STOCK MARKET PREDICTION USING SENTIMENT AND TIME SERIES ANALYSIS IN R

NAGARAJ Karunashree
VYAS Aryan
SZEPEK Maria Sofie

Description:

This project studies the possibilities of forecasting stock market prices of firms using the sentiments captured via web scrapping. We have experimented with stock market price of Tesla using sentiment analysis and ARIMA model. An accuracy analysis was also carried out with a R- sqaured value of each of the model to evaluate how each of them faired in the forecasting. The aim is to help reduce participants in loss while investing using the twitter data. The stock data was pulled of the Yahoo Finance API. The sentiments were obtained off the sentences of tweets from twitter. Results obtained has proved that the ARIMA model has good R-Squared value for short term prediction.

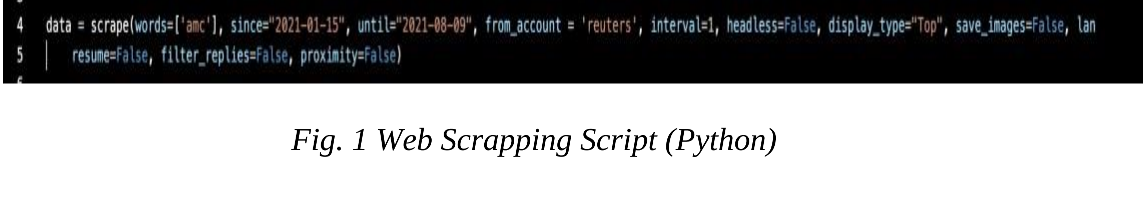
The aim of this project is to forecast the price of the Tesla stock by using time series analysis by using the ARIMA model. It uses the advance computing of computers as a tool to extract data from the web and then parse it on tokens for further computation. The parsed text is further analyzed for sentiments thereby giving each a numeric value ranging from -1 to 1. The penultimate goal is to forecast the stock price of the previously mentioned firms by using the tweets and news data sentiments with the already existing livestock price. For this purpose, we have used historical data of opening price, closing price and highest volume traded for Tesla and Moderna. The stock price is affected by various factors like current trade scenarios, people's liking, the company's performance, pandemics, etc.

Understanding the Objective

The aim of our project is to build a model that predicts the stock price of a given firm while it extracts words from twitter on related topics. Our goal is neither to make billions off the system nor waste billions too. But the objective is to help stock market investors by giving them a direction in taking a decision or not. Whether to buy/hold/sell a stock by providing the result in terms of visualizations.

Data Collection

This is the process where we used a python script to scrape data off twitter. A sample of the python script used to scrape data is shown below. Also, we used Yahoo finance to get the stock price data for the corresponding interval of time.



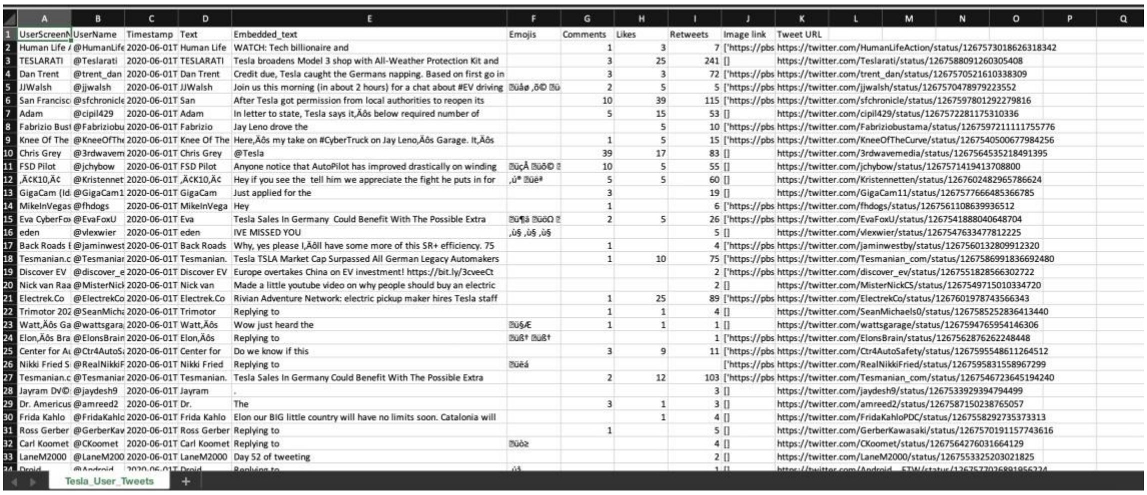
```
from Scweet.scweet import scrape
from Scweet.user import get_user_information, get_users_following, get_users_followers

data = scrape(words=['amc'], since="2021-01-15", until="2021-08-09", from_account = 'reuters', interval=1, headless=False, display_type="Top", save_images=False, lan
    resume=False, filter_replies=False, proximity=False)
```

Fig. 1 Web Scrapping Script (Python)

Data Pre-Processing

This is the stage where the acquired data is processed into final datasets to work on. Cleaning the dataset is the focus in this stage. The overview of the dataset is shown below. There are many columns out of which we used the “Embedded text” column as the main feature for sentiment analysis as that contained the tweets of people. The data pre-processing was also used on the stock price data to make it ready and be combined into a value vector.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	UserScreenN	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets	Image link	Tweet URL						
2	Human Life /	@HumanLife	2020-06-01T	Human Life	WATCH: Tech billionaire and		1	3	7	['https://pbs	https://twitter.com/HumanLifeAction/status/1267573018626318342						
3	TESLARATI	@Teslarati	2020-06-01T	TESLARATI	Tesla broadens Model 3 shop with All-Weather Protection Kit and		3	25	241	[]	https://twitter.com/Teslarati/status/1267588091260305408						
4	Dan Trent	@trent_dan	2020-06-01T	Dan Trent	Credit due, Tesla caught the Germans napping. Based on first go in		3	3	72	['https://pbs	https://twitter.com/trent_dan/status/1267570521610338309						
5	JJWalsh	@jjwalsh	2020-06-01T	JJWalsh	Join us this morning (in about 2 hours) for a chat about #EV driving	□üåø ,õ© □ü	2	5	5	['https://pbs	https://twitter.com/jjwalsh/status/1267570478979223552						
6	San Francisc	@sfchronicle	2020-06-01T	San	After Tesla got permission from local authorities to reopen its		10	39	115	['https://pbs	https://twitter.com/sfchronicle/status/1267597801292279816						
7	Adam	@cipil429	2020-06-01T	Adam	In letter to state, Tesla says it,Äôs below required number of		5	15	53	[]	https://twitter.com/cipil429/status/1267572281175310336						
8	Fabrizio Bust	@Fabriziobu	2020-06-01T	Fabrizio	Jay Leno drove the			5	10	['https://pbs	https://twitter.com/Fabriziobustama/status/1267597211111755776						
9	Knee Of The	@KneeOfThe	2020-06-01T	Knee Of The	Here,Äôs my take on #CyberTruck on Jay Leno,Äôs Garage. It,Äôs		1	5	15	['https://pbs	https://twitter.com/KneeOfTheCurve/status/1267540500677984256						
10	Chris Grey	@3rdwavem	2020-06-01T	Chris Grey	@Tesla		39	17	83	[]	https://twitter.com/3rdwavemedia/status/1267564535218491395						
11	FSD Pilot	@jchybow	2020-06-01T	FSD Pilot	Anyone notice that AutoPilot has improved drastically on winding	□üçÅ □üõ© □	10	5	55	[]	https://twitter.com/jchybow/status/1267571419413708800						
12	,Ä¢K10,Ä¢	@Kristennet	2020-06-01T	,Ä¢K10,Ä¢	Hey if you see the tell him we appreciate the fight he puts in for	,ù* □üë*	5	5	60	[]	https://twitter.com/Kristennetten/status/1267602482965786624						
13	GigaCam (Id	@GigaCam1	2020-06-01T	GigaCam	Just applied for the		3		19	[]	https://twitter.com/GigaCam11/status/1267577666485366785						
14	MikeInVegas	@fhdogs	2020-06-01T	MikeInVega	Hey		1		6	['https://pbs	https://twitter.com/fhdogs/status/1267561108639936512						
15	Eva CyberFo›	@EvaFoxU	2020-06-01T	Eva	Tesla Sales In Germany Could Benefit With The Possible Extra	□ü¶ä □üöΩ □	2	5	26	['https://pbs	https://twitter.com/EvaFoxU/status/1267541888040648704						
16	eden	@vlexwier	2020-06-01T	eden	IVE MISSED YOU	,ù§ ,ù§ ,ù§			5	[]	https://twitter.com/vlexwier/status/1267547633477812225						
17	Back Roads I	@jaminwest	2020-06-01T	Back Roads	Why, yes please I,Äôll have some more of this SR+ efficiency. 75		1		4	['https://pbs	https://twitter.com/jaminwestby/status/1267560132809912320						
18	Tesmanian.c	@Tesmaniar	2020-06-01T	Tesmanian.	Tesla TSLA Market Cap Surpassed All German Legacy Automakers		1	10	75	['https://pbs	https://twitter.com/Tesmanian_com/status/1267586991836692480						
19	Discover EV	@discover_e	2020-06-01T	Discover EV	Europe overtakes China on EV investment! https://bit.ly/3cveeCt				2	['https://pbs	https://twitter.com/discover_ev/status/1267551828566302722						
20	Nick van Raa	@MisterNick	2020-06-01T	Nick van	Made a little youtube video on why people should buy an electric				2	[]	https://twitter.com/MisterNickCS/status/1267549715010334720						
21	Electrek.Co	@ElectrekCo	2020-06-01T	Electrek.Co	Rivian Adventure Network: electric pickup maker hires Tesla staff		1	25	89	['https://pbs	https://twitter.com/ElectrekCo/status/1267601978743566343						
22	Trimotor 202	@SeanMicha	2020-06-01T	Trimotor	Replying to		1	1	4	[]	https://twitter.com/SeanMichaels0/status/1267585252836413440						
23	Watt,Äôs Ga	@wattsgara	2020-06-01T	Watt,Äôs	Wow just heard the	□ü§Æ	1	1	1	[]	https://twitter.com/wattsgarage/status/1267594765954146306						
24	Elon,Äôs Bra	@ElonsBrain	2020-06-01T	Elon,Äôs	Replying to	□üß† □üß†			1	['https://pbs	https://twitter.com/ElonsBrain/status/1267562876262248448						
25	Center for Au	@Ctr4AutoS	2020-06-01T	Center for	Do we know if this		3	9	11	['https://pbs	https://twitter.com/Ctr4AutoSafety/status/1267595548611264512						
26	Nikki Fried S	@RealNikkiF	2020-06-01T	Nikki Fried	Replying to	□üëá				['https://pbs	https://twitter.com/RealNikkiFried/status/1267595831558967299						
27	Tesmanian.c	@Tesmaniar	2020-06-01T	Tesmanian.	Tesla Sales In Germany Could Benefit With The Possible Extra		2	12	103	['https://pbs	https://twitter.com/Tesmanian_com/status/1267546723645194240						
28	Jayram DV©	@jaydesh9	2020-06-01T	Jayram	.				3	[]	https://twitter.com/jaydesh9/status/1267533929394794499						
29	Dr. Americus	@amreed2	2020-06-01T	Dr.	The		3	1	3	[]	https://twitter.com/amreed2/status/1267587150238765057						
30	Frida Kahlo	@FridaKahlo	2020-06-01T	Frida Kahlo	Elon our BIG little country will have no limits soon. Catalonia will			1	4	[]	https://twitter.com/FridaKahloPDC/status/1267558292735373313						
31	Ross Gerber	@GerberKav	2020-06-01T	Ross Gerber	Replying to		1		5	[]	https://twitter.com/GerberKawasaki/status/1267570191157743616						
32	Carl Koomet	@CKoomet	2020-06-01T	Carl Koomet	Replying to	□üò≥			4	[]	https://twitter.com/CKoomet/status/1267564276031664129						
33	LaneM2000	@LaneM200	2020-06-01T	LaneM2000	Day 52 of tweeting				2	[]	https://twitter.com/LaneM2000/status/1267553325203021825						
34	Droid	@Android	2020-06-01T	Droid	Replying to	,ù³			1	[]	https://twitter.com/Android__FTW/status/1267577026891956224						

Fig. 2 Tweets data (TESLA)

Data Preprocessing

To process the data, we use the ARIMA(p,d,q) model. Generally, stock investors use the auto regressive and moving average models to forecast the future trends. Highlights here would be estimation, forecasting and identification. These steps are repeated recursively until an optimal model is identified for prediction. R language provides auto.arima() method to forecast the time series data according to ARIMA(p,d,q) model.

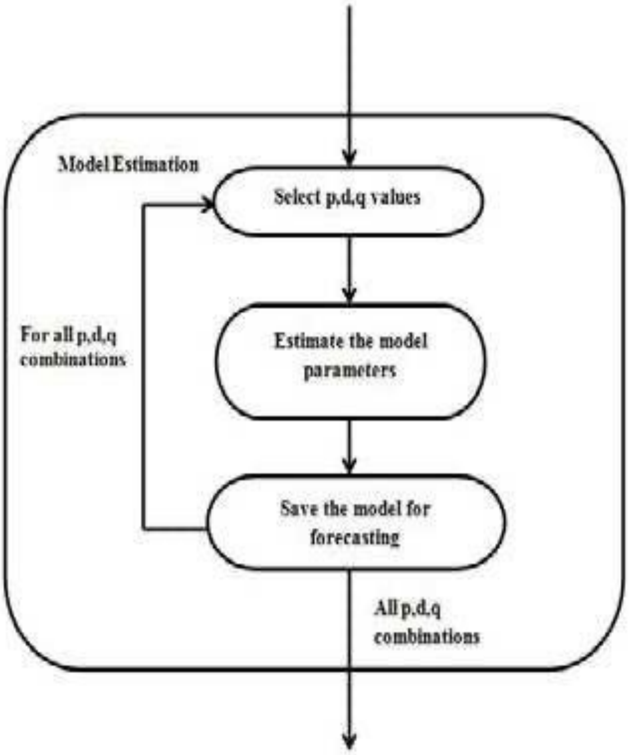
Forecasting Results

The process of predicting the future by relying upon the past and current data is called as forecasting. Various prediction techniques are used by stock analysts to predict the future stock trends value. The ‘forecast’ package offered in R was used to predict the future trends which took in values off the sentiment score and past historical stock price data. The “SAS” package of R was used for sentiment analysis thereby giving each sentence a sentiment score. All of these were fed into the ARIMA model which then forecasted the results for time series predictions. It also offered exponential smoothing and space models.

View and Analyze Results

This is done to evaluate and view the outcome of the model. Screenshots of the evaluation and results are provided further in the section. Investors can view the results and graphs with a comparison view and then invest in the stock. They can use this as an “assistance” to buy/sell/hold a particular stock.

AIRMA MODEL


Model Estimation
Select p,d,q values
For all p,d,q combinations
Estimate the model parameters
Save the model for forecasting
All p,d,q combinations

In ARIMA model, the identification is to be accomplished using auto co-relation function and partial auto co-relation function in order to identify p, d and q standards. For any realistic time, sequence generally p, d and q values vary between 0 and 2, but model estimation is executed for all probable combinations of p, d and q values.

ARIMA() Function in R –

Arima() function automates the inclusion of a constant. By default, d = 0 or the value of d = 1. A consant will be included if it improved the AICc value; for d > 1 the constant is always omitted. If allowdrift= FALSE is specified, then the constant is only allowed when d = 0.
In ARIMA model, the future value of a variable is a linear combination of past values and past errors, expressed as follows:

$$Y_t=\phi_0+\phi_1Y_{t-1}+\phi_2Y_{t-2}+...+\phi_pY_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-...-\theta_q\varepsilon_{t-q}$$

Fig. 4 Arima function Math

Where, Y_t is the actual value and $\square_t$ is random error at t, $\square_I$ and $\square_j$ are the coefficients, p and q are integers which are often referred to as autoregressive and moving average reciprocal.

Results and Conclusion

Word cloud

The Word Cloud for Tesla tweets shows us the most common words used in users’ tweets. We then calculate the sentiment score of these words thereby giving each a value ranging between 1 to -1. 1 being highly positive and -1 being highly negative.

```
tweet_Table <- tweet_data %>%
  unnest_tokens(word, stripped_text)
data(stop_words)
tweet_Table <- tweet_Table %>%
  anti_join(stop_words)
tweet_Table <- tweet_Table %>%
  count(word, sort = TRUE)
tweet_Table <-tweet_Table %>%
  filter(!word %in% c('replying', 'tesla', 'andothers', 'model', 'tsla', 'car', 'cars',
                      'company', 'dont', 'teslas', 'im', 'didnt', 'musk', 'elon', 'gif',
                      'time', 'day', 'battery', 'stock', 'drive', 'people', 'ive', 'doesnt',
                      'cybertruck', 'texas', 'aug', 'jul', 'range', 'hey', 'nikola',
                      'hes', 'thread', 'solar', 'vehicle', 'electric', 'ev', 'lol',
                      'driving', 'autopilot', 'world', 'guy', 'month', 'fsd', 'app',
                      'energy', 'video', 'money', 'jun', 'service', 'tesmaniancom',
                      'companies', 'youtubecom', 'supercharger', 'delivery',
                      'factory', 'youre', 'price', 'truck', 'berlin', 'california',
                      'china', 'austin', 'ill', 'wait', 'call', 'jul', 'vehicles', 'update', 'share',
                      'home', 'yeah', 'giga', 'ceo', 'told', 'guys', 'evs', 'road', 'tech',
                      'week', 'tslaq', 'theyre', 'lot', 'twitter', 'answer'))
wordcloud2(tweet_Table,size=0.7,color=’random-light’, background Color="black")
```

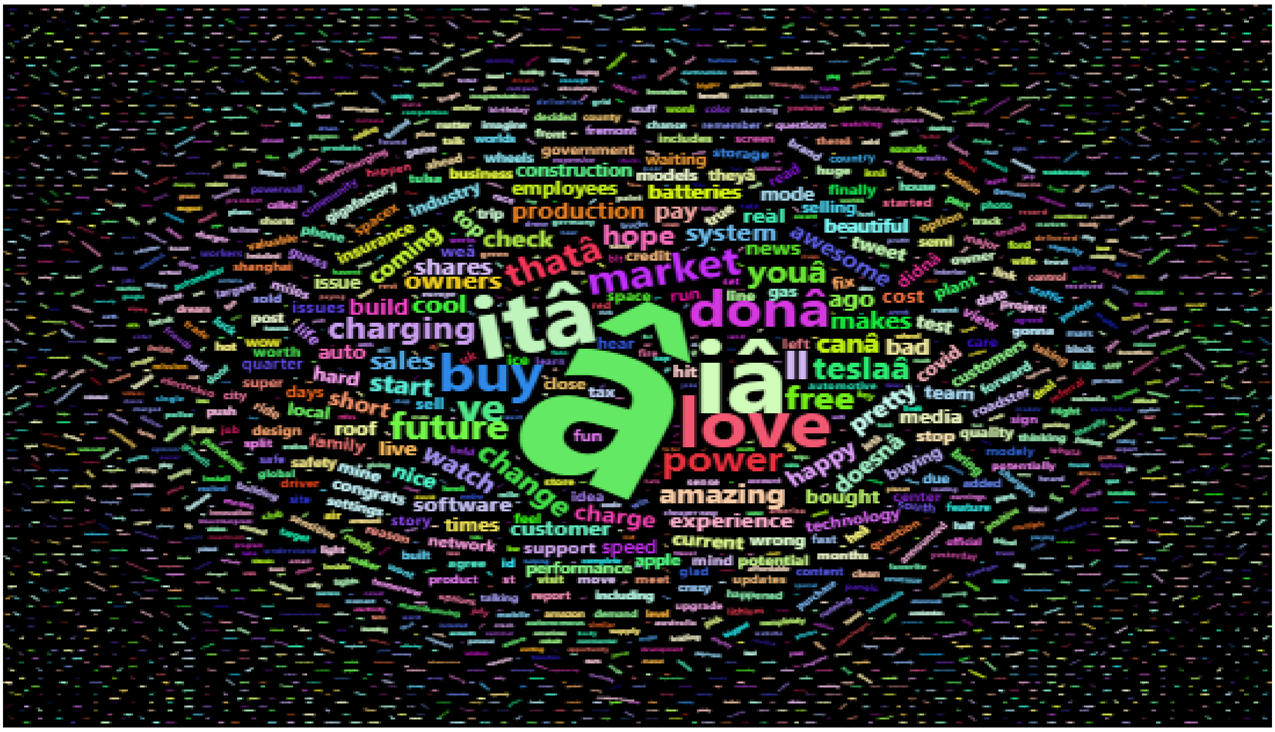


Fig. 5 Word Cloud (Tesla User Tweets)

Time Series Plot of the sentiment Time series

Plotting the sentiment score w.r.t. the price shows us the correlation of how the stock price of TESLA has been greatly affected by the tweets.

```
plot_data <- subset (sentimentr_news_data, select = c(1))
plot_data <- cbind(plot_data, news_sentiment = sentimentr_news_data$Sentiment,
                   tweet_sentiment = sentimentr_tweet_data$Sentiment,
                   price = price_data$Returns)
rownames(plot_data) <- plot_data$Date
# Converting data into time-series
news_sentiment <- ts(plot_data$news_sentiment,start=decimal_date(ymd("2020-06-07")),
frequency=52)
tweet_sentiment <- ts(plot_data$tweet_sentiment,start=decimal_date(ymd("2020-06-07")),
frequency=52)
price <- ts(plot_data$price,start=decimal_date(ymd("2020-06-07")), frequency=52)
VAR_data <- cbind(news_sentiment, tweet_sentiment, price)
# Plotting time-series data
forecast::autoplot(VAR_data) +
  ggtitle("Time Series Plot of the Sentiment Time-Series") +
  theme(plot.title = element_text(hjust = 0.5))
```

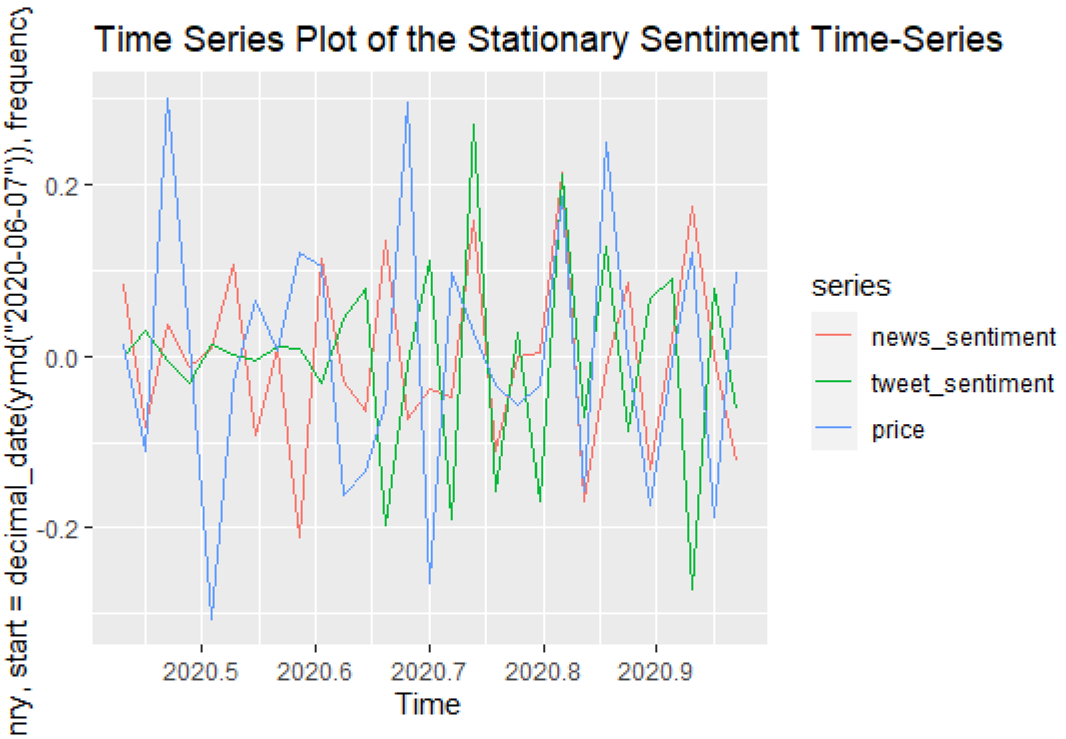


Fig 6 Graphing the tweets and price (TESLA)

Choosing Best lag for our data

choosing best lag for our data using Vector Auto-regressions VAR.

```
# Choosing best lag for our data
apply(VAR_data, 2, adf.test)
stnry = diffM(VAR_data)
apply(stnry, 2, adf.test)
plot.ts(stnry)
autoplot(ts(stnry, start=decimal_date(ymd("2020-06-07")), frequency=52)) +
  ggtitle("Time Series Plot of the Stationary Sentiment Time-Series")
VARselect(stnry, type = "none", lag.max = 6) # Highest lag order
var.a <- vars::VAR(stnry, lag.max = 6, ic = "AIC", type = "none")
```

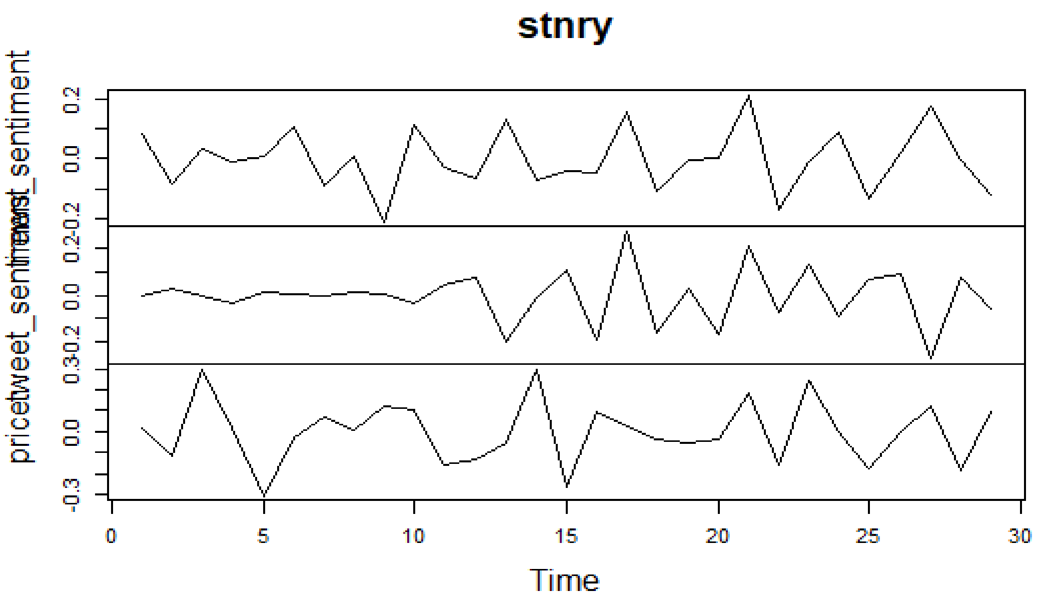


Fig 7 Plot Vector Auto-regressions

Forecasting News and tweet sentiments and the price

The above figure shows us how the news and tweets of users for the word “Tesla” will be forecasted continuing the current and previous trends.

```
# Forecasting the time-series for 8 weeks ahead
fcast = predict(var.a, n.ahead = 8)
par(mar = c(2.5,2.5,2.5,2.5))
plot(fcast)
news_sentiment = fcast$fcst[1];
tweet_sentiment = fcast$fcst[2];
price = fcast$fcst[3];
x = news_sentiment$news_sentiment[, 1]
y = tweet_sentiment$tweet_sentiment[, 1]
z = price$price[, 1]
tail(VAR_data)
x = cumsum(x) + 0.17856389
y = cumsum(y) + (-0.003998803)
z = cumsum(z) + 0.04988139
par(mar = c(2.5,2.5,2.5,2.5))
```

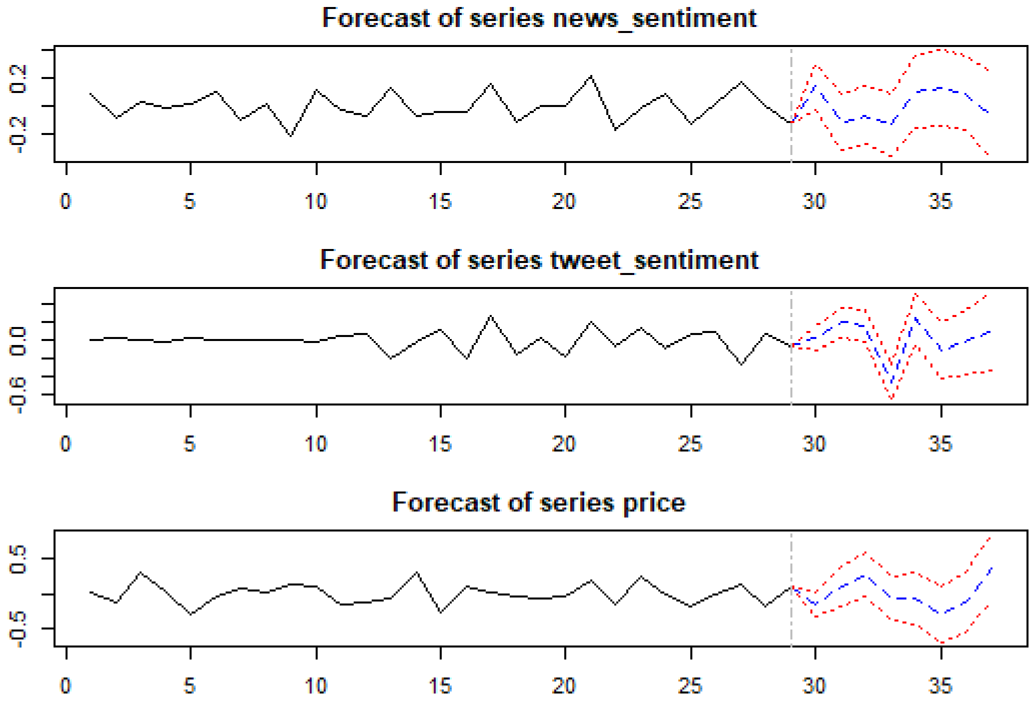
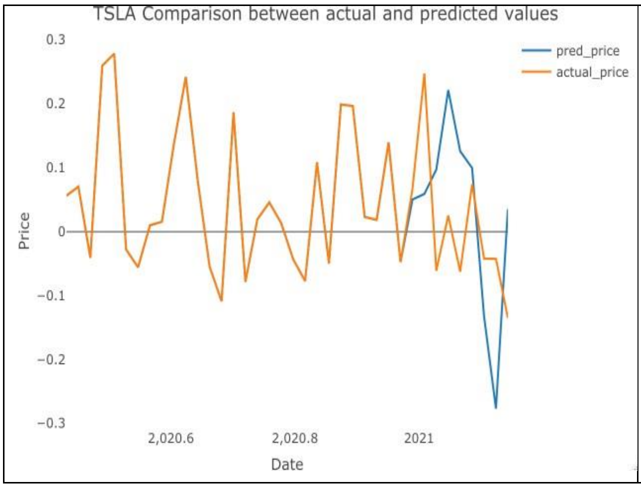


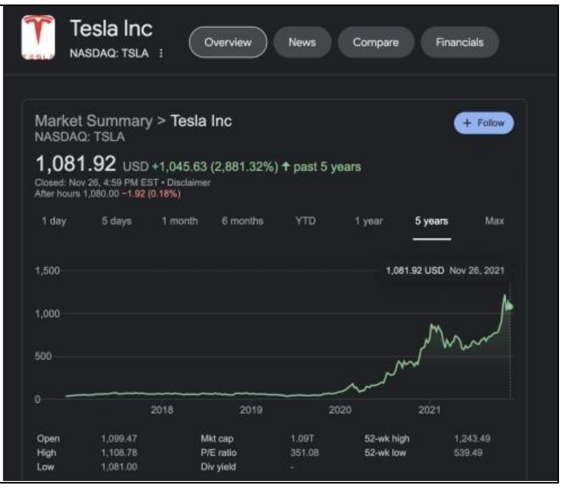
Fig 8 Forecasting News and tweet sentiments and the price

price stitched with predicted price

As we can see that on the (LHS), the blue line indicates the predicted price of TESLA's stock by our model is close to the real time stock price of TESLA (RHS). Post 2021 start, the model was given to predict the stock and it did perfectly.

```
pred_price = ts(c(VAR_data[,3], z), start=decimal_date(ymd("2020-06-07")),
frequency=52)
pred_price_datframe <- as.data.frame(pred_price[1:38])
colnames(pred_price_datframe) <- c("z")
a = zoo(pred_price[1:38])
xyplot(a, grid=TRUE, panel = function(x, y, ...){
  panel.xyplot(x, y, col="red", ...)
  grid.clip(x = unit(31, "native"), just=c("right"))
  panel.xyplot(x, y, col="green", ...) })
```





Market Summary > Tesla Inc
NASDAQ: TSLA

1,081.92 USD +1,045.63 (2,881.32%) ↑ past 5 years
Closed: Nov 26, 4:59 PM EST • Disclaimer
After hours 1,080.00 −1.92 (0.18%)

Open	1,099.47	Mkt cap	1.09T	52-wk high	1,243.49
High	1,108.78	P/E ratio	351.08	52-wk low	539.49
Low	1,081.00	Div yield	-		

Conclusion

In this project we attempted to predict the stock market for TESLA using the historical stock price data and encompassing the impact of sentiments using tweets and news headlines of the same. The experimental results obtained demonstrated the potential of the ARIMA model in short term prediction. This could guide the investors in investing wisely on whether buy/sell/hold that stock.

The factors affecting the R-Squared value are regional investing trends, pandemics, stock’s current value, etc.

Investors can take the risk of investment easily when they have an idea about the future value of the stocks.