Advances in Management & Applied Economics, Vol. 11, No. 6, 2021, 139-158
ISSN: 1792-7544 (print version), 1792-7552(online)
https://doi.org/10.47260/amae/1168
Scientific Press International Limited

Using Textual and Economic Features to Predict the RMB Exchange Rate

Yi-Chen Chung[1], Hsien-Ming Chou[2], Chih-Neng Hung[3] and Chihli Hung[4]*

Abstract

This research proposes an integrated framework for the use of textual and economic features to predict the exchange rate of the TWD (Taiwan dollar) against the RMB (Chinese Renminbi). The exchange rate is affected by the current economic situation and expectations for the future economic climate. Exchange rate forecasting studies focus mainly on overall economic indices and the actual exchange rate, but overlook the influence of news. This research considers both textual and economic factors and builds three basic prediction models, i.e. multiple linear regression (MLR), support vector regression (SVR), and Gaussian process regression (GPR) for the prediction of the RMB exchange rate. In addition to the three basic prediction models, this research uses ensemble learning and feature selection techniques to improve prediction performance. Our experiments demonstrate that textual features also play an important role in predicting the RMB exchange rate. The SVR model is shown to outperform the other models and the MLR model is shown to perform worst. The ensemble of three basic models performs better than its individual counterparts. Finally, the models which use feature selection techniques demonstrate improved results in general, and different feature selection techniques are shown to be more suitable for different prediction models.

JEL classification numbers: D80, F31, F47.
Keywords: Exchange rate prediction, Text mining, Ensemble learning, Time series forecasting.

[1] Department of Information Management, Chung Yuan Christian University, Taiwan.
[2] Department of Information Management, Chung Yuan Christian University, Taiwan.
[3] Department of Green Energy and Environment Resources, Chang Jung Christian University, Taiwan.
[4] Department of Information Management, Chung Yuan Christian University, Taiwan.
* Corresponding Author

Article Info: *Received:* October 6, 2021. *Revised*: October 19, 2021.
Published online: October 24, 2021.

1. Introduction

Exchange rate fluctuations may have a significant impact on a country's international trade, flow of capital, stock market (Ruan et al., 2017), and even industrial growth (Dai et al., 2020). Due to the numerous explicit and implicit factors that can influence exchange rate fluctuations, accurate forecasting of the exchange rate time series is a substantial task for both researchers and practitioners (Bilson and Marston, 1984). The exchange rate determination theory (Mussa, 1984) analyzes what factors determine and affect the exchange rate. These include the international indebtedness theory, the purchasing power parity theory, the interest rate parity theory, the balance of payments theory, and the asset price theory. However, there is no overarching exchange rate determination theory.

Most exchange rate forecasting studies focus on overall economic indices and the exchange rate itself (e.g. Dai and Xiao, 2005; Lin et al., 2012; Wang et al., 2019; Ren et al., 2019; He et al., 2018; Chou et al., 2019; Zhou et al., 2020). Mussa (1984) considered that the exchange rate is affected by the current economic situation and expectations of future economic situation. Economic time series data and news contain not only impacts upon, but possible causes of events (Kumar and Ravi, 2016). Due to the rapid development of information technology, a vast number of unstructured documents have been distributed across the internet. Text mining technologies have been used to exploit this huge corpus for tasks such as information retrieval, sentiment analysis, bankruptcy prediction, stock trading forecasting, market segmentation, etc. (Chan and Chong, 2017; Kumar and Ravi, 2016; Hung, 2017), but until now have only been used to a limited extent for exchange rate prediction.

The rapid development of China's economy has increased the international influence of the Renminbi (RMB, Chinese yuan) significantly (Batten and Szilagyi, 2016; Chen et al., 2018). In particular, the RMB has been included in the International Monetary Fund (IMF) special drawing rights (SDR) basket of currencies since October, 2016. China is one of Taiwan's most important trading partners, so variations in the RMB exchange rate could have a significant effect on Taiwan's import and export trade and capital flows. Thus, this study takes the exchange rate of the Taiwan dollar (TWD) against the RMB as an example, and selects not only relevant economic factors but also news from China and Taiwan to develop a prediction model for the RMB exchange rate. Conversely, most existing exchange rate forecasting studies utilize only a single prediction technology, thus limiting their validity and objectivity. This research uses an ensemble learning technology to obtain improved prediction results. We integrate three kinds of predictive models, give corresponding weights to each, propose a multi-model prediction framework, and finally compare a single model with our proposed integrated model for RMB exchange rate prediction.

The remainder of the paper is organized as follows. Section 2 presents a brief literature review. Section 3 proposes the methodology, which includes preprocessing of documents, filtering by sentiments, filtering by term frequency, integrating features, selecting features, building prediction models, and evaluation. The experiment design and results are shown in Section 4. The conclusion and areas for possible further work are presented in Section 5.

2. Literature Review

In essence, exchange rate forecasting is a time series prediction task. The methods for time series prediction can be broadly divided into three groups, namely statistical techniques, machine learning techniques and hybrid approaches. The first group consists of moving average (MA), linear regression (LR), multiple linear regression (MLR), logistic regression, autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), correlation analysis (CA), common correlated effect (CCE), the autoregressive conditional heteroskedasticity model (ARCH), and the generalized-ARCH model (GARCH), etc. (Bowerman et al., 2005). For example, Ren et al. (2019) used CCE to estimate the monetary model in growth rates for the prediction of exchange rates. Their models outperform models based on the random walk theory. Zhou et al. (2020) employed a GARCH-MIDAS model to investigate the impact of economic policy uncertainty between China and the United States on RMB exchange rate volatility. Their GARCH-MIDAS model outperforms the traditional GARCH-type models.

Due to the vast volumes of data available on the internet and the rapid development of computer hardware and software, machine learning and data mining are becoming increasingly important. Data mining is a technique for retrieving important information hidden in a large quantities of data (Fayyad et al., 1996; Hand et al., 2001). Machine learning uses machine simulation approaches for mining data. There are many techniques which can be used for time series prediction (Palit and Popovic, 2005), such as the multilayer perceptron neural network (MLP), recurrent neural network (RNN), self-organizing map (SOM), support vector regression (SVR), genetic algorithm (GA), Gaussian process regression (GPR), fuzzy logic, etc. For example, Liu et al. (2020) explored the application of GPR in forecasting the volatility of foreign exchange returns and found that this approach has shown great potential. Bo and Chi (2009) adopted MLP and RNN to predict the RMB exchange rate, and Ye (2012) found that MLP is effective in forecasting the RMB exchange rate. SVR has many applications in time series forecasting due to its robust performance (Yuan, 2013; Lin et al. 2012). Fu (2010) used SVR to predict the exchange rate of the euro against the US dollar. Chaos-theory based SVR has been used to predict foreign exchange rates (Huang et al., 2010). Fu et al. (2019) demonstrated that their proposed evolutionary SVR outperforms MLP models in RMB exchange rate prediction. From their experiments, Yang et al. (2007) concluded that SVR has better generalization ability than artificial neural networks for time series prediction.

Some studies have adopted a hybrid approach to pursue a better performance. There are many different ways in this group. For example, de O. Santos Júnior et al. (2019) used a hybrid system of ARIMA, SVR, MLP that searches for a suitable function for time series forecasting. Babu and Reddy (2014) used a hybrid ARIMA-ANN (artificial neural network) model for forecasting time series data and the hybrid model has a better performance. Zainuddin et al. (2019) applied hybrid bootstrap models to improve the performance of ARIMA-ANN hybrid models for time series forecasting.

Generalized data mining techniques includes text mining, i.e. searching or retrieving for useful information in a large amount of text. Text mining has been widely used in business information analysis (Kumar and Ravi, 2016). For example, Fung et al. (2003) predicted stock price changes based on news articles by excavating multiple time series simultaneously. Schumaker et al. (2012) explored how the sentiment of financial news articles influences the stock market. Yu et al. (2013) studied the impact of social and traditional media and the interrelatedness of short-term corporate stock market performance using news articles about the online stock market. Studying how news can directly or indirectly affect currency prices through order flows, Evans and Lyons (2008) found that news can explain more than 30% of the daily price difference. Jin et al. (2013) used news and linear regression models to predict changes in the foreign exchange market. Chatrath et al. (2014) used news to predict increases or decreases in currency and found that news could account for around 22-56% of currency changes.

Although many recent studies have attempted to forecast the RMB exchange rate, they have focused primarily on overall economic indices. This research proposes instead to use text mining technologies to analyze positive and negative news, and to combine the results with specific economic indices to improve the performance of RMB forecasting.

3. Methodology

This research uses an integrated framework of news and economics to analyze and predict the exchange rate of the TWD against the RMB. As the exchange rate is affected by both countries, the news and economic indices are based on data from Taiwan and China. The framework consists of seven modules, which are preprocessing of documents, filtering by sentiments, filtering by term frequency, integrating indices, selecting features, building prediction models, and evaluation. The conceptual structure of the proposed model is shown in Figure 1. Details of these modules are described as follows.

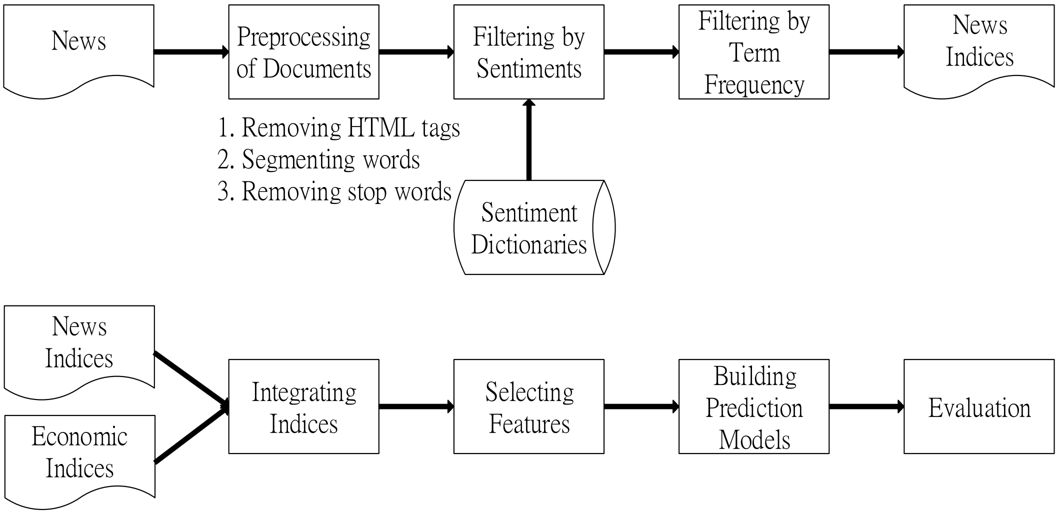

Figure 1: The conceptual structure of the proposed model

3.1 Preprocessing of Documents

The preprocessing of documents includes the three steps shown in Figure 1, i.e. removing HTML tags, segmenting words, removing stop words. Taiwan and China news is collected from Liberty Times Net (https://ltn.com.tw) and Baidu news (https://news.baidu.com) respectively, based on a query, (RMB exchange rate). In this module, the first step once the news documents are retrieved from the internet is the removal of HTML tags. As the meaning of a single Chinese character can be ambiguous (Zhang et al., 2004), we use a renowned Chinese open source word segmentation system, i.e. Jieba (https://github.com/fxsjy/jieba) for tokenization of Chinese words. One advantage of using the Jieba Chinese word segmentation system is that it can modify the reference words for the word segmentation. Next, this model removes stop words by use of a stop word list, which contains 767 words.

3.2 Filtering by Sentiments

A method that extracts positive or negative sentiment orientation from textual information is called sentiment analysis (Nasukawa and Yi, 2003) or opinion mining (Dave et al., 2003). Many research works such as (Boudt et al., 2019; Brandt and Gao, 2019; Chen et al., 2019; Feuerriegel and Prendinger, 2016; Hung 2017; Kazmaier and van Vuuren, 2020; Lee and Chen, 2020) have shown that sentiments in documents influence consumer purchasing decisions, trading markets, exchange-traded fund returns, etc. A document is made up of sentences and a sentence is made up of words. Thus, sentiments of a document come from its words. After the removal of stop words, we rely on two pre-defined sentiment dictionaries, i.e. NTUSD (National Taiwan University Semantic Dictionary) (Chen et al., 2019) and HowNet (Fu et al., 2017) to capture a given word's sentiment orientation. NTUSD is considered to be a fundamental sentiment dictionary. We then remove those words whose sentiment polarity conflicts between the two dictionaries. For those words that are not in the sentiment dictionaries, a human expert is deployed to assign their sentiment orientations. As the same news may be interpreted differently by each party in a transaction, this study considers positive and negative emotions based on news from the viewpoints of both parties.

3.3 Filtering by Term Frequency

In the module for filtering by sentiments, we have collected positive and negative sentiment words from both Taiwanese and Chinese news. Although the stop words have been removed, there are still many sentiment words that may not be closely related to the TWD-RMB exchange rate. This research uses a straight dimensionality reduction approach, which keeps the 20 most frequent sentiment words on each sentiment polarity for Taiwan and China, as this dimensionality reduction technique has been shown to be equally as effective as other more complex dimensionality reduction techniques (Schütze and Silverstein, 1997). Other parameters for the number of the most frequent sentiment words, such as 10 and 30 words, have been tried and have produced similar results. Therefore, we have 80 sentiment words in total as news indices. Each news index is represented by its term frequency.

3.4 Integrating Features

Through the above steps, we have obtained 20 positive and 20 negative words from China and Taiwan news respectively, a total of 80 sentiment words. The economic indices are also divided into Taiwan indices and China indices. Based on the indices suggested in the literature, Taiwan indices include the following twelve: gross domestic product, economic growth rate, industrial production index, unemployment rate, labor force participation rate, export value, import value, consumer price index, amount of currency in circulation, inflation rate, index of consumer confidence, ratio of exchange rate volatility. China indices include the following six: gross domestic product, economic growth rate, consumer price index, export value, import value, amount of currency in circulation. Finally, the 80 textual features and 18 economic indices are normalized.

3.5 Selecting Features

In prediction models, there is sometimes an implicit relationship between input features and prediction results. Feature selection, also known as variable selection or attribute selection, is a technique for reducing the number of input features when developing predictive models (Guyon and Elisseeff, 2003). The selection of the most important features can reduce the dimensions of the model, and may also improve the prediction performance. In this research, we have 98 input features in total, which include 18 economic indices and 80 word occurrences. We apply two feature selection techniques, namely Spearman's correlation analysis and stepwise regression analysis.

Both Pearson correlation analysis and Spearman's correlation analysis are techniques that can be used to express the strength and direction of the relationship between variables (Akoglu, 2018). The result will always be between +1 and -1. A positive value indicates a positive correlation, and a negative value indicates a negative correlation. The greater the absolute value of the correlation, the greater the correlation between the two variables. The Pearson correlation analysis is implemented under an assumption that the data is normally distributed. Unlike the Pearson correlation analysis, Spearman's correlation analysis can be used to deal with the relationship between an ordinal scale variable and a non-nominal scale variable, which is more suitable for this research. The correlation coefficient can be treated as a correlation threshold. Features which fall below a specific threshold such as the moderate correlation coefficient, 0.4, are considered to be irrelevant features, and are then filtered out. Conversely, stepwise regression is a linear multivariate regression method that uses multiple linear regressions to gradually delete unimportant variables. At a preassigned significance level, such as 5%, the important variables of the stepwise regression model can be systematically determined.

3.6 Building Prediction Models

This research uses multiple linear regression, support vector regression, and Gaussian process regression to track fluctuations in the RMB exchange rate. Multiple linear regression (MLR) is a method for finding the relationship between independent and dependent variables (Freedman, 2009). It is a commonly used model for time series prediction. A general form of MLR is shown in (1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{1}$$

where y is dependent variable, $x_i \ldots x_k$ are independent variables, β_0 is the y-intercept, β_i determines the contribution of the independent variable x_i, k denotes the number of independent variables, and ϵ is the residuals.

Support vector machine (SVM) is a classification algorithm widely used in data mining (Cortes and Vapnik, 1995). Vapnik (1995) and Drucker et al. (1997) extended the concept of support vectors to regression problems.

Suppose the training data $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where $x_i \in \mathcal{R}^d, y_i \in \mathcal{R}$. (x_i, y_i) is the pair of input and target values and N is the number of training samples. The goal is to find a function *f(x)* that has at most ε deviation from the actually obtained targets y_i for all training data and at the same time is as flat as possible.

The function *f(x)* can be described as (2).

$$f(x) = w \cdot x + b, \tag{2}$$

where $w \in \mathcal{R}^d,\ b \in \mathcal{R}.\ w$ is the weight vector $and\ b$ is the bias. Flatness means to pursue a small w and can be treated as a convex optimization problem in (3).

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 \tag{3}$$

$$\text{subject to} \quad \|y_i - (w \cdot x_i - b)\| \le \varepsilon, \text{ where } \varepsilon \ge 0$$

However, it is necessary to introduce slack variables ξ_i, ξ_i^* to deal with infeasible constraints of the optimization problem. Therefore, Equation (3) is re-written as Equation (4) and *C* is a constant, which controls fitness.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{4}$$

$$\text{subject to} \quad \begin{cases} y_i - w \cdot x_i - b \le \varepsilon + \xi_i \\ w \cdot x_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^*, C \ge 0 \end{cases}$$

Then the Lagrange function, *L*, is introduced to exploit quadratic optimization problem, which is shown in (5).

$$\begin{aligned} L = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ & - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + w \cdot x_i + b) \\ & - \sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* + y_i - w \cdot x_i - b) \end{aligned} \tag{5}$$

The detailed process of mathematical derivation of the solution can be referred to (Vapnik, 1995) and the regression version of SVM is shown in (6).

$$f(x) = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b, \tag{6}$$

where $0 \leq \alpha_i, \alpha_i^* \leq C$ and $K(x_i, x_j)$ is a kernel function. We use the polynomial kernel function (7) in this research as it performs well.

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^p, \tag{7}$$

where *p* is the degree of the polynomial kernel. The inner product can be treated as a measure of similarity between two vectors.

That is, the kernel function is a method for computing similarity in the higher dimensional feature space using the original attribute set. Support vector regression (SVR) has found extensive application, not only in data mining, but also in business studies (Fu, 2010; Huang et al., 2010; Colombo and Pelagatti, 2020). Many research works have demonstrated that the performance of SVR is robust and thus it has been used in this research.

Machine learning can be treated as learning a function from examples. Gaussian process regression (GPR) uses the Bayesian machine learning approach, which is based on a prior distribution over the space of functions (Rasmussen and Williams, 2006). In comparison with traditional machine learning technologies, such as artificial neural networks, GPR has the advantage of learning more easily, and is also used in this research (Rasmussen, 2003). GPR brings together work in the statistical and machine learning communities. GPR is defined as a collection of random variables and every sub-collection of which has a joint Gaussian distribution. For the training data, *D*, a general Gaussian process is shown in (8).

$$f(X) \sim gp(m(x), k(x, x')), \tag{8}$$

where $m(x)$ is a mean function of x shown in (9) and $k(x, x')$ is a covariance function, which is shown in (10).

$$m(x) = \mathrm{E}[f(x)] \tag{9}$$

$$k(x, x') = \mathrm{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{10}$$

This research uses MLR, SVR and GPR as the basic prediction models. An ensemble learning system improves the performance of a single classifier by creating multiple classifiers and combining these classifiers to achieve a more robust performance (Hung and Chen, 2009; Tsai and Hung, 2014). Ensemble learning is a strategy whereby multiple weak single classifiers are combined in order to improve overall classification performance.

The task in this research is a time series prediction, so resampling techniques cannot be used. This research uses a weighted average method to construct an ensemble of three basic methods and its equation is shown in (11).

$$y = a \times MLR + b \times SVR + c \times GPR \tag{11}$$

where a, b, c are weighted values and y is the prediction result.

Based on two kinds of features (i.e., text and economic indices), two feature selection approaches (i.e., Spearman's correlation analysis and stepwise regression) and two kinds of prediction algorithms (i.e., individual and ensemble prediction models), we develop six prediction models, as follows:

1. Individual model using economic features.
2. Individual model using textual features.
3. Individual model using economic and textual features.
4. Ensemble model using economic and textual features.
5. Feature selection by the technique of Spearman's correlation analysis.
6. Feature selection by the technique of stepwise regression analysis.

3.7 Evaluation

This research uses 18 economic indices and 80 text indices to predict the exchange rate of the TWD against the RMB. Based on monthly data, we collect data from June 2013 to June 2018 as the training set, and data from July 2018 to September 2018 as the test set or prediction target. In terms of evaluation measures, this research uses mean absolute error (MAE), mean square error (MSE), mean absolute percent error (MAPE), and root mean square error (RMSE), which are commonly used in the field of time series prediction and are shown in (12)-(15) respectively.

$$MAE = \frac{1}{T}\sum_{i=1}^{T} |y_i - \hat{y}_i|, \tag{12}$$

$$MSE = \frac{1}{T}\sum_{i=1}^{T} (y_i - \hat{y}_i)^2, \tag{13}$$

$$MAPE = \frac{1}{T}\sum_{i=1}^{T} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%, \tag{14}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{T} (y_i - \hat{y}_i)^2}{T}}, \tag{15}$$

where T is the number of samples in the test set, y_i is the actual value and $\hat{y}_i$ is the forecasting value.

4. Experiments

In this research, we build our prediction models based on Waikato environment for knowledge analysis (WEKA, https://www.cs.waikato.ac.nz/ml/weka/), which is a renowned machine learning workbench. Our methodology proposes six experimental cases described as follows.

4.1 Case 1-Individual Model Using Economic Features

This case is the benchmarking model, which presents the prediction performance of three individual prediction methods (i.e. MLR, SVR, GPR) based on 18 economic features. The training set is used to build the prediction models and the test set is the prediction target. Table 1 shows the prediction performance for individual models. The SVR method presents the best prediction results evaluated by all evaluation measures, followed by Gaussian process regression. The multiple linear regression method produces the poorest prediction performance.

Table 1: Prediction performance for individual models based on 18 economic features

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2085	0.2149	0.2109
2018/8	0.2228	0.2085	0.2150	0.2110
2018/9	0.2226	0.2085	0.2151	0.2109
MAE		0.012367	0.005867	0.009933
MSE		0.000160	4.08E-05	0.000105
MAPE		0.055860	0.026430	0.044843
RMSE		0.012636	0.006387	0.010260

4.2 Case 2-Individual Model Using Textual Features

Like Case 1, this case builds three individual prediction models, but in contrast to Case 1 which uses economic features, this case uses 80 textual features to predict the RMB exchange rate. Table 2 shows that the SVR method outperforms the other two methods, and the multiple linear regression method produces the poorest prediction performance. In comparison with the benchmarking models in Case 1, apart from the multiple linear regression method, the SVR and GPR methods using textual input features outperform those which use economic features evaluated by all evaluation criteria.

Table 2: Prediction performance for individual models based on 80 textual features

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2086	0.2154	0.2113
2018/8	0.2228	0.2082	0.2151	0.2113
2018/9	0.2226	0.2077	0.2152	0.2111
MAE		0.012700	0.005633	0.009633
MSE		0.000170	3.91E-05	9.98E-05
MAPE		0.057354	0.025364	0.043481
RMSE		0.013027	0.006253	0.009988

4.3 Case 3-Individual Model Using Economic and Textual Features

This case combines economic features with textual features to predict the RMB exchange rate. There are two integrating approaches. One is an average approach and the other is concatenation. For the average approach, the integrated output value is the average value of the output values from prediction models using textual features and those using economic features respectively. This approach is straightforward and it is not necessary to rebuild the prediction models. Table 3 shows that the SVR method produces the strongest prediction performance and the multiple linear regression method has the poorest prediction performance. In comparison with Cases 1 and 2, the prediction performance of this case lies between them.

Table 3: Prediction performance for individual models based on average integration from output of models using economic and textual features

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2086	0.2152	0.2111
2018/8	0.2228	0.2084	0.2151	0.2112
2018/9	0.2226	0.2081	0.2152	0.2110
MAE		0.012500	0.005700	0.009767
MSE		0.000164	3.93E-05	0.000102
MAPE		0.056455	0.025671	0.044087
RMSE		0.012801	0.006273	0.010105

For the concatenation approach, 80 textual features and 18 economic features are treated as input features, making 98 input features in total. Three basic prediction methods are applied for the prediction targets. Table 4 shows that the SVR method produces the best prediction results evaluated by all evaluation measures, followed by the Gaussian process regression method. The multiple linear regression method gives the poorest prediction performance. In comparison with Cases 1 and 2, the prediction performance of this case is the worst. The reason for this may be that not all features have a predictive ability, so the selection of more significant features may improve prediction performance.

Table 4: Prediction performance for individual models based on concatenation of economic and textual features

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2081	0.2116	0.2084
2018/8	0.2228	0.2081	0.2116	0.2084
2018/9	0.2226	0.2081	0.2116	0.2083
MAE		0.012767	0.009267	0.012500
MSE		0.000170	9.26E-05	0.000163
MAPE		0.057672	0.041823	0.056463
RMSE		0.013028	0.009623	0.012771

4.4 Case 4-Ensemble Model Using Economic and Textual Features

This case uses the same concatenation of input features as used in Case 3 and seeks a better prediction performance by using an ensemble learning technique. A weighted average method is used to construct an ensemble of three basic methods (i.e., MLR, SVR, GPR). As the prediction targets are the RMB exchange rates in July, August and September of 2018, we take the RMB exchange rates in April, May and June of 2018 as the temporary prediction targets in order to obtain the three parameters (i.e. *a*, *b*, *c*) in (11). Thus, for this temporary prediction, data from June 2013 to March 2018 as the training set, and data from April 2018 to June 2018 as the test set. Table 5 shows the prediction results for the temporary prediction targets and Table 6 shows the values for parameters of *a*, *b*, and *c* in (11).

Table 5: Prediction results for temporary prediction targets based on 98 input features

Month	Target	MLR	SVR	GPR
2018/4	0.2137	0.2078	0.2127	0.2095
2018/5	0.2137	0.2077	0.2127	0.2093
2018/6	0.2172	0.2078	0.2124	0.2089
MAE		0.007100	0.002267	0.005633
MSE		5.31E-05	8.35E-06	3.53E-05
MAPE		0.032988	0.010486	0.026152
RMSE		0.007284	0.002889	0.005941

Table 6: Parameters for the ensemble of MLR, SVR and GPR methods

a	b	c
8.535866982	-3.130799683	-4.267933491

The weighted parameters in (11) have been established. Table 7 shows the prediction performance for the ensemble model based on concatenation of economic and textual features. Compared with the prediction performance of the individual models, the ensemble model achieves the best prediction performance evaluated by all criteria. These results outperform the benchmarking models (i.e. economic models) in Case 1, textual models in Case 2, and the two integrated models in Case 3.

Table 7: Prediction performance for the ensemble model based on economic and textual features

Date	Target	Ensemble
2018/7	0.2172	0.224399
2018/8	0.2228	0.224399
2018/9	0.2226	0.224826
MAE		0.003675
MSE		1.98E-05
MAPE		0.016775
RMSE		0.004448

4.5 Case 5-Feature Selection technique using Spearman's Correlation Analysis

The aim of this case is to analyze the representativeness of the 98 input features using Spearman's correlation analysis. The correlation coefficient threshold is set at 0.4 as this is a moderate correlation (Akoglu, 2018). Table 8 shows the prediction performance when using Spearman's correlation feature selection. Compared with the models in Cases 2 and 3, those in Case 5 achieve better prediction results when Spearman's correlation analysis is used as the feature selection technique. Compared with the models in Case 1, except for the multiple linear regression method, the SVR and GPR methods outperform those which do not use the feature selection technique.

Table 8: Prediction performance using Spearman's correlation feature selection

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2084	0.2153	0.2114
2018/8	0.2228	0.2084	0.2155	0.2116
2018/9	0.2226	0.2084	0.2155	0.2115
MAE		0.012467	0.005433	0.009367
MSE		0.000162	3.58E-05	9.41E-05
MAPE		0.056313	0.024469	0.042279
RMSE		0.012734	0.005981	0.009700

4.6 Case 6-Feature Selection by the Stepwise Regression Analysis technique

The aim of this case is to analyze the representativeness of the 98 input features using stepwise regression analysis. Based on a preassigned 5% significance level, this case shows the prediction performance of the three basic models in Table 9. Compared with those models which do not use any feature selection technique in Case 3, except for the SVR method, this feature selection technique can produce an improvement in prediction performance. In comparison with the models in Case 5, which use the feature selection technique of Spearman's correlation analysis, the models in Case 6 perform better while using the MLR and GPR methods and worse while using the SVR method.

Table 9: Prediction performance using stepwise regression feature selection

Month	Target	MLR	SVR	GPR
2018/7	0.2172	0.2085	0.2141	0.2117
2018/8	0.2228	0.2085	0.2142	0.2119
2018/9	0.2226	0.2085	0.2140	0.2119
MAE		0.012367	0.006767	0.009033
MSE		0.000160	5.25E-05	8.78E-05
MAPE		0.055860	0.030502	0.040771
RMSE		0.012636	0.007246	0.009373

5. Conclusions and Further Work

This research proposes a novel framework to predict the exchange rate of the RMB based on features of news and economics. Three methods (i.e., multiple linear regression (MLR), support vector regression (SVR), and Gaussian process regression (GPR) are used in this research as basic prediction models. We then use an ensemble learning method to improve the prediction performance. Two feature selection approaches (i.e., Spearman's correlation analysis and stepwise regression) are then used to select significant features. The results of the experiments demonstrate that textual features also play an important role in predicting the RMB exchange rate. In terms of the three basic models, SVR outperforms the other models. The ensemble of three basic models performs better than the individual models. The models which use the feature selection technique produce better prediction performance than the models which do not use any feature selection technique in general. Finally, in terms of the feature selection technique, the performance of SVR models which use Spearman's correlation analysis is stronger than the models which use stepwise regression. Conversely, stepwise regression is more effective than Spearman’s correlation analysis when GPR and MLR models are used.

Several directions for further work are considered. For example, many time-based deep learning neural networks such as recurrent neural networks (RNN), long short-term memory (LSTM), and gated recurrent unit (GRU) models have shown potential in various fields (Zhang et al., 2018) and could thus be evaluated in future work. The method of selecting news related to the RMB exchange rate could be expanded from specific keywords to domain articles. More specifically, the practicality of query expansion technology (Vechtomova and Wang, 2006) in the field of information retrieval is one possible direction for further evaluation. This research uses Chinese words as the word processing unit. Future research could also use phrases, sentences, paragraphs or documents as the word processing unit. The focus of this research is to distinguish between positive and negative words with regard to the exchange rate. One possible interesting research direction is to label words based on exchange rate trends (such as rising and falling) in order to track time series more accurately. Finally, in addition to Chinese news, future research could include cross-language (such as Chinese and English) related news to test the effectiveness in predicting the RMB exchange rate.

References

[1] Akoglu, H. (2018). User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, Vol. 18, No. 3, pp. 91-93.

[2] Babu, C.N. and Reddy, B.E. (2014). A moving-average filter based hybrid ARIMA-ANN model for forecasting time series data. Applied Soft Computing, Vol. 23, pp. 27-38.

[3] Batten, J.A. and Szilagyi, P.G. (2016). The internationalisation of the RMB: new starts, jumps and tipping points. Emerging Markets Review, Vol. 28, pp. 221-238.

[4] Bilson, J.O. and Marston, R.C. (1984). Exchange Rate Theory and Practice, University of Chicago Press.

[5] Bo, S. and Chi, X. (2009). RMB exchange rate forecasting in the context of the financial crisis. Systems Engineering- Theory & Practice, Vol. 29, No. 12, pp. 53-64.

[6] Boudt, K., Neely, C., Sercu, P. and Wauters, M. (2019). The response of multinationals' foreign exchange rate exposure to macroeconomic news. Journal of International Money and Finance, Vol. 94, pp. 32-47.

[7] Bowerman, B.L., O'Connell, R.T. and Koehler, A.B. (2005). Forecasting, Time Series, and Regression: An Applied Approach. Belmont, CA: Cengage Learning Thomson Learning.

[8] Brandt, M.W. and Gao, L. (2019). Macro fundamentals or geopolitical events? a textual analysis of news events for crude oil. Journal of Empirical Finance, Vol. 51, pp. 64-94.

[9] Chan, S. and Chong, M. (2017). Sentiment analysis in financial texts. Decision Support Systems, Vol. 94, pp. 53-64.

[10] Chatrath, A., Miao, H., Ramchander, S. and Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. Journal of International of Money and Finance, Vol. 40, pp.42-62.

[11] Chen, M.-Y., Liao, C.-H. and Hsieh, R.-P. (2019). Modeling public mood and emotion: stock market trend prediction with anticipatory computing approach. Computers in Human Behavior, Vol. 101, pp. 402-408.

[12] Chen, P.-F., Zeng, J.-H. and Lee, C.-C. (2018). Renminbi exchange rate assessment and competitors' exports: new perspective. China Economic Review, Vol. 50, pp. 187-205.

[13] Chou, H.-M., Li, K.-C. and Pi, S.-M. (2019). Multinational effects of foreign exchange rate in stock index with classification models for medium-term investment. Advances in Management and Applied Economics, Vol. 9, No. 3, pp. 43-53.

[14] Colombo, E. and Pelagatti, M. (2020). Statistical learning and exchange rate forecasting. International Journal of Forecasting, Vol. 36, pp. 1260-1289.

[15] Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. Machine Learning, Vol. 20, No.3, pp.273-297.

[16] Dai, X.-F. and Xiao, Q.-X. (2005). Time series analysis applied in prediction of RMB's exchange rate. Journal of University of Shanghai for Science and Technology, Vol. 27, pp. 341-344.

[17] Dai, Z., Zhu, H. and Dong, X. (2020). Forecasting Chinese industry return volatilities with RMB/USD exchange rate. Physica A: Statistical Mechanics and its Applications, Vol. 539.

[18] Dave, K., Lawrence, S. and Pennock, D.M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. Proceedings of International Conference on World Wide Web.

[19] de O. Santos Júnior, D.S., de Oliveria, J.F.L. and de Mattos Neto, P.S.G. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowledge-Based Systems, Vol. 175, pp. 72-86.

[20] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J. and Vapnik. V. (1997). Support vector regression machines. Advances in Neural Information Processing Systems, Vol. 9, pp. 155-161.

[21] Evans, M.D.D., Lyons and R.K. (2008). How is macro news transmitted to exchange rates? Journal of Financial Economics, Vol. 88, pp. 26-50.

[22] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, Vol. 17, No. 3, pp. 37-54.

[23] Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies. Decision Support Systems, Vol. 90, pp. 65-74.

[24] Freedman, D.A. (2009). Statistical Models: Theory and Practice. Cambridge, UK: University Press.

[25] Fu, C. (2010). Forecasting exchange rate with EMD-based support vector regression. Proceedings of International Conference on Management and Service Science (MASS), Wuhan, China.

[26] Fu, S., Li, Y., Sun, S. and Li, H. (2019). Evolutionary support vector machine for RMB exchange rate forecasting. Physica A: Statistical Mechanics and its Applications, Vol. 521, pp. 692-704.

[27] Fu, X., Liu, W., Xu, Y. and Cui, L. (2017). Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis. Neurocomputing, Vol. 241, pp. 18-27.

[28] Fung, G.P.C., Yu, J.X. and Lam, W. (2003). Stock prediction: integrating text mining approach using real-time news. Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering, Hong Kong, China, pp. 395–402.

[29] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

[30] Hand, D.J., Smyth, P. and Mannila, H. (2001). Principles of Data Mining, Cambridge, MA: MIT Press. 2001.

[31] He, K., Chen, Y. and Tso, G.K.F. (2018). Forecasting exchange rate using variational mode decomposition and entropy theory. Physica A: Statistical Mechanics and its Applications, Vol. 510, pp. 15-25.

[32] Huang, S.-C., Chuang, P.J., Wu, C.-F. and Lai, H.-J. (2010). Chaos-based support vector regressions for exchange rate forecasting. Expert Systems with Applications, Vol. 37. No. 12, pp. 8590-8598.

[33] Hung, C. (2017). Word of mouth quality classification based on contextual sentiment lexicons. Information Processing and Management, Vol. 53, No.4, pp. 751-763.

[34] Hung, C. and Chen, J.-H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. Expert Systems with Applications, Vol. 36, No. 3, pp. 5297-5303.

[35] Jin, F., Self, N., Saraf, P., Butler, P., Wang, W. and Ramakrishnan, N. (2013). Forex-foreteller: currency trend modeling using news articles. Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1470-1473.

[36] Kazmaier, J. and van Vuuren, J.H. (2020). A generic framework for sentiment analysis: leveraging opinion-bearing data to inform decision making. Decision Support Systems, Vol. 135, No. 113304, pp. 1-21.

[37] Kumar, B.S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain. Knowledge-Based Systems, Vol. 114, pp. 128-147.

[38] Lee, C.-C. and Chen, M.-P. (2020). Happiness sentiments and the prediction of cross-border country exchange-traded fund returns. The North American Journal of Economics and Finance, Vol. 54, No. 101254, pp. 1-25.

[39] Lin, C.-S., Chiu, S.-H. and Lin, T.-Y. (2012). Empirical mode decomposition-based least squares support vector regression for foreign exchange rate forecasting. Economic Modelling, Vol. 29, No.6, pp. 2583-2590.

[40] Liu, B., Kiskin, I. and Roberts, S. (2020). An overview of Gaussian process regression for volatility forecasting. Proceedings of 2020 International Conference on Artificial Intelligence in Information and Communication, pp. 681-686.

[41] Mussa, M. (1984). The theory of exchange rate determination. Bilson, J.O., Marston, R.C. (eds.), Exchange Rate Theory and Practice, University of Chicago Press, pp. 13-78.

[42] Nasukawa, T. and Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing. Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture.

[43] Palit, A.K. and Popovic, D. (2005). Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications, London: Springer-Verlag London Limited, 2005.

[44] Rasmussen, C.E. (2003). Gaussian processes in machine learning. Lecture Notes in Computer Science, Vol. 3176, pp. 63-71.

[45] Rasmussen, C.E. and Williams, C.K.I. (2006). Gaussian Processes for Machine Learning, the MIT Press.

[46] Ren, Y., Wang, Q. and Zhang, X. (2019). Short-term exchange rate predictability. Finance Research Letters, Vol. 28, pp. 148-152.

[47] Ruan, Q., Yang, B. and Ma, G. (2017). Detrended cross-correlation analysis on RMB exchange rate and Hang Seng China Enterprises Index. Physica A: Statistical Mechanics and its Applications, Vol. 468, pp. 91-108.

[48] Schumaker, R.P., Zhang, Y., Huang, C.-N. and Chen, H. (2012). Evaluating sentiment in financial news articles. Decision Support Systems, Vol. 53, No. 3, pp. 458-464.

[49] Schütze, H. and Silverstein, C. (1997). A comparison of projections for efficient document clustering. Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 74-81.

[50] Tsai, C.-F. and Hung, C. (2014). Modeling credit scoring using neural network ensembles. Kybernetes, Vol. 43, No. 7, pp. 1114-1123.

[51] Vapnik, V.N. (1995). The Nature of Statistical Learning Theory, Springer, New York.

[52] Vechtomova, O. and Wang, Y. (2006). A study of the effect of term proximity on query expansion. Journal of Information Science, Vol. 32, No. 4, pp. 324-333.

[53] Wang, K., Chang, M., Wang, W., Wang, G. and Pan, W. (2019). Predictions models of Taiwan dollar to US dollar and RMB exchange rate based on modified PSO and GRNN. Cluster Computing, pp. 1-12.

[54] Yang, J.-F., Zhai, Y.-J., Xu, D.-P. and Han, P. (2007). SMO algorithm applied in time series model building and forecast. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, pp. 2395-2400.

[55] Ye, S. (2012). RMB exchange rate forecast approach based on BP neural network. Physics Procedia, Vol. 33, pp. 287-293.

[56] Yu, L.-C., Wu, J.-L., Chang, P. -C. and Chu, H.-S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowledge-Based Systems, Vol. 41, pp. 89-97.

[57] Yu, Y., Duan, W. and Cao, Q. (2013). The impact of social and conventional media on firm equity value: a sentiment analysis approach. Decision Support Systems, Vol. 55, pp. 919-926.

[58] Yuan, Y. (2013). Forecasting the movement direction of exchange rate with polynomial smooth support vector machine. Mathematical and Computer Modelling, Vol. 57, pp. 932-944.

[59] Zainuddin, N.H., Lola, M.S., Djauhari, M.A., Yusof, F., Ramlee, M.N.A., Deraman, A., Ibrahim, Y. and Abdullah, M.T. (2019). Improvement of time forecasting models using a novel hybridization of bootstrap and double bootstrap artificial neural networks. Applied Soft Computing, Vol. 84, No. 105676.

[60] Zhang, L., Wang, S. and Liu, B. (2018). Deep learning for sentiment analysis: a survey. arXiv: 1801.07883.

[61] Zhang, M.-Y., Lu, Z.-D. and Zou, C.-Y. (2004). A Chinese word segmentation based on language situation in processing ambiguous words. Information Sciences, Vol. 162, No. 3-4, pp. 275-285.

[62] Zhou, Z., Fu, Z., Jiang, Y., Zeng, X. and Lin, L. (2020). Can economic policy uncertainty predict exchange rate volatility? New evidence from the GARCH-MIDAS model. Finance Research Letters, Vol. 34, No. 101258.