Research on Stock Price Forecast Based on News Sentiment Analysis

——A Case Study of Alibaba

Lingling Zhang[1], Saiji Fu[2], Bochen Li[3]

[1, 2, 3]University of Chinese Academy of Sciences, 100190, China
zll933@163.com

Abstract. Based on the media news of Alibaba and improvement of L&M dictionary, this study transforms unstructured text into structured news sentiment through dictionary matching. By employing data of Alibaba's opening price, closing price, maximum price, minimum price and volume in Thomson Reuters database, we build a fifth-order VAR model with lags. The AR test indicates the stability of VAR model. In a further step, the results of Granger causality tests, impulse response function and variance decomposition show that VAR model is successful to forecast variables *dopen*, *dmax* and *dmin*. What's more, news sentiment contributes to the prediction of all these three variables. At last, MAPE reveals *dopen*, *dmax* and *dmin* can be used in the out-sample forecast. We take *dopen* sequence for example, document how to predict the movement and rise of opening price by using the value and slope of *dopen*.

Keywords: News sentiment, dictionary matching, stock price forecast.

1 Introduction

As one of the most common sources of daily life information, it is unavoidable for media news to be decision-making basis for individuals, institutions and markets. Nevertheless, even in the recognition of the vital position of news, it can be difficult for investors to screen out effective information and make investment plan to maximize profits. Recently, more and more investors' and financial analysts' attentions have been paid on news sentiment. In May 2017, in the Global Artificial Intelligence Technology Conference (GAITC), held in the National Convention Center, it is proposed that AI will play an increasingly crucial role in the financial field in future. And text mining is going to has a promising application prospects. However, manually extracting news sentiment from news text turns out to be difficult and time-consuming.

At present, the sentiment analysis in financial mainly includes two aspects, investor sentiment and text sentiment. Nevertheless, most of Chinese scholars' researches are focused on text sentiment. With the rapid development of Internet and AI, struc-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

tural data analysis is far from enough to meet the need of people's daily life. Hence, the sentiment analysis of news text in this study is of great implication.

The effective source of information is the guarantee of text sentiment analysis. Kearney and Liu (2013) summarize various information sources, including public corporate disclosures, media news and Internet postings [1]. Dictionary matching and machine learning are the common methods of text sentiment analysis, with its own pros and cons. Dictionary matching [2-6] is relatively simple, but the subjectivity of the artificial dictionary is larger and the accuracy is limited. On the contrary, machine learning [7-10] is able to avoid subjective problems and improve accuracy, but it comes with a higher cost and much more work. In domestic study, public sentiment analysis is getting more and more popular. However, Chinese dictionaries, especially in specific areas, have not been established. Most of scholars rely on Cnki Dictionry, which is not suitable for financial analysis. Additionally, unstructured data as such Microblog and comments [11] are often utilized in domestic public sentiment analysis, which is too subjective consciousness compared with media news. Thus, immense volume of data is required to match the professional and literal dictionary. As a result, foreign dictionary turns out to be more mature and suitable, together with a wide use of English language, dictionary matching has gained its popularity. Words in dictionary matching are divided into three categories: positive, negative and neutral. It is worth of noting that constructing or selecting a sentiment dictionary that is applicable to financial study. What’s more, designing an appropriate weighting scheme has been a breakthrough in text sentiment analysis.

The stock market is closely concerned by investors. The study of the stock price forecast has also become a heated and difficult problem in recent years. At present, econometric analysis [12-16] in stock price prediction model has been very mature, such as linear regression model, vector autoregressive model, Markov chain model, BP neural network model, GARCH model [15-20]. In spite of this, unstructured data is not fully utilized, resulting the inability for pure mathematical model to achieve accurate forecast of stock market. Therefore, it provides a new method of combing quantitative news sentiment with traditional mathematical model.

The rest of paper is organized as follows. In section 2, we construct a VAR model based on news sentiment analysis. In section 3, we conduct a series of empirical tests, including data processing, unit root test, Granger causality test, impulse response function analysis and variance decomposition. In section 4, we test the forecast effect of in-static and out-static sample. Finally, in section 5, we conclude and give future work of our research.

2 Construction of VAR Model

2.1 News Sentiment Analysis

This article mainly uses the news released by the media as the source of information. In order to ensure more comprehensive information contained in the news, this article takes Alibaba as an example, using Gooseeker software to capture press release date,

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

news content and news links of 4569 news from 12 news reports including Sina Finance, China Daily, PR Newswire, The Dow Jones Network, Economic Times, Seeking Alpha, etc. The frequency of the data is based on the day, from September 19, 2014 (the day that Alibaba listed). As a representative of unstructured data, news needs to be processed through the process of Fig.1 [4].

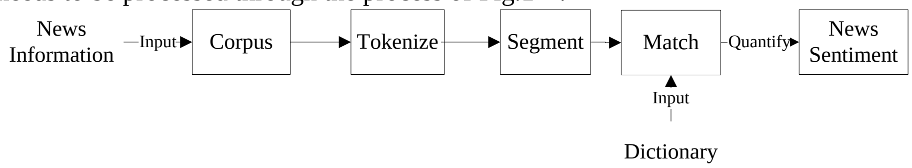

Fig. 1. Main Process of News Sentiment.

Among the process, 1) corpus, namely the collection of news, needs to be further processed in order to become useful information; 2) tokenize, that is the secondary processing of the corpus. This article combines the regular expression module in Python with Excel to remove the collection of non-essential characters in corpus; 3) segment is transforming a string into single words according to a certain characteristic; 4) match is the key means to complete the word and dictionary matching, which can be considered as the transition from unstructured data to structured data. This paper chooses the L & M dictionary as matching dictionary. This dictionary contains a number of positive and negative words, and is more suitable for the field of finance and economics. For example, "tax" is considered as a negative vocabulary in other dictionaries while a neutral vocabulary in L & M dictionary [1]. This dictionary consists of words with the same root but different meanings and different roots but the same meaning. For instance, the word "care" and "careless" have the same stem, but the meaning is exactly the opposite. The word "gram" and "grammar" also have the same root, with irrelevant meaning as well. Currently, some scholars adopt the method of stem and root matching, which will cause the problem of low accuracy. In view of the root matching will bring statistical error to some extent, this paper sacrifices matching efficiency in exchange for a higher match accuracy by treating words with the same root as different words and making the L&M dictionary a regular one dimensional array. Through matching, this article statistics the frequency of positive words and negative words appearing in each piece of news respectively, and imports the matching result into Mysql database; 5) Quantification is the destination of unstructured data into structured data. This paper defines the result of quantification as sentiment. The choice of the quantification formula is directly related to the forecast effect of the stock price in the later period. Therefore, it is very important to select a reasonable formula.

Due to the impact of the event itself, there will be the same source of different news reports and different sources of the same report. For the former, it may be necessary to sum the word frequency to quantify the text; for the latter, averaging the word frequency may be more appropriate. In order to avoid the tedious work-load of above two methods, this paper adopts the sampling method for approximate treatment. That is to say, if the sampling results show that most of the news comes from

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

different events, then all news of the same day is regarded as different events, otherwise, it is regarded as the same event. Based on the above factors, this article selects formula (1) and use of SQL statements to quantify the news sentiment. The advantage of this formula is that regardless of whether the news of the same day is eventually treated as the same event or different event, the result is the same. At present, the formula is also quite popular with scholars [21].

$$S' = \frac{\sum PF/n - \sum NF/n}{\sum PF/n + \sum NF/n} = S = \frac{\sum PF - \sum NF}{\sum PF + \sum NF} \quad (1)$$

In formula（1）, *S* denotes the sentiment values calculated by adding up, *S'* represents the sentiment values by averaging. When $S(S')>0$, the sentiment demonstrates positive, investors may be optimistic about the situation on the day, on the contrary, the sentiment takes on negative, investors may be pessimistic. *PF* indicates the frequency of positive words appearing on a particular day's news, and *NF* indicates the frequency of negative words appearing on a particular day's news.

2.2 Construction of Stock Price Forecasting Model

The stock market, as an active zone for investors, is often regarded as a barometer of economic activity and plays a decisive role in the development of the national economy. Choosing and building a reasonable stock price forecasting model is of great significance to all countries, enterprises and individuals. Based on the literature of stock price forecast, this paper summarizes the variables commonly used in predecessors' stock price forecasting, including the three categories of technical indicators, macroeconomic variables and stock price raw data[11,22-24]. Among them, the adoption of technical indicators combined with the original data is popular, and the forecast results are often satisfactory. However, the effective market hypothesis put forward by Eugene Fama in 1970 holds that all valuable information has been timely, accurately and fully reflected in the stock price movements. Even though the theory is still controversial, it can be thought that the past transaction information affects the investor sentiment on the one hand. On the other hand, the investor sentiment also indicates the volatility of the future stock market. That is, the original stock price data not only contains the information needed by investors, but also by the external sentiment. Based on this, this article assumes that the combination of raw data and sentiment value of stock price can predict the trend of future stock price. In summary, this article initially identifies the variables in the model as follows: closing price (*close*), opening price(*open*), minimum(*min*), maximum(*max*), trading volume(*volume*) and news sentiment(*sentiment*).

Considering the significant time series features and the lasting effects of each variable, this paper determines to construct a time series model. However, for the commonly used time series models such as AR (p), MA (p), and ARMA (p), the model for solving the univariate problem is served in spite of the lag effect. Taking all factors into consideration, this article focuses on the VAR (p) model. VAR model is often

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

used to predict interconnected time-series systems and to analyze the dynamic impact of stochastic disturbances on the variable system, thus explaining the impact of various economic shocks on the formation of economic variables. At present, VAR model is widely sought after by many economists. Its general form can be expressed as formula (2).

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t \qquad t = 1, 2, \cdots, T \tag{2}$$

Where Y_t is an n-dimensional endogenous variable, $t \in$ T, α_i ($i \in$ N, $0 \le i \le p$) is the parameter matrix to be estimated, εt is an n-dimensional random vector, E(ε_t)=0, p denotes the lag order. Equation (2) can be called VAR (p) model.
Ignoring the constant term, Eq. (2) can be abbreviated as Eq. (3).

$$A(L)Y_t = \varepsilon_t \tag{3}$$

Among them, $A(L) = I_n - \alpha_1 L - \alpha_2 L_2 - \cdots - \alpha_p L_p$, $A(L) \in R^{nxn}$, L is a lag operator. The formula (3) is generally called the unrestricted vector autoregressive model [25].
In summary, the preliminary non-restrictive VAR (2) model to be established in this paper is shown in Eq. (4).

$$\begin{pmatrix} close_t \\ open_t \\ min_t \\ max_t \\ volume_t \\ sentiment_t \end{pmatrix} = \alpha_0 + \alpha_1 \begin{pmatrix} close_{t-1} \\ open_{t-1} \\ min_{t-1} \\ max_{t-1} \\ volume_{t-1} \\ sentiment_{t-1} \end{pmatrix} + \alpha_2 \begin{pmatrix} close_{t-2} \\ open_{t-2} \\ min_{t-2} \\ max_{t-2} \\ volume_{t-2} \\ sentiment_{t-2} \end{pmatrix} + \cdots \alpha_p \begin{pmatrix} close_{t-p} \\ open_{t-p} \\ \min_{t-p} \\ \max_{t-p} \\ volume_{t-p} \\ sentiment_{t-p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \\ \varepsilon_{5t} \\ \varepsilon_{6t} \end{pmatrix} \tag{4}$$

3 Empirical Test of VAR Model

3.1 Data Source and Processing of Stock Price

The stock data in this article is sourced from the Thomson Reuters database. We extract opening price, closing price, the maximum price, the minimum price and trading volume from the database for a total of 633 trading days from September 19, 2014 (listed) to March 24, 2017. The data frequency is the day. In the meantime, in order to test the final out-of-sample prediction effect of the model, this paper specifically selects a total of 575 transaction days from September 19, 2014 to December 30, 2016 as sample data to input into the models, and the remaining data in total of 57 from January 3, 2017 to March 24, 2017 are reserved for the test data to test the model. Eviews9.0 is selected as the measurement software of this article.

In data processing, the six variables of the model are standardized to eliminate the dimensional difference between the variables. Generally believed that the absolute value of more than 3 can be considered as abnormal values after the standardization of the data. The results show that the trading volume data on the day of Sept. 19, 2014

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

is close to 17 and much higher than 3 after standardization, which is attributable to the noticeably higher number of news media coverage on the listing day that leads to the overwhelming reaction of the public and the abnormal trading volume. In order to avoid the large error brought to the model by the extreme trading volume on the listing day, this paper excludes the data on the date of listing before the model is constructed, and keeps the stock price data and sentiment values of the remaining 574 trading days.

3.2 Unit Root Test of VAR Model

The application of VAR model requires that the sequence be stable, otherwise, it is easy to produce false regression [12]. For example, wrong conclusion may be made within are two variables with no economic relationship. However, the sequences encountered in real life are often non-stationary, which need to be differenced to obtain the smooth sequence. In order to eliminate the phenomenon of pseudo-regression, we use the ADF test to test the sequence of model variables. The results are shown in Table 1.

Table 1. T ADF Test Results.

Variables	Test Statistics	1% Threshold	5% Threshold	10% Threshold	P Value	Stable or not
volume	-12.54943	-3.974123	-3.417668	-3.131264	0.0000	Yes
sentiment	-19.04287	-3.974123	-3.417668	-3.131264	0.0000	Yes
dclose	-22.44971	-3.974152	-3.417681	-3.131272	0.0000	Yes
dopen	-26.25389	-3.974152	-3.417681	-3.131272	0.0000	Yes
dmax	-22.09662	-3.974152	-3.417681	-3.131272	0.0000	Yes
dmin	-22.26319	-3.974152	-3.417681	-3.131272	0.0000	Yes

The results show that volume and sentiment are I (0) processes, close, open, max and min are I (1) processes, denoted as *dclose*, *dopen*, *dmax* and *dmin* respectively. There is a clear mapping between close and *dclose*. When *dclose*> 0, it can be inferred that today's closing price is higher than the closing price yesterday, on the contrary, the closing price today is lower than yesterday's closing price, the remaining variables are the same to be obtained. Finally, the six stationary sequences of *dclose*, *dopen*, *dmax*, *dmin*, *volume* and *sentiment* are added to the VAR model. Taking lag 2 as an example, the transition from formula (4) to formula (5) is made.

$$\begin{pmatrix} dclose_t \\ dopen_t \\ dmin_t \\ dmax_t \\ volume_t \\ sentiment_t \end{pmatrix} = \alpha_0 + \alpha_1 \begin{pmatrix} dclose_{t-1} \\ dopen_{t-1} \\ dmin_{t-1} \\ dmax_{t-1} \\ volume_{t-1} \\ sentiment_{t-1} \end{pmatrix} + \alpha_2 \begin{pmatrix} dclose_{t-2} \\ dopen_{t-2} \\ dmin_{t-2} \\ dmax_{t-2} \\ volume_{t-2} \\ sentiment_{t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \\ \varepsilon_{5t} \\ \varepsilon_{6t} \end{pmatrix} \tag{5}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

3.3 Determination of Lag Period in VAR Model

The determination of lag order is directly related to the quality of the model. On the one hand, the larger the lag order, the more realistic and comprehensive the information reflected. On the other hand, an excessively large lag order will lead to a decrease of the freedom degree of the model and an increase of the estimated parameters, thereby increasing the error and decreasing the prediction accuracy. Based on this, the proper lagging order plays a decisive role. In this paper, the 8-order lag test is carried in VAR(2) model by Eviews9.0, the results shown in Table 2.

Table 2. Lag Period test results.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	522.8801	NA	6.49e-09	-1.826431	-1.780439	-1.808481
1	1098.340	1136.687	9.64e-10	-3.732651	-3.410706	-3.606999
2	1227.335	252.0639	6.94e-10	-4.061255	-3.463357*	-3.827900
3	1321.375	181.7657	5.65e-10	-4.266342	-3.392491	-3.925286*
4	1382.570	116.9841	5.17e-10	-4.355370	-3.205566	-3.906612
5	1426.921	83.84496	5.02e-10*	-4.384881*	-2.959124	-3.828421
6	1461.213	64.09933	5.06e-10	-4.378843	-2.677133	-3.714681
7	1489.105	51.54672	5.21e-10	-4.350195	-2.372532	-3.578331
8	1526.057	67.50489*	5.19e-10	-4.353557	-2.099941	-3.473991

According to the principle of asterisk at most, it is determined that the model is optimal for 5 lags, so the VAR (5) model is established as Eq. (6).

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_3 Y_{t-3} + \alpha_4 Y_{t-4} + \alpha_5 Y_{t-5} + \varepsilon_t \tag{6}$$

Among them, $Y = \begin{pmatrix} dclose \\ dopen \\ dmin \\ dmax \\ volume \\ sentiment \end{pmatrix}, \alpha_0 = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{pmatrix}, \varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \varepsilon_{4t} \\ \varepsilon_{5t} \\ \varepsilon_{6t} \end{pmatrix}$

The results of the VAR model can be estimated by OLS.

The AR test is used to determine the stability of the VAR (5) model, as shown in Fig. 2, all the characteristic roots of the model fall within the unit circle, indicating that the model is stable.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

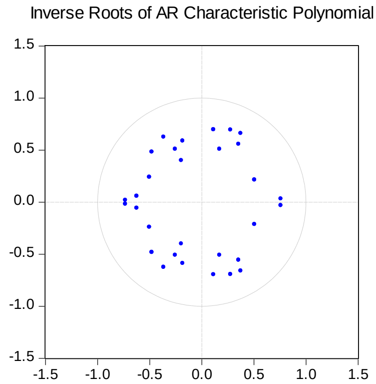

Fig. 2. Discrimination of Model Stability.

3.4 Empirical Analysis of VAR(5) Model

Even the stability of the VAR model is indicated in the above analysis, it still is unlikely to explain the whether and to what extent does the news sentiment contribute to the model. Therefore, we use the Granger causality tests, impulse response function and variance decomposition analysis to analyze the model in a further step.

(1) Granger Causality Tests

The causality test for the time series data of 6 variables in this study is conducted using "Granger Causality Test", respectively. Table 3 summarizes the test results where the P value is less than 0.05. The P value for variable *dopen* is 0.0000, pointing out that variable *dopen* has significant impact on the lagged items *dclose*, *dmax*, *dmin*, volume and sentiment. That is to say, variables *dclose*, *dmax*, *dmin* and *volum*e can be capitalized to forecast *dopen*. Also, *dmax* and *dmin* have significant impact on the lagged items of the rest of variables.

Table 3. Granger Causality Test Results.

Variables	**H0**	**Chi 2**	**Prob > Chi 2**	**Accept the H0 or not**
dopen	*dclose, dmax, dmin, volume* and *sentiment* do not casue *dopen*	957.2198	0.0000	No
dmax	*dclose, dopen, dmin, volume* and *sentiment* do not casue *dmax*	224.9506	0.0000	No
dmin	*dclose, dopen, dmax, volume* and *sentiment* do not casue *dmin*	242.9907	0.0000	No

(2) Impulse Response Function

Based on the stability of model, the impulse response function explains the response of an endogenous variable to one of the innovations. It traces the effects on present and future values of the endogenous variable of one standard deviation shock to one of the innovations. According to Granger Causality test, we examine the response of variables *dopen*, *dmax* and *dmin* to residual disturbance.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

1) The Response of Variable *dopen*

It can be seen from the Fig. 3 that variable, the shock of one standard deviation at the current period has a strong impact on variable *dopen*, which begins to fluctuate around 0 since period 3, nearly vanishing at period 9. Likewise, given an unexpected shock in *dclose*, *dopen* will initially increase and starts to fall afterwards, fluctuating around 0. This response has acted in line with the shock of itself, converging to 0 at period 9. The relationship between sequences *dopen* and *dmin*, *dmax* and *volume* is not significant. With the existence of lags, the effect on the sequence is also small, exhibiting a fluctuating trend till period 9. In line with that, the lag also exists in the response of *dopen* to *sentiment* at current period. The link between *sentiment* and *dope*n can be quite complex as it can either be positive or negative, which gradually disappears at period 8.

Hence, we can draw the conclusion that except *dopen* itself, only variables *dclose* and *sentiment* have a significant influence on *dopen*.

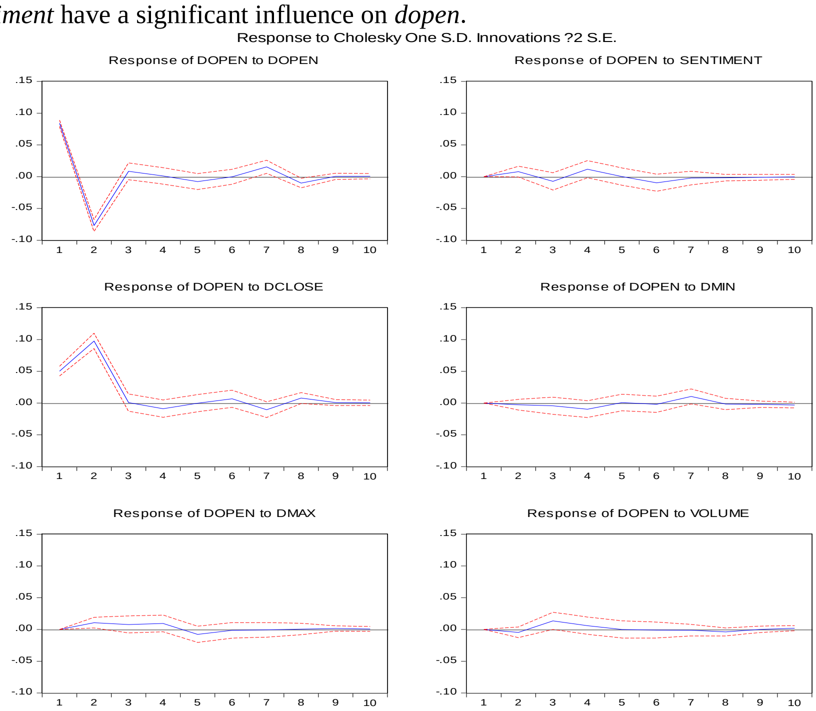

Fig. 3. Response of Variable *dopen* to System Variables.

2) The Response of Variable *dmax*

Due to limited space, the figure of the response of variable *dmax* is not shown here. However, the result depicts that the link between *dmax* and *dmax* presents up-and-down trends till period 3. Like the response of *dopen* to *dopen*, the trend gets close to 0 then. At current period, the variable *dclose* has an even stronger shock to *dmax* than *dmax* itself. The impact gets weaker since period 2, almost decreasing to 0 since period 3. In the event of a one standard deviation shock in *dopen*, *dmax* will decrease up

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

until period 2, after which it will increase. The *dmax* will decrease again up until period 4. It takes about 9 periods for *dmax* to fully become stable. Finally, the result obtained from the IRF suggests a 1 period lag time facing a one standard deviation shock of *sentimen*t, then rising and falling, gradually showing no response till period 9.

Accordingly, except *dmax* itself, only variables *dclose*, *dopen* and *sentiment* have a significant influence on *dmax*. The degrees of impact are in their stated order.

3) The Response of Variable *dmin*

Due to limited space, the figure of the response of variable *dmin* is not shown here. However, the result describes that variable *min* will be positively affected by *dclose*, *dopen* and *dmin* at current period. These influences then start to decline and get close to 0 since period 3. The lags exist in the response to *dmax*, *volume* and *sentiment*, especially the *sentimen*t. It takes about 7 periods these 3 variables to fully become stable. In particular, *volume* has a general positive effect on the sequences.

Therefore, variable *dmin* is only affected significantly by *dclose*, *dopen* and *dmin*. The relationships between *dmin* and the rest of variables are not significant.

In this paper, we focus on how the news sentiment effects stock price. As what have been stated above, variable *sentiment* can make contribution to the forecast of *dopen*, *dmax* and *dmin*. In particular, the *dopen* and *dmax* have more significant influence on sentiment, compared to *dmin*. In the meantime, *dopen*, *dmax* and *dmin* are first order difference sequence of *open*, *max* and *min*, respectively. It is easy to find out that there turns out to be a corresponding relationship between difference sequence and original sequence. Taking the *dopen* for example, if *dopen*>0, it means the opening price has the tendency to climb. And a larger slope leads to higher price, and vice versa. In line with *dopen*, the value and slope of first order difference sequence of *dmax* and *dmin* also enable us to predict the trend of original sequence, determining investor's expectation.

(3) Variance Decomposition Analysis

In order to discover how does every structural shock contribute to the change of variable, we adopt *Relative Variance Contribution Rate (RVC)* to examine a relationship between variable *j* and the response of variable *i*. Based on the results of Granger causality tests and impulse response function, we will pay our attention on the decomposition analysis of *dopen*, *dmax* and *dmin* from period 1 to 10.

Firstly, we run the analysis with variable *dopen*. The result shows that variables *dclose* and *dopen* contribute most to *dopen*, next are *sentiment* and *dmax,* whereas *dmin* and *volume* barely have no impact on the forecast of *dopen*, in accordance with the result of impulse response function.

Secondly, Variance decomposition of variable *dmax* presents that our finding further confirms the earlier impulse response function: one standard deviation shock of *dmax* makes the greatest contribution the *dmax*, then are the *dclose*, *dopen* and *sentiment*. Particularly, the effect of *sentiment* is small at first, and becomes larger as the time goes by.

Finally, the result of variance decomposition of *dmin* shows that the effects of six variables on *dmin* last for 10 periods. The variables making the largest contribution is *dmin* and *dclose*. Also there are similar but non-trivial responses of *dmin* to the rest of

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

variables. The influence of *sentiment* on *dmin* is small in the initial stage, after which it will increase.

Due to limited space, the result of variance decomposition tables is not shown here. It can be concluded that the results of variance decomposition of *dopen, dmax* and *dmin* are essentially in agreement with the results of previous impulse response function. News sentiment variable *sentiment* has significant effect on all three variables. The impacts of *sentiment* on *dmax* and *dmin* are small in the initial stage, after which it will become greater. Our conclusion is consistent with Larkin and Ryan (2008), which documents that news is successfully able to predict stock price movement, although the predictive movement only accounts for 1.1% of whole movement [25].

4 Discussion on Forecast Effect of VAR(5) Model

4. 1 Forecast Effect of In-Static Sample

Even though news sentiment can be used to forecast stock price, the forecasting effect remains unknown. We adopt 575 samples of variable *dopen* to achieve in-sample forecast. Sample 250 – 400 from 22/04/2016 – 17/09/2015 is randomly chosen to present a clearer observation. Fig. 4 reveals the comparison between the actual value sequence (in solid line) and forecast value sequence (in dashed line).

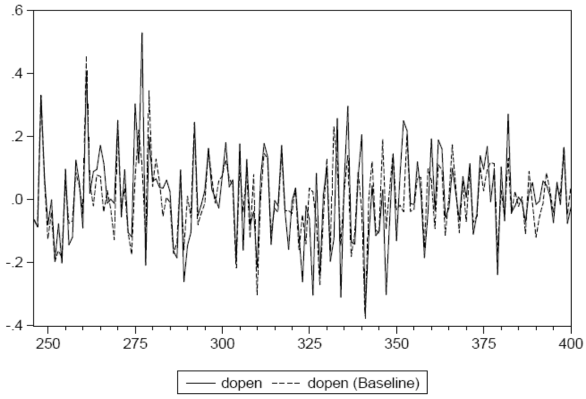

Fig. 4. Forecast Result of In-Static Sample of Variable *dopen*.

In a further step, *mean absolute percentile error* (*MAPE*) is used to evaluate the in-sample forecasting accuracy. The *MAPE* of *dopen, dmax* and *dmin* are all less than 10 (2.12,2.48 and 5.33, respectively), enabling extrapolation forecasts of these three variables.

4.2 Forecast Effect of Out-Static Sample

Fig. 5 depicts comparison between actual value sequence (in solid line) and forecast value sequence (in dashed line), using the samples from 576 to 632, which date from

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

03/01/2017 to 24/03/2017. The out-sample prediction is generally satisfactory, where the forecast sequence is nearly line with original sequence. Even the specific abnormal data indicates the correct movement.

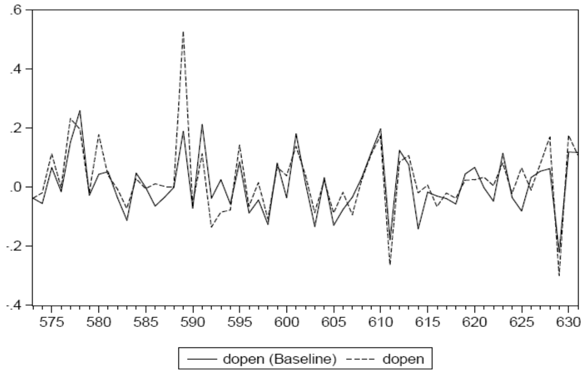

Fig. 5. Forecast Result of Out-Static Sample of Variable *dopen*.

The VAR(5) model is proved to be effective to forecast variable *dopen* by using either in or out sample data. It is well-known that the opening price acts as a signal for stock market, indicating investor's expectation. A high opening price means investors are optimistic about stock price, resulting in a promising development of market. Nevertheless, it can be harder for profit taking or arbitrage when the price goes too high; A low opening price express the possibility that market is going to be bad or whipsawed, requiring combing with the specific situation to make prediction; A price closed to the previous session's closing price shows no obvious rise and fall. Hence, a thorough understanding of opening price is of great importance for investors. Impulse response function above is suggested to forecast the movement of opening price, by giving a look at the value and slope of variable *dopen* sequence. By this way investor's expectation can be further revised. Variable *dmax* and *dmin* can also be predicted by conducting the same method. A wide discrepancy illustrates an active stock market and a greater profit opportunity, and vice versa.

5 Conclusion and Future work

In this study, we have proposed a forecast model to predict news sentiment around stock price. Base on dictionary matching, unstructured news text is transformed into structured news sentiment. We build a fifth-order VAR model with lags using the data of original stock price, including opening price, closing price, maximum price, minimum price and volume of transaction. Granger causality tests, impulse response function and variance decomposition analysis are employed to analyze the data of Alibaba news and its stock transaction. The result identifies the ability of VAR model to forecast variable *dopen, dmax* and *dmin*. In other words, news sentiment makes contribution to predict all these three variables. What's more, variable *dopen* is used to examine the predict effect of VAR model. The forecast sequence is accordance with origi-

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

nal sequence, successfully to reflect the sequence general movement. However, due to the complexity of stock market, limited ability of author, more explanatory variables need to be concerned in the model, enhancing investor's decision in a further step.

References

1. Kearney, C. and Liu, S. (2013). "Textual Sentiment Analysis in Finance: A Survey of Methods and Models", Financial Analysis, 33(3), pp.171-185.
2. Tetlock, P. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", The Journal of Finance, 62(3), pp.1139–1168.
3. Tetlock, P., Saar-Tsechansky, M. and Macskassy, S. (2008). "More than Words: Quantifying Language to Measure Firms' Fundamentals", The Journal of Finance, 63(3), pp.1437–1467.
4. Chowdhury, S. G., Routh, S., Chakrabarti, S. (2014). "News Analytics and Sentiment Analysis to Predict Stock Price Trends", International Journal of Computer Science and Information Technologies, 5(3), pp.3595-3604.
5. Loughran, T., Mcdonald, B. (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", The Journal of Finance, 66(1), pp.35–65.
6. Ferguson, N. J., Philip, D., Lam, H. Y. T., Guo J. (2015). "Media Content and Stock Returns: The Predictive Power of Press", Multinational Finance Journal, 19(1/1), pp.1-31.
7. Schumaker, R. P., Zhang, Y., Huang, C. N., Chen, H. (2012). "Evaluating Sentiment in Financial News Articles", Decision Support Systems, 53(3), pp.458-464.
8. Schumaker, R. P., Chen, H. (2009). "A Quantitative Stock Prediction System Based on Financial News", Information Processing and Management, 45(5), pp.571-583.
9. Feng, L. I. (2010). "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve [1] Bayesian Machine Learning Approach", Journal of Accounting Research, 48(5), pp.1049–1102.
10. Sehgal, V., Song, C. (2007). "SOPS: Stock Prediction Using Web Sentiment", ICDM Workshops, IEEE.
11. Zhu, M. J., Jiang, H. X., Xu, W. (2016). "Stock Price Prediction Based on the Emotion and Communication Effect of Financial Micro-blog", Journal of Shandong University (Natural Science), 51(11), pp.13-25.
12. Cao, Y. B. (2014). "Study on the Influence of Open Market Operation on Stock Price – An Empirical Analysis Based on VAR Model", Economic Forum, (7), pp.88-94.
13. Liu, L. (2006). "A Research on the Relationship between Stock Price and Macroeconomic Variables Based on Vector Autoregression Model", Hunan University.
14. Yu, Z. J., Yang S L. (2013). "A Model for Stock Price Forecasting Based on Error Correction", Chinese Journal of Management Science, pp.1-5.
15. Xu, F. (2006). "GARCH Model of Stock Price Prediction", Statistics and Decision, (18), pp.107-109.
16. Chen, Z. X., He, X. W., Geng, Y. X. (2008). "Macroeconomic Variables Predict Stock Market Volatility", International Institute of Applied Statistics Studies, pp.1-4.
17. Xu, W., Li, Y. J. (2015). "Quantitative Analysis of the Impact of Industry and Stock News on Stock Price", Money China, (20), pp.31-32.
18. Sun, Q., Zhao, X. F. (2008). "Prediction and Analysis of Stock Price Based on Multi-objective Weighted Markov Chain", Journal of Nanjing University of Technology (Natural Science Edition), 30(3), pp.89-92.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33

19. Xu, X.J., Yan, G.F., (2011). "Analysis of stock price trend based on BP neural network", Zhejiang Finance, (11), pp.57-59.
20. Peng, Z. X., Xia, L. T. (2004). "Markov Chain and its Application on Analysis of Stock Market", Mathematica Applicata, (S2), pp.159-163.
21. Gao, T. M. (2009), "Method and Modeling of Econometric Analysis: Application and Example of EViews", Beijing: Tsinghua University Press.
22. Chen, X. H., Peng, Y. L., Tian, M. Y. (2016). "Stock Price and Volume Forecast Based on Investor Sentiment", Journal of Systems Science and Mathematical Sciences, 36(12), pp.2294-2306.
23. Zhang, S. J., Cheng, G. S., Cai J H, Yang J W. (2013). "Stock Price Prediction Based on Network Public Opinion and Support Vector Machine", Mathematics in Practice and Theory, 43(24), pp.33-40.
24. Xie, G. Q. (2012). "Stock Price Prediction Based on Support Vector Regression Machine", Computer Simulation, (4), pp.379-382.
25. Larkin, F., Ryan, C. (2008). "Good News: Using News Feeds with Genetic Programming to Predict Stock Prices", Genetic Programming, Berlin Heidelberg: Springer, 49-60.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_33