National College of Ireland

Impact Analysis of Market Sentiments, Gold and Crude oil prices on DOW30 stocks.

MSc Research Project
MSc. In Data Analytics

Yash Mehta
Student ID: x18179916

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Yash Nilesh Mehta
Student ID:	X18179916
Programme:	MSc. In Data Analytics **Year:** 2019-20
Module:	Research in Computing
Supervisor:	Mr. Hicham Rifai
Submission Due Date:	27th September 2020.
Project Title:	"Impact Analysis of Market sentiments, Gold and Crude oil prices on DOW30 stocks"
Word Count:	6614 **Page Count:** 20

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 bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yash Nilesh Mehta

Date: 27th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

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

Impact Analysis of Market sentiments, Gold and Crude oil prices on DOW30 stocks

Yash Nilesh Mehta
X18179916

Abstract

Data Analytics and AI has established its roots in stock market trading for past few years. Automated trading developed using AI is capable of taking multiple factors as input to finally predict the stock movement. But researchers are still working on defining the sum of exact factors that influences the stock market. Till date many researches have proved that human sentiments are significantly causal to stock index movement and can help to predict stock prices and trends. This research aims at developing a machine learning model to predict the future trend of DOW30 stocks using multiple factors like market sentiments from twitter and news, gold prices and crude oil prices. Datasets used in this research have been collected from various sources like Yahoo finance for stocks data, Kaggle for news sentiments data and Twitter analysis dataset collected from one of the previous research works. Granger causality test has been implemented to identify the factors that are significant in predicting future trend of stocks. Multiple machine learning models like Vector Auto Regression (VAR), SVM and kNN has been implemented. SVM and kNN's performance has been evaluated using various factors like accuracy, precision, recall, and f1 score. Whereas, VAR's performance has been evaluated using MAPE. The models have also been tested using transfer learning technique and the results showed that the VAR model performed more consistently.

Keywords: *DOW30, Stock Market, Market Sentiments, Crude Oil and Gold prices, Forecasting, stock index prediction, kNN, VAR, SVM, Transfer Learning.*

1 Introduction

The Stock markets are very uncertain in nature which makes it very tough to predict its movement, hence it becomes difficult for traders to identify stocks which will indeed give profitable returns. Due to many such challenges like stock market's behavior and identifying the factors that can cause movement of particular stock index, it becomes very difficult yet interesting task to make a reliable prediction model for stock market.

From the previous researches done in this domain, it can be seen that researchers have concentrated either on prices of gold, oil and other commodities or historical data to make a prediction on stock market movement. But it has also been proved that human sentiments play a huge role in predicting future trend of stocks.

1.1 Background and Motivation

Till date many researches have been performed to understand the behavior of stock market and predict its movement using machine learning and Deep learning techniques. In one of the researches, (Moghar, 2020) have tried to predict the stock market index using LSTM model, where the major objective of research was to identify ideal number of epochs with which the model performs best. Other researches like (Pyo et al., 2017) and (Patel et al., 2015a) have tried to predict the future movement of stock market by implementing and comparing multiple machine learning techniques like ANN, SVM, Random forest, Naïve Bayes to identify the best performing model for forecasting. (Patel et al., 2015b) and (Banik, Khan and Anwer, 2014) have used various hybrid machine learning techniques for predicting future trends of stock market. But all these researches are focused on using historical data and stock market factors, whereas many researches like (Nofsinger, 2005) and (Bollen, Mao and Zeng, 2011) have proved that human sentiments also play a very significant role in predicting stock market trend. Hence this research focuses on testing the significance of multiple factors like human sentiments, Gold prices and Crude oil prices for predicting the stock market trend. Secondly, building multiple machine learning models like VAR, SVM, and kNN using most significant factors to predict the trend of stock market.

1.2 Research question

"Does public sentiments, gold and crude oil prices have significant causality on stock market movement? If yes then, to what extent can these factors help in predicting future trend of stock market?"

1.3 Research Objectives

Objective1 – To discuss and review previous researches done in this research area.

Objective2 – To identify factors causing significant causality on DOW30 stocks.

Objective3 – Implementing multiple machine learning models to predict the stock market trend using multiple causal factors, and evaluating these models.

Sub-Objective (a) – The data needs to be cleaned and transformed

Sub-Objective (b) – Building machine learning models to predict future trend of DJIA stocks

Sub-Objective (c) – Evaluating performance of models using corresponding evaluation tests.

1.4 Outline of the paper

Rest of the report is structured as follows: Section 2 discusses some of the previous research works done in this field and explains different methods and machine learning models used by these research works to predict the stock market. Section 3 discusses and explains datasets and methodology that has been used for achieving objectives of this research. Section 4 describes design and implementation of this study. Section 5 details the evaluation of models used to define performance of each implemented model. Section 6 provides final conclusion of this research and also enlists future work that can be done on this study to improve the usability.

2 Literature Review

2.1 Introduction

This section provides support for defining the research goals by performing technical review of various researches previously done in the same domain/area of this research. This section is further divided into four subsections, where the first section discusses various machine learning and deep learning models used by previous researches for purpose of predicting stock market index. The second subsection reviews various researches done to identify the impact of Gold and Crude oil on movement of stock prices. Final subsection provides an overview of various papers which explains the importance of sentiment analysis in field of finance.

2.2 Stock prediction using machine learning and historical data.

Many researchers have tried to predict stock market index using various machine learning models, but few models in specific like Recurrent Neural Network (RNN), Artificial Neural Network (ANN) and SVM have always showed good performance and accuracy, and hence these particular models are widely used in researches. One of these researches done by (Patel *et al.,* 2015a) implemented and compared prediction performances of ANN, naïve bayes, SVM and Random Forest for predicting future trends of stocks. Similarly (Pyo *et al.,* 2017) used ANN and SVM to forecast the future trends of Korean stock market. The process and techniques implemented by (Patel *et al.*, 2015b) is different as the author has developed the study in two stages: In first stage, study uses Support Vector Regression (SVR), ANN and Random Forest to make predictions on stocks. In second stage, the authors have built various hybrid machine learning models by fusing SVR, ANN and Random Forest. Results showed that hybrid models performed more accurately and precisely.

2.3 Impact of Gold and Oil prices on Stock Markets

Researches are being carried out to find various factors that are significant in predicting stock market. Many researches have used multiple factors like technical indicators of stocks, volume of trade, and many more. But there are many researches that have tried to

find impact of commodity prices like Gold and Crude oil on prediction of stock prices. One of such research (Coronado, Jiménez-rodríguez and Rojas, 2018) performed an empirical study of finding the correlation between Crude oil, Gold and stock Market using Granger Causality test. The study found evidence that there is a bi-directional relationship between commodity (Gold and Crude oil) prices and stock prices. This means that the change in prices of stocks can be predicted by analyzing change in commodity prices and vice versa. Another study done by (Vveinhardt et al., 2017) provides an evidence that Gold and oil prices have an asymmetric influence on stock prices. This study used two different techniques i.e. Granger Causality and Bound testing (ARDL) approach and showed that there was no long-term relationship between stock prices and commodity prices, but presence of some relationship for short term period was seen.

2.4 Importance of Sentiments analysis in stock market prediction

Many researches have previously analyzed causality of market sentiments on movement of stock prices and have resulted that market sentiments have a significant impact on stock indexes and can help in predicting future prices of stocks.

(Bollen, Mao and Zeng, 2011) has used Granger Causality test to find the causality of market sentiments or social mood on changes in stock prices. The test showed that public sentiments are significantly causal to stock prices and can help in predicting the future trends of stocks. For purpose of sentiment analysis, study implements GPOMS and Opinion finder. A fuzzy Neural Network was built to predict the changes in closing values of stocks using twitter sentiments. Study resulted in accuracy of 86.7% for predicting the future values of stocks. Major drawback of study is that it performs sentiment analysis only on one language of tweets.

Another famous study done in this field (Nofsinger, 2005) works on finding the impact of social mood on stock market and gets two main outcomes: 1) If the public mood is optimistic than the prices of stocks increases and if public mood is pessimistic than values of stocks decreases. 2) A stock market bubble can be created due to extremeness in public mood, due to which the stocks become either overvalued or undervalued. (Porras, 2019) analyzes the impact of social sentiments on volume of stock index. Study successfully fulfilled its main objective by proving a positive relationship between social mood and volume of stock index. Though this study was successful in finding correlation between social mood and stock index, it failed to predict the future stock prices using public sentiments.

(See-to and Yang, 2017) focuses on finding impact of dispersion of market sentiments on returns and volatility of stocks. Authors employed opinion mining techniques and Naïve Bayes classifier to classify the tweets and calculate the dispersion of sentiments. Study successfully fulfilled its objective by proving that volatility of stocks, stock prices and market sentiments are positively correlated to each other.

Work done by (Birbeck and Cliff, 2019) is slightly different than all the studies reviewed,

as authors have used unique approach of using stock prices for sentiment analysis. Study has incorporated the logic that if price of stock rises than the public mood is positive and if the prices of stocks decreases than public mood is negative. Using this logic, sentiment analysis is performed and dataset is built where the tweets are labelled as buy/sell instead of positive/negative. Finally, a Bayesian classifier is used to make predictions for stocks.

One of the base papers (Jain, 2019) considered for this study, has used Granger Causality test to analyze the impact of twitter sentiments, Gold prices, Crude oil prices and Forex rates on stock prices. Study found that twitter sentiments and volume of tweets had significant impact on stock prices whereas, Gold, Crude oil and Forex rates are not significantly causal on stock prices. The study successfully fulfilled its objective by identifying causal factors and building various machine learning models viz. VAR, LSTM, ARIMA and SARIMAX to predict the stock market movement.

When it comes to sentiment analysis majority of researchers prefer twitter data or other social media data. But news sentiments also play an important role in sentiment analysis. Many researches have used news data to predict the stock market index. One of such researches (Yu *et al.*, 2015), authors have used a new technique in which they consider the decay of sentiments and flow of news to calculate the impact of news. Study showed that news sentiments could not help in increasing the prediction capability of market volatility.

Work done by (Kalra and Prasad, 2019) used sentiments calculated from news articles, to make predictions on movement of stock market. Naïve Bayes classifier is used to classify the news articles as negative or positive. Various machine learning models i.e. SVM, kNN, Naïve Bayes and Neural Networks are implemented and compared for purpose of stock price prediction using news sentiments. Results concluded that kNN model performed the best.

(Sirimevan et al., 2019) used both twitter and news sentiments to make predictions on stock prices. Study implements LSTM and an ensembled machine learning technique for purpose of making predictions on stock market index using public sentiments, and historical financial data. Study was able to make predictions with high accuracy using ensemble method.

Table 1 provides overview of few of the relevant researches reviewed above. It also compares the researches with study carried out here.

Table 1 Comparison of related works

Research Paper	Historical stocks data	Twitter Sentiments	News Sentiments	Transfer Learning	Approach
(Patel *et al.*, 2015a)	Yes	No	No	No	ANN, Naïve Bayes, SVM, Random Forest.
(Bollen, Mao and Zeng, 2011)	Yes	Yes	No	No	Granger Causality, Fuzzy Neural Network
(See-to and Yang, 2017)	Yes	Yes	No	No	Linear Regression, Support Vector Regression.
(Birbeck and Cliff, 2019)	Yes	Yes	No	No	Bayesian Classifier
(Jain, 2019)	Yes	Yes	No	No	Granger Causality, VAR, ARIMA, SARIMAX, LSTM
(Kalra and Prasad, 2019)	Yes	No	Yes	No	SVM, kNN, Naïve Bayes and Neural Networks
(Sirimevan *et al.*, 2019)	Yes	Yes	Yes	No	LSTM, ensemble machine learning.
This study	Yes	Yes	Yes	Yes	Granger Causality, VAR, SVM, kNN.

Based on the detailed review of researches done in this section, it was easy to select the models and techniques for this study. Also, none of the researches compared in table 1 have used transfer learning technique to check the robustness of models. Stock market is very unstable and frequently changing also the sentiments varies very frequently, hence it is important for model to be stable and maintain the prediction accuracy. Hence this study has incorporated transfer learning technique to check cross-data performance of models.

3 Research Methodology

This research follows CRISP-DM methodology which is ideally preferred by researchers. This methodology has various stages and further sections of this research explains briefly all the stages of CRISP-DM methodology followed in this study. Selection of research methodology is inspired by (Jain, 2019).

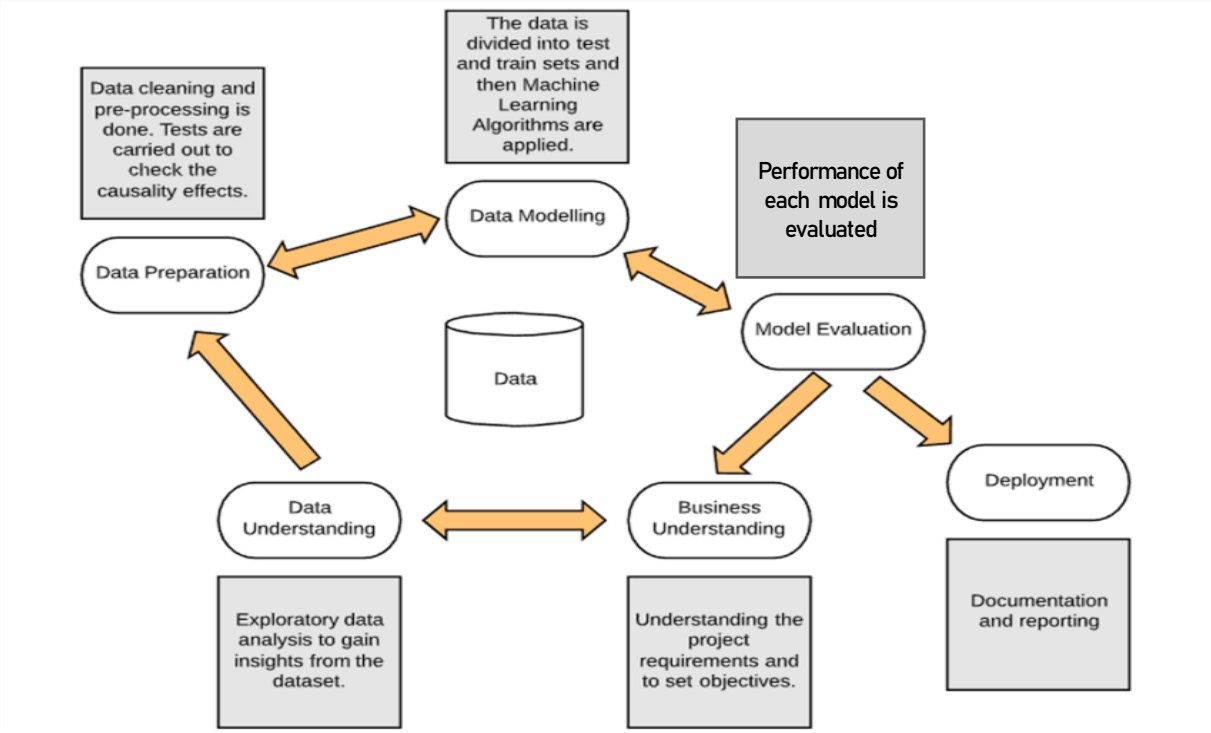

Fig. 1 DJIA prediction methodology

3.1 Business understanding

Stock market is an enormous domain and hence the area of this research i.e. predicting stock market trend using various factors is also very wide. Due to limited time available for this research work, only DJIA (Dow Jones Industrial Average) stocks have been considered. DJIA is an index for 30 large publicly owned blue chip companies. Also, to set the objective of this research, gold prices and crude oil prices have been studied. Impact of public sentiments on DJIA stocks is also studied by performing sentiment analysis on news data and twitter tweets which are targeted for DOW30.

3.2 Data understanding

After understanding the requirements and setting up the objectives for research, next step is to gather and understand the data. To meet the objectives of this research, multiple datasets are required. These datasets are as follows: DJIA stocks data, Crude oil prices, Gold prices, News sentiments and Twitter sentiments datasets. These datasets are further explained in detail in following section.

3.2.1 Dataset Background

The DJIA stocks, crude oil prices and gold prices datasets are publicly available and are extracted from yahoo finance[1] in a csv format.

[1] https://finance.yahoo.com/

Table 2. Overview of datasets

```
dfCrude.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2650 entries, 2010-01-04 to 2020-07-10
Data columns (total 6 columns):
Open         2650 non-null float64
High         2650 non-null float64
Low          2650 non-null float64
Close        2650 non-null float64
Adj Close    2650 non-null float64
Volume       2650 non-null int64
dtypes: float64(5), int64(1)
memory usage: 224.9 KB
```

Fig.2 Overview of Crude oil dataset

```
dfGold.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2649 entries, 2010-01-04 to 2020-07-10
Data columns (total 6 columns):
Open         2649 non-null float64
High         2649 non-null float64
Low          2649 non-null float64
Close        2649 non-null float64
Adj Close    2649 non-null float64
Volume       2649 non-null int64
dtypes: float64(5), int64(1)
memory usage: 224.9 KB
```

Fig.3 Overview of Gold dataset

```
df.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1387 entries, 2013-06-28 to 2018-12-31
Data columns (total 6 columns):
Open         1387 non-null float64
High         1387 non-null float64
Low          1387 non-null float64
Close        1387 non-null float64
Adj Close    1387 non-null float64
Volume       1387 non-null int64
dtypes: float64(5), int64(1)
memory usage: 75.9 KB
```

Fig.4 Overview of DOW30 dataset

```
twitter_data.info(verbose=True)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 163050 entries, 0 to 163049
Data columns (total 5 columns):
Date          163050 non-null datetime64[ns]
likes         163050 non-null int64
content       163050 non-null object
language      163050 non-null object
Sentiments    163050 non-null object
dtypes: datetime64[ns](1), int64(1), object(3)
memory usage: 6.2+ MB
```

Fig.5 Overview of Twitter dataset

```
dfnews.info()

<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 1989 entries, 2008-08-08 to 2016-07-01
Data columns (total 26 columns):
Label    1989 non-null int64
Top1     1989 non-null object
Top2     1989 non-null object
Top3     1989 non-null object
Top4     1989 non-null object
Top5     1989 non-null object
Top6     1989 non-null object
Top7     1989 non-null object
Top8     1989 non-null object
Top9     1989 non-null object
Top10    1989 non-null object
Top11    1989 non-null object
Top12    1989 non-null object
Top13    1989 non-null object
Top14    1989 non-null object
Top15    1989 non-null object
Top16    1989 non-null object
Top17    1989 non-null object
Top18    1989 non-null object
Top19    1989 non-null object
Top20    1989 non-null object
Top21    1989 non-null object
Top22    1989 non-null object
Top23    1988 non-null object
Top24    1986 non-null object
Top25    1986 non-null object
dtypes: int64(1), object(25)
memory usage: 499.6+ KB
```

Fig.6 Overview of news dataset

These datasets have details about DJIA stocks from year 2013 to 2018, crude oil prices and gold prices from year 2010 to 2018. These datasets have 7 columns each, which provides the values of opening price, highest price, lowest price, closing price, adjacent closing price and trading volume for each date of DJIA stocks, crude oil and gold prices respectively. DJIA stocks dataset has 1387 rows, Gold dataset has 2649 rows and Crude oil dataset has 2650 rows.

The twitter sentiments dataset has been collected from one of the previous research works done in similar area (Jain, 2019). Originally, author has scrapped the twitter tweets and performed text analysis on collected tweets dataset to determine the sentiment of tweet

i.e. negative or positive. This dataset has total 1,63,050 tweets from year 2013 to 2018 as shown in figure 5. This dataset has 5 columns in total which describes the number of likes on each tweet, original content of the tweet, language and sentiment of each tweet i.e. either positive or negative.

The news dataset is downloaded from a publicly open platform [Kaggle][2] in csv format. This dataset has total 1989 news articles from year 2008 to 2016. There are 26 columns in total which describes the sentiment of articles on each date i.e. negative or positive, and top 1 to top 25 columns are the actual news headlines for corresponding date.

3.3 Data preparation

3.3.1 Data cleaning

Before implementing any models or performing the causality tests, all the datasets were checked thoroughly for any missing values or NA values. It is very important to clean the dataset because noise in data can lead to errors in results. Cleaning the NA values and noise enhances the performance of models, which results in achieving optimum outcomes. In this stage, firstly the check for missing values was performed using simple isna() function in python shown in figure 7.

```
df.isna().sum()
Open          0
High          0
Low           0
Close         0
Adj Close     0
Volume        0
dtype: int64
```

Fig. 7 Checking for NA values

There are no NA values in any of the datasets. The datasets were also combined wherever required. Depending on the requirement, the datasets were merged either using python code i.e. df.merge() or using Microsoft excel itself. Positive ratio of tweets is calculated in python before calculating the causality of each factor on stocks. For calculating the positive ratio, tweets were firstly grouped by the means of date and then total number of positive and negative tweets for each day were calculated, later the number of positive tweets on each day were divided by total number of tweets on that day. This positive ratio of tweets was then stored into a csv format using df.to_csv function in python. After calculating the positive ratio, overall sentiments for each date is determined. This sentiments determination is done using simple logic of maximum sentiments on each date i.e. either positive or negative. For example, if there are 3 positive tweets and 1 negative tweet on a particular date, then the sentiment assigned to that date will be positive. Trend of stocks i.e. either uptrend or downtrend is also calculated using simple for loop in

[2] https://www.kaggle.com/aaron7sun/stocknews

python where, if the closing price of stocks for corresponding date is higher than that of previous date than there is an uptrend for that particular date.

3.3.2 Exploratory Data Analysis

It is very crucial to gain insights from data and understand the data, and for achieving this Exploratory data analysis is performed. Firstly, the time series chart for closing values of DJIA stocks, gold prices and Crude oil prices are plotted to check the seasonality as shown in the Figure 8,9 and 10 below:

Fig. 8 Closing price for DJIA

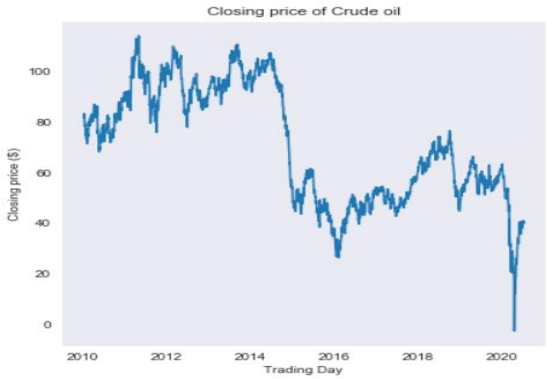

Fig. 9 Closing price of Crudeoil

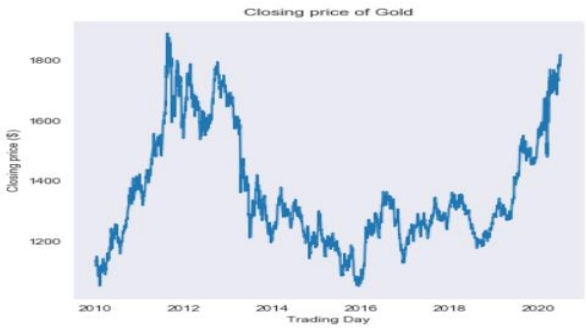

Fig. 10 Closing price of Gold

As it is seen in above images that due to unpredictable behavior of stock market it is hard to determine the seasonality. Also, the DJIA stocks has down trend in the end which means the value is DJIA stocks is reducing further. Gold and Crude oil have an upward trend in the end which shows that prices are increasing. In the second stage, boxplots are used to determine the outliers in closing values of DJIA, Gold and Crude oil.

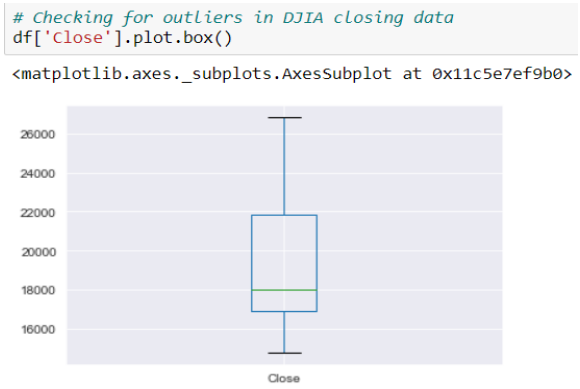

Fig. 11 Boxplot for DJIA close value

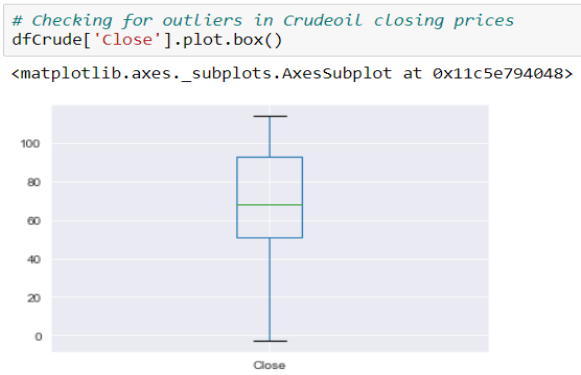

Fig. 12 Boxplot for Crudeoil close value

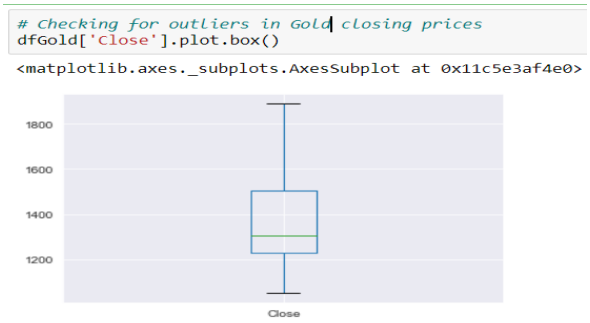

Fig. 13 Boxplot for Gold closing values

Figure 11, 12 and 13 shows that there are no outliers in any of the datasets. Exploring the datasets, it is clear that no further processing is required on the datasets.

3.4 Modelling

It is very crucial to implement the valid model for the purpose of making predictions. By reading various previous researches done in similar field, multiple models are identified that performs good in predicting stock index using sentiments and time series data. Also, the datasets used in this research are time series multivariate datasets, hence the models have to be selected by considering their performance for this type of datasets. Finally, after reviewing the two base papers considered for this research, three best performing models were selected. Selection of VAR model is inspired from (Jain, 2019), selection of kNN and SVM model is inspired from (Kalra and Prasad, 2019).

3.5 Evaluation

To check the performance of model, it is important to evaluate models using corresponding evaluation matrix. In this research, the datasets are split into train and test data in ratio 80:20 respectively using train_test_split() function in python. Prediction results obtained after testing the models are then evaluated using same evaluation factors used in base papers from where model selection is inspired. VAR model is evaluated using error function called MAPE. SVM and kNN models are evaluated using four evaluation matrices namely Accuracy, Precision, Recall and F1 score.

4 Design Specification and Implementation

4.1 System Design

This section explains the techniques i.e. Granger Causality Test and machine learning models i.e. VAR, SVM and kNN used in this study.

4.1.1 Granger Causality Test

As discussed in section 3.2, this research uses 5 datasets in total. As one of the objectives of this research is to identify factors that are significantly causal to DOW30 stocks data, this objective is achieved using Granger Causality test. This test helps in identifying that whether one time series dataset is significant in predicting another time series data[3]. Basically, Granger Causality tests is a type of hypothesis test, where if p-value is less than 0.05 then the hypothesis can be rejected.

4.1.2 Support Vector Machines (SVM)

SVM is a supervised machine learning based classification algorithm. SVM classifier works on a simple logic of dividing the N-dimensional space into N parts by a hyperplane, where N is nothing but the number of features that needs to be classified[4]. This algorithm uses support vectors which are nothing but data points that are used to identify optimal position of hyperplane[4]. Figure 14 shows the block diagram of SVM.

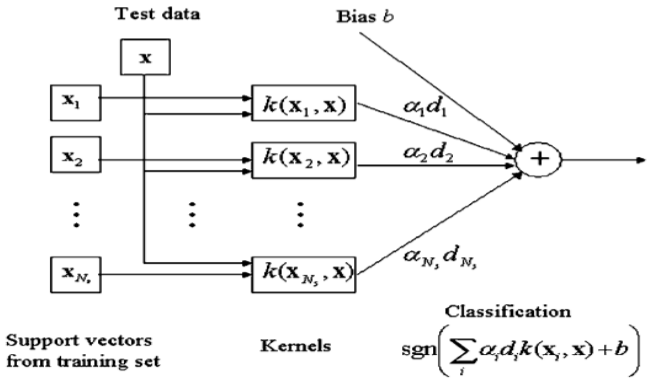

Fig. 14 Block diagram of SVM classifier for a two-class case (Gu, 2009).

In above image, x is an input vector from test data, x1 to x_{Ns} are support vectors derived from training data, αd are the weights added to reduce the error (Gu, 2009). $k(x_1,x)$ to $k(x_N,x)$ are the kernel functions. For this research the features that needs to be classified by SVM are two viz. uptrend or downtrend. So, while working on the training dataset SVM algorithm tries to create an optimal hyperplane, and then performs classification on test data.

4.1.3 K-Nearest Neighbor (kNN)

[3] https://www.sciencedirect.com/topics/social-sciences/granger-causality-test
[4] https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

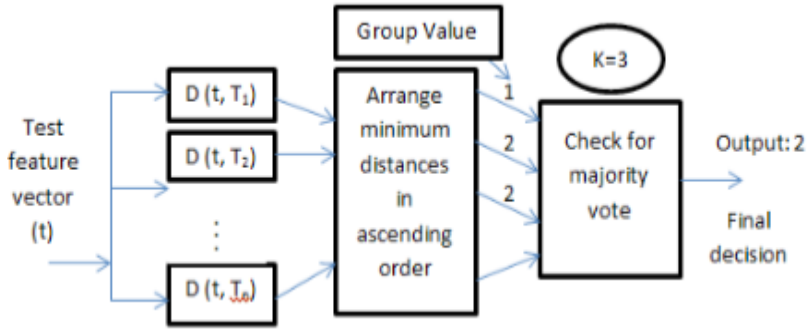

Fig. 15 Block diagram of kNN classifier (Gupta *et al.*, 2016)

kNN is also a supervised machine learning based algorithm and unlike SVM, kNN can work as both regression and classification algorithm. kNN works on an assumption that similar data points will always be close to each other[5]. Process flow of kNN is as follows: Firstly, algorithm calculates distance for ach datapoint, then it sorts the distances in ascending order. Later, for each datapoint k number of samples are taken from sorted list and the given datapoint is classified by taking mode of k samples[5].

4.1.4 Vector Auto Regression (VAR)

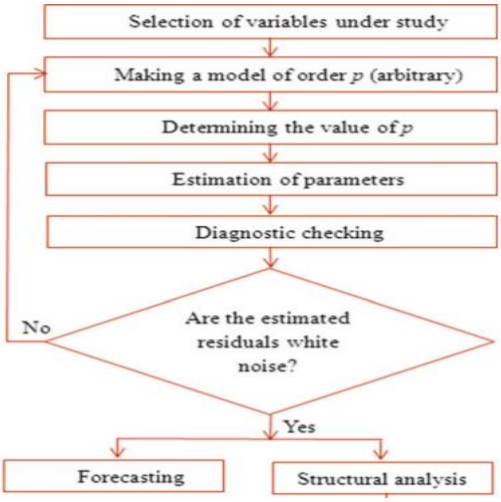

Fig. 16 VAR modelling (Rahman, 2014)

VAR has proved to be one of the best models for making predictions on multivariate time series (Jain,2019). VAR is usually used to identify linear correlation between multiple time series datasets. Mathematical equations for VAR model is as follows:

$$Y_t = A_o + \sum_{i-1}^{p} A_i\, y_{t-1} + u_t \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 1$$

[5] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

Where,
yt = Column vector
Ai = unknown coefficient matrix,
Ao = deterministic constant,
ut = error column vector

First step in VAR modelling is to select the variables that are endogenous. Then, the VAR model uses lagged values of all the variables fitted to model and an error term to calculate an equation for each input variable. Later, an estimator function like least square estimator is used to make estimations for parameters in model. In the next step, various checks like normality and autocorrelation are performed to diagnose the model. Then, stability of model is checked after which finally the forecasting is done. Basically, VAR uses past values of a variable over a pre-defined sample period to describe movement of that variable.

4.2System Implementation

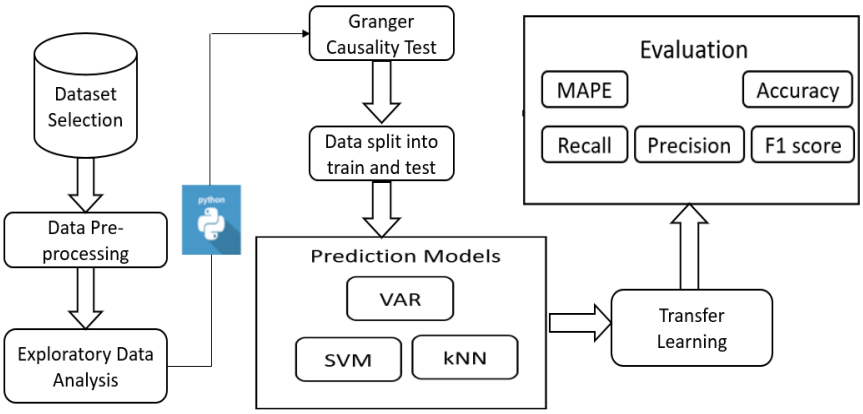

Fig. 17 Process Flow Diagram

Above sections have detailed the purpose and objectives of this research and also explained the process flow that has been carried out. This section particularly dives deep into the steps taken to implement all the selected machine learning models. It also discusses the steps taken to perform Granger causality tests on all the datasets. This research work is implemented on Jupyter notebook using Python. Python has a very big community and can easily get online support if needed. Also, Python has big amount of easily available libraries for almost every kind of operation. The implementation process was initiated by performing various data preparation techniques explained in section 3.3. In the next stage, Granger causality test is performed on the datasets and later, all three machine learning models are implemented. The process flow followed for this research is shown in figure.17. Firstly, the task of data selection and data gathering was performed as explained in section 3.2. In the second stage, various pre-processing techniques as explained in section 3.3 were applied onto the datasets to bring the datasets into required format. Later, exploratory data analysis (EDA) was performed on the datasets to gather useful insights of data. Steps taken for EDA are explained briefly in section 3.3.2. In the next step, Granger Causality test was performed on all the datasets to check the causality of each factor on stock price, and also to check the significance in predicting future trend

of stocks. This test was performed using 'grangercausalitytest' function available in 'statsmodels.tsa.stattools' library. Figure 18 shows python code used for calculating Granger Causality test. Results obtained from Granger Causality tests are evaluated briefly in section 5.

```
In [6]: #importing stock market and Twitter positive ratio datasets
        dfPosRatio = pd.read_excel(r'C:/Users/yash8/Desktop/Research project/Datasets/stocks_Posratio.xlsx', index_col = 0)

        #Calculating Causality
        from statsmodels.tsa.stattools import grangercausalitytests
        grangercausalitytests(dfPosRatio[['Twitter_pos_ratio','Close']], maxlag=3)
```

Fig. 18 Granger Causality test

After identifying significant factors, datasets are combined as per the requirements and resultant dataset is split into train and test data for model fitting. Later, all the three models are trained using combination of DOW30 stocks trend and Twitter sentiments data. These trained models are than used to make predictions on test dataset. The final step is to analyse the performance of models using transfer learning technique, where the pre-trained models are tested on combination of DOW30 stocks trend and news sentiments dataset. Performance evaluation of models is explained in section 5.

5 Evaluation

The main focus of this research was to identify some of the factors that are significantly causal in predicting DOW30 stock price movement, and also to test predictive performance and robustness of three machine learning models i.e. VAR, SVM and KNN. For calculating the causality of various factors on stocks data, Granger causality test is used. Three different machine learning models are used for the purpose of making predictions on stocks movement. Finally, transfer learning technique is used where pre-trained models are tested on a new dataset. In this section, results of Granger causality tests and machine learning models are evaluated.

5.1 Experiment 1: Using Granger Causality test to identify significant factors.

To meet one of the primary objectives of this research i.e. identifying significant factors to build a prediction model for DOW30 stocks, Granger Causality test is used. Table 2 shows the results of Granger Causality tests of each factor against DOW30 price.

Table 3. Results of Granger Causality

	DOW30 Stock Price		
	P value (1st Lag)	P value (2nd Lag)	P value (3rd Lag)
Twitter positive ratio	0.0005	0.0028	0.0041

News Sentiments	0.00	0.00	0.00
Crude oil prices	0.83	0.30	0.49
Gold prices	0.89	0.96	0.98

Table 2 provides the p-values for 3 lags for twitter positive ratio data, News sentiments data, Crude oil prices and Gold prices data. From the results of Granger Causality tests shown in table 2, it can be seen that only Twitter positive ratio dataset and news dataset have p-value less than 0.05, and Crude oil and Gold prices have p-value greater than 0.05. This means that out of the four factors, only Twitter sentiments and news sentiments have significant causality on stock prices of DOW30. Hence all the machine learning models are build using these two factors.

5.2 Experiment 2: Training and testing machine learning models on Twitter positive ratio dataset

5.2.1 Evaluating SVM and kNN

Firstly, SVM and KNN models are trained and tested using combination of twitter positive ratio and stock trend dataset. Here after training, the models are making predictions on test dataset, and the prediction performances of models are evaluated using Accuracy, Precision, Recall and F1 score. Results obtained in this experiment are shown in table 3:

Table 4. Performance of SVM and KNN

	SVM	KNN
Accuracy	0.51	0.58
Precision	0.26	0.55
Recall	0.51	0.58
F1	0.35	0.53

Table 3 compares evaluation scores of SVM and kNN model. From the results, it can be seen that there is no huge difference between accuracy of SVM and kNN. Overall, it can be said that kNN performed slightly better than SVM with accuracy of 58% which is higher than that of SVM i.e. 51%. Also, precision, recall and F1 scores of kNN are higher than that of SVM

5.2.2 Evaluating VAR

In this step, VAR model is trained and tested on combination of twitter positive ratio and stock trend dataset. Firstly, the model was trained using above mentioned multi-variate data and then the predictions were made on test data. The prediction performance of VAR model was evaluated using Mean Absolute Percentage Error (MAPE). As discussed in section 2 and section 4.1.4, VAR model is best suited for multi-variate data. It showed really good performance with MAPE of only 5.50%. Figure 19 shows comparison of actual vs predicted values:

Fig. 19 Actual vs Predicted prices

5.3 Experiment 3: Testing pre-trained machine learning models on News sentiments data (Transfer learning).

In this section, all the three machine learning models which are already trained using combination of stock trend and Twitter positive ratio dataset, are tested using totally new dataset which is combination of news sentiments and stock trends.

5.3.1 Evaluating SVM and kNN

Here the prediction performance of models is analyzed by making predictions on totally different dataset than the one used to train the model. Evaluation matrices used are same as section 5.2.1. Results obtained in this experiment are shown in table 4:

Table 5. Performance of SVM and KNN

	SVM	KNN
Accuracy	0.44	0.38
Precision	0.22	0.34
Recall	0.47	0.38
F1	0.30	0.32

Table 4 compares evaluation scores of pre-trained SVM and kNN models when tested with new dataset. As it can be seen from the results that SVM has an accuracy of 0.44 and recall of 0.47. So, there is no drastic difference in performance scores of SVM model tested on new dataset vs SVM model tested on the same dataset used for training. Accuracy of kNN reduced from 0.58 to 0.38. Hence it is proved that SVM is more robust model compared to kNN when it comes to predicting multi-variate time series.

5.3.2 Evaluating VAR

In section 5.2.2, VAR model is evaluated by dividing same dataset for testing and training. Here, The VAR model which is already trained using twitter positive ratio data is tested on totally new dataset i.e. news sentiments. Model showed great consistency in performance and proved that it is robust and reliable model for stock market prediction. The MAPE of model was just 2.77% which is even lesser than that achieved while testing model with same dataset as training. Figure 20 shows comparison of actual vs predicted values:

Fig. 20 Actual vs Predicted close

6 Discussion and Conclusion

After performing detailed analysis of previous researches done in same domain, the models and techniques to be used for this research were decided as explained in section 2 and section 3.4.

Using Granger Causality test, it was found that market sentiments whether it be news or twitter, play a significant role in predicting movement of stock market. Whereas, commodity prices like Gold and Crude oil prices are not causal to stock market movement. Hence, one of the main objectives of research i.e. to identify factors that are causal to stock market movement is also fulfilled.

Looking at the evaluation scores of models, it can be said that VAR model has outperformed the tests with MAPE error of 6.18% while training and testing on same dataset, and error score (MAPE) of only 2.77% when tested with new dataset. Therefore, we can say that VAR model is robust and perfectly suitable for stock market prediction. SVM and kNN models showed an accuracy of 0.51 and 0.59 respectively while training and testing on same dataset. Whereas, after testing these models on new dataset SVM performed better than kNN with accuracy of 0.44 compared to kNN's accuracy of 0.38. It can be seen that drop in kNN's accuracy is greater than SVM. Therefore, it can be said that SVM is more reliable model when it comes to cross-dataset predictions. Here, major focus is kept on accuracy and error score because, objective of this research is to know how accurately can applied machine models predict the movement of stocks.

In all, the research has successfully completed all its objectives. Also, it successfully answers the research question by proving that market sentiments have a significant impact on changes in stock prices, and can play a vital role in predicting future trend of stocks. Whereas, crude oil and gold prices are not causal enough to help make predictions on stock movement.

One major limitation of this study is that the final data frames obtained after performing pre-processing on datasets explained in section 3.3 have very low number of rows, due to which the models are not getting enough data to train accurately. Also, the twitter sentiments and news sentiments datasets have articles/tweets only in English language. To understand the market sentiments more accurately it is important to avoid filtering of languages. Because

processing more languages will give more accurate data and hence more accurate predictions can be made.

6.1 Future Work

This research has only considered 4 factors/sources i.e. Twitter Sentiments, News sentiments, Gold prices and Crude oil prices for checking causality on stock prices. For future work more factors like trading volume, volume of tweets, sentiments from blogs, and other social media platforms can also be considered to get more accurate analysis and prediction of stock market. Also, it has been proved in some of the researches that factors like volume of retweets and likes on the tweet also play important role in identifying the market sentiments towards the stocks. Hence, these factors can also be included in future study.

7 Acknowledgement

I would like to express my gratitude towards my mentor - Prof. Hicham Rifai, who has provided great extent of guidance and support throughout this study. I would also like to acknowledge Smit Jain (Jain, 2019) for providing his support and letting me use the twitter sentiments dataset from his study. Finally, I would like to thank my friends and colleagues who kept me motivated and supported me throughout my study.

References

Birbeck, E. and Cliff, D. (2019) 'Using Stock Prices as Ground Truth in Sentiment Analysis to Generate Profitable Trading Signals', *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*. IEEE, pp. 1868–1875. doi: 10.1109/SSCI.2018.8628841.

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.

Coronado, S., Jiménez-rodríguez, R. and Rojas, O. (2018) 'An Empirical Analysis of the Relationships between Crude Oil , Gold and Stock Markets', 39, pp. 193–208.

Gu, I. Y. H. (2009) 'Wood defect classification based on image analysis and support vector machines', (November). doi: 10.1007/s00226-009-0287-9.

Gupta, S. *et al.* (2016) 'Segmentation , Feature Extraction and Classification of Astrocytoma in MR Images', (October). doi: 10.17485/ijst/2016/v9i36/102154.

Jain, S. (2019) *Analysing effect of Twitter , Oil Prices , Gold Prices and Foreign Exchange on S & P500 Using Machine Learning MSc in Data Analytics Smit Jain National College of Ireland*

Kalra, S. and Prasad, J. S. (2019) 'Efficacy of News Sentiment for Stock Market Prediction'. IEEE, pp. 491–496.

Moghar, A. (2020) 'ScienceDirect ScienceDirect April Using Stock Market Prediction LSTM

Neural Network a LSTM Recurrent Stock Market Prediction Using', *Procedia Computer Science*. Elsevier B.V., 170, pp. 1168–1173. doi: 10.1016/j.procs.2020.03.049.

Nofsinger, J. R. (2005) 'Social Mood and Financial Economics', 6(3), pp. 144–160.

Patel, J. *et al.* (2015a) 'Expert Systems with Applications Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques', *EXPERT SYSTEMS WITH APPLICATIONS*. Elsevier Ltd, 42(1), pp. 259–268. doi: 10.1016/j.eswa.2014.07.040.

Patel, J. *et al.* (2015b) '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.

Porras, G. (2019) 'Social Mood Impact on Financial Decision Making: A Study of Twitter Sentiment on Stock Index Volume'.

Pyo, S. *et al.* (2017) 'Predictability of machine learning techniques to forecast the trends of market index prices : Hypothesis testing for the Korean stock markets', pp. 1–18.

Rahman, M. A. I. (2014) *Tanvir Islam · Prashant K . Srivastava Computational Intelligence Techniques in Earth and Environmental Sciences*. doi: 10.1007/978-94-017-8642-3.

See-to, E. W. K. and Yang, Y. (2017) 'Market sentiment dispersion and its effects on stock return and volatility'. Electronic Markets, pp. 283–296. doi: 10.1007/s12525-017-0254-5.

Sirimevan, N. *et al.* (2019) 'Stock Market Prediction Using Machine Learning Techniques', pp. 0–5.

Vveinhardt, J. *et al.* (2017) 'Asymmetric influence of oil and gold prices on Baltic and South Asian stock markets : Evidence from Johansen cointegration and ARDL approach', 22(4), pp. 422–438.

Yu, X. *et al.* (2015) 'An Impact Measure for News: Its Use in Daily Trading Strategies', *SSRN Electronic Journal*. doi: 10.2139/ssrn.2702032.