FORESEE THE MOBILE PHONE SALES USING HYBRID DATA BASED PREDICTION MODEL

K.R.Prasanna Kumar[1], C.Pavithra Dharshini [2], V.Nivedha[3], R.Santhiya [4], K.Logeswaran[5], A.P.Ponselvakumar[6]

[1] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*
krprasanname@gmail.com

[2] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*
pavithradharshini1399@gmail.com

[3] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*
nivedhaanand19@gmail.com

[4] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*

santhiyajaya987@gmail.com

[5] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*

[6] *Department of Information Technology, Kongu Engineering College, Tamil Nadu, India.*

ABSTRACT:

In past decades, usually people purchase electronic products or gadgets at nearby retail stores or from direct brand showrooms. The manufacturers collect feedback from the customers via sales points, calls, messages, emails and feedback forms during service. The customer feedback plays a vital role in improving the product quality as well as to know the need of the customers. These reviews may not reach the new customers and as well as the originality of the reviews are not ensured. In recent days of thriving information technology, because of huge arrival of shopping portals like Flipkart, Amazon and so on people started to buy products via these portals. These portals beyond sharing the product information also allows the buyers to share their feedback as well as the experience with the purchased product. New buyers do read those online reviews, comments and compare dozens of stores and products before deciding to purchase a product. These customer comments also serve as a source for the companies to predict the sales of their present product and tentative prediction of future sales. By collecting these reviews and stock market values would help the companies to make an estimation of sales take place in the near future. In this paper, Hybrid Data based Prediction Model (HDPM) is proposed which includes sentiment scores of customer reviews and stock market values to improve the forecasting accuracy.

KEY TERMS: Predict, Forecasting, Mobile sales, Hybrid data, Multiple linear regression.

I. INTRODUCTION

Personal recommendations and word of mouth references were once the primary way for customers to collect the information and performance about the particular product. With the growth of online shopping, customers have become increasingly keen to share their opinions and feelings in the online shopping portals and also in the social media platforms [1, 13]. Beyond the online shopping portals customers convey their voice to the world by sharing their reviews in social media portals like Twitter, Facebook, YouTube and so on [6, 29]. So most of the consumers trust online reviews rather than personal recommendations and most of the people will hesitate to buy a product that has more negative comments and rating. The online customer comments contain plenty of knowledge regarding merchandise and services. Since people's direct opinions will reflect the pros and cons of a product, understanding the views and emotions in user comments is very important. Mining user reviews is of great significance which not only helps the potential users for making them purchasing decisions but also enable the companies to get their product feedback and to predict the future sales [2, 27]. Sales prediction helps the companies to gauge the demand of their products and to make a strategic plan to increase their growth and quality of the products.

The mobile phone sales market has always been very competitive, and recently with the growing variety of new products it is very challenging to predict which brand will dominate the market. Having an estimate of the product sale before the release of the product would provide the company with prior knowledge of its profit and also help them decide the quantity of the release in different regions based on the request. Influences of social media comments play an important role in sales forecasting [3, 31]. The information collected in the form of reviews shows the consumer's opinion towards the product and it acts as the major input for sales prediction [4].

Social media platforms and e-commerce platforms are the emerging approach to obtain information by analyzing user comments [5]. Thus, sales prediction is performed using sentiment analysis of customer reviews. Besides online comments, stock market values also have influences on purchasing power of mobile phones. The closing values of a company in a time period is used to predict the market's movement over time. Previous studies on sales forecasting stated that univariate data containing stock market values gives less forecasting accuracy. In this paper, both time series models and multivariate regression models are used to predict the mobile phone sales of the particular brand. Stock market values forecasting is performed using time series analysis. Time series analysis is the method for analyzing historical data to forecast values based on previous observed values [17]. In multivariate regression analysis, two or more variables are used to predict the value of one variable. In this study, customer comments and stock markets values are employed to predict sales using multivariate regression model.

II. RELATED WORK

Ping-Feng Pai *et al*. proposed a set of multivariate regression techniques such as time series models and and Least Square Support Vector Regression (LSSVR) models to predict the

monthly total vehicle sales. The customer reviews from twitter and the stock market values are considered as the input for the forecasting using LSSVR. The time series data models are constructed based on the stock market values [6]. From the tweets, the senti-scores are calculated which used to denote the positive and negative sentiment information. Hybrid data with deseasonalizing procedures are employed to compare the forecasting accuracy of monthly total vehicle sales [6]. The results clearly show the hybrid approach provides more accuracy when compared to the traditional approach.

Nandal *et al*. provided a survey on sentiment analysis and opinion mining [27]. The aspect level sentiment analysis is performed based on the mining of reviews [7]. The web page information is extracted using web crawler API, which extract information from the web pages. The classifiers which can be utilized for performing classification are Decision trees, Naive Bayes (NB), Support Vector Machine and K-Nearest Neighbor [7]. The results are evaluated in terms of precision and accuracy.

Guixian Xu *et al*. proposed the topics-based sentiment analysis method for big data. Integration of topic semantic information into text representation is done through a neural network model [8]. The attention mechanisms introduced into the neural network. This mechanism compares the sentences with the relevant context. Context-aware vector is introduced to calculate the weight of each word. In addition, to make the model more adaptable, the method of sentiment dictionary tagging is used to obtain the training data. The proposed model can effectively improve the accuracy of sentiment analysis [8].

Liu *et al*. proposed a sentiment analysis method based on Transfer Learning (TL). In the industry [9], user preference for different commodities are obtained through the product reviews. The implicit emotion in the text are effectively mined, which can help enterprises or organizations to make an effective future decision. The explosive growth of data undoubtedly brings more opportunities and challenges to sentiment analysis. TL algorithms employed are based on the Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Hybrid Neural Network (HNN) models [9]. The applications of TL put forward to the development trend of sentiment analysis.

Jagdale *et al*. [10] proposed a dictionary-based approach using NB and Support Vector Machine (SVM) algorithms. In the dictionary based approach, a small set of opinion words from twitter are collected manually as a seed. Seed is a potentially used information that generates pseudo random number. The different datasets are collected for different events from Twitter [10]. Then well-known dictionaries or thesaurus are used to expand the set of opinion words. Opinion mining is performed where the sentiscore of the texts are obtained.

Elwalda *et al.* developed a model called Perceived Derived Attributes (PDA) [11]. PDA clearly depicts the importance of the customer reviews in product selection. The online shoppers usually trust the e-commerce portal who provides customer reviews importantly for customers who checks the reviews frequently before placing the purchasing [11]. The analysis clearly

shows the online customer reviews which in turn affects the shopper's decision to buy products can influence sales results.

Fan *et al*. deployed the novel model that combines the Bass/Norton model and sentiment analysis to predict sales for automotive industry [12]. The historical sales data and online user reviews are considered for predicting the future sales of the product [12]. The results clearly showed the integrated novel model generated high forecasting accuracy than other forecasting models.

III. PROPOSED METHODOLOGY

In this paper, HDPM is proposed to predict the mobile phone sales using stock market values and customer reviews. HDPM consists of both time series models and multivariate regression model. The time series models include naive forecasting method, exponential smoothing technique, Auto Regressive Integrated Moving Average (ARIMA) model. Time series models are used to deal with univariate data (stock market values) [13]. Multivariate regression model includes Multiple Linear Regression (MLR) model which is used to deal with multivariate regression data or hybrid data [6, 16]. The hybrid data or multivariate regression data consists of stock market values and numeric sentiment scores of online customer reviews. SentiStrength detection technique is used to find the sentiment scores of online customer reviews [14, 25].

A. Naive forecasting method

The naive forecasting method is the estimation technique during which the last period's actual value is used as the current period's forecast value, without adjusting them or attempting to determine causal factors. The formula is expressed in the equation (1) [6, 18, 19,31].

$$Z_t = P_{t-1} \quad (1)$$

where Z_t denotes the calculated forecast value at time t, and P_{t-1} indicates the actual previous value at time (t-1) [6].

B. Exponential smoothing method

The exponential smoothing method is the statistic forecasting method for univariate data that uses linear combination of past forecast values. This method combines error, trend and seasonal component in the calculation. It minimizes the information storage requirements. The formula for calculating forecast value is represented in the equation (2) [6, 20, 21].

$$Z_t = Z_{t-1} + \alpha(P_{t-1} - Z_{t-1}) = \alpha P_{t-1} + (1-\alpha)P_{t-1} \quad (2)$$

where α is the smoothing factor $0 < \alpha < 1$, Z_t is the Forecast value, Z_{t-1} is the Previous forecast value, and P_{t-1} is the Actual previous period [6].

C. ARIMA model

The ARIMA model is a time series analysis model that uses statistical data to predict future scope of a product. It predicts the future sales value by finding the differences in the time series values instead of actual values. It eliminated the noise or irregularity attached in a time series [17]. ARIMA model is a combination of two models Auto Regressive (AR) and Moving Average (MA) and the binding part is the integration part (I). AR model is the correlation between the previous time period to the current. ARIMA model has three parameters, auto regressive lags (p), order of differentiation (d) and moving average (q). Partial Autocorrelation (PA) graph is used to predict the value of p and Auto Correlation (CF) graph is used to predict the value of q. The formula for calculating AR(p) is expressed by the equation (3) [6, 15, 22].

$$V_t = X + \sum_{i=1}^{p} z_i V_{t-i} + e_t \quad (3)$$

where p denotes the order, X is the constant, e_t indicates the error at time t, and z_i is the coefficient of autoregressive V_{t-i}. The error accumulation at the autoregressive is calculated using the equation (4) [6].

$$V_t = Y - \sum_{i=1}^{q} y_i e_{t-i} + e_t \quad (4)$$

where q is the order of differentiation, Y indicates the mean of the series and y_i is the coefficient of e_{t-i}. ARIMA (p, q) is expressed by the equation (5) [6].

$$V_t = X + \sum_{i=1}^{p} z_i V_{t-i} + Y - \sum_{i=1}^{q} y_i e_{t-i} + e_t \quad (5)$$

D. MLR model

MLR model is the most typical kind of regression analysis. MLR is used when prediction of a variable (dependent variable or criterion variable) is based on the value of two or more variables and correlation is analyzed. The multiple regression is employed to clarify the relationship between one continuous variable and two or more independent variables. The independent variables may be continuous or categorical. The formula for calculating values with k predictor values and q response is given in the equation (6) [6, 23, 24].

$$y = \beta 0 + \beta 1 p_1 + \beta 2 p_2 + \cdots + \beta k p_k + \varepsilon \quad (6)$$

where β0, β1, β2…, βk are regression coefficients.

IV.DATA COLLECTION:

In this paper, two kinds of time series data were used to predict the mobile sales of the brands Apple and Samsung. The first dataset is the historical monthly stock market values of the brands were obtained from the Yahoo Finance website for the period from April 2017 to January 2020. The dataset contains information like date, open, high, low, close and volume. The second

one consists of the customer reviews of the products for the considered brands which taken for the e-commerce store Amazon from www.kaggle.com. The collected reviews are preprocessed by eliminating the stop words like punctuations, comma and so on also the texts with exactly same contents are removed. The noises like website references and special characters are deleted [29]. The SentiStrength approach [25] [26] is used to calculate the sentiment scores for the preprocessed reviews. The lexicon-based classifier is used to compares text against the sentiments and assign scores from -0.5 to +5.0. The positive number indicates a favorable attitude whereas a negative number indicates negative sentiments.

The dataset is then divided into training dataset and testing dataset [6]. Training dataset is used to train the model for performing different actions. Test dataset is used to see how well the machine can predict new values based on its training. The period of training dataset is from April 2017 to April 2019 and testing data is from May 2019 to January 2020.

V. RESULTS:

Prediction results of monthly total mobile phone sales using HDPM are illustrated in this section. The forecasting performance is measured by calculating error percentage of individual models. Mean Absolute Percentage Error (MAPE), Weighted Mean Absolute Percentage Error (WAPE) [30] and Normalized Mean Absolute Error (NMAE) [28] are used to calculated forecasting error.

The MAPE is a measure of forecasting accuracy in statistics. This measures the accuracy as an error percentage. The MAPE value is calculated using the formula (9) [6].

$$\text{MAPE}(\%) = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{Y_t - F_t}{Y_t}\right| \quad (9)$$

The WAPE is a measure of forecast error. It controls the infinite error problem of MAPE. The WAPE value is calculated using the formula (10) [6].

$$\text{WAPE}(\%) = 100\frac{\sum_{t=1}^{N}|F_t - Y_t|}{\sum_{t=1}^{N}Y_t} \quad (10)$$

The NMAE is the normalized absolute error where average of mean error is normalized. The NMAE value is calculated using the formula (11) [6].

$$\text{NMAE} = \frac{1}{Y_h - Y_l}\left[\frac{1}{N}\sum_{t=1}^{N}|Y_t - F_t|\right] \quad (11)$$

where N indicates the total number of prediction period, Y_t denotes the actual value at time t, F_t is the calculated forecast value at time t, Y_h represents the highest actual value, and the lowest actual value is indicated as Y_l.

TABLE 1. Values of MAPE, WAPE and NMAE to predict the total Apple mobile phone sales.

Models	**Naïve Model**	**Exponential Smoothing Model**	**ARIMA Model**	**MLR model**
MAPE	6.95%	8.95%	10.87%	4.96%
WAPE	2.61%	8.21%	8.70%	1.15%
NMAE	0.18	0.34	0.37	0.07

TABLE 2. Values of MAPE, WAPE and NMAE to predict the total Samsung mobile phone sales.

Models	**Naïve Model**	**Exponential Smoothing Model**	**ARIMA Model**	**MLR model**
MAPE	16.58%	19%	18%	9%
WAPE	7.46%	9.3%	8.45%	6.6%
NMAE	0.2	0.3	0.26	0.18

Table 1 illustrates the forecasting error percentage calculated using MAPE, WAPE and normalized absolute error calculated using NMAE for Apple mobile phone sales. Here the α value is taken as 0.2 for exponential smoothing model and (p, d, q) value as (1,1,1) for ARIMA model. The time series models obtain an average MAPE of 8.92% and MLR model obtains 4.96%. The approximate error percentage variation between time series models and MLR models are from 3% to 6%. Similarly, time series models obtain an average normalized absolute error of 0.19 which is higher than the MLR model. The overall result analysis shows that MLR model gives less forecasting error than its competing time series models.

Table 2 illustrates the forecasting error percentage calculated using MAPE, WAPE and normalized absolute error calculated using NMAE for Samsung mobile phone sales. The α value is taken as 0.2 for exponential smoothing model and (p, d, q) value as (2,2,2) for ARIMA model. The time series models obtain an average MAPE of 18% and MLR model obtains 9%. The

approximate error percentage variation between time series models and MLR models are from 8% to 11%. Similarly, time series models obtain an average normalized absolute error of 0.25 which is higher than the MLR model. It can be seen that MLR model performs better than other time series models as it gives less forecasting error.

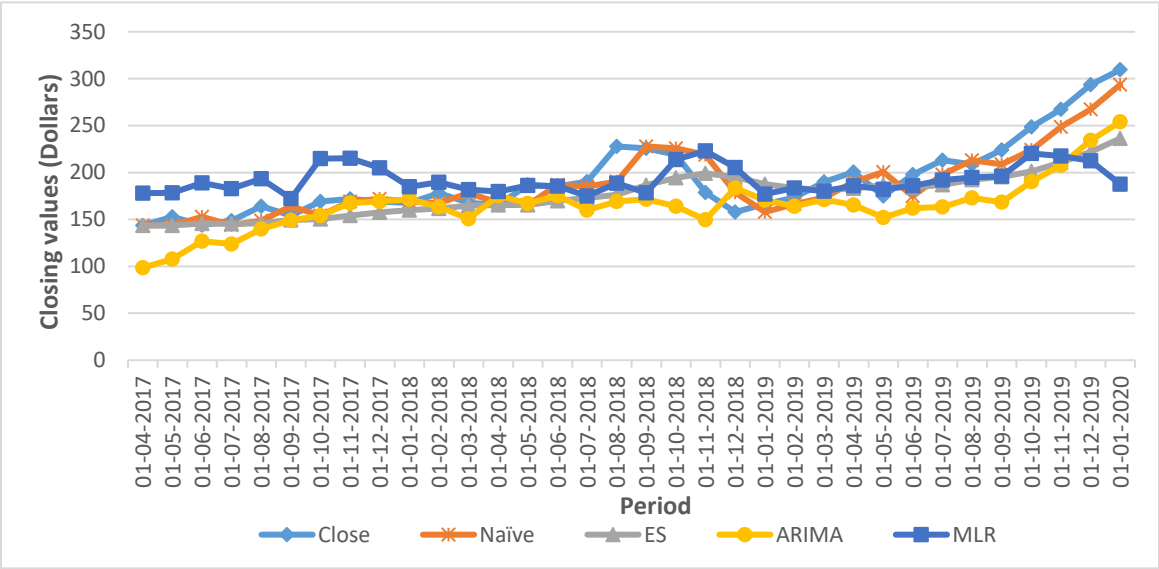

FIGURE 1. Illustrates the actual and predicted monthly sales of Apple mobile phones using HDPM.

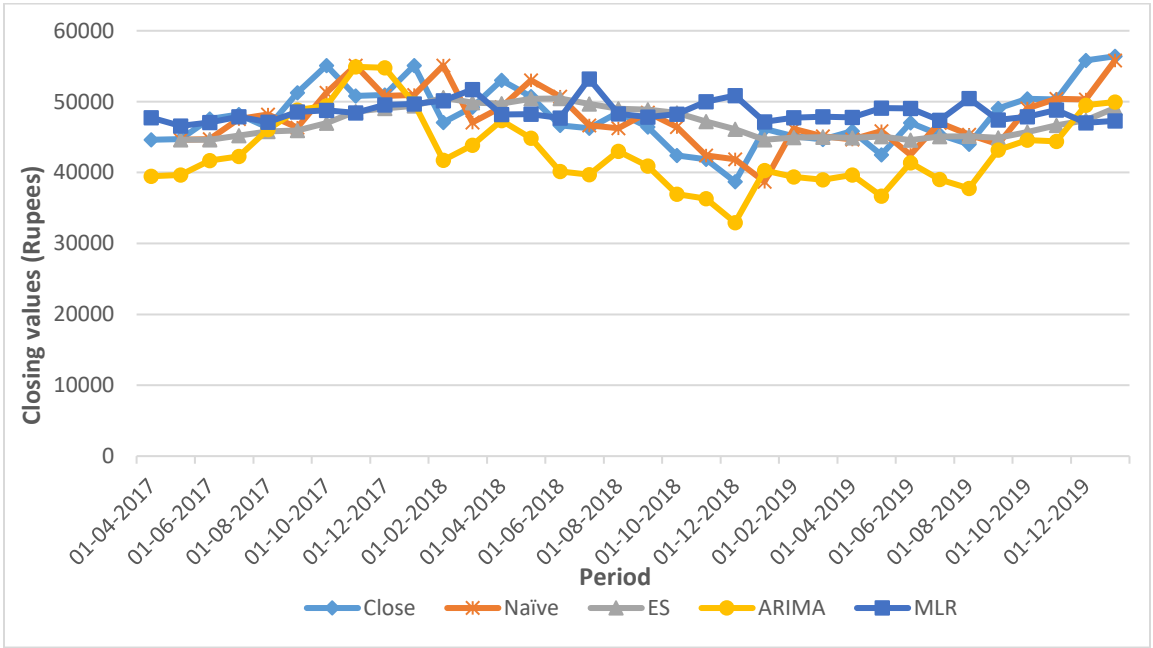

FIGURE 2. Illustrates the actual and predicted monthly sales of Samsung mobile phones using HDPM.

Figure 1 shows the comparison of HDPM predicted mobile phone sales for Apple brand and actual values. Figure 2 illustrates the comparison of actual values and predicted sales values of Samsung mobile phones using HDPM. The results clearly show the hybrid data based approach provides better accuracy than traditional single data based prediction techniques.

VI. CONCLUSION

The HDPM is proposed with an objective to increase accuracy for short term forecasting using multivariate regression data. The numerical results show that forecasting mobile phones sales by MLR model obtains more accurate forecasting results than other forecasting models. It clearly shows the use of hybrid data or multivariate data which includes sentiment analysis of consumer comments and stock values would decrease the forecasting error percentage. In the future, social media data such as Twitter, Facebook and YouTube are may be considered for forecasting.

REFERENCES:

[1] Z. Xiang, Q. Du, Y. Ma, and W. Fan, ‘‘A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism,’’ Tourism Manage., vol. 58, pp. 51–65, Feb. 2017.

[2] I. Erkan and C. Evans, ‘‘The influence of eWOM in social media on consumers’ purchase intentions: An extended approach to information adoption,’’ Comput. Hum. Behav., vol. 61, pp. 47–55, Aug. 2016.

[3] N. Kim and W. Kim, ``Do your social media lead you to make social deal purchases? Consumer-generated social referrals for sales via social commerce," Int. J. Inform. Manage., vol. 39, pp. 38_48, 2018.

[4] A. Alalwan, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, ‘‘Social media in marketing: A review and analysis of the existing literature,’’ Telematics Inform., vol. 34, no. 7, pp. 1177–1190, 2017.

[5] S. E. Shukri, R. I. Yaghi, I. Aljarah, and H. Alsawalqah, ‘‘Twitter sentiment analysis: A case study in the automotive industry,’’ in Proc. IEEE. JordanConf. Appl. Electr. Eng. Comput. Technol. (AEECT), Amman, Jordan,Nov. 2015, pp. 1–5.

[6] P. Pai and C. Liu, "Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values”, IEEE Access, vol. 6, pp. 57655-57662, 2018.

[7] Neha Nandal, Jyoti Pruthi and Amit Choudhar,” Aspect Based Sentiment Analysis Approaches with Mining of Reviews: A Comparative Study”, vol.7, Issue-6, March. 2019.

[8] Guixian Xu1, Ziheng Yu1, Zhan Chen1, Xiaoyu Qiu2, and Haishen Yao1,"Sensitive Information Topics-Based SentimentAnalysis Method for Big Data", Aug.2019.

[9] Ruijun Liu 1,2, Yuqian Shi 1,3, Changjiang Ji 2, and Ming Jia 1, "A Survey of Sentiment Analysis Based on Transfer Learning", vol. 7, 2019.

[10] Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Deshmukh," Sentiment Analysis on Product Reviews Using Machine Learning Techniques ", Springer Nature Singapore Pte Ltd. 2019.

[11] A. Elwalda, K. Lü, and M. Ali, ``Perceived derived attributes of online customer reviews," Comput. Hum. Behav., vol. 56, pp. 306_319, Mar. 2016.

[12] Z. Fan, Y. J. Che, and Z. Y. Chen, ``Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis," J. Bus. Res., vol. 74, pp. 90_100, May 2017.

[13] S. C. Ludvigson and C.Steindel, ''How important is the stock market effect on consumption?'' Econ. Policy Rev.-Federal Reserve Bank New York, vol. 5, no. 2, pp. 29–51, 1999.

[14] S. E. Shukri, R. I. Yaghi, I. Aljarah, and H. Alsawalqah, ''Twitter sentiment analysis: A case study in the automotive industry,'' in Proc. IEEE. Jordan Conf. Appl. Electr. Eng. Comput. Technol. (AEECT), Amman, Jordan, Nov. 2015, pp. 1–5.

[15] D. Fantazzini and Z. Toktamysova, ''Forecasting German car sales using Google data and multivariate models,'' Int. J. Prod. Econ., vol. 170, pp. 97–135, Dec. 2015.

[16] A. Sa-ngasoongsong, S. T. Bukkapatnam, P. S. Iyer, and R. P. Suresh, ''Multi-step sales forecasting in automotive industry based on structural relationship identification,'' Int. J. Prod. Econ., vol. 140, no. 2, pp. 875–887, 2012.

[17] M. Hur, P. Kang, and S. Cho, ''Box-office forecasting based on sentiments of movie reviews and Independent subspace method,'' Inform. Sci., vol. 372, pp. 608–624, Dec. 2016.

[18] F. M. De and X. Yao, ``Short-term load forecasting with neural network ensembles: A comparative study," IEEE Comput. Intell. Mag., vol. 6, no. 3, pp. 47_56, Aug. 2011.

[19] B. Render, J. R. M. Stair, M. E. Hanna, and S. H. Trevor, Quantitative Analysis for Management, 12th ed. Upper Saddle River, NJ, USA: Pearson,2015.

[20] R. G. Brown, Statistical Forecasting for Inventory Control. New York, NY, USA: McGraw-Hill, 1959.

[21] D. Trigg and A. Leach, ``Exponential smoothing with an adaptive response rate," J. Oper. Res. Soc., vol. 18, no. 1, pp. 53_59, 1967.

[22] G. E. Box and G. M. Jenkins, Time Series Analysis: Forecasting and Control, 5th ed. San Francisco, CA, USA: Holden-Day, 1976.

[23] F. Wijnhoven and O. Plant, ‘‘Sentiment analysis and Google trends data for predicting car sales,’’ in Proc. 38th Int. Conf. Inf. Syst., Seoul, South Korea, 2017, pp. 1–16.

[24] T. Geva, G. Oestreicher-Singer, N. Efron, and Y. Shimshoni, ‘‘Using forums and search for sales prediction of high-involvement products,’’ MIS Quart., vol. 41, no. 1, pp. 65–82, 2017.

[25] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, ``Sentiment strength detection in short informal text," J. Assoc. Inf. Sci. Tech., vol. 61, no. 12, pp. 2544_2558, 2010.

[26] M. Thelwall, K. Buckley, and G. Paltoglou, ``Sentiment strength detection for the social Web," J. Assoc. Inf. Sci. Tech., vol. 63, no. 1, pp. 163_173, 2012.

[27] Logeswaran, K, Suresh, P, Savitha, S, Prasanna Kumar. K.R, Ponselvakumar, A. P. and Kannan, A. R. "Data Driven Diagnosis of Cervical Cancer using Association Rule Mining with Trivial Rule Expulsion Approach", International Journal on Emerging Technologies, 11(2): 110–115, 2020.

[28] N. Oliveira, P. Cortez, and N. Areal, ``The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices," Expert Syst. Appl., vol. 73, pp. 125_144, May 2017.

[29] K.R.Prasanna Kumar, M.Pranesh, T.G.Rakhul Raahje, P.Vignesh, Dr.K.Kousalya, K.Logeswaran, "Factual Product Recommendation System by Eliminating Fake Reviews using Machine Learning Techniques", International Journal of Advanced Science and Technology, Vol. 29, No. 7s, (2020), pp. 2729-2735.

[30] M. Beladev, L. Rokach, and B. Shapira, ``Recommender systems for product bundling," Knowl. -Based Syst., vol. 111, pp. 193_206, Nov. 2016.

[31] Logeswaran, K., P. Suresh, S. Savitha, and Prasanna Kumar.K.R. "Optimization of Evolutionary Algorithm Using Machine Learning Techniques for Pattern Mining in Transactional Database." In Handbook of Research on Applications and Implementations of Machine Learning Techniques, pp. 173-200. IGI Global, 2020.