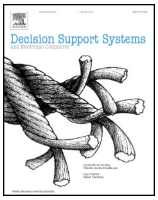
Decision Support Systems 61 (2014) 93–105

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Media-aware quantitative trading based on public Web information

Qing Li [a], Tiejun Wang [a], Qixu Gong [a], Yuanzhu Chen [b], Zhangxi Lin [c], Sa-kwang Song [d,*]

[a] *Southwestern University of Finance and Economics, China*
[b] *Memorial University of Newfoundland, Canada*
[c] *Texas Tech University, USA*
[d] *Korea Institute of Science and Technology Information, Republic of Korea*

ARTICLE INFO

Article history:
Received 10 January 2013
Received in revised form 27 December 2013
Accepted 30 January 2014
Available online 8 February 2014

Keywords:
Stock market
Sentiment analysis
Web information
Trading strategy

ABSTRACT

Recent studies in behavioral finance discover that emotional impulses of stock investors affect stock prices. The challenge lies in how to quantify such sentiment to predict stock market movements. In this article, we propose a media-aware quantitative trading strategy utilizing sentiment information of Web media. This is achieved by capturing public mood from interactive behaviors of investors in social media and studying the impact of firm-specific news sentiment on stocks along with such public mood. Our experiments on the CSI 100 stocks during a three-month period show that a predictive performance in closeness to the actual future stock price is 0.612 in terms of root mean squared error, the same direction of price movement as the future price is 55.08%, and a simulation trading return is up to 166.11%.

1. Introduction

Since Keynes put forward the concept of "animal spirits" in the 1930s, a large number of researchers have explored to understand the determinants of movements in stock market prices [20]. In traditional financial theory, a stock price is always driven by "unemotional" investors to equal the firm's rational present value of expected future cash flows. In other words, investors of the stock market are "rational" and they efficiently respond to new information regarding the stock market products. Investors' decisions in the market fully reflect the effects of any information revealed. This Efficient Market Hypothesis (EMH) contains three different levels of information sharing: the weak form, the semi-strong form, and the strong form [14]. Inspired by the weak form of EMH, numerous attempts have been made to measure movements of stock markets using quantitative information related to firm fundamentals [2,14,26]. Unfortunately, some studies show that substantial stock market movements cannot be captured ideally by the quantitative measures of firms' fundamentals [9,19]. This is because the actual market is not as efficient as explained by the EMH and investors are inevitably emotional.

In contrast, recent studies in behavioral finance show that emotion does influence investment decisions. In particular, investors are not as aggressive in forcing prices according to fundamentals as traditional financial theories would suggest [34], and a belief about future cash flows and investment risks is not only justified by the cold facts at hand [11]. These findings are consistent with examples such as the frustrating news about Steve Jobs' health caused the lower stock return of Apple, Inc. in 2003 and 2009, although the fundamentals of the firm were healthy. In factuality, investment decisions can be affected by emotional impulses of investors. Thus, we hypothesize that the tones of newly-released news articles influence investors to buy or sell. It would be interesting to quantify such financial news sentiment to investigate its impact on stock markets.

With the technological advancement fertilizing vibrant creation, sharing, and collaboration among Web users, the impact of media on stock markets has been increasingly prominent. Specifically, traditional news has evolved into various forms of social media including blogs, tweets/micro-blogs, discussion boards, and social news. With such broad communication channels, investors can rapidly reach more valuable and timely information. Meanwhile, the adaption of user engagement in social media effectively magnifies the information in the news via comments, votes, and so forth. With such rapid information influx, decisions of investors tend to be influenced by emotion of peers and the public. This may well lead to a herd behavior in investment.

In this article, we propose and implement a media-aware trading strategy to study the impacts of Web information on stock markets. It has the following unique features that have not been attempted in previous work.

- It is the first attempt to study the combined effect of Web news and social media on stock markets, particularly at the individual stock level. Our experiments on the CSI 100 stocks show that the amplification of social media on Web news contributes to the performance of stock predictions.
- To successfully measure news sentiment and capture public moods on investment, we propose an innovative approach to automatically

[*] Corresponding author.
E-mail address: esmallj@kisti.re.kr (S. Song).

0167-9236/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2014.01.013

- extract finance-oriented sentiment words from the Web to construct a financial sentiment dictionary.

The rest of this article is organized as follows. We first briefly describe related research in Section 2. The design details for our media-aware quantitative trading strategy are presented in Section 3. We then implement a trader with such principles and test its performance using real stock data from the Shanghai Stock Exchange and the Shenzhen Stock Exchange (Section 4). This paper is concluded with speculation on how the current prototype can be further improved in Section 5.

2. Related work

Observing fluctuations of stock prices with news feeds, some economists have explored the power of verbal information on stock markets. The research can be traced back to the work of Cutler, Poterba, and Summers in 1989 [9], which on one hand showed that there was no direct link between news and stock returns at that time. On the other hand, later evidence for the impacts of news on stock markets has been discovered in current markets. For instances, Das and Chen [10] extracted investor sentiment from stock message boards, and found that it was related to stock index, volumes and volatility. This work is more about finding such links than providing prediction. Veronesi [39] showed in theory that stock prices overreacted to bad news in good times and underreacted to good news in bad times using a rational expectations equilibrium model of asset prices. Chan [7] empirically examined monthly stock returns following public news and found that stocks with bad public news displayed a negative drift for up to 12 months and less drift for stocks with good news. Tetlock et al. [35,36] suggested that news itself had limited and short-lived predictive power on future stock prices. The fraction of negative words in firm-specific news stories can be used to forecast low firm earnings. Vega [38] further verified the media influence by revealing that stocks associated with private information experience low or insignificant drift while stocks associated with publicly-available news experience significant drift.

All of these researchers focus on the influence of media on stock markets. However, it is questionable whether any of these methods is efficient to quantify the media influence for econometric analysis. Two general approaches are adopted in these studies. One is to treat the number of firm-specific news feeds as an independent variable [7,38], the other is to compute a sentiment indicator based upon the percentage of the positive or the negative words in an article [35,36]. Both approaches, to a degree, capture the linguistic power of news, and consequentially weaken or even distort the impact of news on stock markets.

While economists are uncovering the relationship between media and stocks, computer scientists working in the areas of artificial intelligence and natural language processing are taking a step further by studying how various media-aware traders make profits in stock markets. A pilot study by Wüthrich et al. [45] attempted to forecast the trends of five major stock market indexes in terms of news articles. Later on, several studies investigated the predictive power of news on a single stock. For instance, Lavrenko et al. [21] proposed the e-analyst system based on the Relevance Language Model (RLM) to associate stock price trends with news stories. Mittermayer [25] developed the NewsCATS system to predict stock price trends for the time immediately after the publication of press releases. Fung, Yu and Lam [15] captured the relationships between Reuters Market news and 33 stocks listed in the Hang Seng Index.

One unique contribution of their work is the consideration of the inter-relationship among different stocks, i.e., it selected and assigned relevant news of other similar firms to the target firm while training the predictive model. Instead of predicting trends in prior research, Schumaker and Chen [30–32] proposed the AZFinText system based on support vector regression (SVR) to predict the +20 minute stock price after a news article was released (i.e., the price with a 20-minute delay). Wang et al. [40] proposed a hybrid stock prediction approach by combining ARIMA and SVR together.

The strategies of this research follow a common paradigm. It starts off with representing a news article as a weighted vector of terms and builds a predictive model to capture the relationship between news and stocks. The stock price trends are estimated with this model on the new arrival of a firm-specific news article. How to quantify a news article is crucial and requires a thoughtful design. Schumaker and Chen [31] experimented on several textual representation approaches, including bag of words, noun phrases, proper nouns, and name entities, and found that representing news with proper nouns is the most efficient. Here, we argue that sentiment words, particularly finance-specific sentiment words, are also valuable to represent a news article since investors are subject to sentiment as observed in behavioral financial research. A crowd's tendency can strengthen or weaken the influence of a news article on investors. Knowing the public mood on a stock is helpful to forecast its price movements. In this article, we capture the public mood on a certain stock from the relevant discussion threads in social media.

Table 1 summarizes previous research on stock analysis using textual information. For each study, the table summarizes its focus, media sources, analysis models, and experiment settings. It also highlights how the research relates to the issues that we aim to address in this article, i.e. quantifying the influence of public moods and news to predict stock market movements.

3. System framework

In this study, we implement a Media-Aware Quantitative Trader, dubbed MAQT, to examine the effectiveness of the proposed trading strategy. The framework of MAQT is sketched in Fig. 1. Essentially, it is a news-driven trading system incorporating public mood. It first represents the firm-specific news articles as weighted vectors of terms, and then captures the investors' sentiment on this stock by analyzing the relevant postings in financial discussion boards. Such information along with stock quotes is fed into the predictive model for training. The stock selection engine selects a number of stocks to predict their future prices using the trained predictive model.

3.1. Media quantification

In this section, we first describe how to represent news articles to quantify their influence on stock markets, and then present the approach to capture the public mood on investing a single stock. To assist sentiment analysis in such media quantification methods, an innovative approach to automatically extract sentiment words in financial domain from the Web is proposed at the end of this section.

3.1.1. Representation of news articles

The basic idea to represent an article in a machine-friendly form is to transform it into a term vector, where each entry is a weighted term in the article. Such a textual representation is called *bag of words* model, which has been used widely in Computer Science areas of natural language processing and information retrieval. Since this approach includes almost all the words [1] in a document, it may not scale very well. Therefore, we use the “important” words in an article to represent it. Essentially, the influence of news articles on stocks originates in two facets:

- *Event*: people tend to adjust their investment strategies if a latest news article conveys some aspects of firms' fundamentals to enrich their knowledge.

[1] A list of so-called stop-words including “the”, “of”, and “at” are removed because they are semantically empty.

Table 1
Comparison of previous research on stock analysis using textual information.

Approach	Focus				Media			Model			Experiment				
	Target	Scale	Source	Market	Text	Emotion	Public mood	Input	Output	Method	Period	Data size	Training & test	Metrics	Cost
Chan [7]	Stock return	Month	DJIPL	NYSE, AMEX, S&P500	No	No	No	News number, economic data	Abnor. return	Regression model	1980–2000	4200 stocks	No	Real value	No
Tetlock et al. [36]	Stock return	Day	DJNS	S&P500	Emotion word	General emotion	No	Word number, economic data	Abnor. return	Regression model	1980–2004	500 stocks	No	Real value	No
Wuthrich et al. [45]	Index	Day	Web news	Five major indexes	Selected words	No	No	News	Index	KNN, regression model	12/06/1997–03/06/1997	No	Three months rolling	Trend	No
Lavrenko et al. [21]	Stock Price	Hour	Biz Yahoo	No	All words	No	No	News	Stock price	Language model	10/15/1999–02/10/2000	127 stocks	90/40 days	Trend	No
Mittermayer and Knolmayer [25]	Stock trend	Minute	PRNews wire	S&P500	Selected words	No	No	News	Trend	KNN, SVM	04/01/2002–12/31/2002	500 stocks	90%/10% stocks	Trend	No
Bollen, Pepe, and Mao [6]	Index	Day	Twitter	NYSE	No	General emotion	Yes	Past DJIA, mood	Present DJIA	SOFNN	02/28/2008–12/19/2008		10/1 months	Real value	No
Wang, Huang, and Wang [40]	Stock price	Quarter	Financial report	No	All words	No	No	News, economic data	Stock price	ARIMA, SVR	1994–2010	6 stocks	20/5 stocks	Real value	No
Schumaker et al. [32]	Stock price	Minute	PRNews Wire	S&P500	Proper nouns	General emotion	No	News, stock price	Future stock price	SVR	10/26/2005–11/28/2005	500 stocks	4/1 weeks	Real value, trend	No
This approach	Stock price	Minute, day, week	Web news	CSI100	Proper nouns, emotion words	Financial emotion	Yes	News, stock price, public mood	Future stock price	SVR	01/01/2011–12/31/2011	100 stocks	9/3 months	Real value, trend	No

- *Emotion*: emotional investors can be affected by optimistic or pessimistic news.

Therefore, we model a news article as a weighted term vector, *V*, with a number of nouns and sentiment terms selected from the article. We believe that the important concepts of firms' fundamentals in a news article can be captured by a set of nouns, and news sentiment is reflected by a set of sentiment terms. Noun detection is a relatively mature technique in natural language processing. Here, we adopt a standard part-of-speech (POS) tagger to extract nouns from news articles. The challenge of sentiment term detection comes from domain-specific sentiment analysis. There are various studies on the detection of general-domain sentiment terms, but they are not applicable here. As per Loughran and McDonald [24], about three-fourths (73.8%) of the negative word counts in the open-domain emotion word list of Harvard-IV-4 [2] are not considered negative in a financial context. Apparently, building a comprehensive finance-specific sentiment word list is of great necessity. We defer the description of how to extract finance-specific sentiment terms to Section 3.1.3.

After extracting nouns and sentiment terms to represent an article as a term vector, the weight of each term indicating its topic importance is measured using the standard TF/IDF weighting schema [4,23].

3.1.2. Detection of public mood

As social creatures, people are generally affected by others in decision making. This phenomenon is particularly serious in the stock investors who work with great expectations and in high pressure [27]. An example is the research on predicting stock movements using Twitter mood [6].

In this study, we analyze firm-specific messages in financial discussion boards to capture public mood in investing stocks. To this goal, we applied our focused Web crawler to download postings from the financial discussion boards of www.sina.com and www.eastmoney.com, two most popular stock discussion forums in China. These data sources provide a solid basis for capturing public mood as both websites each have over 20 million independent visitors per day, and they produce a large number of postings and votes. In addition, each traded firm has its own discussion section on both websites. It is easy to capture the crowd's tendency to invest a single stock with this data source.

Since the influence of public sentiment is waning but last for several days [36], such continuing impact should be considered in quantification. Here, we measure the crowd's mood on a stock from two aspects, i.e., optimism and pessimism. The optimistic mood on a stock, *s*, on a daily basis is measured as,

$$M_s^+ = \sum_{i=0}^{\tau} \frac{P_i}{L_i} \times e^{-i/\beta}, \tag{1}$$

where P_i is the number of the positive words in the discussion threads about stock *s* on the *i*-th day after the news breaks, and L_i is the total number of the words in these discussions. τ is the number of the past days that we consider their sentiment continuing influences, and β is a constant to tune the scale of time attenuation, set to 20 to simulate the attenuation for the number of work days in a month. Similarly, the pessimistic mood of a stock *s* is measured as

$$M_s^- = \sum_{i=0}^{\tau} \frac{N_i}{L_i} \times e^{-i/\beta}, \tag{2}$$

where N_i is the number of the negative sentiment words in the firm-specific discussion threads on the *i*-th day.

[2] http://www.wjh.harvard.edu/~inquirer/homecat.htm.

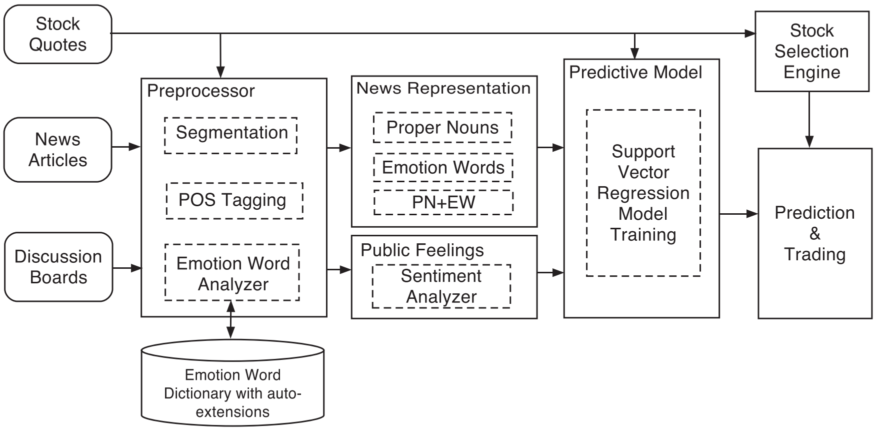


Fig. 1. Design scheme.

3.1.3. Sentiment word detection

Both news representation and public mood detection rely on the technique of judging the sentiment-polarity of a word. A related problem of this issue is sentiment analysis. It maps a given piece of text, such as a document, sentence, or lexicon, to a label drawn from a pre-specified finite set using various supervised or unsupervised machine learning techniques including SVM, Naïve Bayesian Networks and Maximum Entropy [12,28]. Most previous research focuses on general-domain opinion analysis. Here, the challenge lies in domain-specific sentiment analysis. The general sentiment word categorization cannot translate effectively into a discipline with its own dialect [1,24]. Specifically here, an emotionless word in the context of finance can express strong sentiment or a typical emotion word is unemotional in the realm of finance. For instance, the word "bull" originally refers to a male bovine animal but indicates good earning returns in finance domain such as "bull stock". Some typical emotion words, such as "crude", "tire", or "capital" are more likely to identify a specific industry segment in financial events than expressing a negative sentiment as suggested by Harvard-IV-4.

A few researchers study the domain-specific sentiment word extraction [1,8]. The basic idea of these studies is to identify domain-specific words in terms of their statistical associations with domain-specific texts that are ready labeled as positive or negative. In this article, we take a step further by using context information of stock markets to extract domain-specific sentiment words. Specifically, finance-specific sentiment words are extracted based on two hypotheses.

- A word is characterized by the immediate context it appears, i.e., the semantic orientation of a word tends to correspond to the semantic orientation of its neighbors in the context [37].
- A firm-specific article with a positive (negative) tone is typically accompanied by the rising (falling) price trend of relevant stocks. This hypothesis is further confirmed by the findings of Tetlock regarding the interaction between stock markets and daily news, particularly bad news [35].

Therefore, we calculate the joint conditional probability of a word with these two hypotheses as follows and select the words with high probabilities. In particular, the positive probability of word w is denoted as

$$\begin{aligned} P^{+}(w) &= P(w|E=+,T=\uparrow) \\ &\approx P(w|T=\uparrow)P(E=+|w,T=\uparrow), \end{aligned} \tag{3}$$

where E denotes the semantic orientation of its neighbors with two values, $+$ and $-$, representing positive and negative emotion, respectively. T denotes the stock price trend with two values, ↑ and ↓, indicating the upward and downward price trends, respectively.

$P(w|T=\uparrow)$ can be simply estimated as

$$P(w|T=\uparrow) = \frac{U(w)}{N(w)}, \tag{4}$$

where $U(w)$ is the number of the documents tagged with upward stock trend containing word w in the training corpus. $N(w)$ is the total number of the documents containing word w in the training corpus.

$P(E=+|w,T=\uparrow)$ can be estimated with the statistical information obtained from the finance-specific sentiment word set [3] and the training corpus where each article is associated with a price trend. Specifically,

$$\begin{aligned} P(E=+|w,T=\uparrow) &= \sum_{i=0}^{M} P(e_i=+|w,T=\uparrow) \\ &\approx \sum_{i=0}^{M} I^{+}(w,e_i), \end{aligned} \tag{5}$$

where e_i is the sentiment word in the paradigm set S, and M is the total number of the positive words in this set. $I^{+}(w, e_i)$ is the statistical association between word w and positive word e_i in the articles with an upward trend which is measured by χ^2. The estimation of $P(w|T=\downarrow)$ and $P(E=-|w,T=\downarrow)$ is quite similar to the description above with the difference that the positive paradigm words are replaced with the negative paradigm words, and the selected documents are associated with a downward trend rather than an upward trend.

3.1.4. Stock trend

To calculate the statistical information for sentiment word detection, it requires to associate firm-specific news with a downward or upward stock trend. Here, the challenge lies in discovering trends from the time series of stock price. There are approaches for trend discovery.

One is based on curve segmentation analysis. The basic idea is to plot (or approximate) the numerical data into a curve and then segment it into a series of straight lines. For instance, Lavrenko et al. [21] segmented price curves into increasing and decreasing trends for stock prediction. Fung, Yu and Lam [15] adopted curve segmentation to identify the "drop", "steady", and "rise" trends of stocks.

The other approach is based on a moving average theory. Moving average is a technique to smooth a series of data points, typically for observing them at a coarser granularity. It is often used in science, engineering, finance, and other domains as a sort of high-pass filter or historical summary. The general approach is to replace a data point of

[3] Here, we chose the Chinese version of Loughran and McDonald Financial Sentiment Dictionary (http://www.nd.edu/~mcdonald/) as our initial paradigm set. More entries detected by the proposed approach are added as new finance-specific sentiment words.

the series with an average of those “nearby”. Depending on the specific variant of the moving average, the size, i.e. the number of data points taken the average over, and position, i.e. one- or two-sided, of the window, along with how weights are assigned to the data points, define how the average is calculated. The simplest case would be an unweighted average over an equal, finite number of data points on either side.

In contrast, a one-sided moving average is often used to summarize real-time data series at the current point of time, where the future remains unknown yet. This particular form of moving average is called simple moving average, and is used extensively in stock analysis [18].

Since the purpose of trend discovery here is to tag a news article with a stock trend in terms of the media influence after the press release, we adopt the simple moving average approach which accommodates such time-lag effect. The intensity of time lag is determined by the window size n in moving average. We defer the setting of this window size n to Section 4.3.

3.1.5. Predictive model

In this study, the function of our predictive model is to capture the relationship between finance indicators and future stock prices. These finance indicators include firm-specific news articles, public mood, and the stock prices at the point of releasing these news articles. The public mood provides a measurement for the recent investment atmosphere, and a firm-specific news article conveys the information of firm's fundamentals and the attitudes of domain experts. These indicators allow us to explore their combined effect on stock movements.

There are variety of machine learning methods for stock market prediction including Relevance Language Model (RLM) [21], Support Vector Machine (SVM) [25] and Naïve Bayesian [33]. However, all of these works focus on the directional movements rather than numerical stock prices. In this study, we adopt the extended SVM, i.e., SVR (Support Vector Regression) Model, which applies a regression technique to SVM to predict the numerical values of future stock prices [31].

4. Experimental evaluation

The ultimate goal of this study is to examine the combined influence of news articles and public moods on stock market movements. We are of particular interest in the following research questions:

- Does the quantitative information of news articles have the ability to impact stock markets? If this is the case, it proves that public information events are subject to differential interpretations by investors. This presents profitable trading opportunities for skilled investors and, therefore, the trades of informed investors should be more profitable with news releases. Otherwise, the public information would reduce asymmetric information and the trades of informed investors should be less profitable [13].
- Do investors react to the sentiment of news and other people? If so, it provides a concrete evidence to support a critical hypothesis in behavioral finance that investor sentiment affects stock prices [5].

Here, we target on the stock markets in Mainland China. There are two independent stock exchanges, i.e., Shanghai Stock Exchange (SSE) and Shenzhen Stock Exchange (SZSE). Prior relevant studies [30,31,35,36] are mainly on stock exchanges in the United States, especially, New York Stock Exchange (NYSE). Due to lack of market makers in Chinese markets, this study on SSE and SZSE provides a unique insight in the relationship of the media and the stock market without the interference from the trades of market makers.

4.1. Experimental settings

Three databases were constructed and maintained for our experimental study and further peer study.

- *Financial News Corpus*: This corpus contains 124, 470 financial news articles related to 100 companies listed in China Securities Index (CSI 100).[4] This corpus was constructed by querying the Baidu and Google search engines with their advanced search functions specifying date range and websites to download news articles with company common names, abbreviations, or stock number IDs. The collected news articles were released from January 1, 2011 to December 31, 2011 in the reputed Chinese financial websites including finance.sina.com.cn, finance.caixin.com, and finance.people.com.cn. We adopted bloom filters to detect and remove the duplicate news articles [17], and only kept the news articles with company name in titles which narrows the retrieved articles to a certain degree [36]. After removing the HTML tags of each news article, the news body and publication date were stored in the database.
- *Financial Discussion Board Corpus*: This corpus contains the discussion threads of the CSI 100 companies from January 1, 2011 to December 31, 2011 from two premier financial discussion boards in China, i.e., www.sina.com and www.eastmoney.com.
- *Stock Transaction Data*: This corpus contains the high-frequency financial data from January 1, 2011 to December 31, 2011 from the China Stock Market Database (CSMD).

It provides intraday transaction information including price, volume, and time at the granularity of second. In this research, we use the first 9-month of data in year 2011 as training corpus, and the last 3-month data for testing. Instead of using all 100 companies, we remove 11 companies due to the inconsistency in the CSI 100 list.[5] In this testing period, the percentage of the upward trend is 46.12%, the percentage of the downward trend is 49.53%, and the rest is still. The standard deviation of the stock prices in this testing period is 27.12.

To gauge how well the proposed trader, MAQT, performs in capturing stock movements, we choose two evaluation metrics as suggested by Schumaker and Chen [30]: directional accuracy and closeness. *Directional accuracy* measures the upward or downward direction of the predicted stock price compared to the actual movement direction of the stock price. Realizing the fact that it may be close in prediction yet predict a wrong movement direction, the *closeness metric* is used in complement to evaluate the difference between the predicted value and the real stock price in terms of Root Mean Squared Errors (RMSEs).

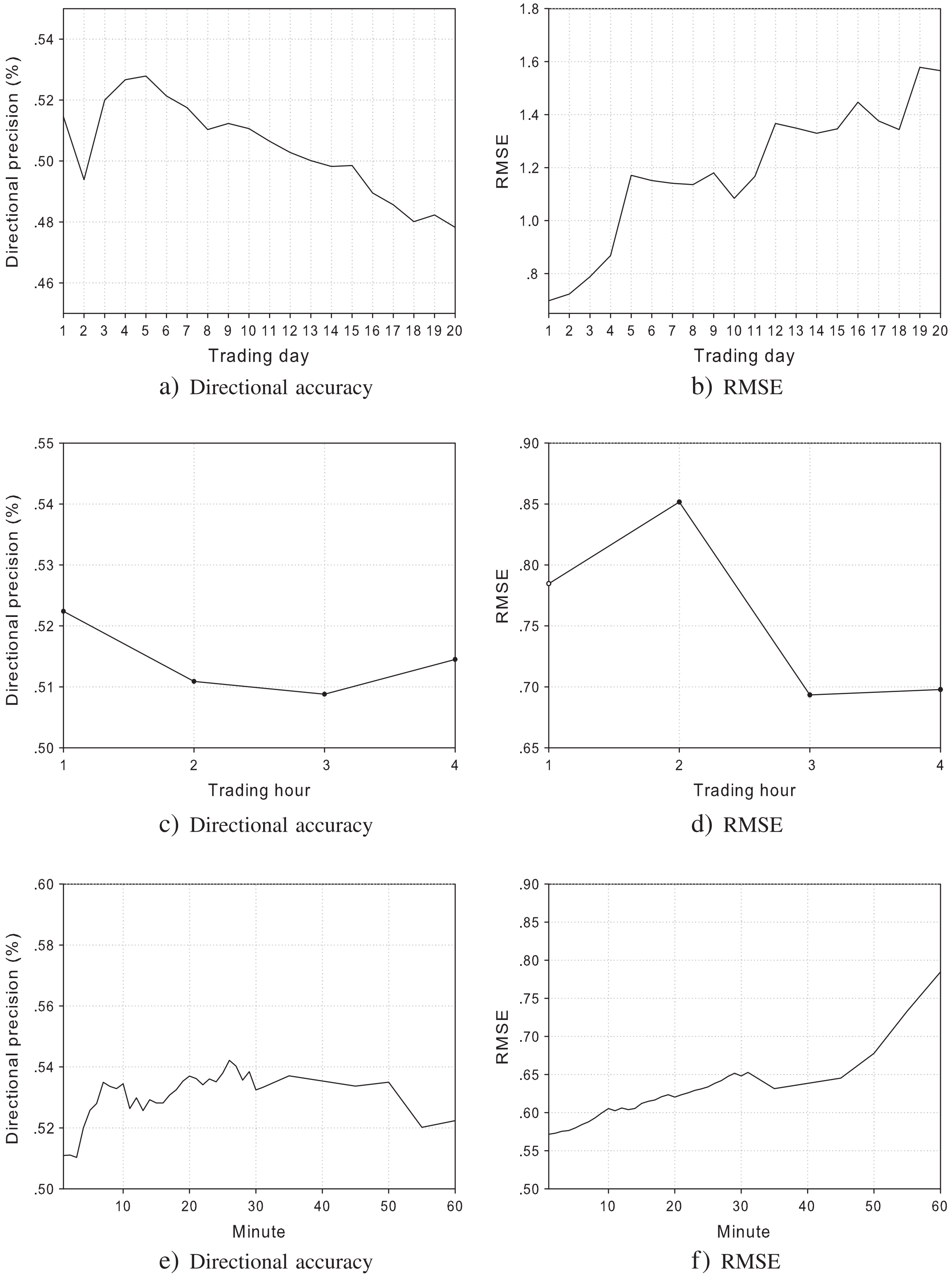
4.2. Time window of the quantitative trader

The predictive model of MAQT is to capture the hidden connections between the input (textual information, public mood, and current stock prices) and the output (future stock prices). Here, we are particularly interested in the outlook time window of the predictive model.

Gidofalvi [16] pointed out that a good outlook time window for stock forecast is about 20 min after the release of the relevant information. Subsequent research [30,31] adopted this window size in their stock prediction studies. However, Chan [7] showed that the influence of news on stock markets could last for several days or even months. To find a proper outlook window for MAQT, we conducted a series of experiments with different window sizes using the first 9-month data of year 2011 for training and the remaining 3-month data for evaluation. Fig. 2 shows the sensitivity results of different outlook window sizes

[4] CSI 100 consists of the largest 100 stocks in mainland China at this point of writing. CSI 100 aims to comprehensively reflect the price fluctuation and performance of the large and influential companies in Shanghai and Shenzhen securities markets.

[5] Since the CSI 100 list is adjusted every half a year, we only experiment on the companies listed in the entire year of 2011.



a) Directional accuracy

b) RMSE

c) Directional accuracy

d) RMSE

e) Directional accuracy

f) RMSE

Fig. 2. Predictive outlook window.

at the scale of minute, hour, and day for 89 listed companies on CSI 100. Note that SSE and SZSE run for 4 h every workday. The trade time is 9:30 to 11:30 am and 1:00 to 3:00 pm.

In Fig. 2(b), (d), (f), we can observe that the predicted price accuracy decreases with time at all outlook window scales. It decreases slowly from 0.5715 to 1.5658 within 20 days in terms of RMSE.

In Fig. 2(a), (c), and (e), the directional accuracy increases and achieves the best performance of 0.5421 at the 26th minute after the news release. It then turns less predicable although there is a small increase at the one-day closure of the markets. At the day level, the directional accuracy also increases and approaches the optimal performance of 0.5279 on the 5th day of the news release, and then becomes less predicable thereafter. It can be observed that there is a good predictive window of around 20 min after the news release.

This finding indeed agrees with the previous research statement that a lag exists between the time that information was introduced and when the stock market would correct itself to an equilibrium [22]. That is, the market could be forecast in short durations after introducing new information. However, it may take days or even longer for investors to fully absorb information as in Chan [7] since the predicted directional performance still keeps increasing for several days after the news release (Fig. 2(a)).

In addition, we determine the optimal predictive outlook window of 26 min for the following experiments in terms of directional accuracy rather than RMSE (Fig. 2(e)). This is because the performance of our investment strategies (Section 4.7) relies on the directional accuracy, i.e. the difference between the predicted future price and the price at the point of news release.

4.3. Emotion words extraction

Previous work [32,35,36] relies on domain-independent sentiment analysis techniques to capture the tone of an article. However, the emotion word list developed for psychology and sociology cannot function well in the realm of finance. To testify this hypothesis, Loughran and MacDonald manually analyzed annual reports of listed companies (Form 10-K) and extracted 354 positive and 2349 negative English words as finance-specific sentiment words [24]. In our study, we proposed an innovative approach to automatically construct a finance-specific sentiment word list and represent the emotions of financial articles with these sentiment words (Section 3.1.3). Specifically, the Chinese translations of the Loughran and McDonald Financial sentiment words were used as our initial paradigm set. More entries were extracted from the training corpus of news articles in terms of their connections with the sentiment words in the initial paradigm set.

A unique feature of our method is that it segments articles by stock trend (up or down) and uses this context-aware information for domain-specific sentiment word extraction. In this work, a simple moving average method is adopted to analyze stock trends (Section 3.1.4). It detects the current trend in terms of the unweighted mean of the previous *n* data points in the series of stock price. Therefore, we carried out a series of experiments to explore the sensitivity of this moving-average windows. Fig. 3 shows the predictive performance using the different emotion word sets that were constructed with the different moving-average window sizes from 1 to 30 days. It can be observed that the predictive performance is not quite sensitive to this window size once we set it to 1–5 days.

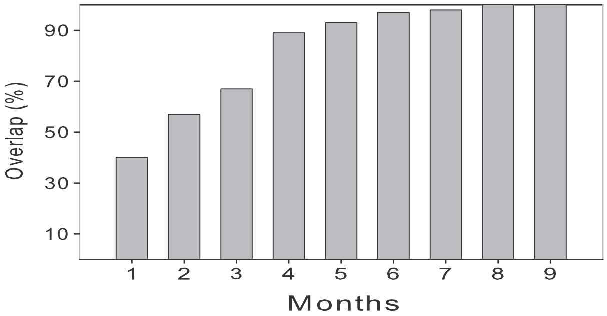


Fig. 4. Sensitivity analysis of sentiment words.

To understand the impact of the training corpus size on sentiment word extraction, we carried out a series of experiments with variable months of news articles released in the first 9 months of the year 2011. In Fig. 4, the *x*-axis depicts the number of the months from which the news articles are selected for constructing sentiment words, and the *y*-axis depicts the percentage overlap between sentiment words for the given number of months indicated and all 9 months of data. It can be observed that similar sentiment words are obtained when more than 8 months of data are utilized to build the sentiment dictionary. Here, we used 9 months of data to construct the finance-specific sentiment word list.

Table 2 is an example of these words and a detailed description of the finance-specific sentiment words can be found in the Online Supplement that accompanies this paper. To our knowledge, this is the largest and the most comprehensive Chinese sentiment word list in finance. The proposed approach is able to extract finance-specific sentiment words in different languages if only the training corpus in the corresponding language is provided.

4.4. News representation

The influence of news on investors comes from two sources, i.e., event and emotion. Therefore, we extracted proper nouns and sentiment words to represent a news article with the assumption that proper nouns reflect a firm's fundamentals and sentiment words convey the optimistic or pessimistic attitudes to news articles.

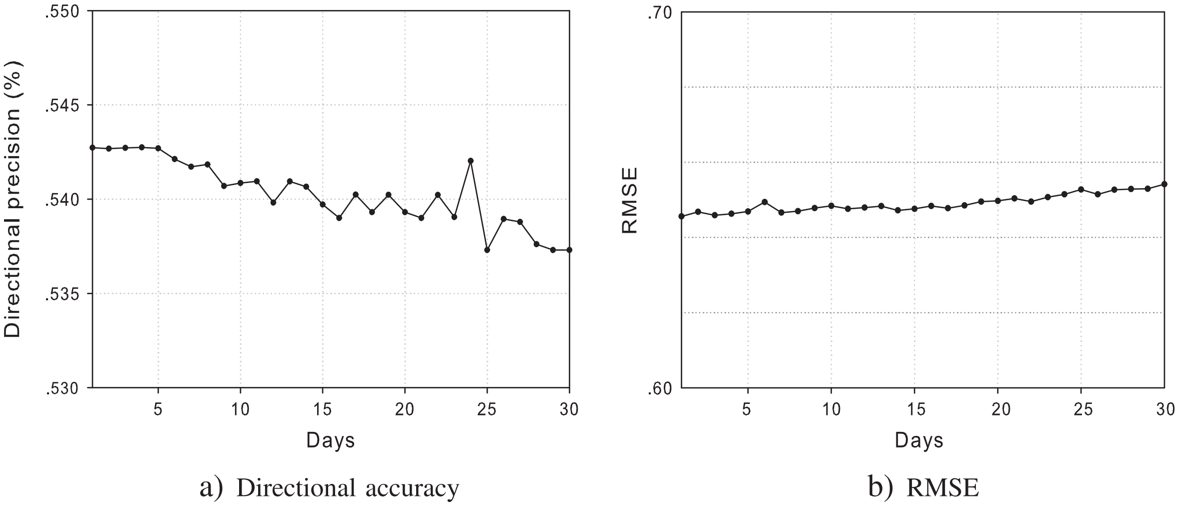


Fig. 3. Sensitivity of moving-average window.

Table 2
Examples of the finance-specific sentiment words.

Positive	Negative
安定 (stable)	套牢 (stuck)
富足 (abundant)	唾弃 (despicable)
蓬勃 (booming)	阻力 (resistance)
公道 (impartial)	下跌 (fall)
周到 (comprehensive)	流出 (exodus)
红火 (prosperous)	阴谋 (conspiracy)
精湛 (proficient)	更低 (lower)
如意 (satisfactory)	涉水 (wading)
昭著 (evident)	衰退 (recession)
强悍 (robust)	威胁 (threat)
高瞻远瞩 (visionary)	动荡 (turmoil)
划时代 (epochal)	低迷 (murky)
榜首 (top)	疲软 (weak)
透明度 (transparent)	破产 (bankrupt)
激增 (ballooned)	赤字 (deficit)
得体 (appropriate)	放松管制 (deregulate)
十全十美 (perfect)	付诸东流 (drain)
始终不渝 (steadfast)	违约 (default)
无可厚非 (legitimate)	惨淡 (gloomy)
家喻户晓 (renowned)	下滑 (slide)

Table 3
Representation.

Method	RMSE	Directional precision
PN	0.6385	54.21%
Harvard	0.6291	49.38%
FS	0.6204	51.36%
PN + FS	0.6121	55.08%

Here, we adopt FudanNLP, a state-of-the-art lexical analysis system for Chinese,[6] as our standard part-of-speech (POS) tagger to extract seven noun categories as our proper noun set from news articles. Sentiment words are detected based on our finance-specific sentiment word list obtained in the previous section.

In our experiments, we study several ways to represent a news article. The predictive performances of these representations are shown in Table 3. *NP* denotes that a news article is represented by a number of weighted proper nouns. *Harvard* is to represent an article by a number of weighted sentiment words from the Harvard-IV-4 list. *FS* means that an article is modeled as a number of weighted sentiment words from the finance-specific sentiment word list.

Before reporting our comparison results using directional precision and RMSE, we provide a significance test to show that the observed differences are not incidental as in Schumaker et al. [32]. Specifically, it takes a pair of equal-sized sets of predicted values on each article in the testing news corpus generated by the targeted and baseline methods, respectively, and assigns a confidence value to the null hypothesis that the values are drawn from the same distribution. If confidence in the hypothesis (reported as a *p*-value) is less than 5%, it typically means that the results of experiments are reliable and convincible.

As we can see, the representation based on finance-specific sentiment words can greatly improve the predictive performance as compared to those based on general sentiment words. In contrast, representing news articles with proper nouns can achieve a good directional prediction but a poor RMSE. In addition, we represent news articles with both finance-specific sentiment words and proper nouns. As shown in the last row of Table 3, this representation achieves the best performance among the four methods. These results are all significant as all targeted values versus baseline's values have *p*-values less than 0.05. Therefore, we favor representing news articles with both proper nouns and finance-specific sentiment words.

4.5. Industry sensitivity

To show the generalizability of the predictive performance over industry, we partitioned the stocks listed in CSI 100 into 6 industry sectors, i.e. manufacturing, finance, mining, transportation, real estate & construction, and service. Fig. 5 shows the predictive performance of each industry sector. It can be observed that these industry sectors are of various predictive performances. This might be caused by the different media coverage of each sector. Therefore, we carried out a series of experiments to show the generalizability over each industry sector. To find the minimum number of industry sectors needed for training, we took all *k*-combinations of six industries as the training data and carry out the predictive experiments $C(6,k)$ times. The industry generalizability is evaluated by the average performance of these combinations. In Fig. 6, the *y*-axis is the average predictive performance, and the *x*-axis is the number of industry sectors, *k*, used for training. We use all 6 industry sectors in our following experiments to achieve the best performance.

4.6. Stock price prediction based on news articles and public moods

In behavioral finance, investors are subject to sentiment in their decision-makings. Information and public mood are two important sources to affect the feelings of investors. In this study, we analyze the firm-specific messages in two premier financial discussion boards from www.sina.com and www.eastmoney.com to capture public mood of each single stock. We carried out a series of experiments to study the influences of such mood (M_s^+ and M_s^-) on stock movements. In particular, we built three models to study their influences. For each stock, Model 1 takes the current stock price, term vectors of relevant news articles, and optimistic public mood on this stock (M_s^+) as the input of the predictive model. Model 2 takes the current stock price, term vectors of news articles, and pessimistic public mood (M_s^-) as the input. In model 3, the input is the current stock price, term vectors of news articles, M_s^+, and M_s^-. The output of these models is the +26 minute stock price (i.e., the price with a 26-minute delay).

From Fig. 7, the pessimistic public mood has a significant contribution in predicting stock movements. Compared to a pessimistic attitude, an optimistic public mood has a limited power in sensing stock movements. These findings are consistent with the prior study that the fraction of negative words in firm-specific news stories forecasts low firm earnings [36]. The joint influence of pessimism and optimism in public mood is noticeable. It can be observed that the impact of the public mood lasts for several days. In our experiments, incorporating the public moods of the recent 3 to 6 days can further increase the stock predictive performance. With a time period less than 3 days or greater than 6 days, the accumulative mood is not sufficient to assist stock predictions.

4.7. Investment experiments

In this section, we describe investment simulation with the proposed MAQT. To gauge the performance of our trader, we compared it to three classic trading strategies, i.e. top-*N*, Simple Moving Average (SMA) [18], and adjusted AZFinText [31,32]. In our simulation, the initial investment budget is RMB10,000 (approximately USD1630), and the investment period is from October 10, 2011 to December 30, 2011,[7] during which the change of the CSI Index was down by 5.21% from 2363 to 2240. In Chinese stock markets, the average trading cost is about 0.2% of the invested value for a selling long transaction and 0.05% for a short selling transaction. Here, we take the assumption of zero transaction cost as previous work [7,21,32,40]. In fact, the

[6] FudanNLP is developed by Fudan University, and accessible at http://code.google.com/p/fudannlp/.

[7] The stock markets in mainland China were closed from October 1 to October 9 because of the National Day holiday and its neighboring weekends.

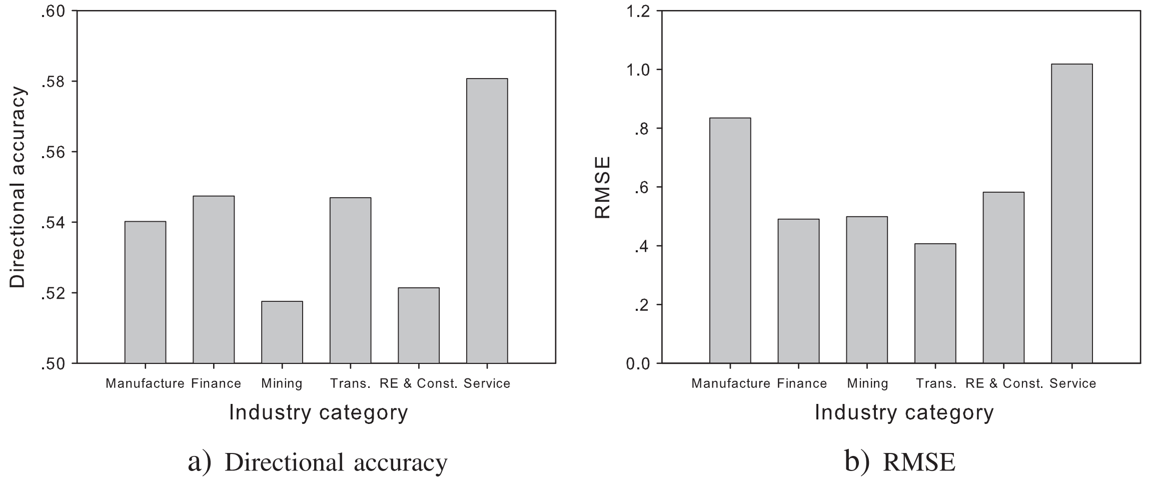


Fig. 5. Industry impact.

transaction costs are effectively absorbed by increasing the volume of each transaction, as long as we are making a profit.

In the top-*N* strategy, we invested in the top-*N* stocks that performed best over the period from January 1, 2011 to September 30, 2011 by buying each at the beginning of October 2011 and selling it at the end of the 3-month evaluation period. In Fig. 8(a), the *y*-axis denotes the RMB value of the trader's portfolio over the 3-month assessment time. The *x*-axis denotes the number of stocks, *N*, which we select to invest in accordance with the performance over the last 9 months in a descending order. When the trader's portfolio includes several stocks, the initial investment budget is equally invested in each stock. It can be observed that all of the top-*N* combinations lost in the assessment period, while the top-30 experienced the smallest loss. To better understand the top-*N* strategy in this period, we further looked into the stock return of each stock in the top-30 after three months with the RMB10,000 startup fund. Only 15 stocks were profitable after three months (Fig. 8(b)).

We adopted the SMA strategy as a benchmark [18]. Different from the top-*N* strategy, which focuses on the long term investment, this strategy is a relatively rapid trading method, which makes minute-by-minute trading decisions in terms of stock trends. Specifically, the investment is triggered when the actual market stock price crosses through the daily moving average of the same stock by a certain margin threshold (or penetration rate), either up or down. Specifically, if the change is upward, stocks should be purchased; if down, stocks should be sold. Note that the performance is sensitive to two parameters, i.e., the window size of the moving average and the penetration rate. Therefore, we experimented with different parameter settings, and found that the best performance was achieved with the moving average window size of 29 days (Fig. 9(a)). As shown in Fig. 9(b), the portfolio returns go up with the increase of the penetration rate, and the best penetration rate for triggering the transaction is over 1% of the invested value. However, according to Fig. 9(c), no investment transaction occurred when the penetration rate was over 1% of the invested value. That is, there is no positive earnings with SMA. As such, we conjecture that SMA is no longer suitable for current stock markets that are usually influenced by social media in an involved and timely fashion.

Different from previous two classic approaches, AZFinText [31] is a media-driven trader. It adopts a SVR model to capture the linkage between financial news and stock prices. To study the impact of news sentiment, Schumaker et al. apply a sentiment analysis tool named OpinionFinder to tell the (positive or negative) tone of a news article, and use such a binary sentiment result alongside the news textual vector as the input of the predictive model [32]. In this study, we use the adjusted AZFinText approach as a benchmark. Lack of Chinese language support in OpinionFinder, we implemented a Chinese sentiment analyzer following the sentiment analysis principles of OpinionFinder [43]. In particular, we took a two-step classification approach for

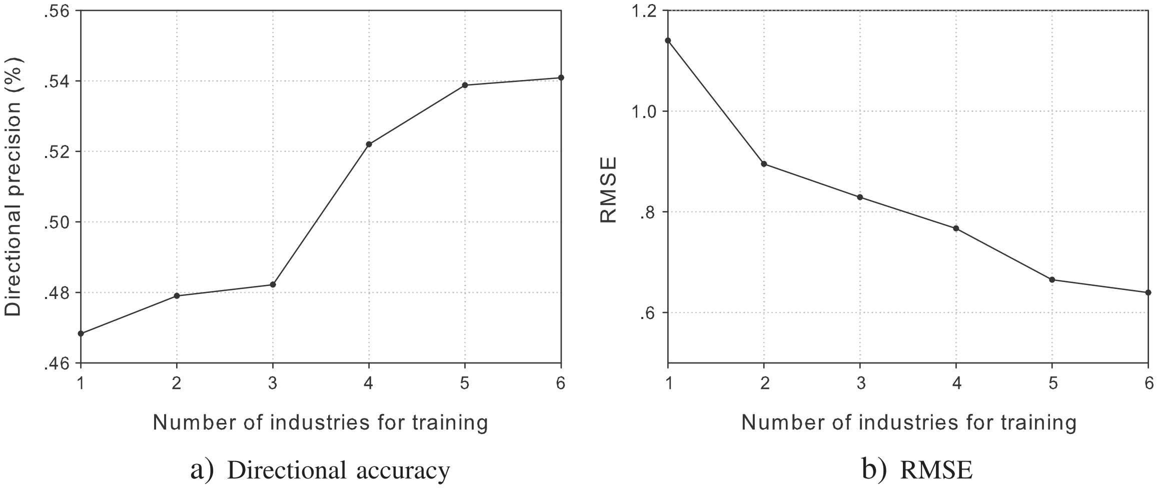


Fig. 6. Industry sensitivity.

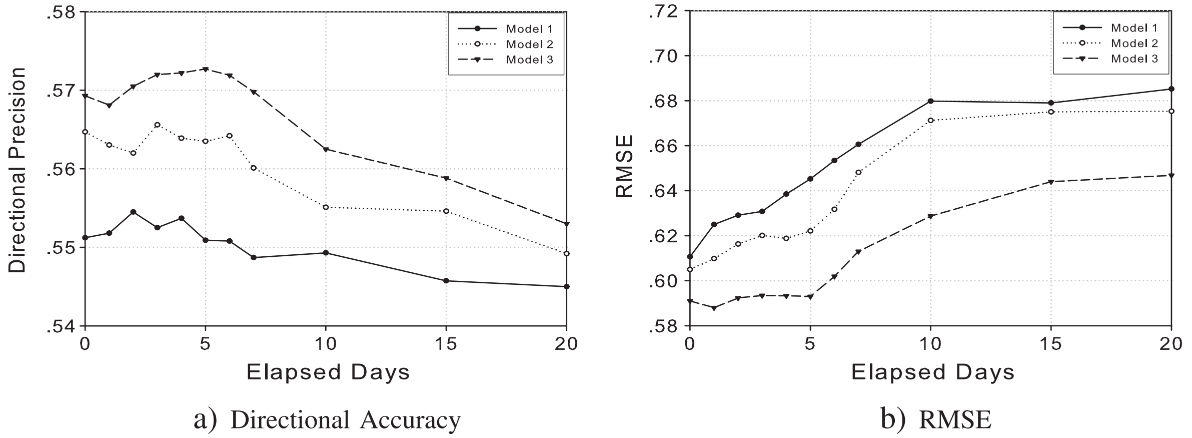


Fig. 7. Public mood.

contextual sentiment analysis as suggested by Wilson et al. [44]. The first classifier focuses on identifying sentiment expressions based on word, modification and sentence features, and the second classifier takes the sentiment expressions and identifies those that are positive and negative in terms of word and polarity features. Both classifiers rely on BoosTexter [29] for boosting. To label the objective patterns for calculating features, we adopted the unsupervised learning approach by Wiebe and Riloff [42] to create objective patterns with unannotated texts. Note that we ignored the document feature indicating topics suggested by Wilson et al., since all our focuses are on the same topic, i.e., economics.

To tune the parameters of the adjusted AZFinText, we first examine the prediction accuracy of the adjusted AZFinText with different outlook window sizes, and find the optimal window size +23 min (Fig. 10(a)). Then, we vary the threshold to trigger transactions within 3-month evaluation period, and find that the idealized returning performance is achieved when we set the threshold to 0.3% of the stock price (Fig. 10(b)).

The proposed MAQT takes a step further by utilizing both Web news and social media. It predicts +26 minute stock price of a firm whenever a firm-specific news is released. We take both short selling and long selling methods to obtain stock returns. In long selling, if the predicted future price is greater than a threshold at the time that article is released, our trader purchases the stock immediately and disposes of the stock in 26 min. The stock return is the difference between the selling and buying prices. In short selling, if the prediction price is less than a threshold at the time that article is released, we borrow the stocks and sell it expecting that it will be cheaper to repurchase 26 min later. The stock return is the stock price at the point of borrowing minus the repurchasing price. To determine a threshold for triggering short selling and selling long, we carried a series of experiments with different percentages of the stock price as the threshold. As shown in Fig. 11, the idealized optimal performance is achieved when we set the threshold to 0.3% of the stock price. (Here, the “oracle” guides the investor to make the maximum possible profit.) To better understand the performance of MAQT, we also provide the theoretical optimal performance of our trading strategy, i.e., the trader had known precisely in advance what the stock change would be when a news article is released. The total earnings for the optimal situation is RMB 2.8941×10^{14}.

Fig. 12 shows the daily stock returns of these four methods over the assessment period. Note that the top-*N* method focuses on the long-term return. Since the top-*N* method does not trade within the assessment time, the daily earning just shows its RMB value of its portfolio on that day. That is, it is the return if the investment period ends at that point. For other rapid trading strategies, each daily earnings is calculated based on the earning performance of the previous trading day. Both media-driven strategies show a great advantage over the two classic methods. Comparing the change of CSI 100 of −5.21% and 103.23% return of AZFinText, the proposed MAQT is quite promising yielding a return of 166.11% in three months. In addition, we used a randomization test [21] to determine if such earnings are statistically significant. Specifically, we conducted 1000 trials of an auto-trader which bought or shorted stocks randomly. Then we compared the earnings of MAQT to the distribution of cumulative earnings from the randomized trials. This auto-trader was set to buy and short particular stocks with the

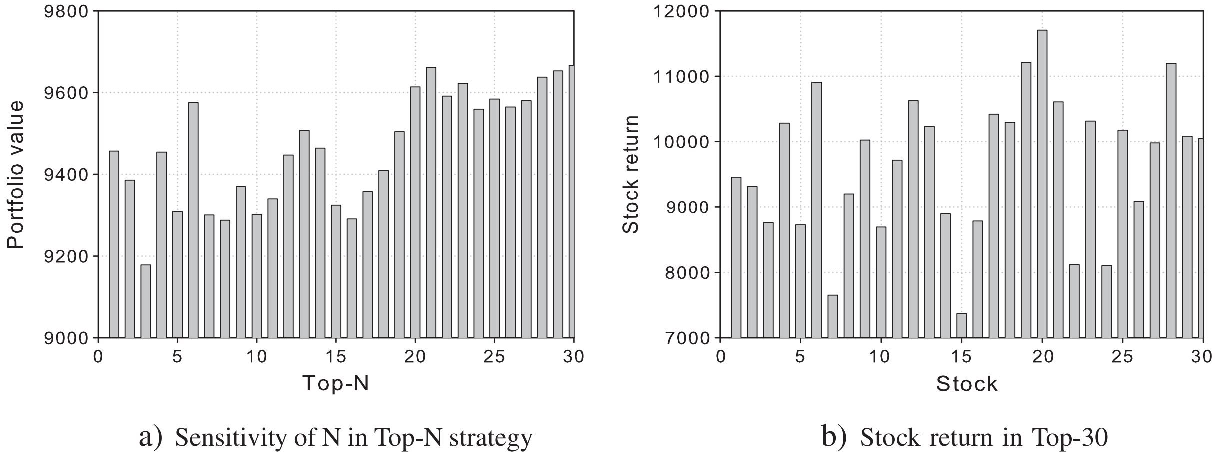
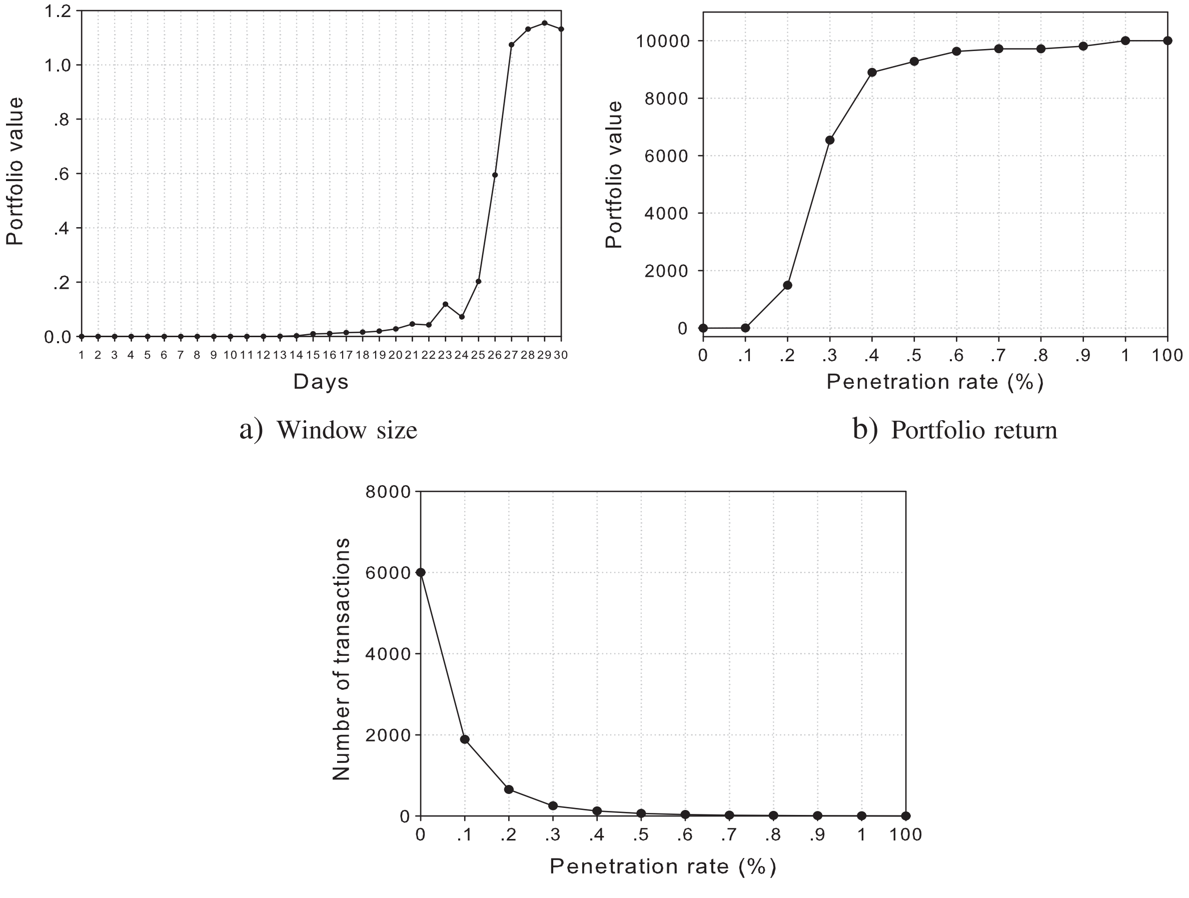


Fig. 8. Top-*N* strategy.



a) Window size

b) Portfolio return

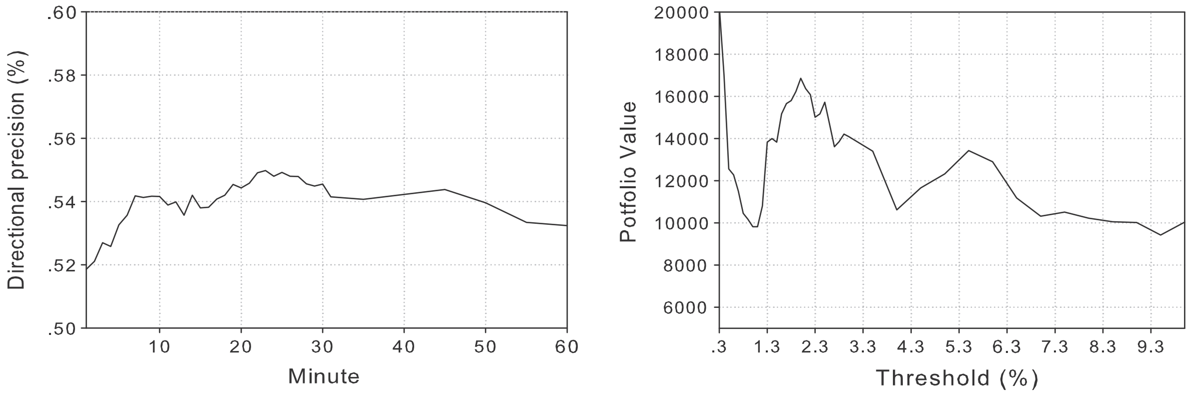
c) Transaction number

Fig. 9. Moving average strategy.

same probability per stock as MAQT, and trading transactions are made at the same times as MAQT (i.e. at the point of releasing a firm-specific news article). The earnings of the auto-trader only beat MAQT 8 times out of the 1000 trials. That is to say, the performance of the media-aware trader is significant at the 1% level.

5. Conclusion and future work

There are several interesting findings in our research. The first is that the media influence of financial news on stocks exists and can be quantified using natural language processing techniques in Computer Science. In particular, the fundamental information of a firm-specific news article can enrich the knowledge of investors and affect their trading activities. Meanwhile, news article sentiment may lead to emotion fluctuations of investors and interfere with their decision making. Both hypotheses are nicely supported by our experimental results that representing news articles with a number of proper nouns and sentiment words provides an effective way to quantify the news for stock prediction. In this research, the sentiment of Web media is quantified by the sentiment words regardless of syntax. It would be interesting



a) Window size

b) Threshold

Fig. 10. Parameter tuning.

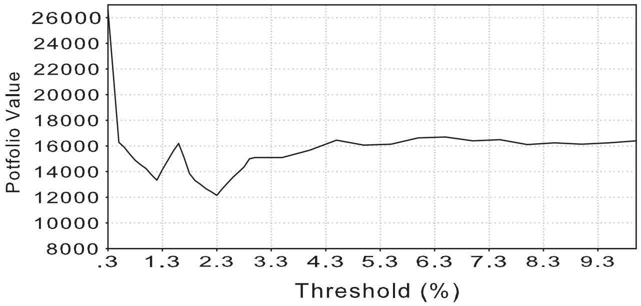


Fig. 11. Threshold for investment.

to explore how the advanced sentiment analysis, especially syntax analysis, could further improve the stock prediction.

In this study, all news articles are treated equally for training the predictive model. However, news articles on different gists and origins may affect the stock market differently. Specifically, certain types of news, such as executive personnel change and new product release, are typically more influential. In fact, Antweiler and Frank [3] show that the topic categories of corporate news including corporate governance, earnings reports, financial issues, operational issues and legal issues, affect investors distinctively. In addition, the article origin, be it official, leaked, or rumored, may have different influences on investors as well [41,46]. Indeed, it would be very interesting to further investigate these factors in combination with the content and sentiment of news.

At the individual level, public sentiment on a stock can affect personal investment decisions. Based on our experimental results, incorporating the public mood of the most recent 3 to 6 days can further increase the stock predictive performance. With the popularity of Web 2.0, there are various types of social media for readers to express their opinions including blogs, tweet/micro-blogs, and social news. An investigation into these new sources to capture the public mood would be imperative.

Acknowledgments

This work has been supported by the Program for New Century Excellent Talents in University in China, National Natural Science Foundation of China (NSFC) (No. 60803106, 61170133, 91218301), the Fundamental Research Funds for the Central Universities (No. JBK120505), the Sichuan Province Science Foundation for Youths, and NSERC Discovery Grant (No. 327667-2010).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.dss.2014.01.013.

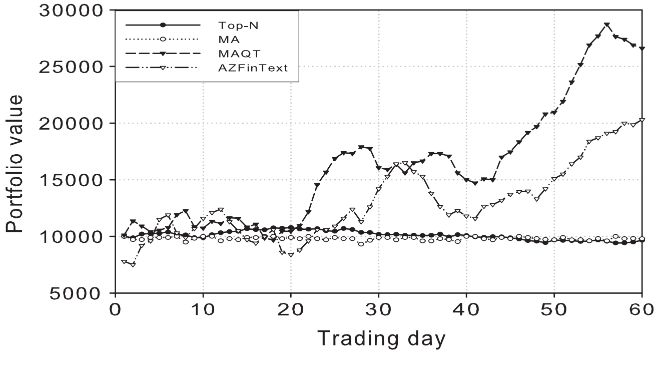


Fig. 12. Comparison.

References

[1] A.S. Abrahams, J. Jiao, G.A. Wang, W.G. Fan, Vehicle defect discovery from social media, Decision Support Systems 54 (1) (2012) 87–97.
[2] T.G. Andersen, T. Bollerslev, F.X. Diebold, C. Vega, Real-time price discovery in global stock, bond and foreign exchange markets, Journal of International Economics 73 (2) (2007) 251–277.
[3] W. Antweiler, M.Z. Frank, Do US stock markets typically overreact to corporate news stories? Working Paper, University of British Columbia, 2006.
[4] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman Publisher, 1999.
[5] N. Barberis, A. Shleifer, R. Vishny, A model of investor sentiment, Journal of Financial Economics 49 (3) (1998) 307–343.
[6] J. Bollen, A. Pepe, H. Mao, Twitter mood predicts the stock market, Journal of Computational Science 2 (1) (2011) 1–8.
[7] W.S. Chan, Stock price reaction to news and no-news: drift and reversal after headlines, Journal of Financial Economics 70 (2) (2003) 223–260.
[8] Y.J. Choi, Y.H. Kim, S.H. Myaeng, Domain-specific sentiment analysis using contextual feature generation, Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, ACM, 2009, pp. 37–44.
[9] D.M. Cutler, J.M. Poterba, L.H. Summers, What moves stock prices? Journal of Portfolio Management 15 (1989) 4–12.
[10] S.R. Das, M.Y. Chen, Yahoo! for amazon: sentiment extraction from small talk on the Web, Management Science 53 (9) (2007) 1375–1388.
[11] J.B. DeLong, A. Shleifer, L.H. Summers, R.J. Waldmann, Noise trader risk in financial markets, Journal of Political Economy 98 (4) (1990) 703–738.
[12] A. Duric, F. Song, Feature selection for sentiment analysis based on content and syntax models, Decision Support Systems 53 (2012) (2012) 704–711.
[13] J.E. Engelberg, A.V. Reed, M.C. Ringgenberg, How are shorts informed?: short sellers, news, and information processing, Journal of Financial Economics 105 (2) (2012) 260–278.
[14] E.F. Fama, K.R. French, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33 (1) (1993) 3–56.
[15] G.P.C. Fung, J.X. Yu, W. Lam, Stock prediction: integrating text mining approach using real-time news, Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering, IEEE, 2003, pp. 395–402.
[16] G. Gidofalvi, Using News Articles to Predict Stock Price Movements, Department of Computer Science and Engineering, University of California, San Diego, 2001.
[17] N. Jain, M. Dahlin, R. Tewari, Using bloom filters to refine Web search results, Proceedings of the 8th International Workshop on the Web and Databases (WebDB), 2005, pp. 25–30.
[18] F.E. James, Monthly moving averages — an effective investment tool? Journal of Financial and Quantitative Analysis 3 (3) (1968) 315–326.
[19] J. Johnson, B. Kitchens, D. Mitra, P. Pathak, Does the market believe in marketing? A text mining based informational value perspective, Proceedings of the 22nd Workshop on Information Technology and Services (WITS), 2012.
[20] J.M. Keynes, The General Theory of Employment, Macmillan, London, 1936.
[21] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Language models for financial news recommendation, Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM), ACM, 2000, pp. 389–396.
[22] B. LeBaron, W.B. Arthur, R. Palmer, Time series properties of an artificial stock market, Journal of Economic Dynamics & Control 23 (9–10) (1999) 1487–1516.
[23] Q. Li, J. Wang, Y.P. Chen, Z. Lin, User comments for news recommendation in forum-based social media, Information Sciences 180 (2010) 4929–4939.
[24] T. Loughran, B. McDonald, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Financial 66 (1) (2012) 35–65.
[25] M.A. Mittermayer, G.F. Knolmayer, Newscats: a news categorization and trading system, Proceedings of the 6th International Conference on Data Mining (ICDM), Ieee, 2006, pp. 1002–1007.
[26] G.V. Nartea, B.D. Ward, H.G. Djajadikerta, Size, BM, and momentum effects and the robustness of the Fama–French three-factor model: evidence from New Zealand, International Journal of Managerial Finance 5 (2) (2009) 179–200.
[27] J.R. Nofsinger, Social mood and financial economics, Journal of Behavioral Finance 6 (3) (2005) 144–160.
[28] B. Pang, L. Lee, Opinion Mining and Sentiment Analysis, Now Publisher, 2008.
[29] R.E. Schapire, Y. Singer, Boostexter: a boosting-based system for text categorization, Machine Learning 39 (2–3) (2000) 135–168.
[30] R.P. Schumaker, H. Chen, A quantitative stock prediction system based on financial news, Information Processing & Management 45 (5) (2009) 571–583.
[31] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Transactions on Information Systems (TOIS) 27 (2) (2009) 12.
[32] R.P. Schumaker, Y.L. Zhang, C.N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems 53 (3) (2012) 458–464.
[33] Y.W. Seo, J.A. Giampapa, K.P. Sycara, Text classification for intelligent agent portfolio management, Proceedings of the 1st International Joint Conference on, Autonomous Agents and Multi-Agent Systems, 2002, pp. 802–803.
[34] A. Shleifer, R.W. Vishny, The limits of arbitrage, Journal of Finance 52 (1) (1997) 35–55.
[35] P.C. Tetlock, Giving content to investor sentiment: the role of media in the stock market, Journal of Finance 62 (3) (2007) 1139–1168.
[36] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, Journal of Finance 63 (3) (2008) 1437–1467.
[37] P.D. Turney, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 21 (4) (2003) 315–346.
[38] C. Vega, Stock price reaction to public and private information, Journal of Financial Economics 82 (1) (2006) 103–133.

[39] P. Veronesi, Stock market overreactions to bad news in good times: a rational expectations equilibrium model, Review of Financial Studies 12 (5) (1999) 975–1007.
[40] B. Wang, H. Huang, X. Wang, A novel text mining approach to financial time series forecasting, Neurocomputing 83 (2012) (2011) 136–145.
[41] G.A. Wang, J. Jiao, A.S. Abrahams, W.G. Fan, Z.J. Zhang, Expertrank: a topic-aware expert finding algorithm for online knowledge communities, Decision Support Systems 54 (3) (2013) 1442–1451.
[42] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, Proceedings of Computational Linguistics and Intelligent Text Processing, 2005, pp. 486–497.
[43] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, OpinionFinder: a system for subjectivity analysis, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP), 2005, pp. 34–35.
[44] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP), 2005, pp. 347–354.
[45] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual Web data, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1998, pp. 2720–2725.
[46] X.L. Zhu, S. Gauch, Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2000, pp. 288–295.

Qing Li is a professor of Southwestern University of Finance and Economics, China. Prior to that he was a post-doctoral Researcher with Arizona State University and Information & Communications University of Korea separately. Li's research interests lie primarily in intelligent information processing and business intelligence. He has published over 40 articles in respected journals and conferences in related areas. He served in the organization and program committees of various international conferences including PACIS 2014, SIGIR 2008, CIKM 2007 and AIRS2005. He received his Ph.D. from Kumoh National Institute of Technology in February 2005, and his M.S. and B.S. degrees from Harbin Engineering University, China.

Tiejun Wang is a Ph.D. student at Southwestern University of Finance and Economics, China. His research interests lie primarily in finance intelligent.

Qixu Gong is a graduate at Southwestern University of Finance and Economics, China. His research interests lie primarily in intelligent information processing and business intelligence.

Yuanzhu Chen is an associate professor in the Department of Computer Science at Memorial University of Newfoundland in St. John's, Newfoundland. He received his Ph.D. from Simon Fraser University in 2004 and B.Sc. from Peking University in 1999. Between 2004 and 2005, he was a post-doctoral researcher at Simon Fraser University. His research interests include graph theory, information retrieval, and wireless sensor networking.

Zhangxi Lin is an associate professor at Texas Tech. University, and also an adjunct professor for Southwestern University of Finance and Economics. He received his Ph.D. in information systems from University of Texas at Austin in 1999, M.S. in Economics from University of Texas at Austin in1996, and M. Eng. in Computer Science from Tsinghua University in 1982. His research interests include management information system and e-Commerce.

Sa-Kwang Song is a senior researcher at Korea Institute of Science and Technology Information (KISTI). He received his Ph.D. in Computer Science from Korea Advanced Institute of Science and Technology (KAIST), Korea, and M.S. degree from Chungnam National University, Korea. His current research interest includes Text Mining, Natural Language Processing (NLP), Information Retrieval, Semantic Web, and Health-IT.