Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 2, September 2020, pp. 1602∼1609
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17i2.pp1602-1609

A computing model for trend analysis in stock data stream classification

Abdul Razak M. S[1], Nirmala C. R[2]
[1]Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, India
[2]Visvesvaraya Technological University, India

Article Info

Article history:

Received Sep 28, 2019
Revised Mar 13, 2020
Accepted Apr 8, 2020

Keywords:

Classification
Data stream
Stock trading
Trend analysis
Data stream

ABSTRACT

For several decades, many statistical and scientific efforts took place for the better analysis or prediction of stock trading. But still it is open to offer new avenues for the scientists to rethink and discover new inferences by adopting latest technological scenarios. In this regard, we are trying to apply classification techniques on stock data stream through feature extraction for the trend analysis. The proposed work is involving k-means for clustering samples into two clusters (the stocks in trend as one cluster and another on as stocks not in trend). The trend analysis is done based on density estimation of the stocks with respect to sectors. A well-known data representation method that is histogram is used to represent the sector which is in trend. This work has been implemented and experimented by considering live NSE (india) data using python and its related tools.

Corresponding Author:

Abdul Razak M. S,
Department of Computer Science and Engineering,
Bapuji Institute of Engineering and Technology, Davangere, India.
Email: msabdulrazak@gmail.com

1. INTRODUCTION

Data stream analysis has opened up new avenues or opportunities for Computer Science and Engineering Scientists. The data stream is a recorded data with respect to time, perhaps it can be regarded as signal. Sometimes the signal may be continuous or discrete. All the parameters apply to signals holds good for data streams. Stock trading and its transactions can generate numerous amount of data with respect to time and hence it can be regarded as a data stream. Data stream classification [1] is an area, enables researchers to identify or extract new features through any acceptable scientific process. Classification techniques on stock data analysis may provide certain inferences. One such inference could be trend of the stock. Indian Stock market has identified eleven major sectors [2] to categorize the stocks. Trend analysis [3] is the process of estimating the entity which is in trend or has grabbed attention among the participating entities. Some of the stocks may be in trend due to several reasons that is season, price, need, dependency, alternate availability, price down in jewelery, currency market and so on.

Data stream analysis [4] is one of the most challenging process in internet applications. Due to its continuous availability and updations an efficient techniques to process and declare inferences are required. Stock market [5] produces enormous amount of data in the repository. The analysis, management of abundant data and producing acceptable results is one of the biggest challenge [6] to the computer scientists because the behavior of the system varies as the new data is added to repository.

Classification model may provide certain revolutionary inferences, perhaps classification on data stream may tend to lot of openings with respect to performance of the classification model. Classification models begins with feature extraction that is properties which may define samples as per the analysis.

The process of feature extraction on data stream and classification of these features has created number of avenues in the research domain. In this paper, we mainly focus on the estimation of the sector (collection of stocks belongs to similar properties), which is in trend for a given time period through classification techniques. Features are extracted by considering live data from the NSE server and clustered. The proposed work has conducted experiments using ANACONDA [7] and Jupyter tools by involving nsepy [8] python package for live data.

2. REVIEW OF LITERATURE

Since, beginning of the stock trading several experiments are in progress to invent required declarations. Author [9] discuss the volatility of Kuala Lampur Composite Index using stochastic volatility (SV) models and Generalized Auto regressive conditional heteroscedasticity (GARCH) models. The model results prove the slight differences in Root Mean Squared error and Moving Average Envelope. Totally 971 daily observations of KLCI Closing price index, from 2nd January 2008 to 10th November 2016, excluding public holidays. SV model is found to be the best based on the lowest RMSE and MAE values. Author [10] involves study of financial market for five different companies from Malaysia namely CIMB, Sime Darby, Axiata, Maybank and Petronas using Machine Learning Algorithms. Two types of experiments were conducted based on the type of data. The first experiment used textual data using financial news involving 6368 articles and classified as positive or negative using SVM. The second experiment used numeric historical data involving 5321 records to predict the stock price is going up or down using Random Forest algorithm. Author [11] has tried to propose an embedded streaming SVM classification architecture for continuous data processing. Paper [12] presents dynamic way of selecting the number of clusters in K-means clustering algorithm. The proposed algorithm is applied for clustering iris dataset and the performance of the algorithm is measured using inter cluster distance and sum of squared error parameters and compared with General K-Means algorithm. Author [13] has proposed a classification model to answer complex question answering process.

Author [14] presents the effects of news in online social media effects purchase of pharmaceutical stocks. The experiment is conducted using Nifty pharma index data and developed sentiment analysis model for pharma stock prediction. The sentiment analysis model achieved an accuracy of 70.59% in predicting daily stock movement. Author [15] presents analysis and prediction of US real time stocks data from yahoo finance using big data analytics. A machine learning model is developed to predict the future crude oil price using the United States Oil fund (USO) data. The model identifies the best features for better oil price prediction. In paper [16] presents an analysis of one year US stock market based on Network approach. The paper addresses the correlation of one stock with other stocks and also identifies the key players in the market based on their number of dependencies.

Author [17] addresses the selection of stock using both technical and fundamental information. A framework is designed to make class predictions for the industrial sector of the Australian stock market. The stock selection, trading strategy outperformed the Australian stock index. The accuracy of the classification models like Decision tree, CHAID tree and Neural network is compared.

3. METHODOLOGY

Figure 1 depicts the methodology of the proposed research work.

3.1. Data collection

Classification would result most useful inferences, these inferences mainly depend on the applicable data which is collected from the environment. This work requires a data stream that is the live data, which is continuous and may fall within some range. The range of data from start time to end time decides the stock and its trend in the market. nsepy is the python package used to access live NSE India stock trading data. This package provides the parameters of each stock namely totalTradedVolume, totalTradedValue, Open, Close, dayHigh, dayLow and so on.

3.2. Feature extraction

The proposed methodology is considered the stock data, which is with respect to time as a discrete signal. Hence all the applicable features corresponding to discrete signals are considered as features in the proposed work. In-spite of many features, only features are considered as per the analysis and which gives

better results and this process is called as feature selection [18]. The range of data from start time to end time decides the stock and its trend in the market. In order to achieve better analysis the type of data and its importance does matter in the classification and conclusion. The importance and its type can be found based on experience of the stock trading or through computing analysis. Stock data have several parameters, namely totalTradedVolume, totalTradedValue, Open, Close, dayHigh, dayLow and so on. Theses parameters are used for further feature extraction.

3.2.1. Standard deviation

Since the model operates on the data which is with respect to time, the amount of standard deviation [19] within the members is essential to estimate. This feature mainly produces the amount of fluctuation among the members of the data stream. Stock data stream is a sequence of values with respect to time or date. This paper has considered five properties from the get_history of numpy package of python. These five properties are Open, Close, Low, High and Volume. For each stock and each property Standard Deviation is estimated.

3.2.2. Kurtosis

K̈urtosis is a measure of the combined weight of a distributionś tails relative to the center of the distribution. This measure may declare the rise in the distribution if the measure turns to positive [20]. Figure 2 clearly depicts the positivity and negativity nature of the measure along with distribution pattern.

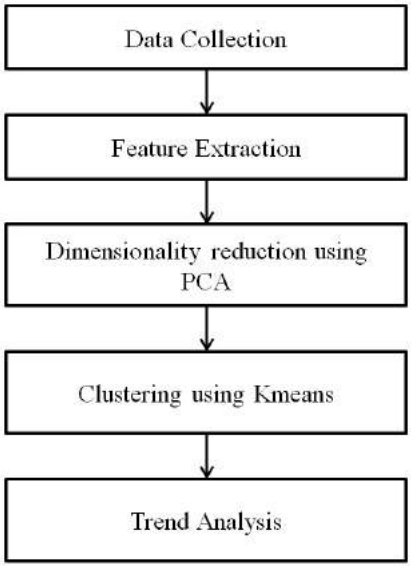

Figure 1. Proposed methodology for trend analysis in stock data stream classification

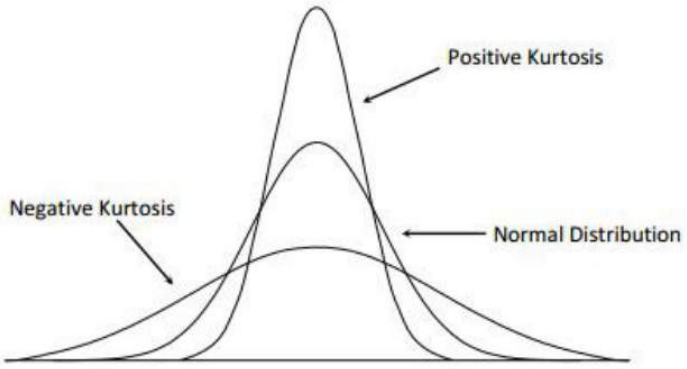

Figure 2. Tailed and centered distribution

3.2.3. Augmented Dickey-Fuller Test [21]

This feature applies to Non-stationary Time variant systems. Stock market data series can be regarded as non-stationary because the mean and variance of the system is varied at any point of time. This model is well suitable for stock market data to analyze the stock trend. This paper uses the unit root with drift test analysis of the Dickey-Fuller test. Unit root or stationarity of the distribution can be estimated using the (1).

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{M} \alpha_i \Delta Y_{t-i} + \mu_t. \quad (1)$$

Figure 3 depicts the data stream which is in trend and the same with drift. This paper uses only deterministic time trend coefficient as a feature for further classification.

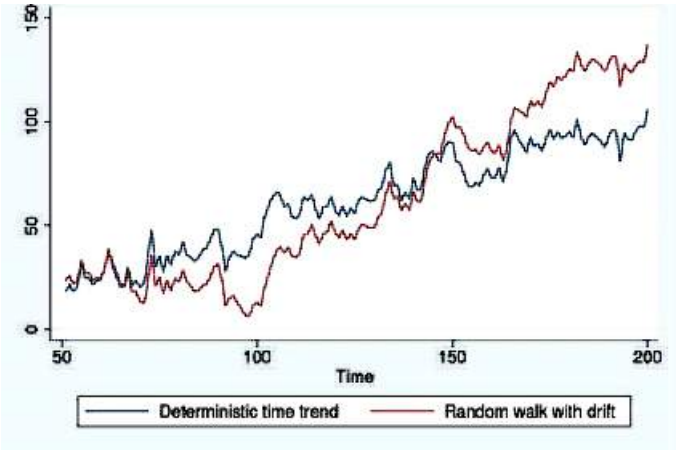

Figure 3. Trend analysis with drift

3.3. Dimensionality reduction using PCA

The proposed methodology extracting around twelve features that is three features from each property (Open, Close, High, Low, Volume) of the stock data. This process is defined around twelve features, sometimes all these features may or may not play an important role in the classification. Hence, dimensionality reduction [22] is one of the techniques to reduce the number of features. This may reduce the complexity of the classification and may improve the process better with meaningful inferences [23]. Principal Component Analysis (PCA) is one of the readily available algorithms for Dimensionality Reduction. The proposed work reduces the twelve features to three features.

3.4. Clustering using K-means clustering algorithm

The proposed work is grouping the available stock features into two clusters using k-means clustering [24] as shown in Figure 4. Where one will be containing the stocks which are in trend and another not.

3.5. Trend analysis

This is the final phase of the methodology, which considers all the samples from cluster 2 (Cluster 2 is assumed as trend cluster, it contains all the samples whose features have given trend coefficients). Distance from origin to the centroid of the cluster declares the selection of the cluster which has stocks in trend. More the distance more will be the trend, this assumption is based on trial and error method. Apply histogram on stock category (stock indices) of the samples. As per the survey, there are eleven stock indices in Indian Stock market. Figure 5 depicts a sample histogram, which declares that banking sector index is in trend compared to all other sectors. The histogram [25] is clearly depicting the status of the sectors in the market. This status indicates that the sector number 5 is in trend. This trend may change as per the market transactions.

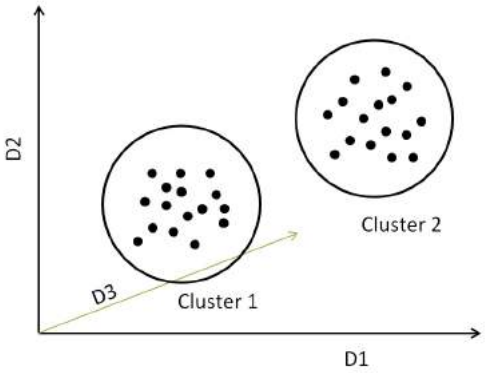

Figure 4. Samples categorized into two clusters

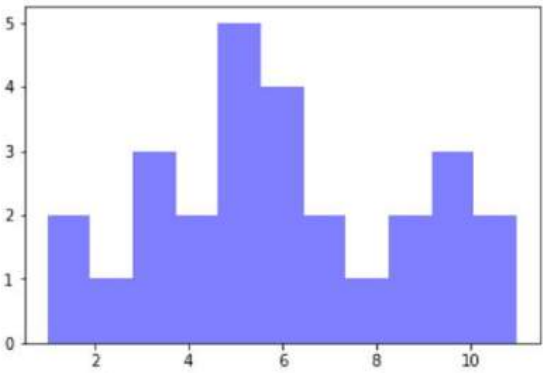

Figure 5. Histogram based on sector indices of samples

4. RESULTS AND DISCUSSIONS

As per NSE (National Stock Exchange) Nifty Auto Index, Nifty Bank Index, Nifty Financial Services Index, Nifty FMCG Index, Nifty IT Index, Nifty Media Index, Nifty Pharma Index, Nifty Private Bank Index, Nifty PSU Bank Index, Nifty Realty Index, Nifty500 Industry Indices are the eleven sector indices. In the proposed work, fifteen stocks have been considered in each sector index for the trend analysis. Figure 6 shows the feature values, extracted from the selected parameters (Open, Close, Low, High and Volume) within a given period.

stock name	sector	openstd	openkurt	openadf	closestd	closekurt	closeadf	highstd	highkurt	highadf	volstd	volkurt	voladf
MARUTI	1	218.3812	-1.27313	-0.96404	227.8751	-1.25752	-1.20263	214.3234	-1.36484	-2.68182	258186.8	-0.1118	-3.52247
BAJAJ-AUTO	1	63.89	-1.14013	-1.48108	62.58151	-0.83865	-2.31299	62.59126	-0.68892	-2.49222	228406.5	8.633229	-8.05329
MOTHERSUMI	1	4.603952	-1.19438	-1.92205	4.368754	-0.8515	-2.03345	4.08587	-1.06117	-1.87724	4339124	-0.33145	-3.6699
EICHERMOT	1	361.3694	-0.56031	-1.80679	343.5338	-0.55712	-1.95993	300.1105	0.450999	1.507105	35280.78	-1.06808	-2.98429
BOSCHLTD	1	406.5581	-0.55787	-4.43087	339.8808	-0.92447	-2.00957	411.4739	-0.4699	-2.44989	5040.081	5.58563	-4.37809
M&M	1	15.59234	-0.76991	-0.74171	14.96106	-0.67234	-2.95921	14.35396	-1.0685	-2.22653	756278.9	-0.63989	-1.75104
TATAMOTORS	1	5.636448	-0.71636	-1.67089	5.206453	-0.83607	-4.64004	5.130288	-0.93161	-1.42931	4303919	-0.70614	-4.7498
HEROMOTOCO	1	81.5398	-1.11621	-2.65431	84.97591	-0.80034	-0.98796	85.86549	-1.25103	-1.80681	368147.1	3.473193	-5.31796
MRF	1	899.43	-1.11772	-4.43422	849.8417	-1.16419	-3.50729	868.2089	-1.09818	-2.14528	2515.16	2.372981	-3.66367
TVSMOTOR	1	22.77802	-1.20876	-1.70061	23.15161	-1.09449	-1.51348	23.52979	-0.88057	-2.18309	883835.1	1.460523	-2.06506
ASHOKLEY	1	2.969427	0.038127	-3.24683	2.697755	-0.69736	-2.00052	2.423338	-0.60059	-1.96336	6970241	2.10426	-18.1294
EXIDEIND	1	4.828073	-0.73882	-2.79914	4.928301	-0.6445	-2.732	4.355131	-0.74854	-2.58257	421386.6	-0.5353	-3.92209
BHARATFORG	1	13.07374	-0.9062	-1.2702	12.121	-0.9577	-1.6645	13.28105	-1.08252	-4.61857	577951.9	4.006156	-3.52824
APOLLOTYRE	1	5.077197	-0.4086	-3.00745	4.640148	0.056319	-2.48407	4.192397	0.680061	-2.90224	7164773	17.22718	-8.6329
AMARAJABAT	1	17.21686	0.582741	-1.67569	15.96625	-0.06506	-1.54237	13.76118	-0.85618	-0.68654	296950.5	-0.29704	-2.54755
HDFCBANK	2	25.72729	-0.82776	-1.60965	24.76981	-0.92744	-1.58487	23.39121	-1.12505	-0.96454	821763.7	1.363624	-3.5939
ICICIBANK	2	9.199001	-1.57836	-1.05453	9.478832	-1.55854	-0.68369	8.852839	-1.60586	-0.6083	4049549	0.57452	-3.15141
AXISBANK	2	17.80271	-0.82941	-2.70621	18.01796	-1.03265	-2.69819	16.89793	-0.93246	-2.38105	3284827	2.106818	-4.08486
RBLBANK	2	24.49846	-0.89014	-3.33579	23.03656	-0.53819	-3.50044	24.39946	-0.63153	-1.69315	648501.8	0.970155	-2.61893
KOTAKBANK	2	19.12602	-0.85514	-2.94891	20.68833	-0.93692	-2.95304	18.46283	-0.63963	-3.0755	756437	0.295309	-2.03236
FEDERALBNK	2	2.127115	-0.84008	-1.8958	2.043064	-1.1592	-1.52611	2.081161	-0.95641	-1.9181	3580000	1.948126	-4.10541
SBIN	2	9.659001	-0.9615	-0.07219	10.28805	-1.21713	0.024072	9.337508	-1.02294	0.244348	7339216	1.711708	-1.8218
INDUSINDBK	2	75.91186	-0.44618	-2.00231	76.78349	-0.29725	0.460776	75.6231	-0.59062	3.610865	2234841	0.410221	-1.51795
YESBANK	2	16.8427	-1.06147	-0.88529	17.56097	-0.71271	0.441967	17.14964	-1.02432	-0.59439	49073354	0.116803	-2.57878

Figure 6. Feature values of selected stocks from sectors

Figure 7 shows the results from PCA (dimensionality reduction), which is applied on features shown in Figure 6. Here the standard deviations, kurtosis and adfs of open, close, high and volume properties into single columns respectively as std, kurt and adf columns. PCA reduces the complexity by extracting necessary features and classification process.

K-means does clustering the given samples into two clusters. The last column of the Figure 7 indicates cluster 1 by 0 and cluster 2 by 1. Figure 8 shows the histogram on the Index column of the Figure 7 by considering only cluster 2 samples.

	sname	index	std	kurt	adf	labels
0	MARUTI	1	-2.979950e+06	-3.134316	-2.810820	0
1	BAJAJ-AUTO	1	-3.009731e+06	5.645651	-2.244067	0
2	MOTHERSUMI	1	1.100987e+06	-3.314986	-1.856616	0
3	EICHERMOT	1	-3.202856e+06	-3.927426	-2.011068	0
4	BOSCHLTD	1	-3.233097e+06	2.633854	0.664823	0
5	M&M	1	-2.481858e+06	-3.602840	-3.051607	0
6	TATAMOTORS	1	1.065782e+06	-3.666301	-2.093312	0
7	HEROMOTOCO	1	-2.869990e+06	0.470658	-1.110267	0
8	MRF	1	-3.235622e+06	-0.632878	0.661722	0
9	TVSMOTOR	1	-2.354302e+06	-1.528942	-2.093764	0
10	ASHOKLEY	1	3.732103e+06	-0.817206	-0.376364	1
11	EXIDEIND	1	-2.816750e+06	-3.477271	-0.968420	0
12	BHARATFORG	1	-2.660185e+06	1.011176	-2.485256	0
13	APOLLOTYRE	1	3.926636e+06	14.354673	-0.707357	1
14	AMARAJABAT	1	-2.941187e+06	-3.182838	-2.127436	0
15	HDFCBANK	2	-2.416373e+06	-1.621927	-2.179545	0
16	ICICIBANK	2	8.114116e+05	-2.486607	-2.745125	0
17	AXISBANK	2	4.669005e+04	-0.874200	-1.061579	0
18	RBLBANK	2	-2.589635e+06	-1.969268	-0.452032	0
19	KOTAKBANK	2	-2.481700e+06	-2.658926	-0.833703	0
20	FEDERALBNK	2	3.418633e+05	-1.039678	-1.879189	0
21	SBIN	2	4.101079e+06	-1.285365	-3.751495	1
22	YESBANK	2	5.101079e+06	-1.385365	-3.851495	1
23	INDUSINDBK	2	-1.003296e+06	-2.501646	-1.857518	0
24	IDFCFIRSTB	2	1.855407e+06	6.292464	2.639913	0
25	PNB	2	5.522142e+06	-2.515992	-2.943419	1
26	BANKBARODA	2	7.504701e+06	-3.560306	-3.153296	1
27	HDFC	3	-1.984585e+06	6.032452	-1.273109	0
28	IBULHSGFIN	3	1.139356e+07	-1.401513	-3.018428	1
29	BAJAJHLDNG	3	-3.217540e+06	-2.172128	-2.108787	0
..	...	...	...	...	...	...
96	FEDERALBNK	8	3.418633e+05	-1.039678	-1.879189	0
97	DCBBANK	8	-2.380482e+06	-2.995978	-1.508862	0

Figure 7. PCA and K-means clustering results

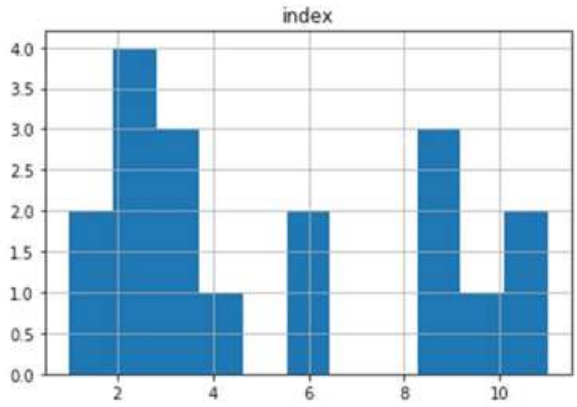

Figure 8. Histogram on sector indices considering only cluster 2 samples

The histogram is clearly declaring that the second sector that is banking sector stocks were in trend during the given period of time in the market. The classification models certainly improves the efficiency of the process which involves large amount of data. The extracted features like Standard Deviation, Kurtosis and Dickey Fuller Test have yielded the result which is acceptable as per the statistics.

5. CONCLUSION

The stock market has tremendous opportunities for businessman, manufacturer, investor and even for data analyst to study the behaviour of environment and society. In this regard, this paper has tried to analyze the stock data stream to estimate trend sector index in the market based on feature extraction and unsupervised clustering (K-means) technique. It has been implemented and demonstrated the results by fetching stock data stream from the server through nsepy package of python. The proposed work has not been considered any performance analysis of the model and the same can be enhanced as a new proposal through big data analytics.

ACKNOWLEDGMENT

Authors acknowledge and thank to Management, Director and Principal of Bapuji Institute of Engineering and Technology, Davangere for providing an opportunity and platform to conduct an experiment and produce meaningful results.

REFERENCES

[1] Nguyen, Hai-Long, Yew-Kwong Woon, and Wee-Keong Ng, "A survey on data stream clustering and classification," *K̈nowledge and information systems*, vol. 45, no. 3, pp. 535-569, 2015.
[2] [Online] Avaible : https://www.niftyindices.com/indices/equity/sectoral-indices.
[3] [Online] Avaible : https://www.investopedia.com/terms/t/trendanalysis.asp.
[4] William McKnight, "Chapter Eight - Data Stream Processing: When Storing the Data Happens Later," *Editor(s): William McKnight, Information Management, Morgan Kaufmann*, pp. 78-85, 2014.
[5] Sachdeva, Akshay, et al., "An Effective Time Series Analysis for Equity Market Prediction Using Deep Learning Model," *International Conference on Data Science and Communication*, 2019.
[6] Shanmugam, D. B., et al., "Data Stream Clustering Challenges and Management System," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 5-6, pp. 2393-2397, 2019
[7] Raschka, Sebastian, and Vahid Mirjalili, "Python machine learning," *Packt Publishing Ltd*, 2017.
[8] [Online] Avaible : https://nsepy.readthedocs.io/en/latest/.
[9] Ezatul Akma Abdullah, and Siti Meriam Zahari, "Modelling volatility of Kuala Lumpur composite index (KLCI) using SV and garch models", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 3, pp. 1087-1094, 2019.
[10] Puteri Hasya Damia Abd Samad, Sofianita Mutalib, "Analytics of stock market prices based on machine learning algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 1050-1058, 2019.
[11] J. Sirkunan, J. Tang, N. Shaikh-Husin, and M. Marsono, "A streaming multi-class support vector machine classification architecture for embedded systems," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 16, no. 2, pp. 1286-1296, 2019.
[12] Md. Zakir Hossain, Md. Nasim Akhtarn, "Ä dynamic K-means clustering for data mining,," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 2, pp. 521-526,2019.
[13] Reddy, A., and Madhavi, K., "Hierarchy based firefly optimized K-means clustering for complex question answering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol.17, no. 1, pp. 264-272, 2020.
[14] Dev Shah, Haruna Isah, Farhana Zulkernine, "P̈redicting the Effects of News Sentiments on the Stock Market,," *IEEE International Conference on Big Data (Big Data)*, 2018.
[15] Zhihao PENG, "S̈tocks Analysis and Prediction Using Big Data Analytics,," *International Conference on Intelligent Transportation, Big Data and Smart City*, 2019.
[16] Susan George,Manoj Changat, "Network approach for stock market data mining and portfolio analysis," *International Conference on Networks and Advances in Computational Technologies*, 2017.
[17] Hargreaves, Carol, and Yi Hao, "Does the use of technical and fundamental analysis improve stock choice: A data mining approach applied to the Australian stock market," *International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, 2012.
[18] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," *IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-6, 2014.

[19] Altman, D. G., and J. M. Bland, "Standard deviations and standard errors," *Bmj*, vol. 331, no. 7521, 2005.
[20] Mardia, Kanti V, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, pp. 519-530, 1970.
[21] Harris, Richard ID, "Testing for unit roots using the augmented Dickey Fuller test: Some issues relating to the size, power and the lag structure of the test," *Economics letters*, vol. 38, no. 4, pp. 381-386, 1992.
[22] Lotlikar, Rohit, and Ravi Kothari, "Adaptive linear dimensionality reduction for classification," *Pattern Recognition*, vol. 33, no. 2 pp. 185-194, 2000.
[23] Cao, L. J., et al., "A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1-2, pp. 321-336, 2003.
[24] Jain, Anil K., "Data clustering: 50 years beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651-666, 2010.
[25] Pizer, Stephen M., et al., "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355-368, 1987.