Jurnal Teknologi

Full Paper

SOCIAL NETWORK NEWS SENTIMENTS AND STOCK PRICE MOVEMENT: A CORRELATION ANALYSIS

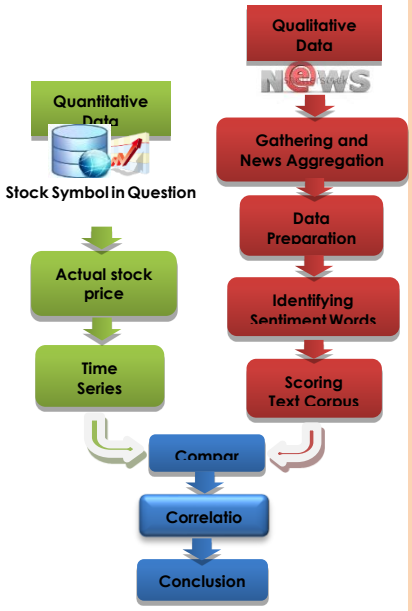
Anupong Sukprasert[a*], Kasturi Kanchymalay[b], Naomie Salim[c], Atif Khan[c]

[a]Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand
[b]Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[c]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Article history
Received
15 May 2015
Received in revised form
1 July 2015
Accepted
11 August 2015

*Corresponding author
Anupong.s@acc.msu.ac.th

Graphical abstract



Abstract

The stock market prediction is one of the most important issues extensively investigated in the existing academic literatures. Researchers have discovered that real–time news has much bearing on the movement of stock prices. Analysts now have to deal with vast amounts of real time, unstructured streaming data due to the advent of electronic and online news sources. This paper aims to investigate the relationship between online news and actual stock price movement. R programming together with R package are applied to capture and analyze the online news data from Yahoo Financial. The data are plotted into graphs to analyze the relationship between the two variables. In addition, to ensure the levels of the relationship, the Pearson's correlation and Spearman's Rank are applied to test whether there is a statistical association between these two variables. This initial analysis of dynamic online news based on sentimental words is relatively constructive.

Keywords: Stock market, news sentiment, correlation analysis

1.0 INTRODUCTION

Stock forecasting is a crucial knowledge that an investor seeks to make profits. The main reason of forecasting in business is to assist investors to decide when to take actions in investing: buying, selling, or holding on to a stock. There are several factors of stock forecasting influencing stock movements. In addition, stock market's engagements are exaggerated through numerous macro-economic elements such as, political situations, organisational' strategies, economic circumstances, investors' prospects, established stockholders' selections, the drive of extra marketplace stock, and consciousness of investors [1]. These factors drive stock prices on an upward, downward or steady.

Stock price trends are determined solely by the interaction between demand and supply. Shifts in demand and supply cause reversals in trends and can be detected in charts. Chart patterns are also likely to be repeated [2]. The shifts of demands and supplies influences the stock price and affects the stock price trends [3]. However, technical analysts believe that the market is always correct. All factors have already been factored into the demand and supply curves and; thus, the trends of the company's stock [4]. The factors mentioned can be found on social media, such as, Twitter, Blogs, Forums, Facebook, and websites such as Google, Yahoo, CNN etc. The collection of information in terms of news retrieved from social media might be useful in analysing the trends of stock market.

In this paper, The R Program, an open-source software is used to build a news engine for collecting and gathering news and actual stock prices from websites. The program can also identify sentiment words and then score headline news to calculate the score of corpus, known as 'NewsSentiment Score'. The remaining contents of this article are structured as follows – Related literatures were discussed in section 2. Next, research methods and procedures of news analysing were proposed. In addition, section 4 described results and discussions. Section 5 concludes this study besides give recommendation for future studies.

2.0 RELATED WORK

Several existing studies related to news sentiment analysis and text analysis have been widely conducted. For example, there are number of studies investigated in the association between news item and its influence on economic indices and oil prices. Correlation of news item and major indices of stock markets DOW JONES, NASDAQ and S&P500 were analysed by the author [5]. The data were collected over a period of ten weeks and the news items were divided into four main groups. The cause of stock price movement was also measured daily by regression model. N. Godbole *et al*. [6] presented news article and blog by extracting sentiment in order to connect entities with sentiment and aggregate, and score each entity. Their work is related to ours in terms of theories and work assessments but differs in the way they used static data while we used dynamic data that capture activities as they unfold. Related theories and work assessments were presented in this study [6]. However, in his studies the static corpus used and not the dynamic corpus type which time for the downloading should be included. The forecasting result also is not as accurate as it is supposed to be in term of the financial indices and stock prices.

J. Leskovec *et al*. [7] focused on extracting sentiment from static corpus of the daily rhythms in the news media. A total of 90 million articles were tracked and a set of novel and persistent temporal patterns presented for the news cycle. The data were analysed by mathematical model using temporal variation that the system exhibits. W.-B. Yu *et al*. [8] presented the use of text mining to explain sentiment of news articles, and also showed the effect of energy demand. News sentiment was quantified and compared with fluctuations in energy demands and prices.

S. Theußl *et al*. [9] employed text mining in R Program to analyse News Sentiment, to find large scale sentiment analysis using static corpus from New York Times. Terms based on their polarity for calculating a sentiment score were also described. Distributed text mining techniques with the Map-Reduce paradigm were described as well. Hofmarcher *et al*. [10] also endorsed a related study. The instruction set to generate sentiment score from static corpus of news articles and studied the relationship of current economic indices. A. Nagar and M. Hahsler [11] used the text mining to aggregate news from different sources and developed news corpus. The natural language processing technique was used to analyse News Sentiment by finding polarity in order to measure the sentiment of the overall news corpus.

3.0 PROPOSED METHOD

In this part, methods and procedures used to collect news from a website are presented. The procedures start from gathering news and aggregating into corpus to form the news sentiment analysis.

3.1 News Gathering and Aggregation

Finance news samples were downloaded from Yahoo Financial, using "tm" Package [12] and we used the R Program to facilitate the use of different plug-ins such as tm.plugin.webmining [13] to capture and manage the real-time news data for further analysis.

3.1.1 Creating News Corpus

As indicated above, the analysis is based on Yahoo Financial and corpus for Apple Inc. of Yahoo Financial was generated for the study. Firstly, we define a stock symbol –"AAPL" for the News Corpus and get it implemented in R code as:

```
>corpus<- WebCorpus(YahooFinanceSource("AAPL"))
```

However, some parts of the news collected may not be required because they are not useful in the analysis and in comparing the news story. Hence, News Aggregation is conducted by obtaining only headlines news in the analysis process since headline news is a short meaningful idea and relevant information about the stocks. Generation of Headline news is achieved with the following R code:

```
> head <- sapply(corpus, function(x) {attr(x, "Heading")})
> mydata <- Corpus(VectorSource(head))
```

3.1.2 Data Preparation

A number of pre-processing activities to remove unwanted data and prepare the text for Natural Language Processing (NLP) were performed. The pre-processing include mapping to lower case, removing punctuation, removing numbers, stripping white spaces, removing stop words, and stemming. The data preparation process is achieved using the following R code: Data Preparation for Headline News corpus

```
mydata <- tm_map(mydata, PlainTextDocument)
mydata <- tm_map(mydata, stripWhitespace)
mydata <- tm_map(mydata, function(x) removeWords(x, stopwords("english")))
mydata <- tm_map(mydata, removeNumbers)
mydata <- tm_map(mydata, removePunctuation)
mydata <- tm_map(mydata, tolower)
mydata <- tm_map(mydata, stemDocument)
```

3.1.3 Identifying Sentiment Words

Sentiment words as proposed by Hu and Liu (XXXX), was identified using a simple algorithm that counts the number of occurrences of "positive" and "negative" words in Headline News. On the basis of counting, scores are assigned to the relevant sentences. The polarity of the sentences are then calculated using "opinion lexicon" published by Hu and Liu [14].

3.2 Scoring Text Corpus

News articles are kept in memory in the form of document-term matrix with the headline news as rows and terms as columns. The Scoring of text corpus for Headlines news is based on the following definitions:

Definitions 1: Headline News is positive if the count of positive words is greater than or equivalent to the count of negative words and vice versa.

$$s_h = \begin{cases} 1, if\ n_p \geq n_n \\ 0, n_p < n_n \end{cases} \quad (1)$$

where s_h is the score of headlines news, n_p is the number of positive words in headline news, n_n is the number of negative words in headline news and the sign function returns the sign of its argument.

For example, the headline news "latest apple inc. smartphone could boost retail electronics sales, analyst says" has 1 Positive term – "boost", and there is no negative terms. So this headline would be a positive sentence. Another example "The iPhone 6 Plus Really Does Have A Bending Problem" Has 1 Negative term – "Problem" There is no positive terms. So this headline would be a negative sentence. We also defined a score for the headline corpus using Definition 2:

Definition 2: Score of corpus is considered by the ratio between the numbers of headline news' positive sign to total number of headline news' sign. For an entire corpus, we count the positive and negative instances and compute the score as:

$$s_c = \frac{\sum_{i=1}^{n} s_{h_i}}{N} \quad (2)$$

where s_c is the score of corpus, N is total number of headline news in corpus. For example, if 5 news headlines are gathered and found that there are 3 positive headlines, the score of corpus should be equivalent to 0.6 and this score is called as NewsSentiment Score.

4.0 RESULT AND DISCUSSION

Results of the suggested method presented by comparing the daily trends of actual stock price movement (increasing and decreasing prices) with the corpus signs of daily news headlines. We analysed 21 days headlines news data for the periods 2nd September 2014 to 30th September 2014.

First, within the same day, we plot the graph by setting the Headline's NewsSentiment Score laps and Apple Inc's stock price movement. For example, on 2nd September, 2014, the NewsSentiment Score has been analysed to compare with the actual stock price. On 3rd September, 2014, trend of hike and drops between Headline's NewsSentiment Score and Apple Inc's stock price movement are compared.

However, there is a need to standardise the two variables into the same unit while the Z-score is applied to help standardise the two variables.

$$Z_{x_t} = \frac{x_t - \bar{x}(x_{t\pm k})}{s(x_{t\pm k})} \quad (3)$$

where $\bar{x}(x_{t\pm k})$ and $s(x_{t\pm k})$ represent the mean and standard deviation of the time series within the period [t-k,t+k]. This normalisation causes the time series to fluctuate around a zero mean and be expressed on a scale of 1 standard deviation. Figure 2 shows the graph plotted.

In addition, the trend change has been analysed to observe the trends of the prices whether it is increasing or decreasing. The formula (4) as shown below is used for trend change calculation. Figure 3 shows graphs plotted.

$$\text{Trend} = \begin{cases} 1, \text{if trend going up} \\ 0, \text{trend going down} \end{cases} \quad (4)$$

Later, the percentage changes have been analysed to observe the trends of stock price by calculating the percentage changes from formula (5). Figure 4 shows graph plotted.

$$\text{Percentage changes} = \left(\frac{\text{New Value} - \text{Old Value}}{|\text{Old Value}|} \right) \text{x } 100\% \quad (5)$$

where |.| referred to as an absolute value, all values are positive.

Finally, percentage change transformed into the form of normalisation by applying the Z-score. Figure 5 shows the graph plotted. The trend of hike and drops between Headline's NewsSentiment Score and Apple Inc's stock price movement are compared as shown in Figure 1-5.

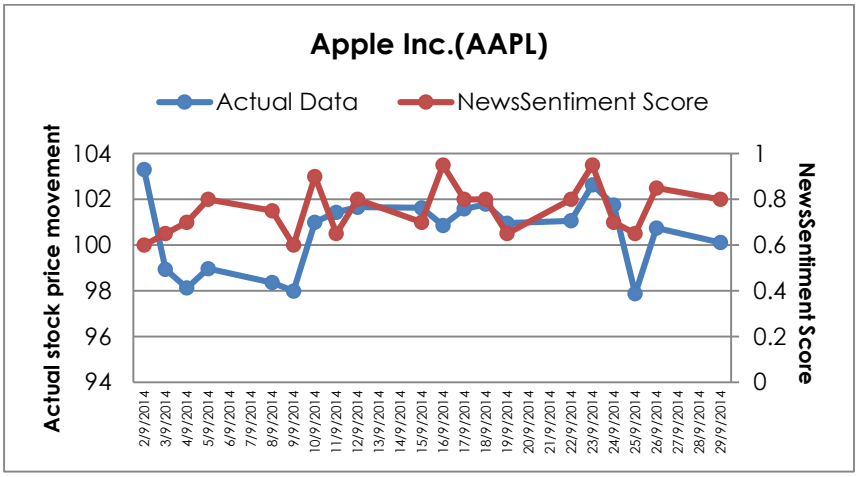


Figure 1 Shows the comparison of the Headline's NewsSentiment Score and Apple Inc's stock price movement (1 day overlap)

From figure 1 we can observe that the NewsSentiment Score is relatively moving in the same direction as the Apple Inc's stock price movement. It can be seen from the above graph that the trends of the two variables are continuously increasing, but at the end of the succession, the trends of the two variables become closer and increase into the same direction. In the overall succesion, two curves are moving in same direction except from period 2/9/2014 to 4/9/2014, and from 15/9/2014 to 17/9/2014 that it does move against another graph. There is also an unexpect sharplydrop at 11/9/2014.

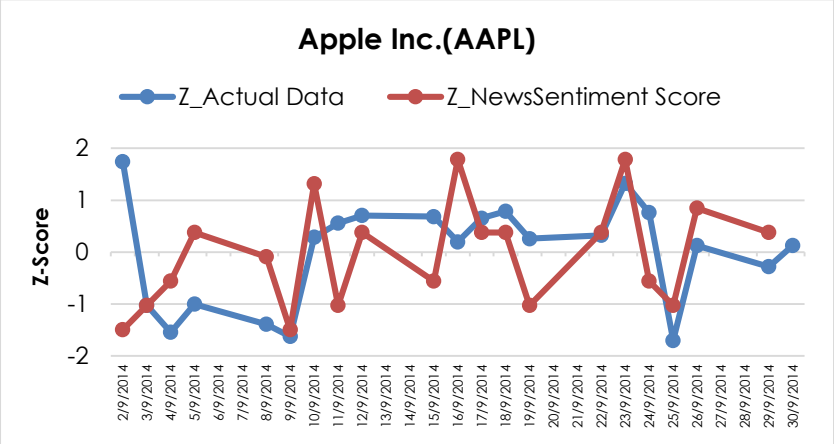


Figure 2 Shows the comparison of the Headline's NewsSentiment Score and Apple Inc's stock price movement (Once all values have been normalised)

Once standardised both Actual Data and NewsSentiment Score into the same unit, we can see that at the beginning of the graph the two variables derive from different position; positive and negative levels. Furthermore, the trends of the graph are moving nearly into the same direction. Finally, we can observe that the trends of the two variables positively increasing into the same direction.

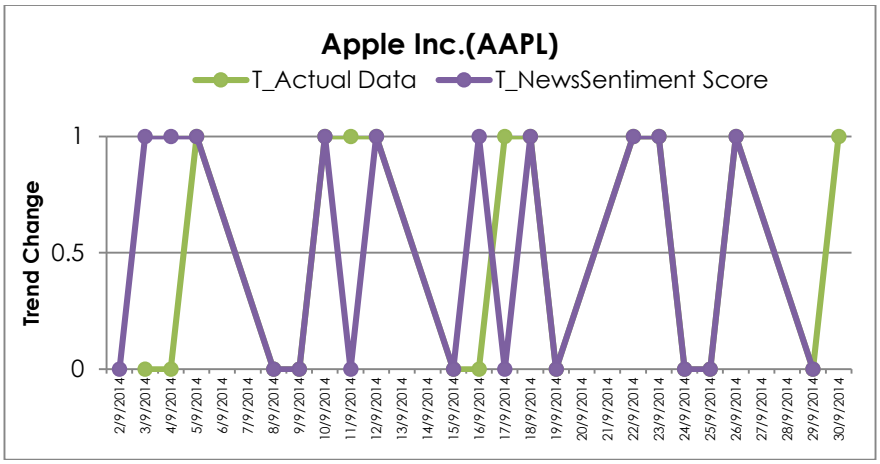


Figure 3 Shows the comparison of the Headline's NewsSentiment Score and Apple Inc's stock price movement (Once calculating trend change)

In Figure 3, once calculating trend change of the Headline's NewsSentiment Score and trend change of Apple Inc's stock price movement. The graph in Figure 3 demonstrates that the two variables derive from the same position, the bottom of the Y axis. The trends of the two variables are gradually moving in the same direction. The graphs are finally manifest relationship under two curves which Trend of actual data change in according to yellow line after some time delay at 4/9/2014 and 16/9/2014. An unexpected change is also occurred at 11/9/2014 and is shown in graph
.

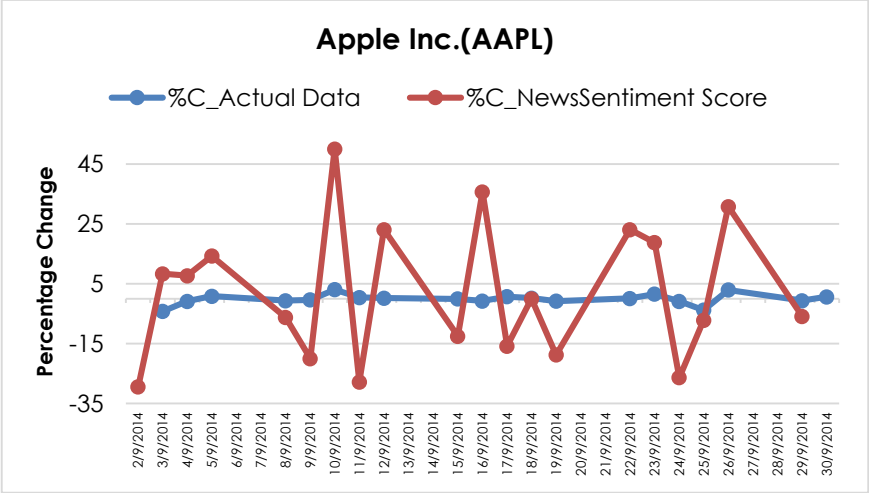


Figure 4 Shows the comparison of the Headline's NewsSentiment Score and Apple Inc's stock price movement (Once calculating percentage change)

Once calculating the percentage change the Headline's NewsSentiment Score and Apple Inc's stock price movement. According to Figure 4. It suggests that the percentage change the Headline's NewsSentiment Score and Apple Inc's stock price movement begins at the negative level. Furthermore, the trends of these two are relatively similar. However, due to the small amount of the percentage change of Apple Inc's stock price movement, it may be obvious to see the differences between the two variables. Hence, the normalization of percentage change has been applied as shown on Figure 5.

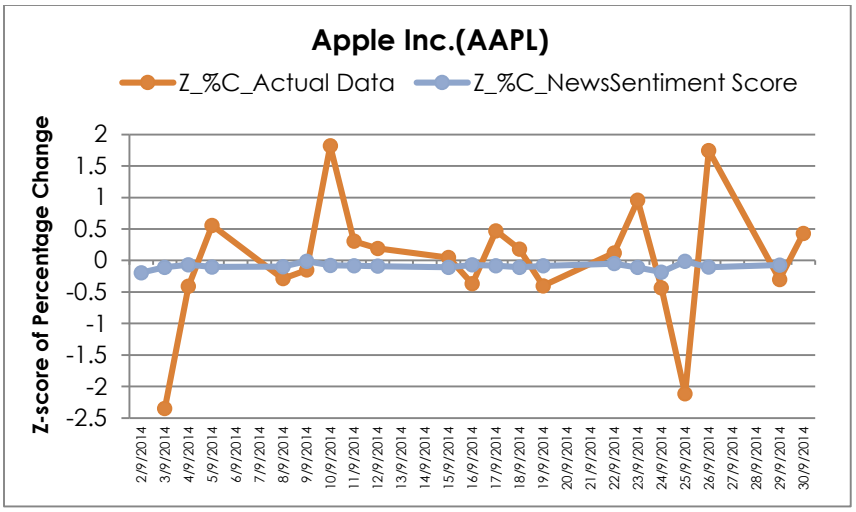


Figure 5 Shows the comparison of the Headline's NewsSentiment Score and Apple Inc's stock price movement (Once calculating normalization of percentage change)

According to Figure 5, the normalization of percentage change the Headline's NewsSentiment Score and Apple Inc's stock price movement starts in the negative level. In addition. There remains small amount of number to indicate the differences between the two variables. However, the trends of these two variables are relatively in the same direction.

Figures 1-5 shows the correlation between the trends of the Headline's NewsSentiment Score and Apple Inc's stock price movements. It obviously shows that when the trend of Headline's NewsSentiment goes up, the trends of Apple Inc's stock price movement also increases. In order to elaborate more, the Pearson's correlation and Spearman's rank have been applied to test the statistical relationship between the two variables between the trends of the Headline's NewsSentiment Score and Apple Inc's stock price movement's. Table 1 shows the results analysed by the Pearson's correlation and the Spearman's rank which support the visual correlation.

Table 1 Correlations table

	Techniques	Headline's NewsSentiment Score			
		Pearson Correlation	Sig. (2-tailed)	Spearman's rho	Sig. (2-tailed)
Apple Inc's stock price movement	1-day lap	0.285	0.223	0.221	0.350
	Standardised	0.285	0.223	0.221	0.350
	Trend Change	0.478*	0.039	0.478*	0.039
	Percentage change	0.408	0.083	0.372	0.117
	Standardised of Percentage change	-0.186	0.445	-0.149	0.542

*. Correlation is significant at the 0.05 level (2-tailed).

5.0 CONCLUSION AND FUTURE WORKS

The significance of this paper is to present the initial results of this study and to highlight the similarities of trend movement between Headline's NewsSentiment score and Apple Inc's stock price movement. The existence of visual correlation shows that Headline's NewsSentiment score can potentially be applied for the prediction of future stock market prices. The correlation between the two variables is not statistically significance may be due to the small number of sample size used. More elaborately, since only one Headline's NewsSentiment score can be calculated per day; up to now 21 samples have been collected. Due to the constraints of the R program, in the data collection procedure, only 20 Headline News, related to stock index, can be collected per day. As J. Hair *et al*. [15] recommend that sample size

is supposed to be larger than 200; hence, future researches may need to consider collecting a larger number of sample size to discover whether or not the results would be varied. The samples in this research are randomly selected from different news websites and sources. Nevertheless, the news has been categorized into various groups such as politics, economics, social, crime, sports, and so forth. More specifically, we are going to classify the news into different categories to match the macroeconomic factors influencing stock market so that this would help to increase the level of accuracy of the forecasting. Furthermore, accuracy can also be improved by collecting stock market terms into the tagging scheme. This scheme can be used along with other techniques to provide a very strong indicator of stock market movements.

Acknowledgement

This research was supported by the ministry of higher education (MOHE) and research management centre (RMC) at Universiti Teknologi Malaysia (UTM) under research university grant category

References

[1] P. Chang, C. Fan, and C. Liu. 2009. Integrating a Piecewise Linear Representation Method and a Neural Network Model for Stock Trading Points Prediction. *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.)*. 39(1): 80-92.

[2] R. D. Edwards, J. Magee, and W. H. C. Bassetti. 2012. *Technical Analysis of Stock Trends*. 9th ed. CRC Press.

[3] R. J. Bauer and J. R. Dahlquist. 1999. *Technical Markets Indicators: Analysis & Performance*.vol. 64. John Wiley & Sons,

[4] C. D. Kirkpatrick II and J. Dahlquist. 2010. *Technical Analysis: The Complete Resource for Financial Market Technicians*. FT Press.

[5] R. Goonatilake and S. Herath. 2007. The Volatility of the Stock Market and News. *Int. Res. J. Financ. Econ*. 3(11): 53-65.

[6] N. Godbole, M. Srinivasaiah, and S. Skiena. 2007. Large-Scale Sentiment Analysis for News and Blogs. *ICWSM*. 7.

[7] J. Leskovec, L. Backstrom, and J. Kleinberg. 2009. Meme-tracking and the Dynamics of the News Cycle. In Proceedings of the 15th ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*. 497-506.

[8] W.-B. Yu, B.-R. Lea, and B. Guruswamy. 2007. A Theoretic Framework Integrating Text Mining and Energy Demand Forecasting. *IJEBM*. 5(3): 211-224.

[9] S. Theußl, I. Feinerer, and K. Hornik. 2009. Distributed Text Mining with tm. In *The R User Conference*.

[10] P. Hofmarcher, S. Theussl, and K. Hornik, 2011. Do Media Sentiments Reflect Economic Indices. *Chinese Bus. Rev*. 10(7): 487-492.

[11] A. Nagar and M. Hahsler. 2012. Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams. In *2012 International Conference on Computer Technology and Science*. 47(Iccts): 91-95.

[12] I. Feinerer. 2014. Introduction to the tm Package Text Mining in R," nd) n. pag. Web. 1-8.

[13] M. Annau. Package 'tm.plugin.webmining': Retrieve structured, textual data from various web sources. 2014. [Online]. Available: http://cran.r-project.org/web/packages/tm.plugin.webmining/tm.plugin.webmining.pdf.

[14] M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04*. 168-177.

[15] J. Hair, R. Anderson, R. Tatham, and W. Black. 1995. Multivariate Data Analysis. 4th Edition with readings. New Jersey: Prentice Hall.