IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 4, 2018, accepted September 28, 2018, date of publication October 4, 2018, date of current version October 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2873730

Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values

PING-FENG PAI, (Senior Member, IEEE), AND CHIA-HSIN LIU
Department of Information Management, National Chi Nan University, Puli 54561, Taiwan

Corresponding author: Ping-Feng Pai (paipf@ncnu.edu.tw)

This work was supported by the Ministry of Science and Technology of the Republic of China, Taiwan, under Contract MOST105-2410-H-260-017-MY2 and Contract MOST 107-2410-H-260-012.

ABSTRACT Owning to the booming of social media, making comments or expressing opinions about merchandises online becomes easier than before. Data from social media might be one of the essential inputs for forecasting sales of vehicles. Besides, some other effects, such as stock market values, have influences on purchasing power of vehicles. In this paper, both multivariate regression models with social media data and stock market values and time series models are employed to predict monthly total vehicle sales. The least squares support vector regression (LSSVR) models are used to deal with multivariate regression data. Three types of data, namely sentiment scores of tweets, stock market values, and hybrid data, are employed in this paper to forecast monthly total vehicle sales in USA. The hybrid data contain both sentiment scores of tweets and stock market values. In addition, seasonal factors of monthly total vehicle sales are employed to deseasonalizing both monthly total vehicle sales and three types of input data. The time series models include the naïve model, the exponential smoothing model, the autoregressive integrated moving average model, the seasonal autoregressive integrated moving average model, and backpropagation neural networks and LSSVR with time series models. The numerical results indicate that using hybrid data with deseasonalizing procedures by the LSSVR models can obtain more accurate results than other models with different data. Thus, both social media data and stock values are essential to forecast monthly total vehicle sales; and deseasonalizing procedures can improve forecasting accuracy in predicting monthly total vehicle sales.

INDEX TERMS Predict, vehicle sales, Twitter, sentiment analysis, stock markets.

I. INTRODUCTION

Before the Internet was developed, transmission of information is mainly through ways, such as leaflets, billboards, television and word of mouth. Recently, the Internet has been growing; and restrictions of space and distance have been reduced. Information of products can be updated at any time by social media. Businesses provide more real-time and convenient services and closely interact with consumers. Due to the rapid flow of messages, the electronic word of mouth and brand image is deeply influenced by social media; and purchase willingness is affected as well. In the study of Bailey *et al.* [1], questionnaires were collected and analyzed by the technology acceptance model to realize causes of consumer consumption by social media. The empirical results showed that perceived ease-of-use and the perceived enjoyment were significantly correlated with the perceived usefulness of social media. Kim and Kim [2] analyzed influences of Facebook's likes sharing or tweets on sales of four companies in Korea, and experiment results indicated that social media sharing can actually increase sales. Chang *et al.* [3] explored the impact of social media on tourism sales by taking a travel company as an example to collect data for international travels. The experiment data contains tourism sales with and without promotions on Facebook. The results revealed that Facebook is a beneficial way to promote tourism sales. Chen and Yin [4] used the Gaussian mixture model and the ordinary least squares method to predict the market of films. The study reported that a competitor's social media commentary significantly influenced box-office sales. Elwalda *et al.* [5] developed a model incorporating the theory of planned behavior into the technology acceptance to explain the effects of online customer reviews on customers' purchasing intentions. The results showed that the perceived usefulness, perceived ease of use and perceived enjoyment of customer reviews significantly affect customer trust and the intention of online purchases, especially for those customers who frequently check comments before purchasing. Therefore, online customer reviews can influence sales results.

Erkan and Evans [6] investigated the impact of the electronic word of mouth on consumers' purchase intentions. An integration model of information adoption and theory of reasoned action was presented in this study. The experiment results exhibited that quality of information, credibility, needs of information and the information adoption of social media are essential factors influencing consumers' willingness to buy products. Alalwan *et al.* [7] reviewed plenty of studies of social media in marketing and pointed out the determination of samples collected to represent opinions on social media is a critical issue. The study showed that consumer behaviors were affected by the information they received. Rui *et al.* [8] employed dynamic panel data model, support vector machines and the Naive Bayesian classifier to make a box-office prediction with word-of-mouth. The study showed that the word-of-mouth of movie sales is related to the number of followers.

Social media is an emerging approach to obtain valuable information by analyzing online information or users' comments. Thus, prediction of sales by sentiment analysis is another essential issue worth studying. Sentiment analysis employs natural language processing, text analysis and linguistics to form valuable insights from data to assist people in making decisions [9]. There are many platforms on the Internet, such as forums, blogs, quizzes, online reviews and social media which deliver or present user-generated content. The online customer reviews refer to user-generated content posted on e-commerce websites or third-party websites that are forms of electronic word of mouth. The online customer reviews are important sources of product-related information for understanding customers' attitudes toward products or services [5], [10]. Lassen *et al.* [11] employed linear regression model with Tweets to predict the quarterly sales of iPhones. The data included the number of Tweets, replies, retweets, subjectivity, the ratio of positivity to negativity, and the ratio of positive tweets. The results showed that social media data from Twitter are helpful to forecast sales of iPhones with a satisfied accuracy. Shukri *et al.* [9] used text mining and sentiment analysis to analyze customer satisfactions of three car brands, namely Mercedes, Audi and BMW. This study utilized the Naive Bayesian classifier to categorize three types of polarity and six different emotions. The experiment results showed that the emotion classifications were consistent with the polarity classification and provide relatively detailed information about the customer satisfaction. In addition, the authors reported that the user satisfaction influenced the spread of opinions on social media. Hur *et al.* [12] employed multiple linear regressions, classification and regression trees, artificial neural networks, and support vector regression models to forecast the number of films audiences. In addition to conventional predictors, the sentiment analysis values of movie reviews were treated as input variables. The numerical results revealed that the use of review sentiments can increase forecasting accuracy. Xiang *et al.* [13] used text analysis to compare three online review platforms in hospitality and tourism. The study showed that differences of information exist among review platforms when the measurements are represented by linguistic and semantic features, positive and negative emotions, ratings, and usefulness. Each platform had different characteristics of comments, reflecting the segmentation of user groups. Thus, to select sources and characteristics of online review platforms is essential when collecting samples of comments.

Sequentially, studies using social media data to predict vehicle sales are depicted as follows. Geva *et al.* [14] employed least-squares linear regression models with Google trends and social media data to forecast monthly car sales. The study showed that predictive models based on data from Google trends data or data from social media are both helpful in forecasting sales of vehicles. The combination of data from Google trends and forum can increase the forecast accuracy of cars with low price brands. While for cars with high price brands, the improvement is not significant. Fantazzini and Toktamysova [15] referred to the methodology of Sa-ngasoongsong *et al.* [16] and further proposed a model using economic variables and Google search data to predict the monthly sales of ten car brands. Both raw data and seasonally adjusted data were applied in this study. The results indicated that the Bayesian vector auto-regressive model performed rather well for all car brands in the short-term and medium-term forecasts. However, in the long-term forecasts for several brands, the simple models including only car sales and Google search data can obtain more accurate forecasting results than the other models. In addition, seasonally adjusted data are suitable for most cases. Wijnhoven and Plant [17] utilized linear regression models to forecast monthly sales of eleven car models. The data contained social media posts and Google trends data. The social media posts included the percentages of negative, positivity to negativity ratio, and total number of monthly social mention volume. The authors reported that social media sentiments had less influence on car sales prediction than the integration of Google Trends data and total number of monthly social mention volume. Fan *et al.* [18] employed comments of online reviews and historical sales data to forecast three generations of car sales. Sentiment indices were collected from the text of online reviews and analyzed by Naive Bayes algorithms incorporated with the imitation coefficient of the Bass/Norton model. The study revealed that the integrated model developed by this study generated more accurate forecasting results than the other models used in this study. Table 1 summarizes studies applying social media data to forecast car sales.

In addition, the stock market values are related to the wealth effect and consumption ability [19], [20]. Two stock market values, Dow Jones Industrial Average (DJIA) and Standard & Poor's 500 Index (S&P 500) were employed as input data to forecast monthly total vehicle sales in this study. The rest of this study is organized as follows. Section 2 presents methodologies utilized in this study. Data collection and the proposed framework

TABLE 1. Summary of studies using social media data to forecast car sales.

1. Study: Geva et al.（2015）[14] 2. Independent variables: Google's comprehensive index of Internet discussion forums, Google search trend data 3. Dependent variables: Monthly sales for 23 car brands in USA from 2007 to 2010 4. Methods: Least-squares linear regression models, Neural Networks, Support Vector Machines , Random Forest 5. Findings: The combination of two types of data can improve forecasting performance.
1. Study: Fantazzini and Toktamysova （2015）[15] 2. Independent variables: Economic variables, Google search data 3. Dependent variables: Monthly sales of ten car brands in Germany during the period from January 2001 to June 2014 4. Methods: Vector Error Correction models, Vector Auto-Regressive models, Bayesian Vector Auto-Regressive models 5. Findings: For the short-term and medium-term predictions, the Bayesian vector auto-regressive model is suitable. In the long-term predictions of several brands, the simple models including only car sales and Google search data are appropriate. Seasonally adjusted data are suitable for most cases.
1. Study: Wijnhoven and Plant（2017）[17] 2. Independent variables: Social media sentiments, Google Trends data 3. Dependent variables: Fifty-two monthly sales for 11 types of cars in the Netherlands during the period from January 2012 to April 2016. 4. Methods: Linear regression models 5. Findings: Integration of social media data and Google Trends data can improve forecasting accuracy.
1. Study: Fan et al. (2017） [18] 2. Independent variables: Online reviews data 3. Dependent variables: Monthly car sales data of Hyundai Elantra in Beijing, China during the period from April 2006 to December 2014 4. Methods: Naive Bayes algorithms, Bass/Norton models 5. Findings: The hybrid model integrating e the Bass/Norton method and sentiment analysis can improve forecasting accuracy.
1. Study: This study 2. Independent variables: Sentiment scores of tweets, Two stock market values 3. Dependent variables: Monthly total vehicle sales in USA from February 2008 to August 2017 4. Methods: Naïve, ES, ARIMA, SARIMA, LSSVR, BPNN, LSSVRTS 5. Findings: Hybrid data with deseasonalizing procedures by the LSSVR models can increase forecasting accuracy.

are introduced in Section 3. Section 4 depicts the numerical results of this study. Conclusions are delivered in Section 5.

II. METHODOLOGY

Seven approaches, namely the naïve method, the exponential smoothing technique, the autoregressive integrated moving average model, the seasonal autoregressive integrated moving average model, backpropagation neural networks, least squares support vector regression with time series models, and the least squares support vector regression were used to forecast monthly total vehicle sales in this study and depicted in this section.

The naïve method treats the actual value of the previous time period as the forecast value of the next time period [21], [22] and can be represented as:

$$\mathrm{F_i} = \mathrm{A_{i-1}} \tag{1}$$

where $\mathrm{F_i}$ is the forecast value at time i, and $\mathrm{A_{i-1}}$ is the actual value at time (i − 1).

The exponential smoothing method [23], [24] is based on the previous forecast and the percentage of forecast error. Due to the simplicity, transparency, and capabilities for capturing various time series data patterns, for many decades, the exponential smoothing technique has been one of the most popular as well as practical ways in time series forecasting. The formula is expressed as follows:

$$\mathrm{F_t} = \mathrm{F_{t-1}} + \alpha(\mathrm{A_{t-1}} - \mathrm{F_{t-1}}) = \alpha\mathrm{A_{t-1}} + (1-\alpha)\mathrm{A_{t-1}} \tag{2}$$

where α, between 0 and 1, is a smoothing constant, $\mathrm{F_i}$ is the forecast value at time i, and $\mathrm{A_{i-1}}$ is the actual value at time (i − 1).

Developed by Box and Jenkins [25], the Autoregressive Integrated Moving Average model has been one of the most popular and powerful tools for dealing with time series prediction problems. The ARIMA model includes two parts, namely the autoregressive and the moving average; and three parameters, the order of the autoregressive (p), the degree of differencing that to make the non-stationary sequence become a stationary sequence (d), and the order of the moving average model (q). The autoregressive model, AR(p), represents the relationship between current and historical values. For a sequence with variables V, the values of the previous period ($\mathrm{V_{t-1}}, \mathrm{V_{t-2}}, \ldots$) are used to forecast the value of the current period ($\mathrm{V_t}$) and expressed by Eq. (3):

$$V_t = A + \sum\nolimits_{i=1}^{p} a_i V_{t-i} + e_t \tag{3}$$

where p is the order, A is a constant, e_t is the error at time t, and a_i is the coefficient of autoregressive $V_{t-\mathrm{i}}$. The moving average model, MA(q), is the error accumulation at the autoregressive stage and illustrated as Eq. (4).

$$V_t = B - \sum\nolimits_{i=1}^{q} b_i e_{t-i} + e_t \tag{4}$$

where q is the order, B is the mean of the series, and b_i is the coefficient of e^{t-i}. The autoregressive moving average model, ARMA(p,q), contains p autoregressive terms and q moving average terms.

The model can be expressed as:

$$V_t = A + \sum\nolimits_{i=1}^{p} a_i V_{t-i} + B - \sum\nolimits_{i=1}^{q} \omega_i e_{t-i} + e_t \tag{5}$$

When the non-stationary data are used, the difference procedure is applied to transform non-stationary data into stationary data. Thus, the notation of ARIMA (p,d,q) is used, where the *d* represents the degree of differencing. Basically, the autocorrelation function and the partial autocorrelation function of the differenced series are employed to determine appropriate values of *p*, *d* and *q*.

When a time series contains seasonal and stochastic variations, variation factors need to be considered when predicting seasonal time series. The seasonal autoregressive integrated moving average model [25], [26] is one of the techniques mostly employed. A seasonal time series, $\{\mathrm{V_t}, \mathrm{t} = 1, 2 \ldots \mathrm{k}\}$, is created by SARIMA (p, d, q)(P, D, Q)$_\mathrm{S}$ process with mean *U* from Box and Jenkins [25] time series model while

$$\varphi_\mathrm{p}(\mathrm{B})\phi_\mathrm{P}(\mathrm{B^s})(1-\mathrm{B})^\mathrm{d}(1-\mathrm{B^s})^\mathrm{D}(\mathrm{V_t}-\mathrm{U}) = \theta_\mathrm{q}(\mathrm{B})\Theta_\mathrm{Q}(\mathrm{B^s})\mathrm{e_t} \tag{6}$$

where *B* is the lag operator; *p*, *d*, *q*, *P*, *D* and *Q* are integers; *p* and *q* are orders of AR and MA models; *P* and *Q* are season orders of AR and MA models; *d* and *D* are the number of regular differences and seasonal differences respectively; and *s* is the seasonal period length; $\varphi_\mathrm{p}(\mathrm{B}) = (1-\varphi_1\mathrm{B}-\varphi_2\mathrm{B}^2-\cdots-\varphi_\mathrm{p}\mathrm{B^p})$ is the regular autoregressive operator with order *p*; and $\theta_\mathrm{q}(\mathrm{B}) = (1-\theta_1\mathrm{B}-\theta_2\mathrm{B}^2-\cdots\theta_\mathrm{q}\mathrm{B^q})$ is the regular moving average operator with order *q*; $\phi_\mathrm{P}(\mathrm{B^s}) = (1-\phi_1\mathrm{B^s}-\phi_2\mathrm{B^{2s}}-\cdots-\phi_\mathrm{P}\mathrm{B^{ps}})$ is the seasonal autoregressive operator with order *P*; and $\theta_\mathrm{Q}(\mathrm{B^S}) = (1-\theta_1\mathrm{B^S}-\theta_2\mathrm{B^{2S}}-\cdots\theta_\mathrm{Q}\mathrm{B^{QS}})$ is the seasonal moving average operator with order *Q*.

Based on the Vapnik–Chervonenkis theory and structural risk minimization principle, the support vector machines [27], [28] has been one of the most influential classification techniques in recent years. For applications in regression, the support vector regression [29]–[31] was developed and has become a prevalent technique in dealing with problems of function approximation and regression estimation. However, both support vector machines and support vector regression face challenges of coping with quadratic functions with high computational complexity during the problem solving processes. By transferring a quadratic programming problem into a linear equation, some modifications of support vector machines and support vector regression models was proposed to moderate the computational complexity. The least square support vector regression (LSSVR) model [32] is one of the modified techniques which can reduce the computation load of the original support vector regression model. For a training data set $\{\mathrm{X_J}, \mathrm{Y_J}\}$, $J = 1, \ldots N$, the LSSVR model can be denoted as Eq. (7) [32], [33]. *W* and δ can be determined by the following optimization function:

$$\begin{aligned}&\text{Min}: \frac{1}{2}\|W\|^2 + \frac{1}{2}\delta\sum\nolimits_{J=1}^{N}\xi_J^2\\ &\text{subject to } Y_J = W^T\cdot\tau(X_J) + b + \xi_J, \quad J = 1,\ldots N\end{aligned} \tag{7}$$

where *W* and is the weighted vector, δ is the penalty factor, *N* is the number of data, ξ_J is the *Jth* error, X_J and Y_J are the *Jth* input and output data set, $\tau(X)$ is a mapping function, and *b* is the bias.

By the Lagrange multipliers technique, the Lagrange form of Eq. (7) is represented as follows,

$$\begin{aligned}\mathrm{L}(\mathrm{W}, \mathrm{b}, \xi, \beta) = {}&\frac{1}{2}\|W\|^2 + \frac{1}{2}\delta\sum\nolimits_{J=1}^{N}\xi_J^2\\ &-\sum\nolimits_{J=1}^{N}\beta\left(W^T\cdot\tau(X_J) + b + \xi_J - Y_J\right)\end{aligned} \tag{8}$$

Where β_J are the Lagrange multipliers.

The solution of the above function can be obtained when all derivatives are equal to zero based on the Karush–Kuhn–Tucker conditions [34]–[36]. Thus, the optimal conditions are derived by finding the partial derivative of Eq. (8) with respect to *w*, *b*, ξ and β. Then, solving the linear equations by the least squares method, the LSSVR model can be illustrated as Eq. (9):

$$Y(X) = \sum\nolimits_{J=1}^{N}\beta_J K(X, X_J) + b \tag{9}$$

$$\mathrm{K}(\mathrm{X}, \mathrm{X_J}) = \tau(\mathrm{X})^\mathrm{T}\cdot\tau\left(\mathrm{X_j}\right)^\mathrm{T} \tag{10}$$

where $\mathrm{K}(\mathrm{X}, \mathrm{X_J})$ is the kernel function fulfilling the Mercer's principle [37]. In this study, the radial basis function with the kernel width σ indicated by Eq. (11) serves as a kernel function.

$$K(X, X_J) = exp(-\frac{\|X - X_J\|^2}{2\sigma^2}) \tag{11}$$

Owing the determination of two LSSVR parameters are very time-consuming, in this study, genetic algorithms [38] were used to obtain appropriate parameters for LSSVR models [39].

III. DATA COLLECTION AND THE PROPOSED FRAMEWORK

Two categories of time series data were used to predict monthly total vehicle sales. The first category is the data set of historical monthly total vehicle sales; the second category is the data set of independent variables including sentiment scores of tweets and two stock market values. The monthly total vehicle sales data from February 2008 to August 2017 can be collected from Bureau of Economic Analysis, U.S. Department of Commerce (https://www.bea.gov/national/#supp). Monthly closing values of Dow Jones Industrial Average and Standard & Poor's 500 Index were obtained from Yahoo Finance (http://finance.yahoo.com/). Using the application programming interface, monthly Tweets were gathered by three keywords, namely "buy car," "buy truck," and "buy vehicle." Totally the number of tweets is around six million. Before texts can be analyzed by the SentiStrength [40], [41], data preprocess procedures have to be conducted. First, only the field of text is left. Secondly, remove texts with exactly the same contents because these texts are usually advertising texts. Thirdly, delete noises such as websites and symbols. The filtered tweets were analyzed by SentiStrength [40], [41] and scores were generated.

The SentiStrength [40], [41] detects positive and negative sentiment strength of text; and assign positive sentiment score from 1 to 5 and negative sentiment score from −1 to −5. The score of 0 does not exit. The score 1 indicates no positivity and −1 represents no negativity. For example, a sentence is "Buying a new car is joyful, but the payment would be appalling." The score of the term "joyful" is 3 and that of the term "appalling" is −4. In this study, one step ahead rolling forecasting was employed for LSSVR models when performing multivariate regression tasks. Sentiment scores of tweets and stock market values of the current month were employed to predict the total vehicle sale values of the next month. Thus, the data collection period of sentiment scores of tweets and stock market values is from January 2008 to July 2017; and that for monthly total vehicle sale values is from February 2008 to August 2017. The periods of training data set, validation data set and testing data set of sentiment scores of tweets and stock market values are from January 2008 to May 2014, from June 2014 to December 2015, and from January 2016 to July 2017; and periods of training data set, validation data set and testing data set of monthly total vehicle sale values are from February 2008 to June 2014, from July 2014 to January 2016, and from February 2016 to August 2017. In addition, for ARIMA, SARIMA, BPNN, and LSSVRTS models, the data from September 2009 to January 2016 were used as training data; and the data from February 2016 to August 2017 were used as testing data set [42]. In this study, the BPNN model includes one hidden layer with 10 hidden nodes, and genetic algorithms [38] were employed to generate suitable parameters for BPNN and LSSVRTS models. Table 2 lists data types with corresponding codes. Table 3 illustrates LSSVR models and data used for predicting the total vehicle sales.

TABLE 2. Data types and corresponding codes.

Data codes	Data descriptions
X1	Sentiment scores of tweets
X2	Stock market values
DX1	Deseasonalized sentiment scores of tweets
DX2	Deseasonalized stock market values
Y	Total vehicle sales
DY	Deseasonalized monthly total vehicle sales

TABLE 3. LSSVR models and data used for predicting monthly total vehicle sales.

LSSVR models	Data used	LSSVR models	Data used
LSSVR1	X1,Y	LSSVR7	DX1,DY
LSSVR2	X2,Y	LSSVR8	DX2,DY
LSSVR3	(X1+X2),Y	LSSVR9	(X1+DX2),DY
LSSVR4	X1,DY	LSSVR10	(X2+DX1),DY
LSSVR5	X2,DY	LSSVR11	(DX1+DX2),DY
LSSVR6	(X1+X2),DY		

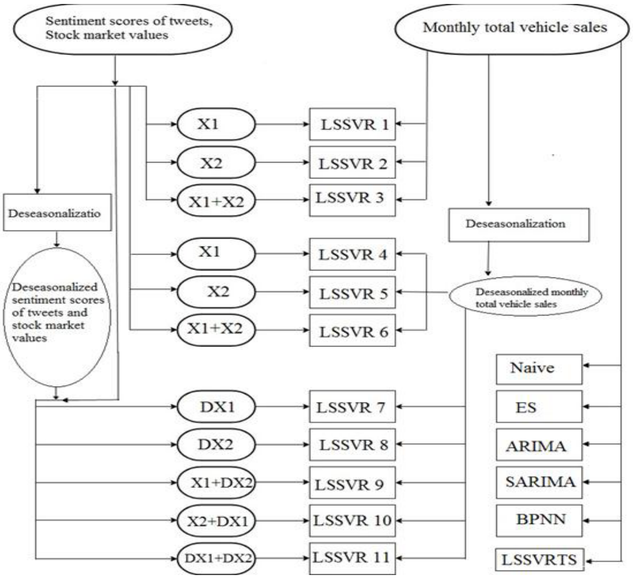

FIGURE 1. The proposed monthly total vehicle sales forecasting framework.

Figure 1 depicts the proposed framework of this study. In this study, both time series models and multivariate regression models were performed to predict monthly total vehicle sales. Four time series models, namely the naïve model, the exponential smoothing model, the autoregressive integrated moving average model, and the seasonal autoregressive integrated moving average model; and one multivariate regression model, namely the least square support vector regression model were employed to predict the monthly total vehicle sales. In addition, deseasonalizing procedures with deseasonalized factors obtained from monthly total vehicle sales data were applied to monthly total vehicle sales data, sentiment scores of tweets and stock market values. Thus, there are totally 11 LSSVR models used to predict the monthly total vehicle sales with different data combinations.

IV. THE NUMERICAL RESULTS

In this section, results of forecasting monthly total vehicle sales by time series models and multivariate regression models with various data types are illustrated. The forecasting performance is measured by MAPE, WAPE [43] and NMAE [44] shown as Eqs.(12-14):

$$\text{MAPE}\,(\%) = \frac{100}{N}\sum\nolimits_{t=1}^{N}\left|\frac{Y_t - F_t}{Y_t}\right| \tag{12}$$

$$\text{WAPE}\,(\%) = 100\frac{\sum_{\mathrm{t}=1}^{\mathrm{N}}|F_t - Y_t|}{\sum_{\mathrm{t}=1}^{\mathrm{N}} Y_t} \tag{13}$$

$$\text{NMAE} = \frac{1}{Y_h - Y_l}\left[\frac{1}{\mathrm{N}}\sum\nolimits_{\mathrm{t}=1}^{\mathrm{N}}|Y_t - F_t|\right] \tag{14}$$

where N is the number of forecasting periods, Y_t is the actual value at period t, and F_t is the forecasting value at period t, Y_{h} is highest actual value, and Y_{l} is the lowest actual value.

TABLE 4. MAPE, WAPE and NMAE values by using time series models to predict monthly total vehicle sales.

Models	Naïve	ES	ARIMA
Parameters	N/A	α=0.6	(p,d,q)=(2,1, 0)
MAPE	8.874%	7.435%	7.121%
WAPE	8.479%	7.222%	6.929%
NMAE	0.228	0.194	0.186
Models	SARIMA	BPNN	LSSVRTS
Parameters	(p,d,q)(P,D,Q)s = (0,1,1)(0,1,0)12	(Learning Rate, Momentum)= (0.497, 0.963)	(δ , σ)= (495.436, 17.742)
MAPE	7.076%	12.99%	10.994%
WAPE	7.041%	0.13%	0.11%
NMAE	0.189	0.35	0.29

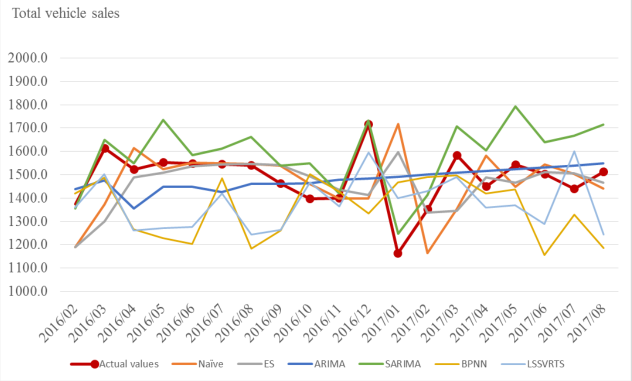

FIGURE 2. Actual and predicted monthly total vehicle sales of six time series models.

TABLE 5. MAPE, WAPE and NMAE values by using LSSVR models with original data of independent variables and original monthly total vehicle sales to predict monthly total vehicle sales.

LSSVR Models	Independent variables	LSSVR parameters (δ , σ)	MAPE WAPE NMAE
LSSVR1	Sentiment scores of tweets	(296.849,12.498)	8.95% 9.071% 0.243
LSSVR2	Stock market values	(8.721,76.004)	8.382% 8.023% 0.215
LSSVR3	Sentiment scores of tweets and stock market values	(7.125,9.792)	6.635% 6.452% 0.173

TABLE 6. MAPE, WAPE and NMAE values by using LSSVR models with original data of independent variables and deseasonalized total vehicle sales to predict monthly total vehicle sales.

LSSVR Models	Independent variables	LSSVR parameters (δ , σ)	MAPE WAPE NMAE
LSSVR4	Sentiment scores of tweets	(495.455, 22.565)	4.912% 4.993% 0.134
LSSVR5	Stock market values	(20.374,93.414)	9.145% 9.22% 0.247
LSSVR6	Sentiment scores of tweets and stock market values	(7.125,9.792)	6.939% 7.097% 0.19

TABLE 7. MAPE, WAPE and NMAE values by using LSSVR models with deseasonalized data of independent variables and deseasonalized monthly total vehicle sales to predict monthly total vehicle sales.

LSSVR Models	Independent variables	LSSVR parameters (δ , σ)	MAPE WAPE NMAE
LSSVR7	Deseasonalized sentiment scores of tweets	(477.840,11.636)	4.671% 4.843% 0.13
LSSVR8	Deseasonalized stock market values	(499.829,398.768)	5.27% 5.303% 0.142
LSSVR9	Sentiment scores of tweets and deseasonalized stock market values	(490.684,209.166)	4.19% 4.231% 0.114
LSSVR10	Stock market values and deseasonalized sentiment scores of tweets	(124.790,35.666)	7.01% 7.122% 0.191
LSSVR11	Deseasonalized sentiment scores of tweets and deseasonalized stock market values	(496.235,169.193)	3.96% 4.008% 0.108

Table 4 illustrates MAPE, WAPE, and NMAE values obtained using time series models to predict total vehicle sales and parameters of models. It can be seen that SARIMA perform the best among four time series models. Figure 2 illustrates point-to-point comparisons of actual and predicted monthly total vehicle sales of four time series models. Tables 5-7 list MAPE WAPE, and NMAE values generated by LSSVR models to predict monthly total vehicle sales with various data combinations and parameters of models. Using deseasonalized sentiment scores of tweets, deseasonalized stock market values, and deseasonalized total vehicle sales to predict monthly total vehicle sales by the LSSVR model outperforms the other models in this study in

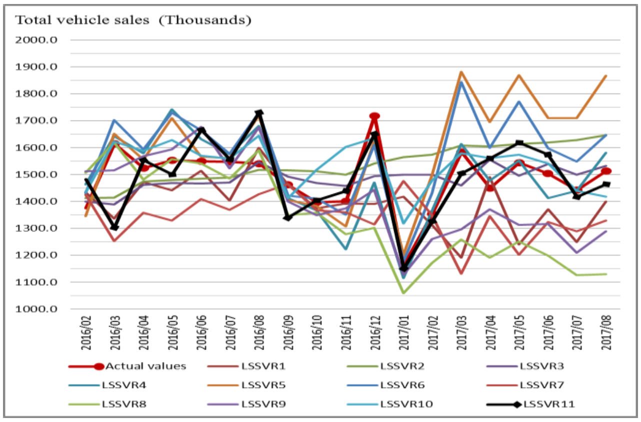

FIGURE 3. Actual and predicted monthly total vehicle sales of 11 LSSVR models.

terms of MAPE values. The point-to-point comparisons of actual and predicted monthly total vehicle sales of 11 LSSVR models with different data types is illustrated in Fig. 3.

V. CONCLUSIONS

This study presented a framework that consists of both time series forecasting models and multivariate regression technique to predict monthly total vehicle sales. Deseasonalizing procedures were employed to deal with different types of data. The numerical results indicate that forecasting vehicle sales by hybrid multivariate regression data with desonalizing procedures can obtain more accurate forecasting results than other forecasting models. The superior forecasting performance could be concluded as follows. First, the use of hybrid data containing sentiment analysis of social media and stock market values can improve the forecasting accuracy. Secondly the deceasonalizing procedures both in condition variables and decision variables do help to increase the prediction performance. For future study, since the determination of keywords for Twitter significantly affects the search results of tweets and have influences on forecasting accuracy, a more systematized technique for selecting proper keywords from Twitter could be a direction for future study. Another possible direction for future study is to employ other social media data, such as Facebook and YouTube to forecast vehicle sales. Finally, the geographical information collection of Twitter possibly could be an essential issue for future study to improve tweets analysis.

REFERENCES

[1] A. A. Bailey, C. M. Bonifield, and A. Arias, "Social media use by young Latin American consumers: An exploration," *J. Retailing Consum. Services*, vol. 43, pp. 10–19, Jul. 2018.

[2] 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.

[3] H.-L. Chang, Y.-C. Chou, D.-Y. Wu, and S.-C. Wu, "Will firm's marketing efforts on owned social media payoff? A quasi-experimental analysis of tourism products," *Decis. Support Syst.*, vol. 107, pp. 13–25, Mar. 2018.

[4] K. Chen and J. Yin, "Information competition in product launch: Evidence from the movie industry," *Electron. Commer., Res. Appl.*, vol. 26, pp. 81–88, Nov./Dec. 2017.

[5] A. Elwalda, K. Lü, and M. Ali, "Perceived derived attributes of online customer reviews," *Comput. Hum. Behav.*, vol. 56, pp. 306–319, Mar. 2016.

[6] 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.

[7] 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.

[8] H. Rui, Y. Liu, and A. Whinston, "Whose and what chatter matters? The effect of tweets on movie sales," *Decis. Support Syst.*, vol. 55, no. 4, pp. 863–870, 2013.

[9] 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.

[10] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," *Comput. Hum. Behav.*, vol. 48, pp. 17–27, Jul. 2015.

[11] N. B. Lassen, R. Madsen, and R. Vatrapu, "Predicting iPhone sales from iPhone tweets," in *Proc. IEEE. 18th Int. Enterprise Distrib. Object Comput. Conf.*, Ulm, Germany, Sep. 2014, pp. 81–90.

[12] 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.

[13] 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.

[14] 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.

[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] 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.

[18] 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.

[19] 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.

[20] F. Milani, "Learning about the interdependence between the macroeconomy and the stock market," *Int. Rev. Econ. Finance*, vol. 49, pp. 223–242, May 2017.

[21] 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.

[22] 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.

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

[24] D. Trigg and A. Leach, "Exponential smoothing with an adaptive response rate," *J. Oper. Res. Soc.*, vol. 18, no. 1, pp. 53–59, 1967.

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

[26] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[29] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in *Proc. 7th IEEE Soc. Workshop Neural Netw. Signal Process.*, Amelia Island, FL, USA, Sep. 1997, pp. 511–520.

[30] K. R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in *Artificial Neural Networks—ICANN*. Berlin, Germany: Springer-Verlag, 1997, pp. 999–1004.

[31] V. Vapnik, S. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 281–287.

[32] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[33] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: Robustness and sparse approximation," *Neurocomputing*, vol. 48, nos. 1–4, pp. 85–105, 2002.

[34] R. Fletcher, *Practical Methods of Optimization*. New York, NY, USA: Wiley, 2013.

[35] W. Karush, "Minima of functions of several variables with inequalities as side constraints," M.S. thesis, Dept. Math., Univ. Chicago, Chicago, IL, USA, 1939.

[36] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proc. 2nd Berkeley Symp. Math. Statist. Probabilities*, 1951, pp. 481–492.

[37] J. Mercer, "Functions of positive and negative type, and their connection the theory of integral equations," *Phil. Trans. R. Soc. London A, Math., Phys. Eng.*, vol. 209, pp. 415–446, Jan. 1909.

[38] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. Ann Arbor, MI, USA: University of Michigan Press, 1975, pp. 439–444.

[39] P.-F. Pai, K.-C. Hung, and K.-P. Lin, "Tourism demand forecasting using novel hybrid system," *Expert. Syst. Appl.*, vol. 41, no. 8, pp. 3691–3702, 2014.

[40] 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.

[41] 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.

[42] M. Štěpnička, P. Cortez, J. P. Donate, and L. Štěpničková, "Forecasting seasonal time series with computational intelligence: On recent methods and the potential of their combinations," *Expert Syst. Appl.*, vol. 40, no. 6, pp. 1981–1992, 2013.

[43] M. Beladev, L. Rokach, and B. Shapira, "Recommender systems for product bundling," *Knowl.-Based Syst.*, vol. 111, pp. 193–206, Nov. 2016.

[44] 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.

PING-FENG PAI (M'04–SM'05) received the Ph.D. degree from the Department of Industrial and Manufacturing Systems Engineering, Kansas State University, Manhattan, KS, USA, in 1998. He is currently a Distinguished Professor with the Department of Information Management, National Chi Nan University, Taiwan. He has authored or co-authored over 50 international journal papers. His research and teaching interests include data mining techniques, neural networks, forecasting, and optimization.

CHIA-HSIN LIU received the bachelor's and master's degrees from the Department of Information Management, National Chi Nan University, Taiwan, in 2016 and 2018, respectively. Her research interests include forecasting and data analysis.