A SENTIMENT-BASED HYBRID MODEL FOR STOCK RETURN FORECASTING

Tiejun Wang, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, sohanter@hotmail.com

Qixu Gong, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, gqx_wolf@163.com

Wang Ren, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, renwangnow@gmail.com

Yanyan Wang, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, wangxyzyy@163.com

Xubin Luo, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, xubinluo@gmail.com

Qing Li, School of Economic Information Engineering, Southwestern University of Finance and Economics, Chengdu, P.R. China, Corresponding Author: liq_t@swufe.edu.cn

Abstract

Real-world financial time series often contain both linear and nonlinear patterns. However, traditional time series analysis models, such as ARIMA, hold the assumption that a linear correlation exists among time series values while leaving nonlinear relation into error terms. Based on financial theories, we argue that investor sentiment is the main contributor to nonlinear pattern of stock time series. Furthermore, we propose a sentiment-based hybrid model (SLNM) to better capture nonlinear information in stock time series. According to the forecasting experimental results, SLNM exhibits the sensitivity to sentiment environments, which in turn supports the argument that investor sentiment is the main source of nonlinear pattern in stock time series. For those stocks that are in top 10 of CAR Ranking List — these stocks are more likely pursued by emotional investors and thus in optimistic sentiment environment, SLNM improves forecasting performance dramatically: Increase Direction Accuracy by 40% and reduce RMSE by 19.3%. While, for those that are in bottom 10 of CAR Ranking List— these stocks defer more emotional investors from further participating in stock trading and thus in pessimistic sentiment environment, SLNM has a fair improvement on performance: Hold the similar Direction Accuracy and reduce RMSE only by 2.5%. All these indicate that investor sentiment play a key role in stock return forecasting. Our work sheds light on the research of sentiment-based prediction models.

Keywords: Stock Return Forecasting, Nonlinear Pattern, Hybrid Model, Investor Sentiment, ARIMA, Support Vector Regression.
1 INTRODUCTION

Stock return forecasting is a key issue in investment and appealing to not only numerous investors but also researchers who specialize in forecasting. However, making accurate prediction is a challenging task because this kind of financial time series is always inherently noisy (Abu-Mostafa and Atiya(1996)). To improve the prediction performance, a huge volume of works have been made mainly in two categories.

The first is traditional econometric model, which is based on the key assumption that future values are linearly related to the past. The representative models of this category are ARIMA\(^1\) by Box et al. (1994), ARCH by Engle (1982) and GARCH by Bollerslev(1986). Though numerous econometric family members have achieved remarkable success, it appears that alternative methodologies in artificial intelligence (AI), i.e, the second category, are fascinating as well. ANN (Artificial Neural Networks) and SVM (Support Vector Machine), which have an inherent advantage in capturing nonlinear relationships, have been proved to be efficient in a bunch of empirical studies (Dahamija and Bhalla(2010), Jiang and He(2012), Lu et al.(2009)).

However, real-world financial time series often contain both linear and nonlinear patterns (Wang et al. (2012)). To overcome the respective limitations of each category mentioned above, hybrid models that combine both econometric method and AI technology have been widely used and achieved fairly good performance. The main reason for adopting hybrid models is one model cannot exactly identify the “true” data generating process of time series while different models may play complementary roles in the approximation of data generating process (Terui and van Dijk(2002)). Moreover, recent studies have also documented that hybrid models can outperform individual models (Khasheii and Bijari(2012), Khan(2011), Chen and Jeong(2010), Pai and Li (2005), Wang et al.(2012)).

When hybrid models beat individual models in smaller MSE\(^2\) (Mean Squared Error), it is widely believed that this “success” is mainly due to “hybrid” whose aim is to capture both linear and nonlinear patterns of time series. However, as far as we know, few existing hybrid models try to investigate what are the main source of linear pattern and what contribute to the nonlinear pattern respectively. We argue that hybrid models deserve the name: “hybrid” only if they make clear what contribute to the different patterns of financial time series. Otherwise, the performance enhancement which seems to come from “hybrid”, actually, comes from unknown.

Rather than leaving the unknowns still “unknown”, we introduce firm-specific fundamentals and investor sentiment to the interpretation of linear and nonlinear patterns in stock time series based on financial theories. Furthermore, we propose a sentiment-based linear-nonlinear model (SLNM) extended from a common linear-nonlinear method firstly introduced by Pai and Lin(2005). We found that SLNM’s performance is sensitive to sentiment environments, which in turn supports that investor sentiment is the main source of nonlinear pattern in stock time series.

In our work, stocks are considered to be in optimistic sentiment environment if they are in top 10 of CAR ranking list (CRL\(^3\)) while those are in pessimistic sentiment environment if they are in bottom 10 of CRL. The experimental results show that the SLNM’s performance is conditional on different sentiment environments. Compared with classical ARMA models, SLNM improves Direction Accuracy by 40% and reduces RMSE by 19.2% on the stocks that are in 10% top of CRL—winner stocks, while for those who are in the 10% bottom of CRL—loser stocks, SLNM has the same Direction Accuracy and a fair decrease of 2.5% in RMSE performance. This indicates that SLNM has better performance improvement on the winners while relatively fair improvement on the losers, which shows that SLNM is sensitive to sentiment environments. The sensitivity to sentiment can be interpreted in the following way. The winner stocks, which have high CARs, are prone to attract more

---

\(^1\) Stock returns series discussed in this paper are all stationary, so we use ARMA model which is a special case of ARIMA.

\(^2\) The plural of MSE means that there are lots of judgements, such as RMSE, MPE, and MAPE.

\(^3\) Stocks are ranked by their past 200 trading days’ cumulative abnormal returns (CARs), which is just the portfolio formation period of a popular strategy: momentum.
emotional investors to participate in the trading; the loser stocks, which have low or negative CARs, will discourage emotional investors and then deter them from further participating in the trading. SLNM whose aim is to better capture nonlinear pattern based on investor sentiment will, therefore, perform conditionally on sentiment environments.

The main contributions of our work are mainly in three-folds:

- Based on financial theories, our paper gives make-sense content to linear and nonlinear patterns of stock series, which is ignored by existing hybrid models. This interpretation sheds light on stock series forecasting in the view of investor sentiment.
- Our paper uses textual analysis method to quantify investor sentiment from daily financial news in terms of the first Chinese Financial Sentiment Dictionary (CFSD)\(^4\). Moreover, we construct a sentiment-based linear-nonlinear hybrid model (SLNM) so as to capture nonlinear pattern in stock return series.
- Different with unconditional forecasting models that aim to predict all stocks in all time periods, our paper conduct a pilot attempt on a conditional forecasting model, which aims to achieve better forecasting performance based on different sentiment environments. Our experimental results indicate that SLNM is a conditional model on investor sentiment.

To illustrate our procedure and explore the performance of our model, Section 2 discusses the role of investor sentiment in stock price movement. Section 3 elaborates our experimental data and forecasting method. Evaluation results will be explained in Section 4. Finally, Section 5 concludes.

## 2 INVESTOR SENTIMENT AND STOCK MARKET

### 2.1 Investor Sentiment, Fundamentals and Price Movement Patterns

In the traditional asset pricing theory, emotion-free investors always force a firm’s stock price to equal the present value of the expected future cash flows, which is the “intrinsic” value or fundamental value (Baker and Wurgler (2007)). Therefore, a firm’s current stock price is the reflection of the rational cash flows which are only determined by the firm’s fundamentals. Actually, Fama(1965) assumes that there are two types of investors in the market: irrational traders who hold random beliefs about future cash flows and rational arbitrageurs who beat with irrational traders and force the prices to their fundamental level. Moreover, in the trading process, those who made mistake in judging “intrinsic” value will lose money to rational arbitrageurs and, eventually, disappear from the market. This implies that irrational traders, or noisy traders later named by Black (1986), cannot influence stock price too much and stock price is just the representation of firms’ intrinsic value, which cannot be influenced by the behaviour of irrational or emotional investors.

However, some anomalies in stock market such as price momentum continuously hunted traditional models in past several decades. Some alternative models based on investor psychology are proposed to accord with anomalies. The most famous one is DSSW (Delong et al. (1990)) which redefines the relationship between two types of traders. Noise Traders, who are subject to sentiment \(^5\) and do not estimate stock value in terms of fundamentals, will create a new risk that deters Rational Arbitrageurs from forcing stock price to its intrinsic value. That means noise traders’ behaviour creates their own live space, by which stock price can diverge from intrinsic values or fundamentals for a long period. Different with Fama’s conclusion, DSSW de facto lays out a scene that a change in sentiment can invoke noisy traders’ irrational trading behaviour, leading to intangible movements in stock price even if the fundamentals do not change in a short run. In other words, stock price series consists of two main driven-ingredients; one is “fundamental”, which determines the intrinsic value of stocks and holds stable in the short run; the other is “investor sentiment”, which contributes to nonlinear price movement such as jumps (Figure 1).

\(^4\) Based on Harvard-IV, Loughran and McDonald (2011), we construct a Chinese Financial Sentiment Dictionary (CFSD), which contains more than 4000 words. We also thank McDonald for making his dictionary public.

\(^5\) Investor sentiment is defined as a belief about future cash flows and investment risks that is not justified by the facts at hand (see Baker et al. (2007)).
Figure 1. Investor sentiment drives price away from the “intrinsic” value

As illustrated in Figure 1, on the one hand, a firm’s fundamental holds stable within days\(^6\), which means that today’s intrinsic value of a firm is almost the same with yesterday. Therefore, the stock series’ smoothness or stability within days can be perfectly simulated by linear assumption. On the other hand, the intangible movement such as sudden jumps can be well captured by nonlinear methodology such as SVM, which has been successfully applied to solve nonlinear or unknown relationship estimations.

2.2 Sentiment Environments and CARs Ranking List

Jegadeesh and Titman (1993) found solid evidence that the cross-section of stock returns can be predicted by their past returns, which means that stock returns exhibit continuation or momentum. Momentum is the tendency of stock returns. Winners that have performed well in the past tend to perform well in the future; Losers that have performed poorly in the past tend to perform poorly in the future.

To investigate rational explanations for momentum, Hong and Stein (1999) propose a theoretical model (HS) from a behaviour finance perspective. In HS model, there exists two group limited rational investors, one is news-watcher who forecast future prices based on information that they privately obtained; another is momentum trader who forecast future prices based on the past prices changes. Furthermore, Constantinos Antoniou et al. (2012) augment HS model to build a tie between investor sentiment and momentum. They argue that good news among losers spread slowly in optimistic period and bad news among winners spread slowly in pessimistic period. Their empirical work shows that momentum is strong only in optimistic period and there is almost no momentum in pessimistic period. Constantinos Antoniou et al. (2013) also discuss noise trading in different sentiment environments. They argue that emotional investors are prone to show up in optimistic environment and leave in pessimistic environment.

In our work, stocks are first ranked by the past cumulative abnormal returns (CARs) and we get a CARs ranking list (CRL). For winner stocks, which are in the top of CRL, will attract more emotional investors to invest more; losers, which are in the bottom of CRL will discourage emotional investors and deter them from further participating in trading. In this paper, therefore, we use the ranking position in CRL to label stocks’ sentiment environment. That is: winners are deemed to be in the optimistic sentiment environment and losers in the pessimistic sentiment environment.

---

\(^6\) Stock prices can hardly predict in a bubble, crash or extreme cases, which are not considered in our paper.
3 FORECASTING METHOD AND EXPERIMENTAL DATA

3.1 Procedure of Sentiment-Based Linear-Nonlinear Model (SLNM) Forecasting

Pai and Lin (2005) propose a novel linear-nonlinear hybrid model (LNM) that aims to capture both linear and nonlinear patterns in financial time series. Our SLNM augments the LNM in SVR\(^7\) training based on Chinese Financial Sentiment Dictionary (CFSD). In SVR training and predicting process, input vector is constructed by sentiment word count extracted from firm-specific financial news as the nonlinear pattern of stock time series is mainly driven by investor sentiment. Input vector of SVR model is also called “sentiment vector”.

In SLNM, stock return series is assumed to contain two parts that represent linear and nonlinear patterns respectively.

\[
R_t = L_t + N_t, \tag{1}
\]

where \(R_t\) is the stock return value, \(L_t\) represents the linear part and \(N_t\) for nonlinear part. There are four main steps in forecasting stock price with SLNM.

- Estimate \(R_t\) by a linear function model: ARMA (Box et al. (1994)) and forecast \(R_t\) by the ARMA model to get \(\hat{L}_t\), i.e. the fitted value of \(R_t\).
- Calculate the residuals \(\epsilon_t = R_t - \hat{L}_t\) and let the residuals \(\epsilon_t\) represent the nonlinear part of \(R_t\).
- Model \(\epsilon_t\) by nonlinear functions, i.e. \(\epsilon_t = f(s_{1,t}, s_{2,t}, \ldots s_{N,t})\), where \(f\) is a nonlinear function trained by SVR and \(s_{1,t}, s_{2,t}, \ldots s_{N,t}\) are the numbers of different sentiment words in total financial news on day \(t\). CFSD contains 4562 sentiment words in total and the dimensions of input sentiment vector is just equal to the number of total sentiment words: 4562.
- Forecast \(\epsilon_t\) by SVRs and get the prediction value: \(\hat{N}_t\) and the final forecasting of stock return is calculated by \(\hat{R}_t = \hat{L}_t + \hat{N}_t\).

3.2 Linear Part Estimation by ARMA and CARs Ranking List Formation

ARMA Model specification period is from the first trading day in January 2008 to the last trading day in February 2012. The average number of observations is about 1,000 and some of them will be less than 1,000 if there are some shutdown days. CARs calculated period are in recent 200 trading days before February 17th 2012\(^8\). Stocks with high (low) CARs will be on the top (bottom) of the ranking list (CRL) and regarded as winner (loser) portfolio. We choose 10\% in top stocks and 10\% in bottom stocks in consideration of our total sample only 100 stocks in CSI-100 (Table 1, Table 2).

<table>
<thead>
<tr>
<th>Stock Code</th>
<th>Observations</th>
<th>ARMA(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>601088</td>
<td>1038</td>
<td>ARMA(1:33,0)</td>
</tr>
<tr>
<td>600010</td>
<td>1014</td>
<td>ARMA(1,1:11)</td>
</tr>
<tr>
<td>601299</td>
<td>543</td>
<td>ARMA(4:9:3)</td>
</tr>
<tr>
<td>600000</td>
<td>1026</td>
<td>ARMA(0,12:29)</td>
</tr>
<tr>
<td>600795</td>
<td>998</td>
<td>ARMA(8:20,0)</td>
</tr>
<tr>
<td>601666</td>
<td>1035</td>
<td>ARMA(1,14:27)</td>
</tr>
<tr>
<td>600489</td>
<td>1032</td>
<td>ARMA(0,1:4:7)</td>
</tr>
<tr>
<td>600188</td>
<td>1024</td>
<td>ARMA(0,1:14)</td>
</tr>
<tr>
<td>601186</td>
<td>986</td>
<td>ARMA(0,1:5:16;20:32)</td>
</tr>
<tr>
<td>000898</td>
<td>1031</td>
<td>ARMA(0,1)</td>
</tr>
</tbody>
</table>

Table 1. 10 winner stocks in past 200 trading days

---

\(^7\) SVR: support vector regression is a non-linear regression function which is an adaption of SVM.

\(^8\) In momentum strategy, there is a gap period after 200 past trading days. In this paper, the gap period is 10 days.
<table>
<thead>
<tr>
<th>Stock Code</th>
<th>Observations</th>
<th>ARMA(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>601699</td>
<td>1031</td>
<td>ARMA(1,0)</td>
</tr>
<tr>
<td>601866</td>
<td>1032</td>
<td>ARMA(2:2)</td>
</tr>
<tr>
<td>101727</td>
<td>802</td>
<td>ARMA(14:22,0)</td>
</tr>
<tr>
<td>600115</td>
<td>985</td>
<td>ARMA(1:9:24,0)</td>
</tr>
<tr>
<td>000629</td>
<td>969</td>
<td>ARMA(26:27,26)</td>
</tr>
<tr>
<td>100825</td>
<td>1026</td>
<td>ARMA(1:8:23,0)</td>
</tr>
<tr>
<td>601919</td>
<td>1033</td>
<td>ARMA(14:24,0)</td>
</tr>
<tr>
<td>000425</td>
<td>987</td>
<td>ARMA(1:2:10:17,0)</td>
</tr>
<tr>
<td>601808</td>
<td>1032</td>
<td>ARMA(3:4,0)</td>
</tr>
<tr>
<td>600111</td>
<td>1032</td>
<td>ARMA(0:4)</td>
</tr>
</tbody>
</table>

Table 2. **10 loser stocks in past 200 trading days**

### 3.3 Construct Sentiment Vector in SVR Training and Forecasting

In SVR training, the key problem is how to find a proper proxy for investor sentiment. One approach is to calculate “indirect” sentiment-related financial variables such as Mutual Fund Flows, Trading Volume, and IPO First-Day Returns, to quantify investor sentiment (Baker and Wurgler (2007)). Another approach is to use “direct” measure of investor behaviour such as surveys or sentiment words in news stories. Some recent studies have found that news related measurement is feasible. Chen and Ghysels (2011) found that average good news reduces next day’s return of stock prices, while both intensive good news and bad news have a severe impact on stock return. Tetlock et al. (2008) adopted the fraction of negative words in firm-specific financial news based on Harvard-IV psychology dictionary (HPD). In the empirical work, he found negative words ratio is negatively related to firm’s next day’s stock return, which means that negative and positive words in financial text can be treated as a proxy for investor sentiment.

In this paper, we follow TetLock et al. (2008) and count sentiment words from specific firms’ daily financial news to construct the sentiment vector in SVRs training and forecasting. For the purpose of research in China, we construct the first Chinese Financial Sentiment Dictionary extended from HPD and financial word lists provided by Loughran and McDonald (2011).

### 3.4 Firm-specific Daily Financial News from the Most Popular Financial Websites

The daily firm-specific financial news is retrieved from the top financial websites ranked by page view index (PVI) from alexa.com, a famous website information provider. PVI measures how pages in one website are reviewed by web surfers and ranks the main financial websites per month. Considering the stability and generalization, we select eight financial websites who have won the top ten at least 10 times in past 15 months ranking (Table 3).

We retrieve daily financial news of 100 firms in CSI-100 Index from January 1\textsuperscript{st} 2012 to May 31\textsuperscript{st} 2012. To obtain the right news stories for specific firms, we use the firm’s common names which are widely used by financial media. In total, we retrieve 76,413 news stories of 100 listed firms between January 1\textsuperscript{st} 2012 and May 31\textsuperscript{st} 2012. To capture the instantaneous impact of invest sentiment on stock close price; we focus on the breaking news of specific firms on a trading day, i.e. the news released between 9:00 a.m. and 14:40 p.m. on trading days in China.

For convenient nonlinear SVRs training, we add up all financial news of a firm between 9:00 a.m. and 14:40 p.m. into a whole financial text. To evaluate investor sentiment impact, we extract only

---

9 News stories are written and edited by journalists or analysts who are investors themselves.

10 In China, there isn’t any institution like Dow Jones news wire that has the high reputation in financial news fields and we have to use a few influential financial websites.

11 We use Baidu Search Engine to retrieve firm specific news from eight top financial websites respectively.
sentiment words in terms of CFSD and the sentiment vector is constructed by the number of different sentiment words. Consequently, the number of dimensions of the sentiment vector in SVRs is equal to the number of sentiment words in CFSD.

<table>
<thead>
<tr>
<th>Website Name</th>
<th>Times at top 10 in recent 15 months ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hexun.com</td>
<td>15</td>
</tr>
<tr>
<td>Eastmony.com</td>
<td>15</td>
</tr>
<tr>
<td>Finance.sina.com.cn</td>
<td>15</td>
</tr>
<tr>
<td>Money.qq.com</td>
<td>15</td>
</tr>
<tr>
<td>Finance.ifeng.com</td>
<td>13</td>
</tr>
<tr>
<td>Jrj.com</td>
<td>13</td>
</tr>
<tr>
<td>Money.163.com</td>
<td>12</td>
</tr>
<tr>
<td>P5w.net</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 3. Financial web site month ranking from August 2011 to October 2012

3.5 Training and Forecasting Nonlinear Parts by SVR

In Section 3.2, we calculate ARMA forecasting errors \( \varepsilon_t \) and regard them as the nonlinear part of stock return series. The errors \( \varepsilon_t \) are trained and predicted by SVR as the followings:

\[
e_{t+1} = f(S_t) = \omega \Phi(s_{1,t}, s_{2,t}, \ldots, s_{N,t}) + b,
\]

where \( e_{t+1} \) is the ARMA forecasting errors, treated as the nonlinear part on day \( t+1 \), \( \Phi \) represents nonlinear mapping between sentiment vector—\( S_t \) and nonlinear parts—\( e_{t+1} \). Sentiment vector’s elements: \( s_{1,t}, s_{2,t}, \ldots, s_{N,t} \) are number of different sentiment words extracted from financial news on day \( t \). \( \omega \) and \( b \) are the coefficients. Gaussian kernel function \( K(x, y) = \exp(- (x - y)^2 / (2\delta)^2) \) is used in our paper.

4 EXPERIMENTAL ANALYSIS

4.1 Performance Criterion

In addition to RMSE (root mean square error), we also introduce Direction Accuracy (DA) to evaluate prediction performance. They are defined as followings:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (R_t - \hat{R}_t)^2},
\]

where \( N \) is the number of total forecasting times on day \( t \), \( R_t \) is actual return of day \( t \), \( \hat{R}_t \) is forecasting return of day \( t \). For SLNM, \( \hat{R}_t \) is the sum of ARMA prediction and the nonlinear prediction by SVRs: \( \hat{R}_t = \hat{L}_t + \hat{N}_t \).

\[
DA = \frac{\text{Number of Right Predictions}}{\text{Number of Total Predictions}} \times 100\%.
\]

In real-world invest strategy, Direction Accuracy (DA) is an important criterion. In the expression of DA, “Number of Right Predictions” mainly has two cases. First, actual return is positive and the prediction of the return is positive too; Second, actual return is negative and the prediction of return is also negative. We do not consider the case that the actual return is zero on day \( t \) and, of course, it is almost impossible that the prediction is zero.
4.2 Experimental Results Analysis

In this experiment, the training period contains about 30 trading days from March 1st 2012 to April 16th, 2012. The prediction period contains three trading days: April 17th, April 18th, and April 19th. Some special samples\(^{12}\), such as temporary close days or companies involved in major reorganizations, are not considered in this experiment.

As illustrated above, SLNM’s forecasting \(\hat{R}_t\) contains two parts:

\[
\hat{R}_t = \hat{L}_t + \hat{N}_t.
\]

\(\hat{L}_t\) is estimated by ARMA model, \(\hat{N}_t\) is predicted by SVR based on financial news released on trading day \(t\). If there is no financial news on a specific day, SLNM model cannot give a prediction of \(\hat{N}_t\) and the forecasting on no news days is just equal to the ARMA forecasting: \(\hat{L}_t\). For the performance comparison on trading day \(t\) between ARMA and SLNM, only the stocks that have at least one piece of financial news can be evaluated. Consequently, the following performance comparisons are conducted on those stocks that have at least one piece of breaking financial news on April 17th, April 18th, or April 19th.

4.2.1 Performance comparison on whole stocks in CSI-100 index

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>DA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA</td>
<td>0.191</td>
<td>71.4%</td>
</tr>
<tr>
<td>SLNM</td>
<td>0.220</td>
<td>72.4%</td>
</tr>
</tbody>
</table>

Table 4. Evaluations of unconditional forecasting

In this subsection, we focus on the comparison of unconditional forecasting performance, which means that all whole stocks in CSI-100 that have breaking news in prediction period are forecasted by ARMA and SLNM. As illustrated in Table 4, though SLNM improves forecasting performance 1% in Direction Accuracy (DA), it still cannot beat ARMA in RMSE on whole stocks, which means unconditional prediction is really hard for any model. In next sub-section, we will pay more attention on conditional prediction model.

4.2.2 Performance comparison on winners and losers

In this experiment, we focus on two special groups: winners that occupied the 10% top of CRL; losers that occupied the 10% bottom of the CRL in 200 past trading days. For winners that have high CARs, SLNM can dramatically enhance forecasting accuracy both in RMSE and DA (Table 5). While for the losers that have low or negative CARs, SLNM has the same DA with ARMA model and a fairly little improvement in RMSE (Table 6).

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>DA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA</td>
<td>0.02170</td>
<td>20%</td>
</tr>
<tr>
<td>SLNM</td>
<td>0.01752</td>
<td>60%</td>
</tr>
</tbody>
</table>

Table 5. Evaluations of winners forecasting.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>DA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARMA</td>
<td>0.01027</td>
<td>75%</td>
</tr>
<tr>
<td>SLNM</td>
<td>0.01001</td>
<td>75%</td>
</tr>
</tbody>
</table>

Table 6. Evaluations of losers forecasting.

\(^{12}\) For example, PingAn Bank is merged with Shenzhen Developing Bank in the first half of 2012, the stock code is 000001.
As illustrated in Table 5, for the stocks with high CARs, SLNM improves DA by 40% and reduces RSME by 19.3%. For the losers, Table 6 shows that SLNM and ARMA are same on DA performance. Although SLNM has better performance on loser stocks, the improvement is very limited: less than 3%, which is quite small compared with the performance on winner stocks.

According to the experimental results above, SLNM performance is conditional on different sentiment environments. SLNM can achieve better performance on the winners while, for the losers, SLNM has limited improvement.

5 CONCLUSIONS

Though traditional econometric time series models have achieved great success in stock return forecasting, they are always accused of lack of attention on nonlinear pattern because real world financial time series often contain both linear and nonlinear patterns. Moreover, existing hybrid models that aim to capture both linear and nonlinear patterns fail to make clear what contribute to linear and nonlinear patterns. Considering the particularity of stock series, we introduce financial theories to give make-sense content to linear and nonlinear patterns. We argue that firm’s fundamentals are the determinations of linear patterns and investor sentiment is the main source of the nonlinear pattern.

Base on the make-sense interpretation, we propose a novel sentiment-based hybrid model (SLNM), which uses daily financial news and SVR to train and forecast nonlinear patterns of stock return series. According to our experiments results, SLNM indicates that it has better forecasting performance on the whole stocks of CSI-100 in Direction Accuracy, not RMSE. Additionally, SLNM’s forecasting performance is conditional on different sentiment environments. For winner stocks, which attract more investor’s attention and thus are deemed to be in optimistic sentiment environment, SLNM can better forecast the nonlinear part of their return series and then, achieve more dramatically improvement in both RMSE and Direction Accuracy. For loser stocks, which deter emotional investors from further participating in the trading and thus are deemed to be in pessimistic sentiment environment, SLNM outperforms ARMA in RMSE and DA by fair improvements.

This shows that SLNM exhibits sensitivity to sentiment environments, which in turn supports the argument that investor sentiment is the main source of nonlinear pattern in stock return series. Our interpretation of stock series and SLNM shed light on forecasting models in different investor sentiment environments.

In the end, there are some additional remarks on stock series forecasting. Different from others, stock series are special financial time series because there are theories in asset pricing, by which stock series can be interpreted meaningfully. However, financial time series forecasting is challenging enough to discourage everyone. We also know that five months are inefficient; we would enlarge our dataset and time range to verify our conclusions in more complicated conditions.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (NSFC) (Grant No. 60803106, 61170133, 60903201), the Fok Ying-Tong Education Foundation, China (Grant No. 121068), and the Program for Key Innovation Research Team in Sichuan Province (2011JTD0028) and Key Innovation Research Team in Southwestern University of Finance and Economics.
References


