Analysing effect of Twitter, Oil Prices, Gold Prices and Foreign Exchange on S&P500 Using Machine Learning

MSc Research Project

MSc in Data Analytics

Smit Jain

x18135340

School of Computing

National College of Ireland

Submitted to:

Dr. Pierpaolo Dondio

National College of Ireland

Project Submission Sheet – 2019/2020

Student Name:	Smit Jain		
Student ID:	18135340		
Programme:	MSc in Data Analytics	**Year:**	2019-2020
Module:	Research project		
Lecturer:	Dr. Pierpaolo Dondio		
Submission Due Date:	12/12/2019		
Project Title:	Analysing effect of Twitter Tweets, Oil Prices, Gold Prices and Foreign Exchange on S&P500 Using Machine Learning.		
Word Count:	8961		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Smit Jain

Date: 11/12/2019

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Abstract

A country's economic status can be judged majorly by the stock market's performance. The performance of the companies registered with the stock market drives the performance of stock market index. Stock market's performance impact the lives of the people, directly or indirectly, who invest in it. Forecasting the stock market is important from investor's point of view to foresee the performance and analysing the trend. To analyse the SP500 index and the impact of other important factors like public sentiment (twitter analysis), oil & gold prices and foreign exchange; machine learning models have been implemented. The datasets used for modelling the datasets have been extracted from different data-sources like data of stock market has been collected from yahoo finance website and twitter tweets have been downloaded by running a command to extract tweets using anaconda prompt. With the help of Granger Causality tests, we have established which factors act as predictors for the stock market index significantly. For the evaluation of the model's performance, different tests have been carried out like MSE, RMSE, MAE & MAPE etc.

Keywords: *SP500, Stock Market, Twitter Sentiment Analysis, Forex, Oil and Gold price, Time-series, Forecasting, RNN*

1 Introduction

The Stock markets experience variations almost every single minute and every single second across the globe. This highly unpredictable nature of the stock markets keeps all the investors on their toes because no one knows how the market is going to behave the next moment. Hence, predicting the stock market is a very challenging yet interesting part. The concept of Stocks and its Market has been in operation since a very long time (decades ago) and all of the countries around the world have their own trading markets to trade and company's own shares are registered.

Highlighting the previous research work done in this field, a major amount of focus can be seen in the determination of trend in the market data based on the historical data. The Standard & Poor's 500 stock index can majorly be influenced by factors like public emotions, commodity prices and financial technical indicators.

1.1 Motivation and Background

To analyse and understand the behavioural pattern of the stock markets, researchers from all across the world have done many types of research work related to the Accounting, Computing, Financial and Statistical fields. By the use of Machine learning methods, (Beyaz *et al.*, 2018) have tried to forecast stock market's performance and have compared different models for forecasting and in the end they have concluded that which model performs better for forecasting. Another approach tried in (Liu *et al.*, 2016; Oliveira, Cortez and Areal, 2017; Weng, Ahmed and Megahed, 2017; Kyun *et al.*, 2019) for the stock market predictions using SVM have used technical indicators. But besides the technical indicators, there are other driving factors like public emotions and foreign exchange which haven't been taken into consideration. Talking about the high political movement in near future – Presidential Elections 2020 in USA, there is a lot of ups and downs estimated in the stock market performance and hence the stock market investors will keep a sharp eye on its performance. Taking the public emotions (via twitter tweets) and other driving factors into consideration, analysing the impact on stock market time series is really beneficial.

1.2 Research question

"To what extent can the stock market index be predicted using Machine Learning techniques, considering the impact of public sentiment, forex rates and oil prices?"

1.3 Research Objective

Objective1 – To review the available literature in relation to the research area.

Objective2 – To design a stock market prediction model and compare the time series with the time series generated by twitter sentiment analysis, forex rates, gold and oil prices and then implement it.

Sub-Objective (a) – The data that is needed for predictions has to be extracted, cleaned and then transformed.

Sub-Objective (b) – The prediction of stock market has to be done by application of various machine learning models.

Sub-Objective (c) – The evaluation of the performance of the models applied using performance evaluation tests.

1.4 Research Contribution

The primary objective of this research work is to predict the stock market's movement with the considerations of public emotions (sentiment analysis on twitter tweets), foreign exchange (EUR-USD), gold price and oil price. After doing an analysis on the causal effect, only factors which are significantly helping to predict the stock market will be taken into consideration.

The investors who invest in stock market will be helped greatly as United States presidential elections is due next year in 2020 and they will be able to make decisions swiftly whether to invest at that time period or not.

The report has been categorized into following heads: the literature review section contains the technical review of the research work done earlier related to this field of study. Here, more than 30 technical papers have been reviewed to get a better understanding around the field of study. After that, the research methodology applied to this research has been discussed. The next section – Implementation consists of all the details about the machine learning models applied for the analysis of the datasets gathered. The project plan has been discussed along with the declaration of the ethics in the next section where details about each step during this research work has been chalked down on a single timeline.

2 Literature Review

2.1 Introduction

The literature review section contains the technical review of the research work done earlier related to this field of study. Here, more than 30 technical papers have been reviewed to get a better and clear understanding around the field of study. Related to the field of research work, this literature review section has been apportioned into four parts, and in each part talk about the research work carried out in the past related only to that field of study. In the first part, the literature review related to the stock markets has been carried out on the past research work. It details out the basic idea about the type of methods and techniques used for stock market analysis and also clarifies the gaps in those research works. The next part consists of the literature review of work done around the analysis of foreign exchange where a lot of work related to the predictions of foreign exchange price has been done in the past. The next part consists of the research work done in the field of sentiment analysis and the techniques and tools used to perform sentiment analysis. The last part consists of research work done in relation to stock market index price, forex and sentiment analysis field.

2.2 Stock markets

The fact that stock market index prices are highly volatile and are subject to change at any point of time, presents a challenge to all the researchers to predict its behaviour. There are thousands of researchers who have spent their valuable time and efforts to research about stock market and predict it. The findings and review of such research work carried out has been given below:

The SVM model has been very famous amongst the researchers for investigating the stock market's movement. Having implemented the supervised learning models like SVM, GDA, Logistic Regression and Naïve Bayes; (Liu *et al.*, 2016) has carried out the research work to predict the stock value. While, (Choudhry and Garg, 2008) have used a hybrid GA-SVM model to predict the movement in the direction of the stock prices. The approach followed by (Liu *et al.,* 2016) has been different as they have taken the dataset from Yahoo finance and have split the dataset in 2 parts where two-third of data have been used for training the model and one-third has been taken for testing the trained model. On the other hand, (Liu *et al.*, 2016) found that SVM with RBF has given most accurate forecasting results. The parameters that they used are quite a few important ones; however, in the conclusion they have mentioned that to improve the performance, time-series analysis could have been done.

Another research work carried out by (Choudhry and Garg, 2008) has developed a model based only on 2 classes. The scope of the research work has been limited to only 3 company stocks from the Indian stock market. Technical indicators have been taken from the stock market and used as input features to the model. However, when an investor is looking at the stock markets at the global level, strategy plays a vital role. In the research work done by (Kyun *et al.*, 2019) stock prices around the globe have been strategically assessed in two ways which are as follows: 1.) Prediction of the stock market direction strategy; 2.) Strategy to allocate regionally. Using the Pearson correlation and VAR model, a network of global instability has been formed with the dataset gathered from 10 countries. The models used for testing various portfolios are Linear Regression, SVM and Random Forest. Rolling data has been used for the purpose of training the models and testing has been performed on the data from year 2005 to 2016. In an attempt to predict the company stock price's percent change for 126 and 258 days ahead, the authors (Beyaz *et al.*, 2018) have used SVR forecasting and NN based models. Using clustering, the P2C (Put-Call) ratio and VIX (Volatility Index) have been calculated which are further used to compare the relative effectiveness in order to understand the moods of the stock market. The results of the work done by (Choudhry and Garg, 2008) have proven that the stock movement prediction has improved by using correlation techniques and GA and the model has performed better overall. Whereas the research work carried out by (Kyun *et al.*, 2019) shows that the medium-term investments have given better results when compared to short-term investments with the use of network indicators. However, the limitation that has been noticed in this research work is that only one measure of connectedness has been used. In the research work done by (Beyaz *et al.*, 2018), the results prove that the forecast for 126 days has been outperformed by 252 days by 50%. And as per the results of the research work by (Beyaz *et al.*, 2018), when technical indicators are used as input to the models, the forecasting increases significantly both collectively and exclusively. However, in this research, only historical data has been taken for forecasting, and no other impacting factors have been taken into consideration.

ANN has been used in the prediction of stock markets by (Yavuz & Ozdemir, 2015; J. Patel, Shah, Thakkar, & Kotecha, 2015; Aldin, Dehnavi, & Entezari, 2012; Shen, Guo, Wu, & Wu, 2011; Yudong & Lenan, 2009; Weng et al., 2017). The authors (Yavuz and Ozdemir, 2015) have attempted to predict the stock market performance for Istanbul only by using ANN (feed forward back propagation) techniques. A similar approach has been followed by (Patel *et al.*, 2015) in order to predict the stock market value (Indian stock market) for 1-10, 15 and 30 days in advance using the Machine Learning techniques. In the first stage, support vector aggregation has been used and subsequently in the second stage, Random Forest, ANN and SVR have been merged with the results of stage 1 which resulted into SVR-RF, SVR-ANN and SVR-SVR models. Using the RSM (Response Surface Methodology), (Yavuz and Ozdemir, 2015) has attempted to find out the best network topology. To train the model (ANN), Matlab R2008b has been used. This model can make predictions in many ways, like predictions can be made one day before the stock market or fifteen days, monthly or yearly in advance. The results of this research suggest that the direction of stock market movement can be foreseen for next day or even next week using the ANN techniques.

ANN model's structure resembles a human brain's structure, and this is one of the reasons why ANN model has gained so high popularity amongst all the researchers and hence the model is named neural network. In this research work, out of all the factors, public opinion or sentiment has not been considered which is one of the major influencing factors. If the predictions made are well in advance, the error value increases significantly; as concluded by the authors (Patel *et al.*, 2015). The performance of one stage prediction models have been outperformed by 2 stage prediction models. The integration of SVR with RF and ANN have performed better, however integration of SVR with SVR has moderately performed. Addition of more statistical parameters could have made a better performance and results better as this research is purely on past data. The TEPIX (Tehran Exchange Price Index) data including the technical indicators has been used to predict the stock market index in the research work done by (Aldin, Dehnavi and Entezari, 2012). For the forecasting, the authors have used ANN techniques. The relationship of the Tehran's stock market index with the technical indicators have been tried to establish by using the three-layer back propagation NN. The results of the research work have been successfully been able to analyse the stock market movement direction appropriately (up to 90%).

The machine learning models' performance is mainly based on the time taken for training and testing the data. Based on the models' performance, training and testing time taken, less computational complexity and prediction accuracy, the authors (Yudong and Lenan, 2009) have compared various machine learning models. Lesser training and testing time mean a better performing model. IBCO technique has been merged with back propagation ANN and this model has performed better than other models. The author in (Weng, Ahmed and Megahed, 2017) have used the stock market data for time series prediction. Dividing the research into three phases, the author has first extracted the data needed in the first phase and then selected the predictor variables in the phase 2 and in the last phase the AI techniques have been applied to predict the stock price movement. The AUC metric generated in the results shows that ANN model has been outperformed by SVM model however only one company stock has been used for prediction.

Another research work carried out for the Shanghai Stock Market by the authors (Shen *et al.*, 2011) has forecasted the performance of stock market performance using a hybrid model. To optimize the RBF (Radial Basis Function), the AFSA (Artificial Fish Swarm Algo.) has been introduced and K-means clustering algorithm is used for improving the efficiency of the model applied above. The conclusion from this research can be drawn out like – RBF and AFSA model together has outperformed all the other models and is quite handy and easy to use; however, the authors haven't focussed on the other factors and only quantitative factors have been considered. Using Ensemble methods have recently gained focus and the authors (Weng *et al.*, 2018) have attempted to predict the stock movement with a data driven approach in 2 phases. In the first phase, the data has been collected using 4 APIs and then in the second phase the machine learning algorithms have been implemented using the ensemble techniques to predict the stock market prices. In the results of this research it has been outlined that the ensemble of the models (BRT and RFR) have performed better than other set of ensembles (SVRE and NNRE)

2.3 Basic models of foreign exchange:

Unlike stock markets, the forex operates 24x7. The prices of different currencies keep fluctuating every time and hence it is one of the toughest challenges to accurately predict the values of currencies. There has been a notable amount of effort spent in this area in attempt to predict the foreign exchange values using machine learning and a review of those researches have been given below.

With the use of general regression NN, an attempt has been made to predict and forecast the forex prices in the work done by (Leung, Chen and Daouk, 2000). The data used for the prediction has been taken from Education dept. of Taiwan. By using the hybrid model techniques, the authors (Ince and Trafalis, 2006) have predicted the forex rates in two stages.

In another attempt to predict forex rates, the authors (Leung, Chen and Daouk, 2000) have forecasted the forex prices using different models and the performances have been compared using a number of statistic techniques and in the conclusion the GRNN model has been found to be the most accurate in prediction of forex rates. (Ince and Trafalis, 2006) used ARIMA model and the number of inputs has been selected in the first phase with co-integration analysis. In the second phase, ANN and SVR has been applied to the models proposed in the phase one. The parametric techniques with the non-parametric techniques have performed better.

2.4 Time-series analysis

Any data set that is having a daily transaction falls under the time-series category. The foreign exchange and stock market are operated daily and this is why most of the researchers use time-series analysis. Now, time-series are of different kinds. The stock market and forex are financial time-series and the authors (Wen *et al.*, 2019) have implemented NN to predict the trends in financial time-series. ARIMA is one of the most widely used models for time-series forecasting. In a similar approach, the author (Zhang, 2003) have implemented a hybrid model of ARIMA and ANN and in the results, it has been observed that the hybrid model has performed better than when individual models have been used. The model used for this research has been implemented in 2 stages. The trend prediction done in the research using the motifs have performed better than the other models have performed. In a similar approach, the authors (Zhang and Berardi, 2001) have used the ensemble model of neural network in order to predict the USD and GBP in a time-series analysis. And the author has concluded that the performance of the ensemble models have performed much better and been able to forecast more accurately in comparison to non-ensemble models.

2.5 Time-series analysis of public sentiment

When it comes to public sentiment, most of the researchers use twitter tweets to analyse the sentiments of the public on any particular day. The reason behind it is strong enough as the twitter data is easily available, and a good volume of tweets can be analysed for each day.

One of the most famous researches done (Bollen, Mao and Zeng, 2011) in the field of stock market prediction has used twitter to do the sentiment analysis of the public mood. They have used GPOMS and Opinion Finder to analyse the tweets downloaded from twitter through the twitter API. After the analysis of the tweets they have segregated the tweets into 6 different emotions like happiness, calmness, anger, etc. The GPOMS tool has become obsolete now and no more used. Similar approach has been tried in (Mittal, 2009) but they developed a new algorithm as they found this approach not very accurate. The Self-Organized Fuzzy Neural Network has been used to predict and forecast the stock price.

2.6 The relationship between stock market and foreign exchange

The relationship between stock market and oil prices has been studied a lot in the recent times and the researchers have done a lot of significant in finding relationship between stock market and forex. In the research work done by (Abul, Haug and Sadorsky, 2012), SVAR model has been used to determine the relationship between stock market and forex rates. Whereas, in the study by (Kang, Ratti and Hwan, 2015), impact of the changes in oil prices have been compared to the volatility in the stock market. The variation in the oil prices can be noticed with respect to the production. If the production shock increases, the oil prices are lowered. And when the there is a shock in economic activity, the oil prices increase. Similarly, with the increase in the price of the stocks, the oil prices also rise. In the research work done (U, 2001), M-VAR has been used and the results tell us that the employment growth and the growth of output are fluctuating with the shocks of the oil prices. The monthly data from the years 1989-1999 has been used to apply empirical analysis. The factors like Interest rates, oil price, Industrial production and industrial employment and Real stock market have been taken in the model implemented. On the other side, (Kang, Ratti and Hwan, 2015) has worked with conditional volatility, implied volatility and realized volatility as volatility of daily measures which are obtained from the price options. The spill-over index, volatility and covariance are statically highly significant when measured and are too large.

The variations in the price causes impact on stock market and this has been observed in the research work carried out by (Cunado and Gracia, 2014), and the countries who trade oil (mostly export) have been taken in to consideration. VEC and VAR models have been implemented for studying the data of 12 European countries. In another research work done by (Abul, Haug and Sadorsky, 2012), the authors have attempted to calculate the impulse response based on the recently developed models and also the models based on standard projection. The models used here outlines that the stock market prices and forex rates (USD) always happen to decrease with increase in the oil price. However, the results of the research (Cunado and Gracia, 2014) shows that European market's behaviour is different and largely depends on the factor which caused the change of the price in oil. The results of the research by (U, 2001) says that price movement of the stock market also sees a depression as the shocks are seen in the oil prices.

The relationship of the stock market and the oil prices in Vietnam has been studied by (Kumar and Narayan, 2010) and empirical model has been implemented. Another research work carried out in a similar fashion by (Mentah *et al.*, 2014) has shown that there is a relationship that exists in the Indian stock market and the prices of crude oil. The oil price is calculated by the trading union in the market prices at spot and the world uses that price decided by the union as a benchmark across the globe. The user sentiment is largely making an impact over the trading decisions. In the research work by (Oliveira, Cortez and Areal, 2017), with the help of the formulas and the KF procedures, the sentiment indicators have been defined. The authors (Kumar and Narayan, 2010) have used the exchange rates to determine the model in order to predict the stock prices. For testing the relationships in the variables, the cointegrations tests have been performed for example Johansen test and structural-break test. The research results show that there exists a correlation in stock prices, oils price and forex rates. Also, from the results it can be established that the stock prices are impacted positively by the oil prices. However, in the research work by (Mentah *et al.*, 2014), it has been concluded that the variations caused in the exchange rates have no impact on either of the stock markets or the oil prices.

The results of the research work carried out by the authors of (Oliveira, Cortez and Areal, 2017) suggest that the sentiment forecasting is not suitable for short-term predictions. The authors of the research work - (Patel, Patel and Patel, 2014) have enlisted all the components of the factors affecting the stock market. Based on the historical data, most of the research work carried in this field has been based on models of time-series and accordingly the predictions have been done. The outcomes of the researches have attained a certain level of prediction accuracy, but as a matter of fact, due to the highly volatile nature of stock markets, it is challenging to predict the exact behaviour of the stock markets.

2.7 Conclusion

In this section, more than 30 technical papers have been reviewed to get a better and clear understanding around the field of study. Related to the field of research work, this literature review section has been divided into four parts, as this is a wide field of study and each part talks about the research work carried out in the past related only to that field of study. Details of a few research papers have been captured in a table (Table 1) and compared with the research work carried out here.

Paper	**Stock prices**	**S&P500**	**Stock Exchange**	**Twitter Sentiment Analysis**	**Approach**
(Liu *et al.*, 2016)	✓	✓	✓	×	SVM
(Wen *et al.*, 2019)	✓	✓	×	×	CNN
(Choudhry and Garg, 2008)	✓	×	×	×	SVM
(Beyaz *et al.*, 2018)	✓	×	×	×	SVR
(Shioda, 2011)	×	×	✓	×	SVM
(Weng *et al.*, 2018)	✓	×	×	×	Ensemble
(Assaf and Alnagi, 2013)	✓	×	×	×	Decision tree
(Ince and Trafalis, 2006)	×	×	✓	×	ARIMA
(Zhang, 2003)	×	×	✓	×	ARIMA
(Yudong and Lenan, 2009)	×	✓	×	×	IBCO-BP
(Aldin, Dehnavi and Entezari, 2012)	×	✓	×	×	ANN
(Oliveira, Cortez and Areal, 2017)	✓	×	×	✓	SVM
Our Model	✓	✓	✓	✓	VAR, LSTM, ARIMA, SARIMAX

Table 1 : comparison of literature review

Based on the review of various literature above, we can now make the decision in an easier and better way to decide the model we are going to use for implementation. The stock markets are mostly impacted by the emotions or the sentiment of the public which has not been covered by most of the researchers, while the other factors have been covered throughout the researches.

3 Research Methodology:

The field of stock markets, forex exchange and oil price is vast and there are countless researches done in relation to these fields and some major contributions have been made to the field of analytics as well. Also, on the contrary, a few authors have concluded that there are certain short-comes in their researches. This gives a motivation to other researchers like me to carry out the tasks related to the future work. The methodology which is most famous amongst the researchers – CRISP-DM (Figure 1) has been used for the implementation of this research work, as inspired from the work done by (Assaf and Alnagi, 2013). Each stage of this methodology has been explained in detail further below.

3.1 Business understanding

For carrying out research in such a wide topic, we need to have a crisp and clear understanding of the basic elements (in this case – stock market, oil prices and forex market) and at the same time keeping in mind the objective of the research topic. Since, the stock markets is a huge and never ending field, and with the time constraints, the scope of this research has been restricted to one particular component of the American stock market – S&P 500 (Standard & Poors 500) which is an index of top 500 companies in American stock market. Other than this, a thorough study has been done on the forex market, oil prices, gold prices and most important – twitter sentiment analysis for understanding the public sentiment. All of these components are studied from American market's point of view.

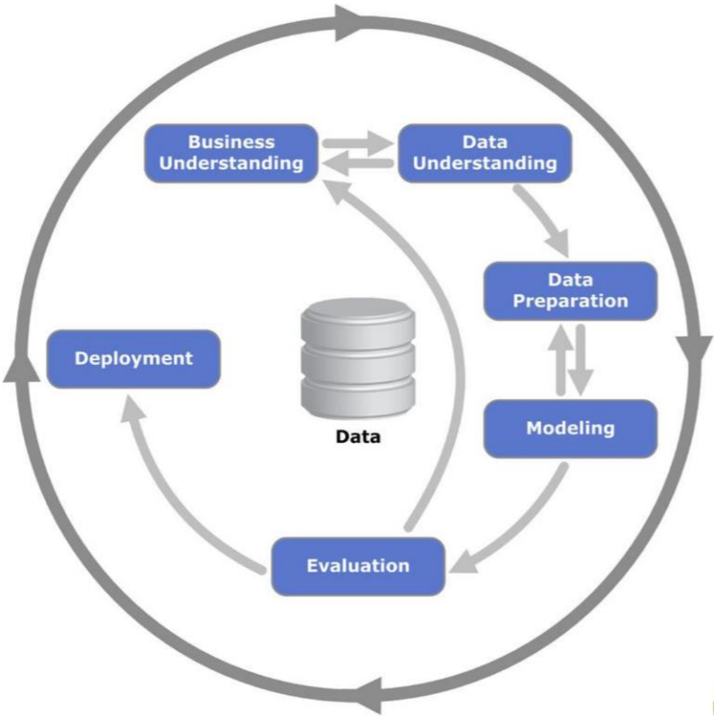


Figure 1: CRISP-DM model

3.2 Data understanding

For carrying out the research work, the first point to work on is gathering the data and then get a thorough understanding of it. There are multiple data sets that will be needed in the implementation of this research work, those are – stock market dataset (S&P 500), sentiment analysis dataset, forex rates dataset (EUR-USD), oil price dataset and gold price dataset. The data needed for the time-series analysis of stock market has been gathered from the publicly available website – yahoo finance and the data has directly been downloaded as a csv format. The dataset for forex rates has been taken from Kaggle which is available in a csv format again. The tweets from twitter has been extracted using anaconda prompt commands and more than 150,000 tweets have been extracted. From a research work done by (Fellers, 2016), it is clear that a relationship has been existing between stock market prices & the forex rates, but the methods used in the research are financial formulas and no time-series analysis has been used. The datasets used in this research have been explained in more details further here:

3.2.1 Dataset Background

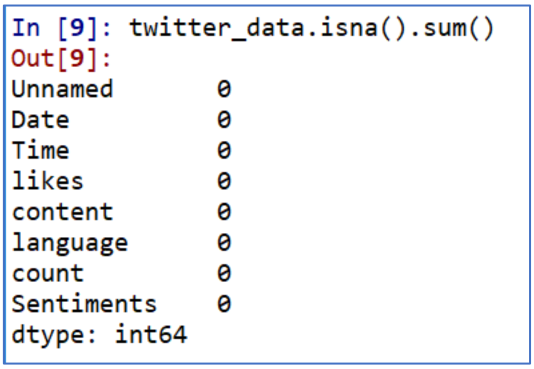
The extracted csv from the yahoo finance website comes with a file with 7 columns having details about each date's open price, close price and highest point reached and the lowest point on that date, then adjusted closing price and the volume of the stocks traded on that date. On the website, all the historical data is available, but we need only about 5 years of data to stay within the scope of our research. The tweets gathered from twitter consists of the date on which the tweet was posted, the content of the tweet and the language it is posted in. For the research, we have taken only English tweets in consideration as of now, but for future work, we can expand the scope and do an analysis of other languages as well because people of all nationalities can be found in America. Based on this, sentiment analysis has been done and positive, neutral and negative tweets have been categorised. Later, a few calculations have been done which are important for the analysis. The crude oil data set has been obtained from Kaggle and prices have been given for each day. Since we are just interested in the price of the oil, only close price of each day has been taken into consideration. The raw datasets gathered are having many unwanted data/columns. We need to perform data cleaning before applying any models to the data set.

3.3 Data preparation

3.3.1 Data cleaning

Before applying any models or doing any computations, the data has to be cleaned thoroughly and, in a machine understandable type. Therefore, here this step is considered to be the most important step as any mistake in this part will lead to erroneous results. The cleaning steps have to be the data where special characters are there, data with symbols or missing data like NA values. All these have to be cleaned and in this research work, python has been used to deal with all the data cleaning and gathering process.

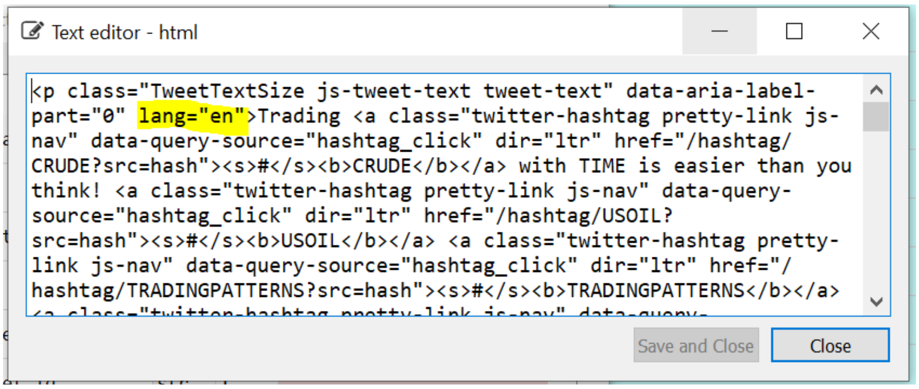
The example of the data cleaning at the very basic level is checking the number of null values in a data set. In python, this can be done very easily just by calling the isna() function after the dataframe name (as shown in Figure 2 below).



```
In [9]: twitter_data.isna().sum()
Out[9]:
Unnamed       0
Date          0
Time          0
likes         0
content       0
language      0
count         0
Sentiments    0
dtype: int64
```

Figure 2: is null code.

The datasets have been cleaned separately and after the cleaning, they all have been merged as and where needed. The stock market dataset has been cleaned in a way that only closing price column was required from the research point of view. Hence all the other columns have been removed in the cleaning before the model is applied. Next, the twitter dataset. The tweets have been first taken from twitter and then imported into python. The tweets have been filtered by the English language and the special characters have been removed. The tweets when extracted from twitter, contains tags (Figure 3) and one of the tags in it is lang. If the value of 'lang' is equal to 'En', the tweet is saved in a list and that list is further used for sentiment analysis.



```
<p class="TweetTextSize js-tweet-text tweet-text" data-aria-label-
part="0" lang="en">Trading <a class="twitter-hashtag pretty-link js-
nav" data-query-source="hashtag_click" dir="ltr" href="/hashtag/
CRUDE?src=hash"><s>#</s><b>CRUDE</b></a> with TIME is easier than you
think! <a class="twitter-hashtag pretty-link js-nav" data-query-
source="hashtag_click" dir="ltr" href="/hashtag/USOIL?
src=hash"><s>#</s><b>USOIL</b></a> <a class="twitter-hashtag pretty-
link js-nav" data-query-source="hashtag_click" dir="ltr" href="/
hashtag/TRADINGPATTERNS?src=hash"><s>#</s><b>TRADINGPATTERNS</b></a>
```

Figure 3 : downloaded tweet content

Before carrying out the Granger Causality test, the positive tweets ratio also has been calculated. This has been achieved by doing the calculations in python and the number of positive tweets on one day has been divided by the total number of positive and negative tweets on that same day. The volume of tweets has also been calculated by counting the total number of tweets each day. This gives us two different datasets of twitter sentiment analysis. One with the ratio of the positive tweets and the other one with the volume of the tweets.

3.3.2 Exploratory Data Analysis

An exploratory data analysis is done when we need to get a basic understanding about the data and gain critical insights of the dataset. The correlation amongst the variables, outliers, trend analysis, and check for class imbalance are a few checks that can be performed at this level. For defining the predictor variables, a few of the researchers have used Random Forest technique to generate the plot for variable importance. On the basis of the exploratory data analysis done, the techniques for the data transformation can be decided and then later on the decision to confirm the models can be made.

3.4 Modelling

For doing the analysis right, the right model has to be applied and it is selected at this stage. From previous literature review, and study of various prediction models, the model has to be decided. Also, the type of data is one of the drivers in choosing the model. If the dataset is having transactions of over a good time-period, then time-series based models have to be applied. The data that has been gathered in the earlier stages is first trained to the model and then the trained model is evaluated with the test data and if the model is able to predict the values of the test data correctly or close to correct, then the model can be said to be a good fit. From previous literature review, it is clear that the model that is going to be implemented for this research work has to be a time-series based model. A few examples of time-series based models are – RNN, VAR, ARIMA and SARIMAX etc. The model implemented in this research has been explained in detail in the implementation section again.

3.5 Evaluation

The evaluation of the model is important to know if the applied model has performed good or not. This can be done by comparing the outcomes of the train data and the test data. All the factors that determine the accuracy of the model have to be assessed carefully and if there are any un-answered issues, they have to be handled.

3.6 Deployment

In this stage, the documentation and reporting work is done, and the knowledge gained must be documented properly. The literature reviewed is properly referenced and each and everything that has been taken from some other work has to be referenced. All the models that have been implemented have to be discussed and enlisted in detail in the report work.

4 Implementation and Evaluation of Stock Market prediction models

4.1 Introduction

The prediction models and techniques carried out in this research work have been explained in detail in this section and is followed by the evaluation techniques. With the data driven approach in mind, the process has been split in to 3 stages, as discussed earlier in the methodology section.

4.2 Process flow diagram

The process that is being followed in the research work has been explained and put up in a flow diagram (Figure 4) below. As mentioned earlier, the process has been split in 3 stages, the first and the foremost stage is the data layer which holds all the data necessary for the models. This is clearly the most important stage as the research work is being carried out with a data driven approach.

The data is first collected from all the sources and then imported into python for the data cleaning purpose. Next, the data that is gathered, is cleaned and exploratory data analysis is carried out and is transformed later to make it ready to be used for modelling. In the next layer, the models are applied, and a business logic is created. In this first, the models that are compatible for the data and which meet the objective of research, are decided and implemented. Next the models implemented are evaluated and performance is measured based on different parameters. The different evaluation techniques to be used are explained further in this report. The final stage of this research is with the client side where graphs and other visual representations are developed to explain the results of the research work carried out.

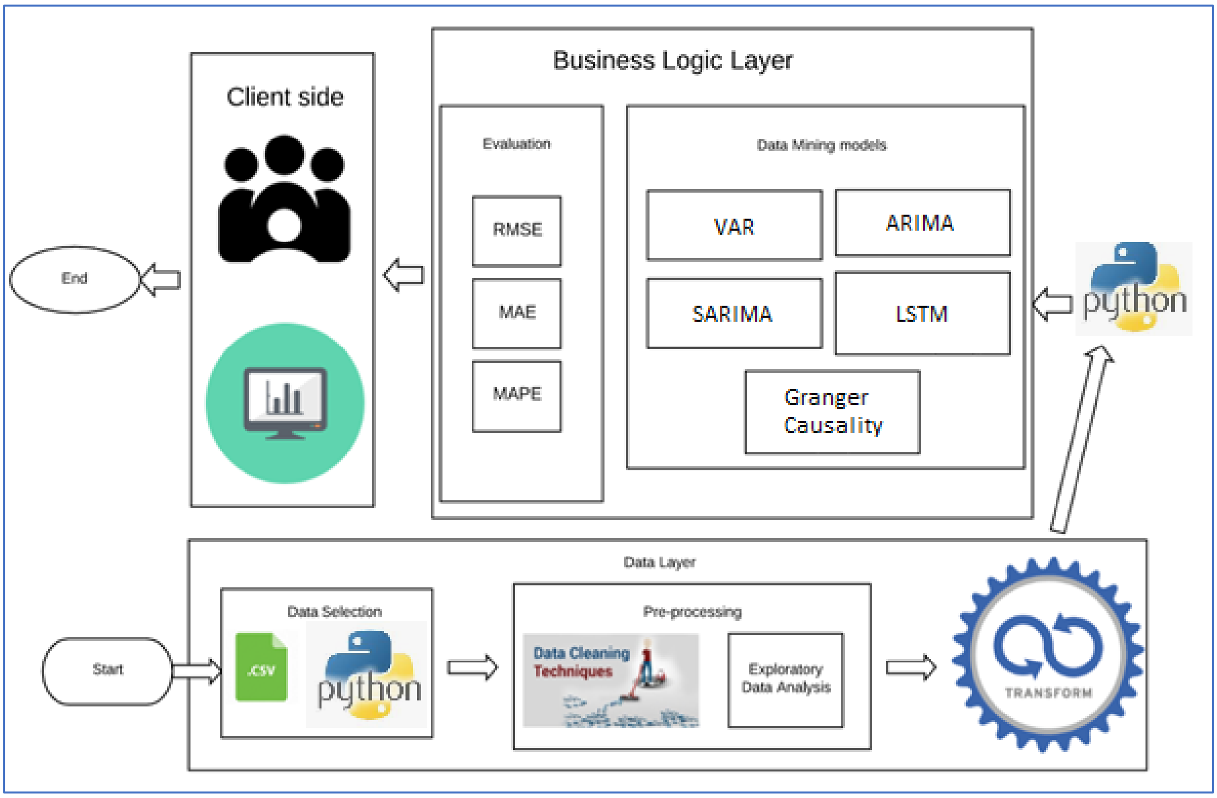


Figure 4: Process Flow Diagram

4.3 Data Models:

The objective of this research paper is to analyse the stock market trends and predict the values based on historical data. And also, to check till what extent the sentiments or public emotions impact the stock market. There have been a lot of researchers who have studied this field and have used n-number of data models including SVM, RF, Neural Networks, etc. However, this research is based on multi-variate data and VAR is the model which is best suitable for such type of data. Other models used here are – ARIMA, SARIMA, LSTM and Granger Causality for checking the causality of the variables in the time series.

4.3.1 Granger Causality test

Now, we have taken many data sets for this research. We have to first check which variables are a causal for the stock market. This can be done with the help of Granger Causality test. In **Granger Causality test**, we can check whether a variable (suppose x1) is a causal for another variable (suppose x2). After this test is done, the datasets which show a causality for the stock market will only be modelled in the VAR. In this research work, after carrying out the Granger Causality test, we found that only the twitter tweets are a causal for stock market prediction because the p value is significant ($p<0.05$) for only that relationship and all other have shown non-significant behaviour. The datasets gathered are for daily transactions of each kind, hence there are two time series that we have to predict.

4.3.2 Vector Auto Regressive (VAR) model

This is one of the most important tools for forecasting for the multi-variate time-series predictions. Since the VAR model has multiple variables, hence it includes multiple equations as well. The VAR model equations are stated as follows:

$$\boxed{y_t = A_0 + \sum_{i-1}^{p} A_i y_{t-i} + u_t,}$$

here, yt - column vector,

A_i - kxk unknown coefficient matrix,

A_o - deterministic constant,

u_t - error column vector.

The error correction:

$$\boxed{\nabla\lambda^{t} = \coprod\lambda^{t-1} + B^{0} + \sum_{b-1}^{i=1}\nabla B^{i}\lambda^{t-i} + \Lambda^{t,}}$$

here, Δ - difference operator,

Y_{t-1} - error correction,

п - speed of adjustment matrix,

B_o - deterministic constant term,

v_t - error column vector.

Based on model proposed above, a multi-variate prediction model has been implemented for this research work. The RMSE value for stock price prediction is 33.29 which is good because the predicted value is in a range more than 2000 and hence the error percent is low. The RMSE for stock prediction is very low near about 0.18 which is again a high-quality prediction.

4.3.3 LSTM

LSTM (Long Short Term Memory) network is a special kind of RNN (Recurrent Neural Network) which is used for long-term dependencies learning. Cells or the memory blocks are the building blocks of a LSTM network. There are two states in a cell which are the cell state and the hidden state. The memory blocks are having gates which remember the things and manipulates them. There are three gates in LSTM – Forget gate, Input gate and the Output gate. In this research the LSTM is used as a linear stack of layers or a sequential model. In the first layer of the LSTM model, 30 memory units have been used and they return sequences. The next layer receives sequences and a dropout is added to each layer of LSTM in order to avoid model overfitting. After implementing the model, the train RMSE is 27.34 and the test

RMSE is 33.64 which is a good score as the stock market's index is more than 2000 and an error around 30 is negligible. LSTM has performed well as per the dataset and is good from stock market's point of view.

4.3.4 ARIMA

The ARIMA model is widely used by researchers for time-series based forecasting. ARIMA is a model used for prediction of future values based on the past values of the own time series. Any time-series which is not having seasonal data can be used in ARIMA model. The main 3 characters of an ARIMA model are p, d and q where p stands for AR terms, q stands for MA term and d stands for difference numbers for converting to stationary time-series.

In the ARIMA model, the time-series is differenced at the least once in order to make it stationary. The equation for ARIMA is:

$$\boxed{Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}}$$

The results for ARIMA in the dataset have been impressive as the MAPE for this model results are 1.59% which means the predicted values are having an error of just 1.59%.

4.3.5 SARIMAX

If the dataset of the time-series shows seasonality, SARIMA can be used. However, if an exogeneous variable can be introduced in the model, SARIMAX can be used. The speciality about SARIMAX model is that it can deal pretty well with the missing values in the datasets. The general equation of a SARIMAX model looks like this:

$$\boxed{\begin{gathered} Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} \\ + \omega_t \end{gathered}}$$

Where, $X_{1,t}$, $X_{2,t}$ … $X_{k,t}$ – k external variables,

β_0, β_1, … β_k – regression coefficients,

ω_t – stochastic residual

The general SARIMAX model comprises of 5 steps which are parameter estimation, model identification, fitness of the model test, including external variables and last forecasting & validation. The mean absolute percentage error for this model after evaluation comes around 3.16%.

4.4 Validation and Evaluation:

The validation of the work done can be done from other researches done related to this field. However, the dataset taken in this research is different from the dataset used by the other researchers as they have taken the values of each stock and here in this research we have used the value of S&P 500 index as a whole and predicted its value. Other than that, in this research, the historical data set has been taken from twitter and then sentiment analysis has been done.

4.4.1 Root Mean Square Error

The standard deviation of prediction errors (residuals) from regression line is known as RMSE. It represents the residual spread. It also represents the density of the data points near the line of best fit. It is widely used in forecasting for the verification of the results of experiments performed.

4.4.2 Mean Absolute Error

The MAE is quite similar to RMSE. The only difference is that in MAE, absolute value of the difference is taken into consideration instead of square of the difference between predicted value and actual value.

$$\boxed{\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}}$$

4.4.3 Mean Absolute Percentage Error

In MAPE, the error value is measured in the terms of percentage.

$$\boxed{\mathrm{MAPE} = \frac{\sum \frac{|A-F|}{A} \times 100}{N}}$$

Here, A – actual value, F – forecast value & N – number of observations.

The evaluation of the results has been done on the basis of the three basic evaluation models which are MSE, RMSE, MAE and MAPE. On the basis of these three, evaluation metrics, the models have been evaluated. The VAR model is applied for multivariate time-series predictions. Since there are many datasets taken into consideration, we perform the **Granger-Causality test** (Table 2) to check the causality of all the variables for predicting the stock market. The results of the Granger Causality for each variable against the stock market price is given below in the table:

	stock market		
	P value (1st Lag)	P value (2nd Lag)	P value (3rd Lag)
twitter volume	0.00*	0.00*	0.00*
twitter posratio	0.04*	0.22	0.32
crude oil	0.79	0.85	0.31
exchange price	0.67	0.72	0.53
gold price	0.57	0.46	0.62

Table 2 : Granger Causality results (P-values)

After analysing the significance of the variables, only volume of the tweets has shown causality for the stock market index predictions. In the table above, asterisk (*) symbol represents that the values are significant. Hence the VAR regression model has been applied only for the stock market vs volume of tweets posted on any date. The root mean squared error for stock market has been captured to be 33.29 and mean absolute error is 26.90 (Figure 5) and these errors are really small in front of the predicted values.

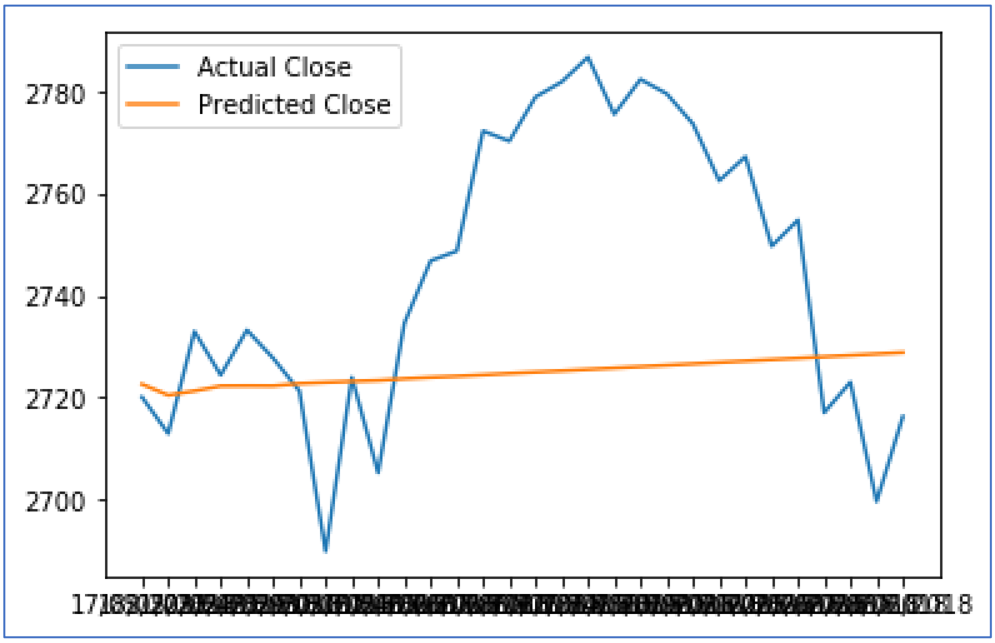


Figure 5 : VAR model results 1

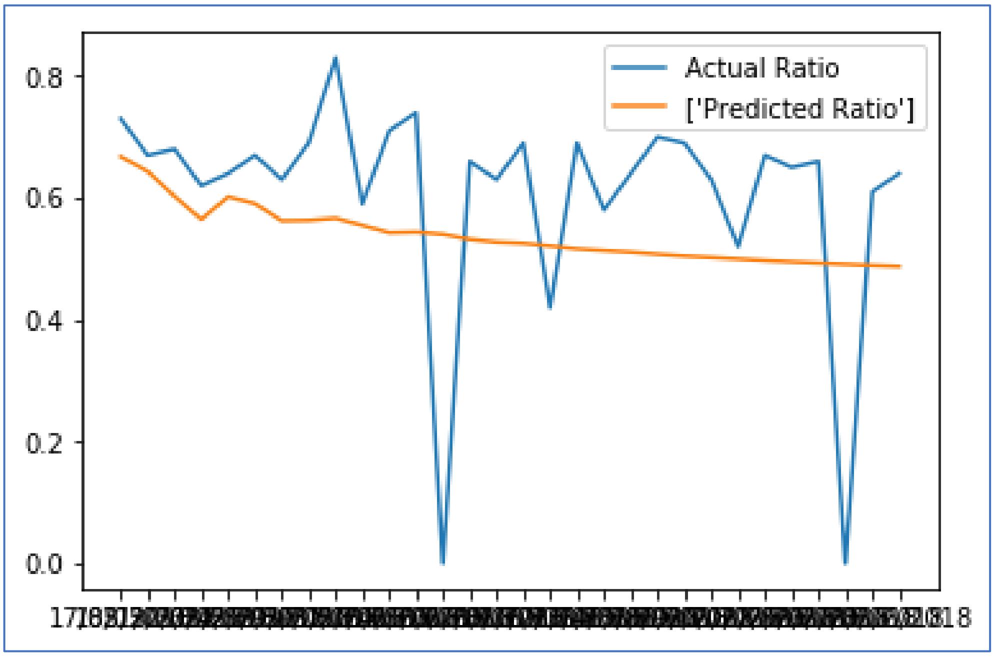


Figure 6 : VAR results 2

On the other hand the root mean squared value for the twitter sentiment analysis has shown great results as well (Figure 6) – 0.18 and the mean absolute error is 0.14. The next model that has been implemented is LSTM and the model gave the RMSE value of 33.69 (Figure 7).

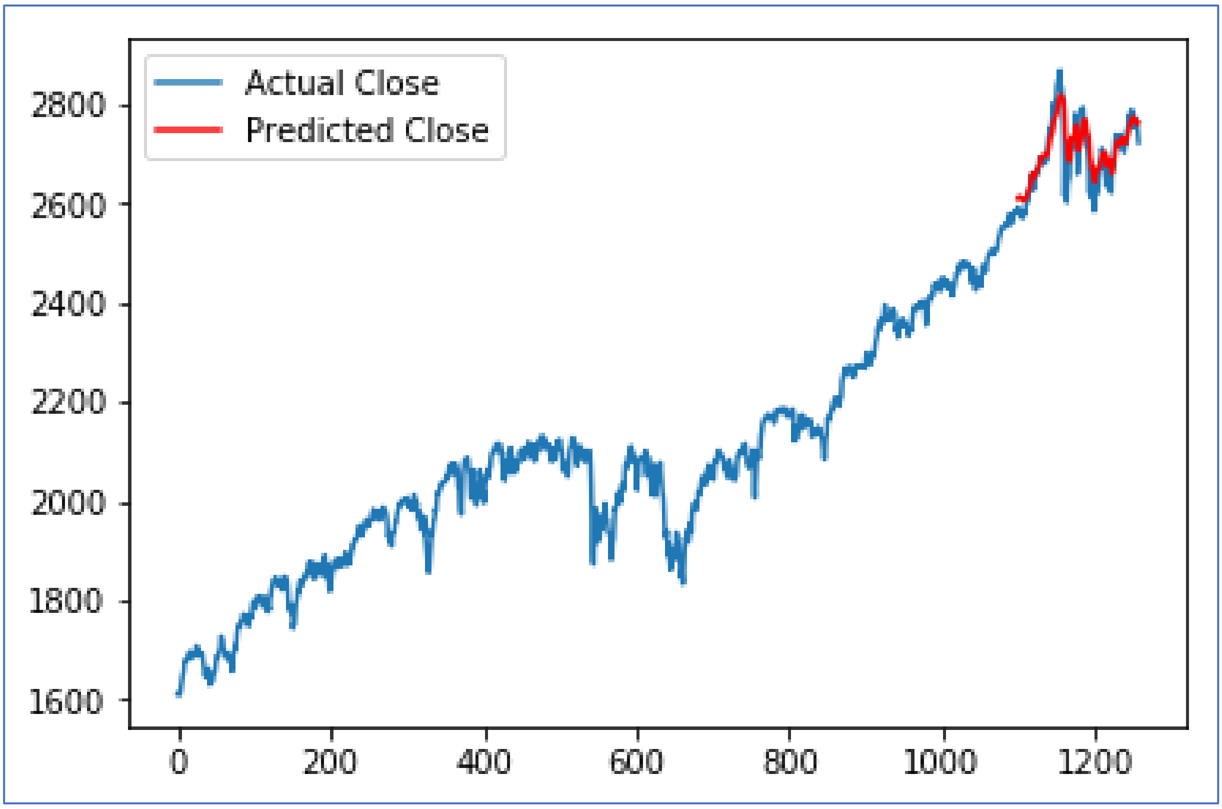


Figure 7 : LSTM result

The LSTM model has performed (Figure 7) a lot better than ARIMA (Figure 8) when compared based on RMSE values because ARIMA RMSE value when calculated comes out to 51.23.

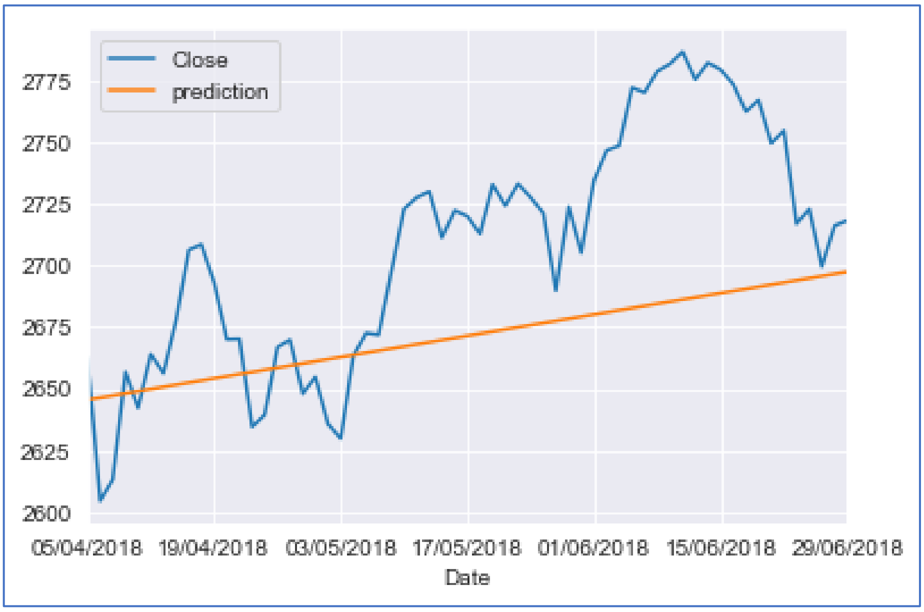


Figure 8 : ARIMA model results

But when the seasonal component is taken into the model, SARIMAX's (Figure 9) mean absolute error value is 3.17% which is high compared to other models and ARIMA's MAE value is 1.59% which is good.

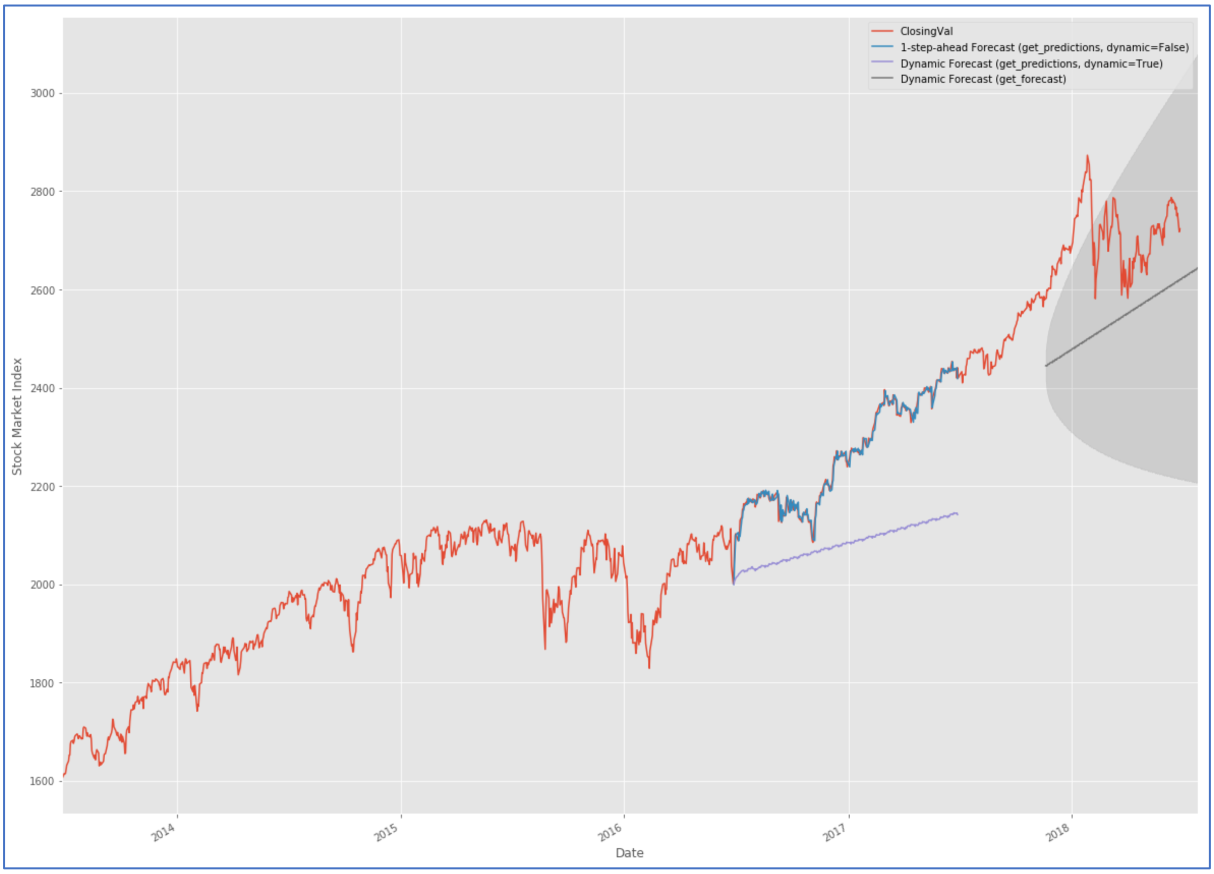


Figure 9 : SARIMAX result

The comparison of all the models based on the evaluation metrics has been described below (Table 3) in the table.

	MSE	RMSE	MAE	MAPE
VAR	1108.84	33.29	26.9	0.99%
LSTM	1135.02	33.69	24.51	0.90%
ARIMA	2625.11	51.23	42.61	1.59%
SARIMAX	12050.84	109.77	85.49	3.16%

Table 3 : evaluation results table

5 Discussion & Conclusion

The results obtained after conducting the research work are satisfactory and as a multivariate prediction model, the VAR has given the best prediction model and the RMSE value is lowest. This model's performance has almost been matched by the model LSTM where the RMSE value is very close to the RMSE of VAR model. The other models implemented in this work are ARIMA and SARIMAX. The ARIMA model's performance has shown higher RMSE value compared to VAR and LSTM but to a surprise, the SARIMAX model has got the maximum error value when compared to all the other models.

This research's prime objective is to analyse and predict the impact of various factors (like public sentiments (twitter sentiment analysis), stock exchange, gold price, oil price) on the stock market index (S&P500) performance in United States. After a careful analysis of all the literatures (previous researches) the models used for the predictions have been analysed and through this the first objective is fully completed and it met the expectations. The researchers have used many models to compare the different prediction models and the comparison has been done using different evaluation techniques.

Since there is a limitation with the number of calls made to the IBM Watson sentiment analysis, the alternate – 'textblob' has been used for doing the twitter sentiment analysis. For the future work, twitter sentiment analysis can be done using the IBM Watson which is known for better results of sentiment analysis and can handle multiple languages easily. This can help predicting the public sentiment more clearly and in return the impact on the stock market can be tracked more accurately.

6 Acknowledgement

I'd like to thank a lot for all the supervision and guidance provided to me by my mentor – Dr. Pierpaolo Dondio and also my parents and friends for keeping me motivated and supporting at all times.

References

Abul, S., Haug, A. A. and Sadorsky, P. (2012) 'Oil prices , exchange rates and emerging stock markets ☆', *Energy Economics*. Elsevier B.V., 34(1), pp. 227–240. doi: 10.1016/j.eneco.2011.10.005.

Aldin, M. M., Dehnavi, H. D. and Entezari, S. (2012) 'Evaluating the Employment of Technical Indicators in Predicting Stock Price Index Variations Using Artificial Neural Networks (Case Study : Tehran Stock Exchange)', 7(15), pp. 25–34. doi: 10.5539/ijbm.v7n15p25.

Assaf, A. A. B. U. and Alnagi, E. (2013) 'Predicting stock prices using data mining techniques 1', pp. 1–8.

Beyaz, E. *et al.* (2018) 'Stock price forecasting incorporating market state', *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, pp. 1614–1619. doi: 10.1109/HPCC/SmartCity/DSS.2018.00263.

Bollen, J., Mao, H. and Zeng, X. (2011) 'Twitter mood predicts the stock market', *Journal of Computational Science*. Elsevier B.V., 2(1), pp. 1–8. doi: 10.1016/j.jocs.2010.12.007.

Choudhry, R. and Garg, K. (2008) 'Choudhry, Garg - 2008 - A Hybrid Machine Learning System for Stock Market Forecasting', 2(3), pp. 315–318.

Cunado, J. and Gracia, F. P. De (2014) 'Oil price shocks and stock market returns : Evidence for some European countries ☆', *Energy Economics*. Elsevier B.V., 42, pp. 365–377. doi: 10.1016/j.eneco.2013.10.017.

Fellers, A. E. (2016) 'Effects of Exchange Rate Changes on S & P 500 Price Movement'.

Ince, H. and Trafalis, T. B. (2006) 'A hybrid model for exchange rate prediction', 42, pp. 1054–1062. doi: 10.1016/j.dss.2005.09.001.

Kang, W., Ratti, R. A. and Hwan, K. (2015) 'Journal of International Financial Markets , Institutions & Money The impact of oil price shocks on the stock market return and volatility relationship', *'Journal of International Financial Markets, Institutions & Money'*. Elsevier B.V., 34, pp. 41–54. doi: 10.1016/j.intfin.2014.11.002.

Kumar, P. and Narayan, S. (2010) 'Modelling the impact of oil prices on Vietnam ' s stock prices', *Applied Energy*. Elsevier Ltd, 87(1), pp. 356–361. doi: 10.1016/j.apenergy.2009.05.037.

Kyun, T. *et al.* (2019) 'Global stock market investment strategies based on financial network indicators using machine learning techniques'. Elsevier Ltd, 117, pp. 228–242. doi: 10.1016/j.eswa.2018.09.005.

Leung, M. T., Chen, A. and Daouk, H. (2000) 'Forecasting exchange rates using general regression neural networks', 27.

Liu, C. *et al.* (2016) 'Forecasting S&amp;P 500 Stock Index Using Statistical Learning Models', *Open Journal of Statistics*, 06(06), pp. 1067–1075. doi: 10.4236/ojs.2016.66086.

Mentah, H. P. *et al.* (2014) 'Crude Oil Price , Exchange Rate and Emerging Stock Market : Evidence from India', 42, pp. 75–87.

Mittal, A. (2009) 'Stock Prediction Using Twitter Sentiment Analysis', (June).

Oliveira, N., Cortez, P. and Areal, N. (2017) 'The impact of microblogging data for stock market prediction : Using Twitter to predict returns , volatility , trading volume and survey sentiment indices', *Expert Systems With Applications*. Elsevier Ltd, 73, pp. 125–144. doi: 10.1016/j.eswa.2016.12.036.

Patel, J. *et al.* (2015) 'Expert Systems with Applications Predicting stock market index using fusion of machine learning techniques', *EXPERT SYSTEMS WITH APPLICATIONS*. Elsevier Ltd, 42(4), pp. 2162–2172. doi: 10.1016/j.eswa.2014.10.031.

Patel, P. P. J., Patel, N. J. and Patel, A. R. (2014) 'Factors affecting Currency Exchange Rate , Economical Formulas and Prediction Models', 3(3), pp. 53–56.

Shen, W. *et al.* (2011) 'Knowledge-Based Systems Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm', *Knowledge-Based Systems*. Elsevier B.V., 24(3), pp. 378–385. doi: 10.1016/j.knosys.2010.11.001.

Shioda, K. (2011) 'Prediction of Foreign Exchange Market States with Support Vector Machine', *2011 10th International Conference on Machine Learning and Applications and Workshops*. IEEE, 1, pp. 327–332. doi: 10.1109/ICMLA.2011.116.

U, E. P. (2001) 'Oil price shocks , stock market , economic activity and employment in Greece ଝ'.

Wen, M. *et al.* (2019) 'Stock market trend prediction using high-order information of time series', *IEEE Access*. IEEE, 7, pp. 28299–28308. doi: 10.1109/ACCESS.2019.2901842.

Weng, B. *et al.* (2018) 'Predicting short-term stock prices using ensemble methods and online data sources', *Expert Systems With Applications*. Elsevier Ltd, 112, pp. 258–273. doi: 10.1016/j.eswa.2018.06.016.

Weng, B., Ahmed, M. A. and Megahed, F. M. (2017) 'Stock market one-day ahead movement prediction using disparate data sources'. Elsevier Ltd, 79, pp. 153–163. doi:

10.1016/j.eswa.2017.02.041.

Yavuz, M. and Ozdemir, N. (2015) ‘Stock Market Index Prediction with Neural Network during Financial Crises : A Review on Bist-100’, (June). doi: 10.18488/journal.89/2015.1.2/89.2.53.67.

Yudong, Z. and Lenan, W. (2009) ‘Expert Systems with Applications Stock market prediction of S & P 500 via combination of improved BCO approach and BP neural network’, *Expert Systems With Applications*. Elsevier Ltd, 36(5), pp. 8849–8854. doi: 10.1016/j.eswa.2008.11.028.

Zhang, G. P. (2003) ‘Time series forecasting using a hybrid ARIMA and neural network model’, 50, pp. 159–175.

Zhang, G. P. and Berardi, V. L. (2001) ‘Time series forecasting with neural network ensembles : an application for exchange rate prediction’, pp. 652–664.