before using applying the sentiment analysis. For data pre-processing similar techniques as done in (Kolasani, S. and Assaf, R. 2020) was applied here as well, that is urls (starting with 'http', 'https', 'www' etc) were replaced with word 'URL', emojis were replaced with their meaning words, @usernames were replace with word 'User', stop words were eliminated and all the non alpha and nonnumeric characters were replaced with space. After data pre-processing sentiment score was obtained for each and every tweet present in the dataset.

The sentiment score from VADER as well as textblob 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.

	Tweet Id	Followers	compound_vader	compound_textblob	weighted_sentiment_vader	weighted_sentiment_textblob
Date						
2016-03-27	7.142169e+17	1249.000000	0.114700	0.000000	84.533900	0.000000
2016-03-28	7.144333e+17	3731.804878	0.074698	0.029437	723.898478	27.971318
2016-03-29	7.147843e+17	4476.458333	0.101644	0.092332	224.874328	569.584933
2016-03-30	7.151729e+17	1897.916667	0.209617	0.071921	206.110193	105.016754
2016-03-31	7.155232e+17	4840.074074	0.082563	0.014863	114.398100	11.711400

Figure 3. Scaled weighted sentiment (in case of both vader and textblob) is constructed as a variable derived from calculated compound sentiment score.

4.2 Combining sentiment scores and adjusted closure of EA stock.

Sentiment scores obtained using Vader and textblob are merged with the second dataset (EA stock market data downloaded from yahoo finance) on the date column. 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.

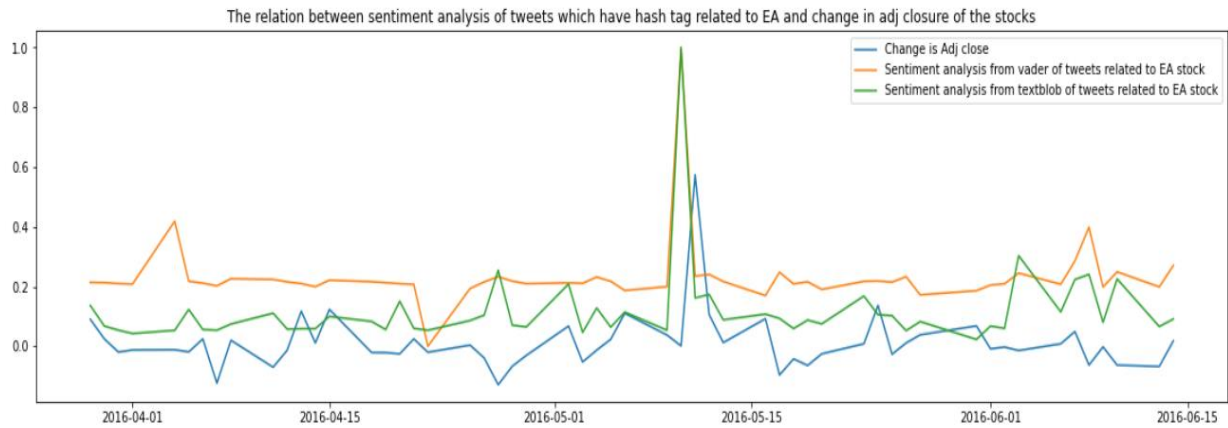


Figure 4. The above plot showcases that the major drops and falls in the sentiment score of related twitter posts are followed by similar raises and drops in stock prices in immediate future.

As we can clearly observe the influence of sentiment of public on the stock prices, this project aims to predict the change in EA stock prices by including the sentiment trend of the twitter posts as an additional feature along with the historical stock prices and compare it with the baseline of predicting of stock price change obtained by using just the historical data.

4.3 Models

Feature set: Date, time related features such as month, week, and day of week in addition to change and adjusted closure price of the stock price observed on the previous 3 days (lag 1, lag 2 and lag 3 of change and adjusted closure) are used as feature set for the baseline model. The features used for baseline model is shown in figure 5. For rest of the models, in addition to above mentioned features weighted sentiment scores obtained from VADER and textblob normalised to range 0 to 1 from 3 previous days are considered as the feature set. . The features used for rest of the models is shown in figure 6.

	change_lag 1	change_lag 2	change_lag 3	adjclose 1	adjclose 2	adjclose 3	month	week	day_of_week
Date									
2016-04-01	-0.019481	0.024676	0.090260	0.276624	0.296104	0.271428	4	13	4
2016-04-04	-0.012337	-0.019481	0.024676	0.264286	0.276624	0.296104	4	14	0
2016-04-05	-0.011689	-0.012337	-0.019481	0.252597	0.264286	0.276624	4	14	1
2016-04-06	-0.018832	-0.011689	-0.012337	0.233766	0.252597	0.264286	4	14	2
2016-04-07	0.024675	-0.018832	-0.011689	0.258441	0.233766	0.252597	4	14	3

Figure 5. Baseline feature set.

	Vader_lag1	Vader_lag2	Vader_lag3	txtlb1_lag1	txtlb1_lag2	txtlb1_lag3	change_lag1	change_lag2	change_lag3	adjc11	adjc12	adjc13	month	week	day_of_week
Date															
2016-04-01	0.209634	0.213004	0.213694	0.053849	0.067648	0.136351	-0.019481	0.024676	0.090260	0.276624	0.296104	0.271428	4	13	4
2016-04-04	0.208101	0.209634	0.213004	0.042725	0.053849	0.067648	-0.012337	-0.019481	0.024676	0.264286	0.276624	0.296104	4	14	0
2016-04-05	0.418465	0.208101	0.209634	0.053300	0.042725	0.053849	-0.011689	-0.012337	-0.019481	0.252597	0.264286	0.276624	4	14	1
2016-04-06	0.217599	0.418465	0.208101	0.123235	0.053300	0.042725	-0.018832	-0.011689	-0.012337	0.233766	0.252597	0.264286	4	14	2
2016-04-07	0.211169	0.217599	0.418465	0.056009	0.123235	0.053300	0.024675	-0.018832	-0.011689	0.258441	0.233766	0.252597	4	14	3

Figure 6. Feature set with sentiment analysis.

Target: The change in the adjusted closure stock price.

Data split:

60% of the data is used as training data, 20% of the data was used as validation data and 20% of the data is used as test data. Validation data is used for selecting the appropriate hyperparameters.

Models used:

Decision tree model with previous 3 days (lag 1, lag 2 and lag 3) of change and adjusted closure as feature set is used as the baseline. Later decision tree (as regressor) model, random forest (as regressor) model, XGBoost (as regressor) model and support vector regressor model were applied for the training data. As the prediction of the change in stock price is a regression problem, here all the models are used as regressors. Most of the models chosen in this project are tree-based models as these tree-based models especially XGBoost are very popular choice for time series forecasting problems.

Hyper-parameter search:

Once the models are trained on the training data, validation data was used to perform hyperparameter search. For hyperparameter search hyper parameter optimization library called hyperopt (Bergstra, J., Yamins, D., & Cox, D. D. 2013) is being used in this project. The 3 steps involved can be condensed as follows: (1) Creating an objective function that can produce the real valued loss for the parameter value it takes as argument. (2) Creating a configuration space where range of possible values for each hyper parameter that needs to be optimized is mentioned. (3) call 'fmin' with chosen search algorithm to search the configuration space by optimizing the objective function.

Hyperopt with respective parameter space for each of models were evaluated and the parameter set providing the least error when evaluated of the validation dataset was chosen as the optimal parameter set. The parameters for which the hyperparameter search was performed is provided in the below table (Table1).

Model	Hyper-parameters	Choice/Range of values in the hyper-parameter
-------	------------------	---

		search space
Baseline: Decision tree regressor	Splitter Max_depth Max_features Criterion Min_samples_split Min_samples_leaf	'best','random' 1 to 15 'sqrt', 'auto', 'log2' 'squared_error','friedman_mse','absolute_error' 5,6,7,8,9 1,2,3,4
Decision tree regressor	Splitter Max_depth Max_features Criterion Min_samples_split Min_samples_leaf	'best','random' 1 to 15 'sqrt', 'auto', 'log2' 'squared_error','friedman_mse','absolute_error' 5,6,7,8,9 1,2,3,4
Random forest regressor	N estimators Max_depth Max_features Bootstrap Min_samples_split Min_samples_leaf	100 to 1000 1 to 15 'sqrt', 'auto' True or False 5,6,7,8,9 1,2,3,4
XGBoost regressor	N estimators Max_depth Learning rate Subsample	100 to 1000 1 to 15 0.0001 to 0.7 0.8 to 1.0
SVM regressor	Kernel Gamma Epsilon Shrinking	'linear', 'poly', 'rbf', 'sigmoid' 'scale','auto' 0.1 to 1 True or False

Table1: Hyperparameter space of all models

Loss metrics:

Mean square error, mean absolute error and root mean squared error are calculated as the loss function. The mathematical representation of these loss functions is given in the figure7. The predictions of this models will be compared and evaluated in the next section.

$$1. RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

$$2. MAE = \frac{1}{n} \sum_{t=1}^n |e_t|$$

$$3. MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$$

where $e_t = Y_t - \tilde{Y}_t$ is the error term, Y_t is the observed value and \tilde{Y}_t is the model prediction.

Figure 7. Mathematical representation of loss metrics (Mishra, P., Yonar, A et al. 2021).

4.4 Evaluations of models.

Model evaluation has been carried out by comparing the loss obtained in the prediction of stock price change (change in the adjusted closure to be specific) in the baseline (model without using sentiment analysis score from twitter posts as one of the features) with the loss obtained by other models using the sentiment score as one of its features. For the loss evaluation MSE, MAE and RMSE are used as loss metrics. This evaluation lets us understand if there is any improvement in the prediction of the stock price change by including the sentiment analysis score as one of the features.

5. Results.

Sentiment analysis:

Obtaining the sentiment analysis score from Vader did not require any data preprocessing. However, for obtaining the sentiment analysis score from textblob, data preprocessing was required, and the preprocessing steps performed are mentioned in section 4.1. Table 2 shows the change in the tweet content after the preprocessing step.

Tweet before data pre-processing	Tweet after pre-processing
#VR hardware is here, but robust game lineup isn't. https://t.co/2himdl5vxl \$EA \$MSFT \$SNE #E3 #tech https://t.co/mHoyRvQthn	vr hardware robust lineup isnt msft sne e3 tech
FXA CurrencyShares British Pound Sterling Trust Summary https://t.co/izmw6jetQr \$FXA \$IVOB \$EA \$HTS #nasdaq #investing	fxa currencyshares british pound sterling trust summary ivob hts nasdaq investing
Listen to \$teelakejake - G\$U\$ \$EA\$UN - 05 Is It Real (prod. Seno) by \$teelakejake #np on #SoundCloud https://t.co/r7t38dy3DF	listen teelakejake gu eaun 05 real prod seno teelakejake np soundcloud
EA Direxion Daily Emerging Markets Bull 3x Shares Day Low https://t.co/HWFklh2Gx8 \$EA \$UA \$FCX \$DIS #nasdaq #share	direxion daily emerging markets bull 3x shares day low ua fcx dis nasdaq share

Table2: Data pre processing

Figure 7 represents the top 1000 most negative words observed in the tweet content. Here one can notice the words such as 'black', 'worried', 'downward', 'breaking' etc contributing to the negative sentiment.

	Min_samples_leaf	2
Random forest regressor	N_estimators Max_depth Max_features Bootstrap Min_samples_split Min_samples_leaf	100 1 'sqrt' False 7 1
XGBoost regressor	N_estimators Max_depth Learning_rate Subsample	600 2 0.302 0.8839716960982944
SVM regressor	Kernel Gamma Epsilon Shrinking	'sigmoid' 'scale' 0.13 False

Table3: Final hyperparameter set.

Mean square error, mean absolute error and root mean square error loss metrics were recorded for predicted values of all models. XGBoost regressor model provided the least error in terms of all 3 loss metrics with a MSE of 0.0015944, MAE of 0.0314637 and RMSE of 0.0399301. Second least errors were obtained by random forest regressor. Figure 6 shows the MSE scores calculated for all models. All the models experimented in this project which had sentiment analysis score as its feature obtained lower prediction errors when compared to baseline model (decision tree model without sentiment analysis scores as its feature). The prediction errors recorded from all the models in this project as displayed in figure 9.

Model	MSE	MAE	RMSE
Baseline Desicion tree	0.004814571525219476	0.06247284912109582	0.06938711353860655
Desicion tree	0.003277858311356068	0.04015157302151003	0.05725258344700323
Random forest	0.001832506760553492	0.03560802133318867	0.04280778855013994
SVM	0.0026474240026304508	0.045068238366188264	0.05145312432331443
XGBoost	0.001594414051038096	0.03146375891167138	0.03993011458834167

Figure 9. MAE, MSE and RMSE scores of all model predictions.

The prediction of change in the adjusted closure price i.e, daily fluctuation in the EA stock price obtained by the models experimented in this project are plotted with actual daily change in the adjusted closure price occurred in the test data in figure 10.

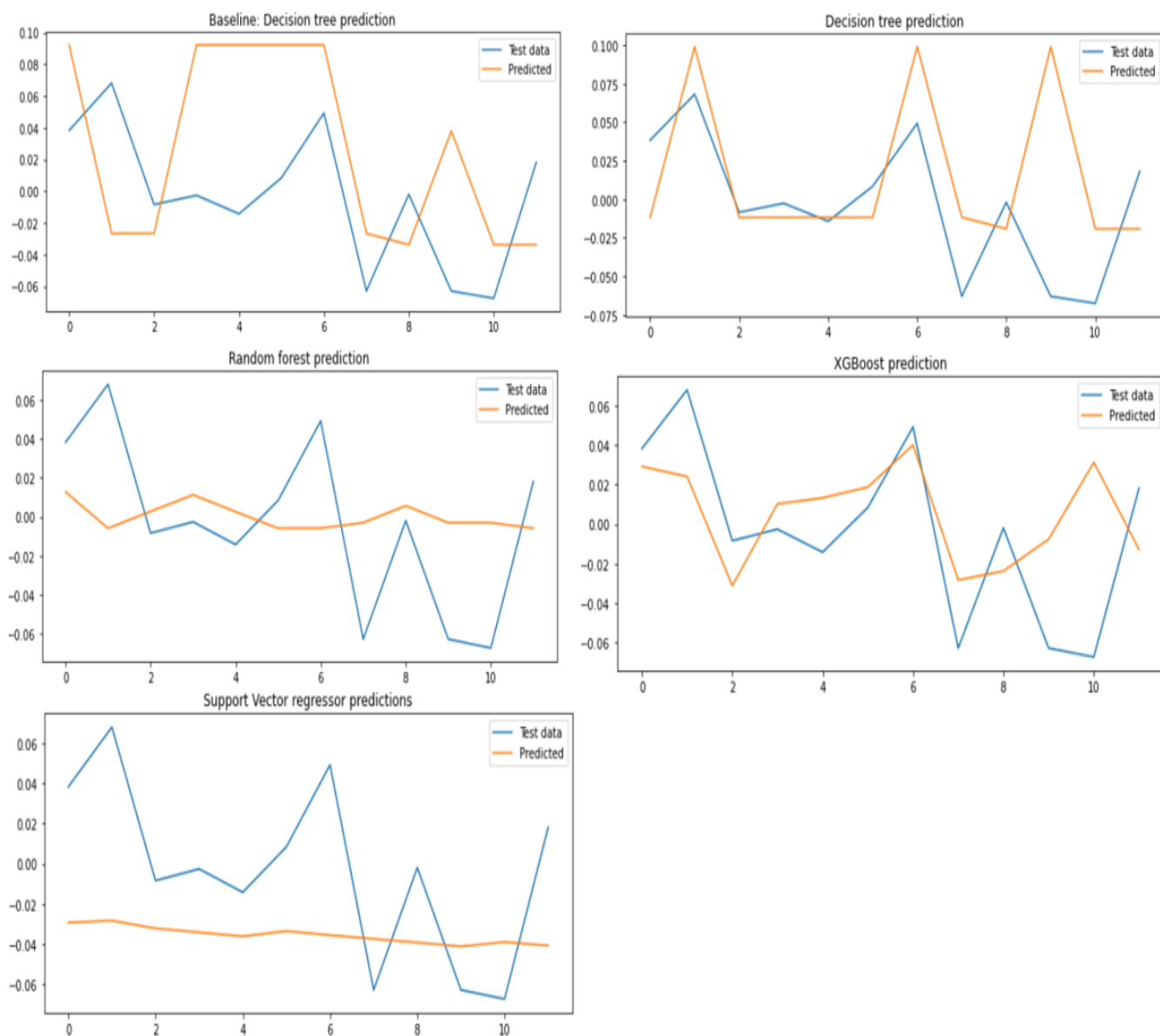


Figure 10. Test data vs. predicted values of change in the adjusted closure price of EA stock from all models.

6. Discussions and conclusion.

The obtained prediction errors in terms MSE, MAE and RMSE presented in the results section shows that the sentiment scores of twitter posts relevant to the stock influences the immediate fluctuations of corporate stock prices. This leads to improvement in the accuracy of prediction of stock price fluctuations in EA stock by using sentiment scores of relevant posts from 3 previous days. There is decrease in the prediction errors obtained in all the models that use sentiment score of twitter data as additional feature when compared to the baseline model which uses only the historical stock prices as its feature. The obtained results are in line with (Pagolu, V. S., Reddy, K. N et al. 2016) where strong correlation between the rise and falls in stock prices with the public sentiments in tweets were observed. It is also inline with reflection of (Zhang, X., Fuehres, H et al. 2011) stating emotional outbursts of any kind leading to spike in the sentiments of the crowd is a predictor on stock price movement on the upcoming day. This can be seen in the obtained results, as both baseline model and the decision tree regressor model use the same machine learning model (decision tree regressor model) for predicting the stock price movement, but due to the additional feature of using the sentiment scores from twitter post, decision tree model has lesser prediction error than the prediction error obtained from baseline model.

Stock price fluctuation prediction from XGBoost regressor model outperforms decision tree, random forest, and support vector regressor model for the dataset used in this study, this can be seen in the lower errors obtained for XGBoost model in the results section. However, since limited number of models and small dataset is considered in this experiment, it is hard to generalize that XGBoost is the best model for this use case. It would be an interesting further study to compare this tree-based model with other more complex deep learning models and transformer-based models.

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. It is hard to account for all the factors that determine the fluctuations of stock price, but even just with historical stock movement there are multiple models in the current state of art such as Long short-term memory (LSTM) (Roondiwala, M., Patel, H., & Varma, S. 2017), transformer-based models (Ding, Q., Wu, S., Sun, H., Guo, J., & Guo, J. 2020) etc to predict stock prices. Even the slight improvement in the accuracy of these prediction can lead to huge monetary gain in the stock trading world. This study shows that there is an improvement in the accuracy of stock price fluctuations predictions (in the models tried out in this project).

As further work, combining the sentiment scores from different social media platforms and public news would most likely improve the prediction accuracy. Fluctuations 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. Sentiment analysis scores from social media and news platform as an additional 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. Only limited number of simple models were tried out in this experiment due to

the limitation of data, more complex deep learning and transformer-based models with a bigger dataset would be ideal to get a concrete picture of the improvement obtained with the addition of sentiment scores.

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.

Amy.(2022).textblob-vs-vader-for-sentiment-analysis-using-python.
<https://pub.towardsai.net/textblob-vs-vader-for-sentiment-analysis-using-python>
76883d40f9ae#

Text blob. (2020). From textblob: <https://textblob.readthedocs.io/en/dev/>

Kolasani, S. and Assaf, R. (2020) Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks. *Journal of Data Analysis and Information Processing*, 8, 309-319. doi: 10.4236/jdaip.2020.84018.

Bergstra, J., Yamins, D., & Cox, D. D. (2013, June). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In Proceedings of the 12th Python in science conference (Vol. 13, p. 20).

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, 26, 55-62.

Roondiwala, M., Patel, H., & Varma, S. (2017). Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*, 6(4), 1754-1756.

Ding, Q., Wu, S., Sun, H., Guo, J., & Guo, J. (2020, January). Hierarchical Multi-Scale Gaussian Transformer for Stock Movement Prediction. In *IJCAI* (pp. 4640-4646).

Mishra, P., Yonar, A., Yonar, H., Kumari, B., Abotaleb, M., Das, S. S., & Patil, S. G. (2021). State of the art in total pulse production in major states of India using ARIMA techniques. *Current Research in Food Science*, 4, 800-806.