Stock Trend Behavior Prediction using Machine Learning Techniques and Trading Simulation

Article history

Received:
02 Dec 2021

Received in revised form:
15 Jan 2022

Accepted:
20 Feb 2022

Published online:
20 May 2022

*Corresponding author
liauchang@graduate.utm.my

Liau Sheau Chang[1], Nilam Nur Binti Amir Sjarif[2], Doris Wong Hooi Ten[3]

[1,2,3]Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia (UTM), Kuala Lumpur 54100 Malaysia
[1]liauchang@graduate.utm.my, [2]nilamnur@utm.my, [3]doriswong@utm.my

Abstract

Due to the choppy fluctuates and uncertainties in the share market, it has been a challenge for financial institution or even investors to be definite with the stock trend. The aim of the paper is to scrutinize different algorithms in data mining to identify the trend of the stock price movement. This will provide contently insights to the investor to make a precise investment and grow their portfolios. Historical price movement are extracted from financial websites. Derived attributes on Simple Moving Average (SMA) with different periods are added as an input parameter. This study proposed a combination of different features to implement with machine learning algorithms which includes k-NN, SVM and J48. The study has achieved high accuracy in stock classification, with 94.872% in k-NN, 94.855% in J48 and 85.257% in SVM. This indicates that for trend movement prediction classification, SVM is the most optimal algorithm to classify the correct trend of the stock movement, followed by k-NN and J48. However, the feature selection is also crucial to have an impactful attribute as the input parameters for better and more accurate predictive analysis. Price movement forecast was also carried out to compare between linear regression, Decision Tree, LSTM and k-NN to be used for future comparison. LSTM is the best algorithm in predicting the stock price with the least RSME indicates that it rhymes closely with the actual stock price movement.

Keywords: *Stock Trend Prediction, Data Mining, Machine Learning, k-NN, SVM, LSTM, ETL, Exponential Moving Averages*

1. Introduction

Over the decades, due to the high rise of inflation rate and globalization effect[1] , investors are looking for alternative investment platforms to secure the funds that have been hard earned. Interest rates in traditional savings and fixed deposit products are no longer lucrative as before. Some countries even demand service charges to secure the savings funds put into the bank and causing the interest to be in negative territory.

Moreover, due to the Covid-19 pandemic that impacted globally since March 2020, there has been a surge in the growth of stock market transactions[1]. Job insecurity and new norm working from home has sped up the consensus to generate passive income from other instrument products. However, some of the investors are investing based on rumors and news which

* *Corresponding author. liauchang@graduate.utm.my*

turn out to be a bad investment over time and causing loss of money. Based on the Efficient Market Hypothesis (EMH) [2], it is impossible to comprehend the market trend over time. This is due to the spread of corresponding news to the public and reflected in the share price in a moment. Financial markets like stock markets are a complex yet chaotic platform, full of uncertainties and influence factors. Political issues, economy policies, corporate actions or even the market theme sentiments all are taking some weightage influencing the stock price movement.

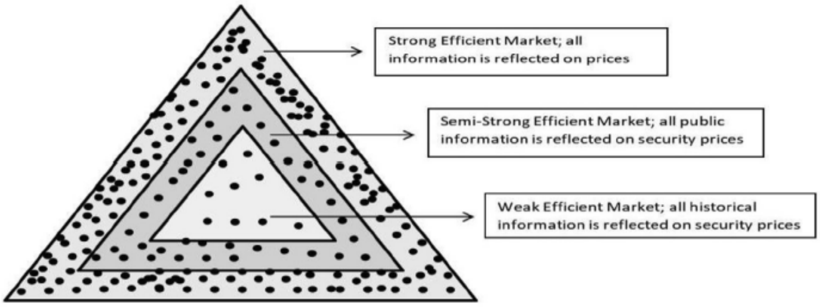

Strong Efficient Market; all information is reflected on prices
Semi-Strong Efficient Market; all public information is reflected on security prices
Weak Efficient Market; all historical information is reflected on security prices

Figure 1. Efficient Market Hypothesis[2]

There have been several strategies adopted in the security market for price movement prediction. The commonly used approaches include (a) fundamental analysis, (b) technical analysis[3], and (c) quantitative analysis. Scrutinizing the company intrinsic value by looking into the financial statements and underlying factors that may impact the current operation and prospects of the businesses is deemed to be fundamental analysis. Some ratios are used as metrics to provide insights into the company economical health and sustainability of company growth[4]. The usage of fundamental analysis will provide a benchmark to predict the trend of a share price movement over a mid to long term period.

Technical analysis on the other hand uses historical data, which includes the price and volume movement for identification of the statistical trend[5]. Price and volume figures are collected and further calculate derived technical indicators for prediction. These indicators presented in graphs illustrate the opportunity window for investment. It is usually used as an indicator for short to mid-term stock price movement trends.

Quantitative trading[6] which uses high-frequency trading and algorithmic trading techniques aiming for short-term periodic investment profit realization. It is normally used by hedge funds and financial institutions which involves huge transactions amount in the financial market. Using disruptive technologies in machine learning and mathematical modelling enables the researchers to create models and back tested the historical market data and iterated for refinement. Once the target result is achieved, the system will be used to automate the trading in the real time market[7].

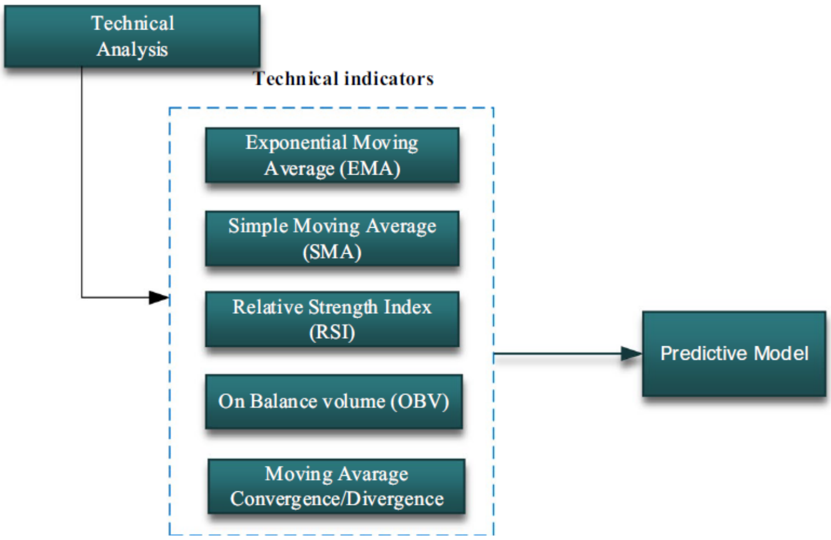

Technical Analysis
Technical indicators
Exponential Moving Average (EMA)
Simple Moving Average (SMA)
Relative Strength Index (RSI)
On Balance volume (OBV)
Moving Avarage Convergence/Divergence
Predictive Model

Figure 2. Technical Indicators for Predictive Model

In the financial domain, dataset from different markets and sources provide different elements. Data that has been extracted normally is huge due to the permutations of attributes[8]. Besides that, the attributes do not always fit the model. All of the data that can be obtained in every day life is in vast quantities. The attributes are required to be filtered based on the impact level. It is not possible to process them manually. This is where the idea of feature selection comes into play. With the enhancement and evolvement of the machine learning techniques will aid in the feature ranking and selection. When there is a huge data set and needs to reduce the number of resources without missing any significant or related information, the feature extraction technique comes in handy. Feature extraction aids in the reduction of duplicate data in a data package[9].

Machine learning is a favorable tool to be used in financial markets for price movement trend prediction research[10]. As part of artificial intelligence domain, machine learning carries out learning procedures based on algorithms without human intervention. There are several common tasks that can be done in machine learning such as classification, regression, clustering, and prediction[11]. Dataset with relevant attributes is provided as a feature input are trained and tested. Classification is a process used to indicate the separation of entities based on significant characteristics that labels onto each of the instances. In the price movement trend prediction model, two labels of classification are introduced, which are uptrend and downtrend. There are some studies on price movement trend prediction model using different algorithms in machine learning context. Support Vector Machine (SVM), Decision Trees, K-Nearest Neighbour (KNN) and Naïve Bayes [12] are the commonly used supervised learning techniques.

In this paper, two approaches are proposed where the dataset are being used for stock trend prediction. The approaches include using data mining techniques for structure data as well as sentimental analysis for unstructured data. Derived columns on price moving averages of different periods are extracted as input features for data mining. Three classification model chosen includes k-NN, SVM and J48. Apart from that, unstructured comments related to selected listed companies are also being used for sentiment analysis. The data are categorized with trends based on some key words in the comments. Both data mining and sentimental analysis will assist in precise stock trend prediction and hence will reduce the risk of investors and to provide better insights for stock recommendation.

This study can be split into four sections, which includes Related Works, Methods & Algorithms, Methodologies, Discussion & Findings and Conclusion. The Related Works describes the previous studies related to stock prediction using data mining techniques; Methods & Algorithms explains on the algorithms used for the study; Methodologies where steps from data preparation till the experiment and findings are discussed; Discussion & Findings display the results from different combination of feature extraction and price forecast using different algorithms; Conclusion summarizes the paper and suggest future research and enhancement that can be done.

2. Related Work

There have been several studies on the prediction of stock trend movement using data mining and text mining. Some of the studies combine both data mining and text mining to conduct a prediction based on the accuracy of classification like [4]. Other focus more on sentiment analysis based on the Twitter feeds using Twitter API like [13] . All the previous studies show that there is strong correlation between the sentiment analysis on Tweeter feeds using text mining and historical price using data mining. With the extensive enhancement on the machine learning algorithms, unstructured data that seems to be complicated has now gradually be easier to be processed for classification, clustering, and regression testing.

Ayman E. Khedr et.al [4] proposed a prediction model by combining both text mining and data mining to run the experiment. Different categories of stocks daily news data from organization, markets and annual report are taken together with historical numeric characteristic during daily floating rate to look insight and predict the stock market behavior. The investigation can be divided into two phases. In the first phase, Naïve Bayes algorithm is implemented to run text mining which is part of the sentimental analysis to identify news polarities. The result show high accuracy range which is 86.21%. For the second stage, K-NN algorithm are been chosen to carry out data mining. Merging of the outcome from previous stage and transform it as input of second stage so that they can proceed the historical numeric data. The result of predict future stocks market trend consider high accuracy which is up to 89.80%. Moreover, it is concluded by the researcher that both Naïve Bayes and K-NN algorithm can perform best prediction in future stocks market price. Hence, both also display a strong correlation in between the news from stocks market and the floating rate of stocks price.

Shila Jawale et.al [5] conducts a study that using Twitter sentimental analysis to foresee the overall of user sentimental who will impact the floating price in India stock market. Random Forest Algorithm have been chosen to run for the data mining. At the beginning, researcher using Python language to collect the raw historical data of the specific organization from Yahoo! Finance. After that implemented Random Forest Algorithm on the datasets for pre-processing. At the same time, the latest tweets from Twitter are grabbed which is relevant to those company and further retrieved the personal detail such as user ID, time, date and located when tweeted, etc. Sentiment analysis[14] had been applied to get the outcome of sentimental value. After that both outcomes are combined to predict the accuracy toward the fluctuation of stock market. Result shows that there is a strong relation of public mood and the floating rate of stocks price. A large-scale that are occupy from the research mentioned that a positive tweet from public good mood will cause a rise of stock price company.

Deep learning has been widely used for financial domain as it is capable for handling huge volume of data while reducing noise and handling nonlinear relationship within the data[15]. In the study, a generator is proposed to mine the data distribution using Long Short-Term Memory (LSTM) and generate predicted data for the closing price of share price. Besides that, a discriminator is proposed using Generative Adversarial Network (GAN) together with Multi-Layer Perceptron (MLP). The discriminator is used to compare the generated data against the real stock data. The prediction model is then evaluated based on some statistical indicators which include Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Average Return (AR) and Root Mean Square Error (RMSE) among different algorithms. GAN algorithm achieves the best result with highest annual return of 75.54% and lowest Mean Absolute Percentage Error which is 1.37%. However, the data points in the input parameter are minimal. More indicators and derived attributes should be computed and extracted as a feature for the model to learn and predict more accurately.

With the evolvement of machine learning and statistical algorithms, Efficient Market Hypothesis is no longer valid for its theory stating that the stock price prediction is impenetrable [16]. In the study, graph theory which grip on Spatio-temporal relationship are employed to the stock price movement for modelling. Two types of graphs are developed, one which comprises the correlation of the historical share price data and the other is the causation-based graph based on the news headlines. Convolutional neural network is used for correlation analysis while traditional machine learning models are used for causation-based graph. Root mean squared error (RMSE), Mean absolute percentage error (MAPE) and mean absolute error (MAE) is used for the performance evaluation. The employment shows that Graph convolutional Network (GCN) gives high accuracy in terms of causal relationship and forecasting. In the study, the GCN model only examines 30 nodes within the graph which is incapable of covering a more complex network. Besides that, the performance of GCN employed to time series forecasting can be scrutinized in future.

3. Method

To experiment the data mining in this research, the total 3 algorithms had been chosen to analysis the accuracy of data. There are K-Nearest Neighbours (k-NN aka IBK), Support Vector Machine (SVM) and Decision Tree J48. The further explanation will be explained in this section for a clear and better understanding on each algorithm.

3.1. K-Nearest Neighbours (K-NN aka IBk)

K-nearest neighbours'(k-NN) [17] algorithm also known as non-parametric instance-based knowledge algorithm (IBk) or lazy algorithm. Non-parametric means the data distribution is not sdefined and the model is only built upon the provided dataset. This is one of the focal points as in the real-world datasets, it does not follow numerical theoretical assumptions. For example, feature space is having N training vectors, k-NN algorithm will identify the k nearest neighbours of a new instance, regardless of the labelling. When the k is set to 1, each training vector will define a Voronoi partition of space.

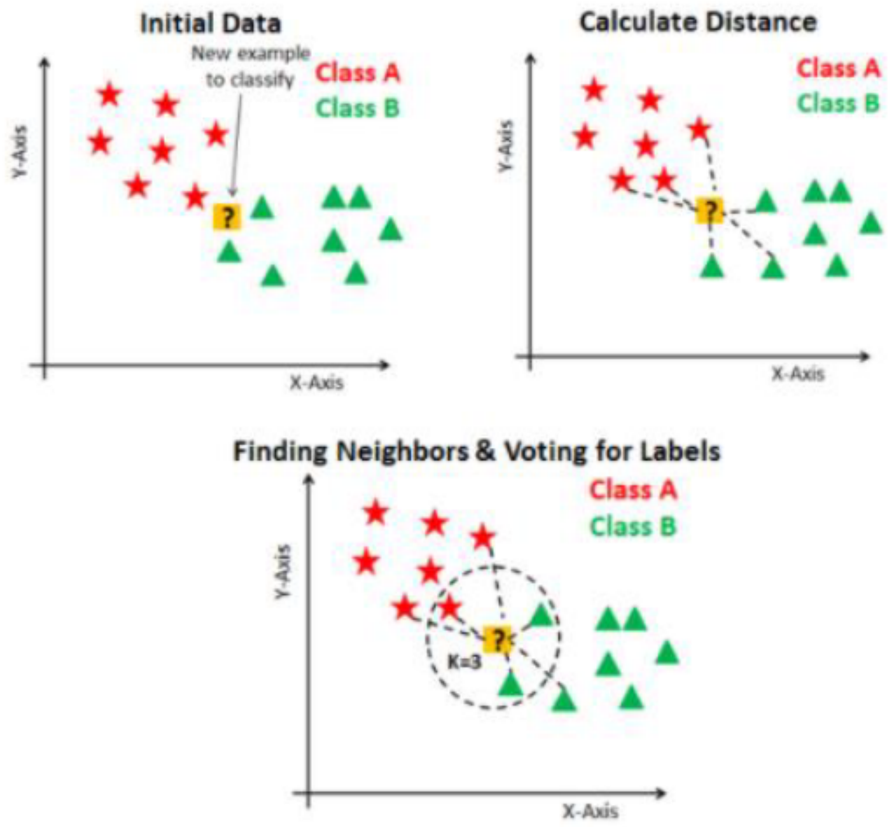


Figure 3. Label voting in k-NN algorithms

Segmentation of a plane into regions adjacent to each of a given group of items is known as Voronoi partition of space. These objects contained weighted points (seeds) in the plane is an ideal segmentation. In summary, seeds within a specific region will comprise all the points of the plane closer to specific seed than to any other. Lazy algorithm does not require any training data and it is only being used in testing phase. One of the drawbacks of k-NN is that it requires a lot of time to scan all the training data and as well to store the training vectors during the testing phase.

The selection of class region for a given data point can be derived from the formula below:

$$R_i = \{x: d(x, x_i) < d(x, x_j), i\ j\} \tag{1}$$

Some remarks to be consider when using k-NN are as follow:

- • For 2 class problem, an odd k value should be taken to prevent tie scenario
- • The k value should not be a multiple of number of classes

3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) [18] are supervised learning models that can be used for linearly separating binary sets and regression of data points. The main function of SVM algorithm is to find a hyperplane to split classes of data vectors with the maximum margin. The data points are classified in an N-dimensional space where the hyperplanes to have maxima distance between any two data points for different classes. Hyperplanes are boundaries used to decide the categorization of the data points. The dimension of the hyperplane also takes in consideration of the number of input features. For example, for a 2-dimensional input, the hyperplane could just be a line, but when it is a 3-dimensional input, then the hyperplane will become a two-dimensional plane.

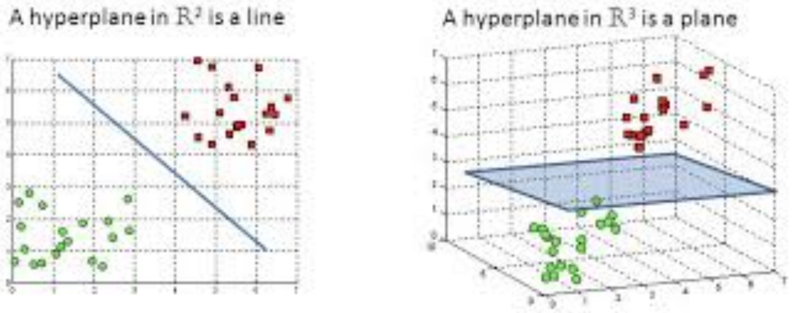


Figure 4. Hyperplanes in 2D and 3D feature space[19]

Support vectors [20] are selected from the classified data points which are closer to the hyperplane. These support vectors form the base for building SVM by maximizing the margin of the classifiers. Maximum margin provided by the hyperplane will be the best selection as it has the broader capacity to generalize data predictions and not overfitting the model to training data. The result will yield better performance on the test data in the same time.

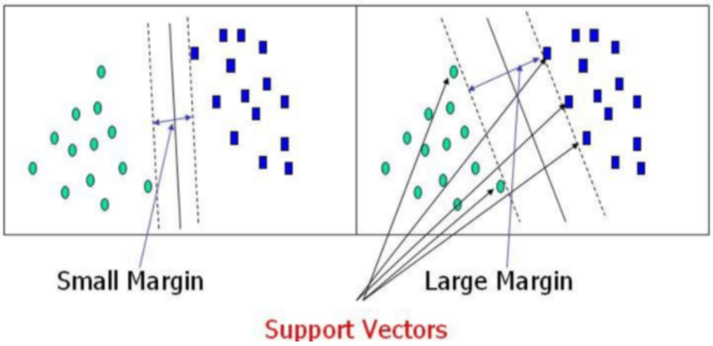


Figure 5. Support Vectors with different size of margin[19]

3.3. Decision Tree J48

J48 also known as C4.5 uses a top-down, recursive, divide-and-conquer strategy [21] to find a good attribute to split on at each stage. Decision tree are built upon based on the training data set which uses the algorithm of ID3 by measuring the entropy of information gained. Selecting attribute for root node where branches are created for each possible attribute value. To get the smallest tress, purest nodes which has the greatest information gain need to be selected. Information gain is measured in bits unit by realization of the value in data attribute from the pool of information gained. Based on the information theory, information is measured by bits using entropy.

$$entropy(p_1, p_2, \dots, p_n) = -p_1 log p_1 - p_2 log p_2 - p_3 log p_3 \tag{2}$$

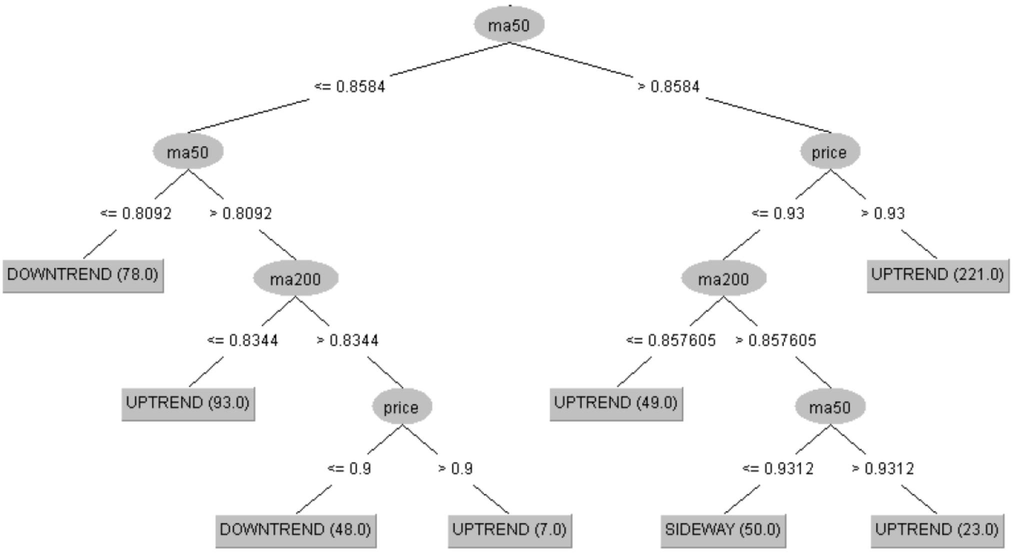


Figure 6. Extract from J48 decision tree

Information gain can be calculated by calculation the entropy of distribution before the split minus the entropy of distribution after it After that, instances will be split into subsets by extending the branches from the node. The attribute with the highest entropy bits is the best choice to be used for extending the branches. The branching continues recursively by using only instances that reach the branch. Process will halt when the criteria where all the instances have the same class.

Pruning in J48 is used with the aim to produce a smaller decision tree and at the same time preventing the overfitting issues to the training data. The core rationale behind is to remove any nodes/branches in such a way not deteriorating the algorithm processing duration.

Some remark on the algorithm is shown as below:

- When the data points are having the same classification, it will create a leaf node to conclude the class selection.
- For each attribute, information gain is calculated based on the entropy difference before and after the split

3.4. Support Vector Regression (SVR)

Bruno et al.[22] in the study scrutinize on the feasibility to implement Support Vector Regression (SVR) in stock price movement prediction in three main market which includes Brazilian, Chinese and America stock market. RMSE and MAPE are used to evaluate in two aspects which include the average returns and volatility of the market based on the standard deviation of the closing market price. There is evidence suggesting that SVR is a good algorithm for the stock market trend prediction context. Besides that, when the volatility of the market is low, the precision of the prediction using SVR increased significantly. However, in the study, there is no real implementation to perform comprehensive algorithm implementation. In addition to that, fundamental factors are not incorporated into the modelling consideration.

3.5. Long Short Term Memory (LSTM)

In financial market, there are a number of indicators as an input feature for price movement prediction[23]. Different equity is dependent on distinct indicators due to the characteristics of the sector itself and the cyclical economical aspects. Chen & Zhou in the study scrutinize the capability of Genetic Algorithms (GA) for feature selection. Ranking features are evaluated based on the weightage and impact to the model. Besides that, enhanced Long Short-Term Memory (LSTM) is used for stock price prediction model. China Construction Bank and CSI 300 stock from Chinese stock market are used for the model. Performance comparisons are done on 8 models with the GA-LSTM combination display the lowest figure on Mean Square Error which is 0.39% on CSI 300 and 0.53% on China Construction Bank. However, the study has limitation on the stock market coverage that only focuses on Chinese Market. Furthermore, the model parameters are manually done using trial and error. Hence a more systematic approach should be used to find the best fit parameters which include the number of attributes that should be used.

4. Methodologies

This section describes the proposed model and methodologies that are implemented for the classification and prediction of the stock price. Figure 3. Shows the proposed model that is used for this study. The objective of the proposed model is to prepare the data collection using web crawler, ETL run to clean and calculate some variables to be used, data pre-processing, machine learning algorithm implementation, analysis on the result, provide insights of the price movement and hence to give recommendation to investor based on the implementation of the machine learning algorithm. This includes 3 main algorithms which are K-nearest neighbours'(k-NN), Support Vector Machine (SVM) and J48 which are originated from C4.5. The different algorithms selection with their own features and characteristics can help to provide better insights and results.

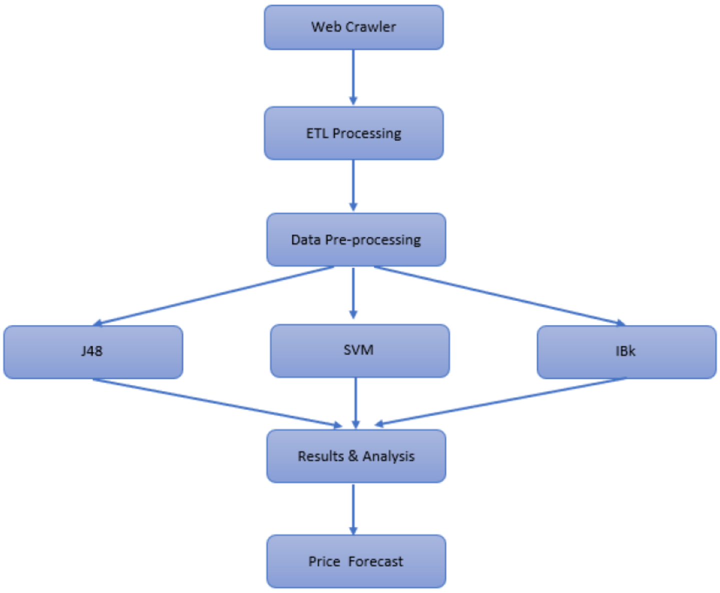


Figure 7. Proposed System for Stock Data Mining and Prediction

4.1. Web crawler

Firstly, a web scraper console application is developed using Microsoft stack IDE, i.e., Visual Studio. **Web crawler** is an automated approach to extract raw data from a webpage. The library used for the web scrapper is Selenium and the programming language used is C#. The historical data of share price movement are copied form the website Investing.com and store in a Comma Separated Value file (CSV). The source columns that are extracted by daily includes the stock name, date, opening, high, low, closing (OHLC) and as well the volume. Currently there are altogether 187 Malaysia listed companies and 17 United States listed companies are crawled for the survey.

4.2. ETL Processing

An ETL system is also developed to process the share price data that has been crawled from the website. All distinct files each representing a stock market counter are processed by daily. The ETL processes is designed in three stages, i.e., Extract, Transform and Load. In Extract phase, all CSV files are loaded into an Initial table without any transformation. Next, in Transform phase, several calculations are carried out to calculate Exponential Moving Average (EMA) [24] of different periods and data cleansing for initial structure formatting. Once the ETL is done, the data will be loaded into Microsoft SQL Server as historical price data repository. In this survey, 3 Malaysian company, i.e. Sapura Energy Bhd, Top Glove Corp Bhd and Magni-Tech Industries Bhd are selected for further in-depth analysis.

4.3. Data Pre-processing

The dataset prepared is further discretized by spotting golden crossover and death cross to indicate the uptrend, sideway and downtrend of the price movement. If the price of stock is above SMA50 and SMA50 is greater than SMA200, it will be labelled with "UPTREND", when the price is between SMA50 and SMA200 while SMA50 is still greater than SMA 200, it will be labelled as "SIDEWAY", and when the SMA50 is crossing SMA200, it will be labelled with "DOWNTREND".

4.4. Training Data with algorithms

The data that has been pre-processed will be used for machine learning implementation and prediction. The three algorithms that are being selected are K-nearest neighbours'(k-NN), Support Vector Machine (SVM) and J48. Each of these algorithms has its own advantages and drawbacks.

4.5. Results and Analysis

Once the data is ready, training and testing on the dataset using different algorithms will be carried out. Results from the algorithm run will be used to validate on its accuracy and precision. Different combination of the derived attributes will also be used to verify the most accurate pair of attributes among the rest.

4.6. Price Forecast

Apart from the classification, price movement prediction by linear regression and k-NN will also be plotted out. It will be useful as a guidance for investor to manage their portfolio and it can also be stored for future comparison.

5. Result and Discussion

This section describes the experiment results and performed prediction to the stock market behavior using 3 algorithms in Weka [25], which includes k-NN, SVM, J48. This experiment is designed in two parts, the first part is to perform prediction model for the stock market behavior to be either uptrend, sideway or downtrend. The second part is used to scrutinize the results of sentiment analysis against the comments in i3investor which can classifies as positive or negative.

In both parts of the experiment, company with different level of price fluctuation are being selected. The three companies are **Sapura Energy Bhd**, **Top Glove Corp Bhd** and **Magni-Tech Industries Bhd**. Three (3) algorithms are performed to calculate the accuracy of the data classification and prediction. On the other hand, the comment text is divided into training and testing sets. The purpose of training set is used to learn the model structure while the testing set is used to confirm the algorithm's accuracy. Comments with positive and negative sentiment classification are used as training data to learn the model. The testing data are expected to be able to classify into the correct classification of positive or negative sentiment.

Prior to applying the technique with the best precision, the proposed methods would be compared to many metrics. Primarily, using WEKA, an uncertainty matrix is used to evaluate a classifier's output by considering the right and incorrect classification rates. The formula for calculating accuracy is as follows: (3).

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Records} \tag{3}$$

Despite consistency, three new criteria for evaluating outcomes are introduced: precision, recall, and F1-score. Out of all the records that are positively estimated, precision counts the cumulative number of true positive records. The relevant formula is as follows: formula (4).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4}$$

Then, out of all the records that are properly labelled as accurate, Recall (also known as Sensitivity) counts the cumulative number of true positive records. The relevant formula is as follows: formula (5).

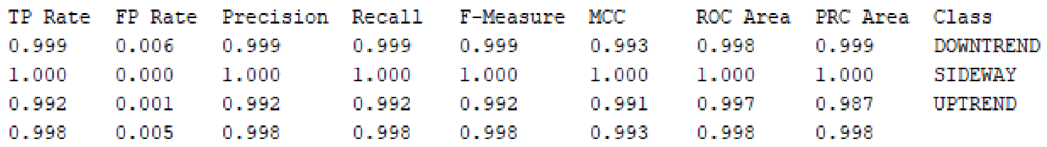
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5}$$

Finally, the F1-Score is a metric that attempts to strike a balance between precision and recall. F1-Score, unlike Accuracy, does not depend on True Negative, which refers to the cumulative number of valid documents that were exactly labelled. The formula for calculating the F1-Score is as follows: (6).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

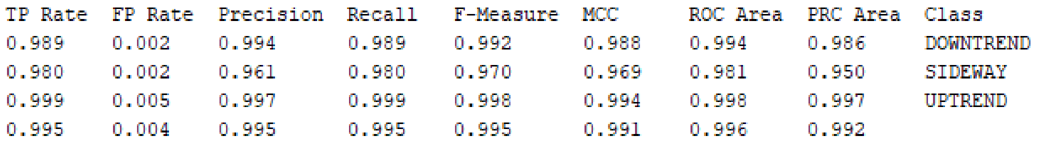
Table 1 shows the accuracy of prediction when KNN algorithm is applied on the price movement of the stocks for a period of 4 years. A couple of evaluation metric are used to measure the effectiveness and validity of the classification model. The evaluation of the k-NN classifiers outperformed other algorithms an having high accuracy and precision rate of more than 98% for trend classification.

Table 1. Results of a k-NN classifier for Sapura Energy Bhd



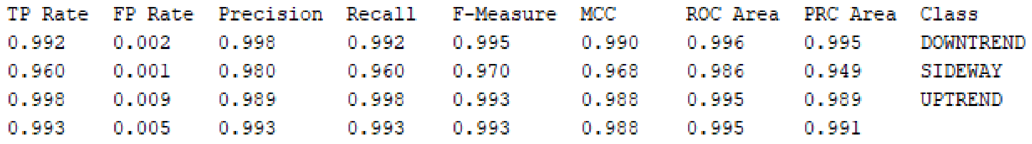
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.999	0.006	0.999	0.999	0.999	0.993	0.998	0.999	DOWNTREND
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	SIDEWAY
0.992	0.001	0.992	0.992	0.992	0.991	0.997	0.987	UPTREND
0.998	0.005	0.998	0.998	0.998	0.993	0.998	0.998	

Table 2. Results of a k-NN classifier for Top Glove Corp Bhd



```
TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.989    0.002    0.994      0.989    0.992      0.988    0.994     0.986     DOWNTREND
0.980    0.002    0.961      0.980    0.970      0.969    0.981     0.950     SIDEWAY
0.999    0.005    0.997      0.999    0.998      0.994    0.998     0.997     UPTREND
0.995    0.004    0.995      0.995    0.995      0.991    0.996     0.992
```

Table 3. Results of a k-NN classifier for Magni-Tech Bhd

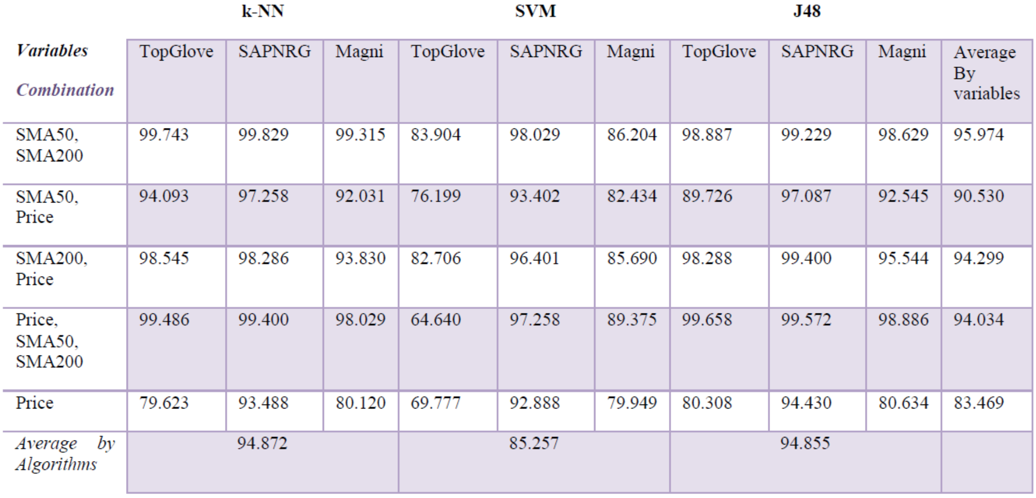


```
TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
0.992    0.002    0.998      0.992    0.995      0.990    0.996     0.995     DOWNTREND
0.960    0.001    0.980      0.960    0.970      0.968    0.986     0.949     SIDEWAY
0.998    0.009    0.989      0.998    0.993      0.988    0.995     0.989     UPTREND
0.993    0.005    0.993      0.993    0.993      0.988    0.995     0.991
```

Table 4. Results of overall model using SMA50 and SMA200 with k-NN Classifier

Measurement	**Companies**					
	Top Glove Corp Bhd		*Sapura Energy Bhd*		*Magni-Tech Industries Bhd*	
accuracy	99.743%	1165	99.829%	1165	99.315%	1159
Inaccurate	0.357%	2	0.171%	3	0.685%	8
Kappa statistic	0.9894		0.9934		0.9873	

Table 5. Results of classification by k-NN, SVM and J48 algorithms



	k-NN			SVM			J48			
Variables *Combination*	TopGlove	SAPNRG	Magni	TopGlove	SAPNRG	Magni	TopGlove	SAPNRG	Magni	Average By variables
SMA50, SMA200	99.743	99.829	99.315	83.904	98.029	86.204	98.887	99.229	98.629	95.974
SMA50, Price	94.093	97.258	92.031	76.199	93.402	82.434	89.726	97.087	92.545	90.530
SMA200, Price	98.545	98.286	93.830	82.706	96.401	85.690	98.288	99.400	95.544	94.299
Price, SMA50, SMA200	99.486	99.400	98.029	64.640	97.258	89.375	99.658	99.572	98.886	94.034
Price	79.623	93.488	80.120	69.777	92.888	79.949	80.308	94.430	80.634	83.469
Average by Algorithms	94.872			85.257			94.855			

The main difference between this study compare to previous research is that we are not only using the price but also include additional derived variable like SMA50 and SMA200. Three variables which includes Price, SMA50 and SMA200 are used to evaluate the impact factor to the algorithms. There are a few combinations of the variables that is used for the experiment to test the accuracy of the algorithms' implementation. Among the five, SMA50 and SMA200 is the best combination for achieving high accuracy of 95.974%, followed by SMA200 and Price which is 94.299%; Price, SMA50 and SMA200 which is 94.034; SMA50 and Price which is 90.530. The lowest accuracy that is achieved is when we provide the stock price solely.

From another perspective, which is by algorithms comparison, k-NN achieved the highest accuracy of 84.872% followed by J48, which is slightly 0.017% lower than k-NN. Out of the 3 algorithms used, SVM display the lowest accuracy which is 85.257. The main reason behind is SVM need more equal datasets in each trend categories to have enough training.

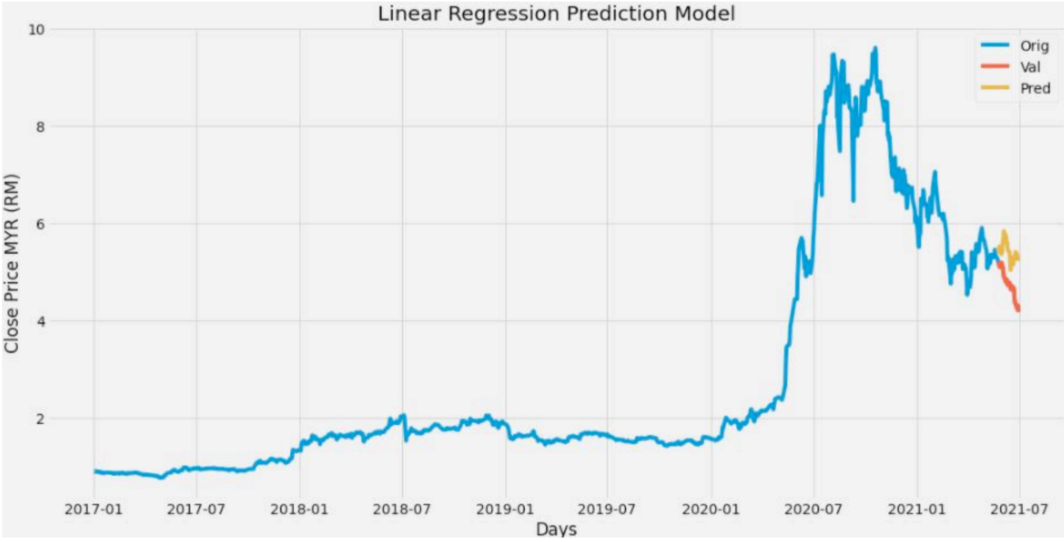


Figure 8. Trend prediction by Linear Regression

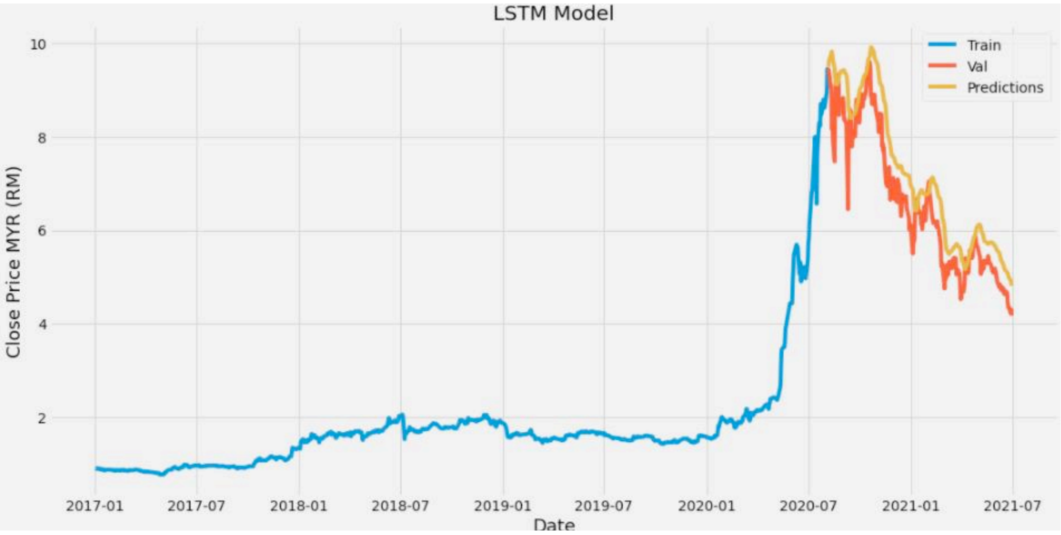


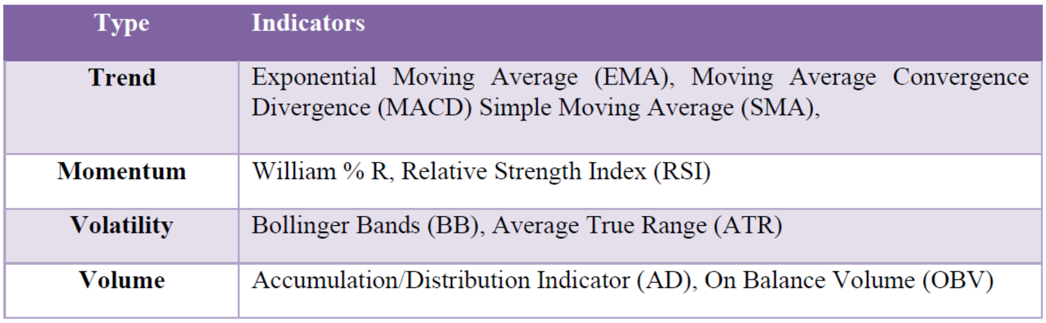
Figure 9. Trend prediction by Long-Short term Memory (LSTM)

Apart from determining the classification of the trend, this study also scrutinize the price prediction by using two algorithms which are Linear Regression 8and Long-Short Term Memory. Figure 6 shows the price trend by using Linear Regression while Figure 9 shows the price trend prediction using LTSM. It is interesting that both the price prediction is having a big difference in terms of the direction. Linear Regression algorithms shows that there is some limitation to predict far future price movement where it move in contradict direction while LSTM can match with the actual close price for the recent days ahead.

6. Conclusion

In this paper, several algorithms applied to the datasets to obtain the classification model and as well to forecasting the stock market trends. The datasets include the historical data for three company with different level of fluctuations and some data mining algorithms. The experimental results were satisfactory and confirm the possibility to use the data mining techniques namely SVM, J48 and KNN for prediction of trend. This will provide some insights for the investors to decide on the holding positions.

To further enhance the accuracy and precision of the implementation, parameters from technical aspect and financial aspect should be included. Technical indicators that can be used are categorized into four, such as trend, momentum, volume, and volatility. Figure shows the indicators that are representation for each of the technical analysis category. Balance Sheet and Profit and Loss report from financial reports by year end and quarter end should also be taken into consideration while preparing the datasets for fundamental screening.



Type	Indicators
Trend	Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD) Simple Moving Average (SMA),
Momentum	William % R, Relative Strength Index (RSI)
Volatility	Bollinger Bands (BB), Average True Range (ATR)
Volume	Accumulation/Distribution Indicator (AD), On Balance Volume (OBV)

Table 6. Indicators for each technical analysis Category[26]

Apart from data mining analysis, sentimental analysis should also suggest being used for further improvement. Sentimental analysis also called as contextual mining. It is a process to identify and extract the nature language of public opinion such as their thought, idea, behavior, assessment, and inspiration toward business product and services from the market. It is a common tool for text classification by using monitor and analyses online conversion to transform gather information whether is positive, negative, or neutral.

Acknowledgments

We are sincerely thankful to University Teknologi Malaysia (UTM) for organizing the Razak Faculty Research Week 2021 to provide a platform to integrate research and education in the field of informatics. It also aspires us to conduct new research discovery and innovation.

References

[1] G. V. Vorontsova, R. M. Ligidov, T. A. Nalchadzhi, I. M. Podkolzina, and G. V. Chepurko, "Problems and perspectives of development of the world financial system in the conditions of globalization," in *International Conference Project "The future of the Global Financial System: Downfall of Harmony"*, 2018: Springer, pp. 862-870.

[2] J. E. Singh, V. Babshetti, and H. Shivaprasad, "Efficient Market Hypothesis to Behavioral Finance: A Review of Rationality to Irrationality," *Materials Today: Proceedings,* 2021.

[3] C. Majaski. "Fundamental vs. Technical Analysis: What's the Difference?" Investopedia. https://www.investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/ (accessed.

[4] H. Ezzeddine and R. R. Achkar, "Ensemble Learning in Stock Market Prediction," in *2021 13th International Conference on Machine Learning and Computing*, 2021, pp. 298-303.

[5] D. Kumar, P. K. Sarangi, and R. Verma, "A systematic review of stock market prediction using machine learning and statistical techniques," *Materials Today: Proceedings,* 2021.

[6] E. P. Chan, *Quantitative trading: how to build your own algorithmic trading business*. John Wiley & Sons, 2021.

[7] J. Tian, Y. Wang, W. Cui, and K. Zhao, "Simulation analysis of financial stock market based on machine learning and GARCH model," *Journal of Intelligent & Fuzzy Systems,* no. Preprint, pp. 1-11, 2021.

[8] M. Ghorbani and E. K. Chong, "Stock price prediction using principal components," *PloS one,* vol. 15, no. 3, p. e0230124, 2020.

[9] A. Thakkar and K. Chaudhari, "Fusion in stock market prediction: a decade survey on the necessity, recent developments, and potential future directions," *Information Fusion,* vol. 65, pp. 95-107, 2021.

[10] J. Ayala, M. García-Torres, J. L. V. Noguera, F. Gómez-Vela, and F. Divina, "Technical analysis strategy optimization using a machine learning approach in stock market indices," *Knowledge-Based Systems,* vol. 225, p. 107119, 2021.

[11] K. Rauniyar, J. A. Khan, and A. Monika, "Review of Different Machine Learning Techniques for Stock Market Prediction," in *Inventive Systems and Control*: Springer, 2021, pp. 715-724.

[12] A. E. Khedr and N. Yaseen, "Predicting stock market behavior using data mining technique and news sentiment analysis," *International Journal of Intelligent Systems and Applications,* vol. 9, no. 7, p. 22, 2017.

[13] S. Nimje, R. Mayya, M. N. A. Baig, and S. Jawale, "Prediction on Stocks Using Data Mining," in *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*, 2020.

[14] N. Jing, Z. Wu, and H. Wang, "A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction," *Expert Systems with Applications,* vol. 178, p. 115019, 2021.

[15] K. Zhang, G. Zhong, J. Dong, S. Wang, and Y. Wang, "Stock Market Prediction Based on Generative Adversarial Network," *Procedia Computer Science,* vol. 147, pp. 400-406, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.procs.2019.01.256.

[16] P. Patil, C.-S. M. Wu, K. Potika, and M. Orang, "Stock market prediction using ensemble of graph theory, machine learning and deep learning models," in *Proceedings of the 3rd International Conference on Software Engineering and Information Management*, 2020, pp. 85-92.

[17] A. Gupta, P. Bhatia, K. Dave, and P. Jain, "Stock market prediction using data mining techniques," in *2nd International Conference on Advances in Science & Technology (ICAST)*, 2019.

[18] X. Li, P. Wu, and W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong," *Information Processing & Management,* vol. 57, no. 5, p. 102212, 2020.

[19] E. Z. Yi. "How to Find Linear (SVMs) and Quadratic Classifiers using MATLAB." Towards Data Science. https://towardsdatascience.com/how-to-find-linear-svms-and-quadratic-classifiers-using-matlab-97ea7550655a (accessed.

[20] P. Dineshkumar and B. Subramani, "SURVEY ON STOCK MARKET PREDICTION TECHNIQUES IN DATA MINING PROCESS," 2020.

[21] X. Huang, Y. R. Huang, M. Wen, A. An, Y. Liu, and J. Poon, "Applying data mining to pseudo-relevance feedback for high performance text retrieval," in *Sixth International Conference on Data Mining (ICDM'06)*, 2006: IEEE, pp. 295-306.

[22] B. M. Henrique, V. A. Sobreiro, and H. Kimura, "Stock price prediction using support vector regression on daily and up to the minute prices," *The Journal of Finance and Data Science,* vol. 4, no. 3, pp. 183-201, 2018/09/01/ 2018, doi: https://doi.org/10.1016/j.jfds.2018.04.003.

[23] S. Chen and C. Zhou, "Stock Prediction Based on Genetic Algorithm Feature Selection and Long Short-Term Memory Neural Network," *IEEE Access,* 2020.

[24] M. C. Angadi and A. P. Kulkarni, "Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R," *International Journal of Advanced Research in Computer Science,* vol. 6, no. 6, 2015.

[25] M. Hernández-Álvarez, E. A. T. Hernández, and S. G. Yoo, "Stock Market Data Prediction Using Machine Learning Techniques," in *International Conference on Information Technology & Systems*, 2019: Springer, pp. 539-547.

[26] K. S. Kannan, P. S. Sekar, M. M. Sathik, and P. Arumugam, "Financial stock market forecast using data mining techniques," in *Proceedings of the International Multiconference of Engineers and computer scientists*, 2010, vol. 1, p. 4.