Information Sciences 278 (2014) 826–840

Contents lists available at ScienceDirect

Information Sciences

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

# The effect of news and public mood on stock movements

Qing Li [a], TieJun Wang [a], Ping Li [a], Ling Liu [a,*], Qixu Gong [a], Yuanzhu Chen [b]

[a] Economic Information Engineering School, Southwestern University of Finance and Economics, China
[b] Department of Computer Science, Memorial University of Newfoundland, Canada

ARTICLE INFO

*Article history:*
Received 26 November 2012
Received in revised form 16 March 2014
Accepted 22 March 2014
Available online 12 April 2014

*Keywords:*
Text mining
Sentiment analysis
News
Social media
Stock market

ABSTRACT

With technological advancements that cultivate vibrant creation, sharing, and collaboration among Web users, investors can rapidly obtain more valuable and timely information. Meanwhile, the adaption of user engagement in media effectively magnifies the information in the news. With such rapid information influx, investor decisions tend to be influenced by peer and public emotions. An effective methodology to quantitatively analyze the mechanism of information percolation and its degree of impact on stock markets has yet to be explored. In this article, we propose a quantitative media-aware trading strategy to investigate the media impact on stock markets. Our main findings are that (1) fundamental information of firm-specific news articles can enrich the knowledge of investors and affect their trading activities; (2) public sentiments cause emotional fluctuations in investors and intervene in their decision making; and (3) the media impact on firms varies according to firm characteristics and article content.



## 1. Introduction

To prove the efficient market hypothesis (EMH) [14], many researchers have been devoted to studying the impact of the information on the movement of stock markets [2,3,7,16,21,23,28,29,38,41,48]. One of the earliest reports, that by Cutler et al. [10], found that macroeconomic performance news could explain approximately one-third of the variance in stock returns. Tetlock et al. [40,41] then showed that general financial news has limited and short-lived predictive power on future stock prices. Recent theoretical studies in behavioral finance have demonstrated that emotion influences investment decisions [11,39]. Such an inference was further confirmed by the findings of Li [23] and Schumaker et al. [36]. The authors discovered that the sentiments contained in financial reports or news articles affect stock returns. To systematically investigate the relationship between Web media and the stock market, we propose an effective methodology to quantitatively analyze the mechanism of information percolation on stocks.

In particular, we first testify that the fundamental information in firm-specific news articles affects the trading activities of investors. This correlation is achieved by representing online financial news that is related with the companies listed in China Securities Index (CSI 100)[1] as weighted term vectors, and by applying a predictive model to analyze the impact of the news on stock movements.

* Corresponding author.
*E-mail address:* lingliu@swufe.edu.cn (L. Liu).
*URL:* http://fife.swufe.edu.cn/BILab/ (L. Liu).

[1] 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 the Shanghai and Shenzhen securities markets.

http://dx.doi.org/10.1016/j.ins.2014.03.096
0020-0255/

Second, we study the sentiment impact on stocks, especially the emotions that the news evokes in the public. In fact, the wide adaption of social media allows readers to have easy access to the opinions or feelings of others via discussion, votes, comments, and similar means. With such a rapid influx of information, investor decisions tend to be influenced by the emotions of peers and the public. Thus, investigating such sentiment percolation and its degree of impact on stocks is important.

A unique contribution of this work is unveiling the black box of the internal functions of sentiments, firm characteristics, and news content on the relationship of Web media and stock markets. To the best of our knowledge, this is the first work to systematically investigate the determinants of Web media on stock markets in a quantitative manner.

The remaining content 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 the trading performance using real stock market data from the Shanghai and Shenzhen Stock Exchanges (Section 4). This paper is concluded with speculation on how the current prototype can be further improved in Section 5.

## 2. Related work

Researchers have explored the power of verbal information on stock markets due to the observation of stock price fluctuations with news feed. Empirical pilot studies have correlated news and stocks. For instance, [44] demonstrated that stock prices overreact to bad news in good times and under-react 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 display a negative drift for up to 12 months; less drift was correlated with stocks with good news. Vega [43] further investigated media influence, reporting that stocks associated with private information experience low or insignificant drifts and that stocks associated with public news only experience significant drifts. By investigating the relationship between "risk sentiment" and stocks, Li [23] discovered that the risk sentiment has a negative prediction on stock returns, i.e., firms with a large increase in risk sentiment suffer negative returns, and vice versa.

With technological advancements fertilizing vibrant creations, sharing, and collaborations among Web users, the adaption of user engagements in social media effectively magnifies the influential power of Web media. Some researchers have focused on predicting the financial performances of the listed firms that utilize social media. For example, Bollen et al. [6] captured the public mood from tweets to forecast stock movements. Luo et al. [27] reported that social media is a significant leading indicator of a number of firm equity values based on the software and hardware industries. Yu et al. [49] suggested that social media attributes have a stronger relationship with stock performances than do conventional media attributes.

Different from empirical studies, experimental economists generate a laboratory-based financial market to explore information availability and trader effectiveness [2]. In particular, [31] created five markets to address the dissemination and aggregation of information, concluding that markets follow rational expectation equilibrium model predictions. Alfarano et al. [2] argued that if the quality of private information is not good, public information has a major effect on stock prices.

Table 1 summarizes some representative research on the influence of news on stock markets. For each study, the table summarizes its focus (firm), media source, and experiment. The firm column compares the selection of equity value, value scale, and focusing market. The media column presents the media source and its metric, i.e., whether the media is quantifiably measured (e.g., number of news articles) or via textual analyses. The experiment column compares the experimental data period and method. The empirical study and laboratory simulation in this research prove the existence of media influence on stock markets. However, a detailed analysis of information percolation and its degree of impact on stock markets has yet to be performed.

How to capture media influence on stocks and bridge such connections remains challenging. Artificial intelligence and natural language processing techniques have been utilized to address these challenges [21,28,33–35,48]. In particular, Wang et al. [45] presented a news article as a term vector using full words and studied the link of news to stocks using a hybrid predictive model. Schumaker and Chen [35] experimented with several textual representative approaches, including bag of words, noun phrases, proper nouns, and name entities, and found that representing news with proper nouns was most efficient. Tetlock et al. [41] analyze the sentiment of news articles and show that the fraction of negative words in firm-specific news stories forecasts low firm earnings. Frank and Antweiler [15] extracted the bullish and bearish sentiments of Yahoo! Finance postings, concluding that the effect of financial discussion boards on stocks is statistically significant. Gilbert and Karahalios [18] reported that an increase of anxiety, worry, and fear emotions produces downward pressure on the S&P 500 index. Table 2 summarizes these studies on media-aware stock analyses, reviewing textual representation, emotion extraction, and analysis model. How each research report relates to the issues that we aim to address in this article (i.e., quantifying the influence of public mood and news to study the impact of Web media on stock markets) is also summarized.

The sentiment analyses presented in previous studies rely on open-domain opinion analyses. The general sentiment word categorization, however, cannot translate effectively into a discipline with its own dialect [1,25]. In this article, to successfully measure news sentiments and capture public moods regarding investments, we propose an innovative algorithm that automatically extracts finance-oriented sentiment words from the Web. Furthermore, a media-aware trading strategy that utilizes finance-oriented sentiments is presented to study the combined effect of Web news and social media on stock

**Table 1**
Literature comparison of the media influence on stock markets.

| Literature | Firm | | | Media | | Experiment | |
|---|---|---|---|---|---|---|---|
| | Equity value | Frequency | Stock selection | Data source | Metrics | Period | Method |
| Chan [7] | Stock return | Monthly | NYSE, AMEX, S&P500 | DJIPL | Quantified metrics | 1980–2000 | Empirical |
| Tetlock et al. [41] | Stock return | Daily | S&P500 | DJNS | Textual analysis | 1980–2004 | Empirical |
| Wüthrich et al. [48] | Index | Daily | Five major Indexes | Web news | Textual analysis | 12/06/1997–03/06/1997 | Experimental |
| Lavrenko et al. [21] | Stock price | Hourly | 127 stocks | Biz Yahoo | Textual analysis | 10/15/1999–02/10/2000 | Experimental |
| Mittermayer and Knolmayer [28] | Stock trend | Per minute | S&P500 | PR news wire | Textual analysis | 04/01/2002–12/31/2002 | Experimental |
| Bollen et al. [6] | Index | Daily | NYSE | Twitter | Textual analysis | 02/28/2008–12/19/2008 | Both |
| Wang et al. [45] | Stock price | Quarterly | 6 Stocks | Financial report | Textual analysis | 1994–2010 | Experimental |
| Schumaker et al. [36] | Stock price | Per minute | S&P500 | PR news wire | Textual analysis | 10/26/2005–11/28/2005 | Experimental |
| Luo et al. [27] | Stock return and risk | Daily | 9 stocks | Lexis/Nexis Web blogs | Quantified metrics | 08/01/2007–07/01/2009 | Empirical |
| Yu et al. [49] | Stock return | Daily | 824 stocks in 6 industries | Twitter, Web news | Both | 07/01/2011–09/30/2011 | Empirical |
| Frank and Antweiler [15] | Stock return | Every 15-min | DJIA,XLK | Yahoo! Finance message board, Raging Bull, WSJ | Textual analysis | 2000 | Empirical |
| Gilbert and Karahalios [18] | Stock price | Daily | S&P500 | LiveJournal posts | Textual analysis | 01/25/2008–06/13/2008, 08/01/2008–09/30/2008, 11/03/2008–12/18/2008 | Empirical |
| Our paper | Stock price | Per minute | CSI100 | Web news, financial discussion board | Textual analysis | 01/01/2011–12/31/2011 | Experimental |

**Table 2**
Core technique comparison in media-driven trading strategies.

| Literature | Text | Emotion | Public mood | Train/test | Metric | Model |
|---|---|---|---|---|---|---|
| Tetlock et al. [41] | Emotion word | General emotion | General market | NA | Real value | Regression |
| Wüthrich et al. [48] | Selected words | General emotion | General market | Three months rolling | Trend | KNN, Regression |
| Lavrenko et al. [21] | All words | General emotion | General market | 90/40 days | Trend | Language model |
| Mittermayer and Knolmayer [28] | Selected words | General emotion | General market | 90%/10% stocks | Trend | KNN, SVM |
| Bollen et al. [6] | | General emotion | Individual stock | 10/1 months | Real value | SOFNN |
| Wang et al. [45] | All words | General emotion | General market | 20/5 stocks | Real value | ARIMA, SVR |
| Schumaker et al. [36] | Proper nouns | General emotion | General market | 4/1 weeks | Real value, trend | SVR |
| Yu et al. [49] | Sentence-level sentiment | General emotion | General market | NA | Real value | NB algorithm |
| Frank and Antweiler [15] | Bullish, bearish | General emotion | General market | NA | Real value, trend | NB algorithm |
| Gilbert and Karahalios [18] | Anxiety, worry, fear | General emotion | General market | NA | real value | Granger causality |
| Our paper | Proper nouns, emotion words | Financial emotion | Individual stock | 9/3 months | Real value, trend | SVR |

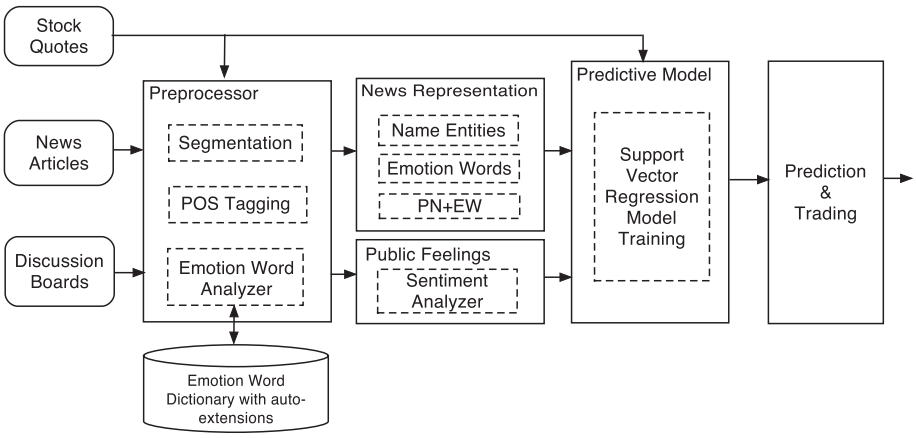



**Fig. 1.** eMAQT: an eMedia-Aware Quantitative Trader.

markets, particularly at the individual stock level. This method allows us to investigate the internal functions of sentiments, firm characteristics, and news content on the relationship between Web media and stock markets.

## 3. Framework of the media-aware quantitative trader

In this work, we study the impact of Web media on stock markets using an electronic-media-aware quantitative trader, termed eMAQT. The framework of eMAQT is illustrated in Fig. 1. eMAQT provides a solid basis for us to determine the internal connections between media and stocks, i.e. how such connections vary according to sentiment, firm characteristics, and news content.

### 3.1. Media quantification

The quantification of textual information remains a challenge for exploring the impact of public Web information on stock movements. Due to differences in the writing styles used, we use different approaches to study the impact of official news articles and public discussions.

#### 3.1.1. Representation of news articles

In behavior finance, the influence of news articles on stocks originates on two facets.

- **Fundamentals:** people tend to adjust their investment strategies in accordance with the latest news articles conveying qualitative and quantitative information, which contribute to the economic well-being and subsequent financial valuation of a company, security, or currency.
- **Emotion:** emotional investors can be affected by the optimistic or pessimistic mood in a news article.

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 news articles can be captured with a set of nouns and that news sentiments can be 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 originates from domain-specific sentiment analyses. Various studies on the detection of general-domain sentiment terms exist; however, they are not applicable here. As per Loughran and McDonald [25], 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. Building a comprehensive, finance-specific sentiment word list is of great apparent necessity We defer the description of how to extract finance-specific sentiment terms to Section 3.2.

After extracting nouns and sentiment words to represent an article as a term vector, the weight of each term indicating its topic importance is measured using standard TF/IDF weighting schema [5][24]. Specifically, we define the weight of term $t$ in document $d$ as shown below

$$w(t,d) = \left(0.5 + 0.5 \times \frac{f(t,d)}{\max_{t'} f(t',d)}\right) \times I(t), \tag{1}$$

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

where $f(t,d)$ denotes the number of occurrences of term $t$ in document $d$, i.e. "term frequency", and $I(t)$ represents the "inverse document frequency" of term $t$ with regard to the training news corpus, and is denoted by

$$I(t) = \log\frac{C}{c(t)}. \tag{2}$$

Here, $C$ defines the corpus size, and $c(t)$ is the number of documents in corpus containing term $t$. After normalization, the final term weight is defined as

$$W(t) = \frac{w(t,d_0)}{\max_{t'} w(t',d_0)}. \tag{3}$$

*3.1.2. Public mood of a stock*

With technological advancements that cultivate vibrant sharing and interactions of users, social media, which include discussion boards, blogs, and microblogs, are becoming increasingly popular and provide a good platform to share investment opinions or feelings on stocks. Because investors are sensitive to the investment decisions of others, quantifying the public mood in social media is important [6]. In this study, we capture the public mood of a stock from firm-specific messages in discussion boards. Specifically, we apply our focused Web crawler to download postings from the two most popular stock discussion forums in China, i.e. www.sina.com and www.eastmoney.com. Because each traded firm has its own discussion section on these websites, we analyze the discussion threads for each firm to capture the public tendency for investment.

Here, we measure the public mood on stock $s$ from two aspects: optimism ($M_s^+$) and pessimism ($M_s^-$). The optimistic mood on stock $s$ is measured as

$$M_s^+ = \sum_{i=0}^{\tau}\sum_{j=0}^{K}\frac{P_{i,j}\times W_j}{l_i}\times T_i, \tag{4}$$

where $P_{i,j}$ denotes the number of the positive words in posting $j$ on the $i$-th day after news release, $L_i$ is the total number of the words in the postings on the $i$-th day, $W_j$ represents the weight of posting $j$, and $T_i$ defines the time factor. The posting weight $W_j$ is applied to differentiate the influence power of postings and eliminate noise. Intuitively, influential postings are those whose contents are read or discussed by a larger number of readers. Therefore, we calculate the weight in terms of clicks:

$$W_j = \frac{c_j}{\max_{t'} c_{t'}}, \tag{5}$$

where $c_j$ is the number of clicks on posting $j$, and $max_{t'}c_{t'}$ denotes the largest number of clicks on the day of posting message $j$. Because the influence of public sentiment wanes but lasts for several days [41], the time factor $T_i$ is defined to tune the influence power with time passed,

$$T_i = e^{-i/\beta}. \tag{6}$$

Here, $\tau$ is the number of the passed days that we consider a sentiment to have continued influence, and $\beta$ represents a constant tuning the time attenuation scale, and is set to 20 to simulate the attenuation for the number of work days in a month.

Similarly, the pessimistic mood of stock $s$ is measured as

$$M_s^- = \sum_{i=0}^{\tau}\sum_{j=0}^{K}\frac{N_{i,j}\times W_j}{l_i}\times T_i, \tag{7}$$

where $N_{i,j}$ defines the number of negative words in posting $j$ on the $i$-th day.

*3.2. Sentiment analyses*

To capture the moods of news stories and public feelings, understanding the emotion polarity of a word with sentiment analyses is critical. Here, the challenge lies on domain-specific sentiment analyses. Most previous work has employed general-domain sentiment analyses. However, general sentiment words may not be emotional in the realm of finance [1,25]. Specifically, some typical sentiment words, such as "crude" or "tire", are more likely to identify with a specific industry in economics rather than to express a negative emotion. In fact, Loughran and McDonald [25] found that approximately three-quarters of all negative words in a general emotion word dictionary (Harvard-IV-4) are not considered negative in a financial context. In addition, an emotionless word can be a sentiment in the realm of finance to some degree. The word "bull" originally refers to a male bovine animal; it also indicates good earning returns in the finance domain, such as "bull stock". To precisely capture such sentiments in the finance domain, we propose an innovative algorithm that automatically extracts finance-oriented sentiment words from the Web. This process is achieved by analyzing the contextual information of words in the realm of finance.

*3.2.1. Finance-oriented sentiment word detection*

A common strategy for sentiment word extraction is to apply various supervised or unsupervised machine learning techniques, including Naïve Bayesian Networks, SVM, and LDA, to label the sentiment polarity of a word in an open-domain training corpus [12,30]. A few researchers have extended such a method to study domain-specific sentiment word extraction [1,9]. Different from general sentiment word detection, these techniques identify domain-specific words in terms of their statistical associations with domain-specific texts that have been previously labeled positive or negative. In this study, we take a step further by incorporating stock market information to enrich the contextual associations of domain-specific sentiment words. Specifically, finance-specific sentiment words are extracted based on two classical hypotheses.

- A word is characterized by its contextual information, i.e., the semantic orientation of a word tends to correspond to the semantic orientation of its neighbors in the textual content [42].
- A firm-specific article with a positive (negative) tone is typically in accordance with the upward (downward) price trend of a relevant stock. This hypothesis is further confirmed by Tetlock's study regarding the interaction between stock movements and daily news, particularly pessimistic news [40].

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

$$\begin{aligned} P^{+}(w) &= P(w|E=+,T=\uparrow) \\ &\approx P(w|T=\uparrow)P(E=+|w,T=\uparrow) \\ &= P(w|T=\uparrow)\sum_{i=0}^{M}P(e_i=+|w,T=\uparrow) \end{aligned} \tag{8}$$

where $E$ denotes the semantic orientation of its neighbors with two values, $+$ and $-$, representing positive and negative emotion, respectively. Here, $T$ defines the stock price trend with two values, $\uparrow$ and $\downarrow$, indicating upward and downward price trends, respectively. $e_i$ is the sentiment word in the paradigm set $S$,[3] and $M$ represents the total number of positive words in the set.

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

$$P(w|T=\uparrow)=\frac{U(w)}{N(w)}, \tag{9}$$

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

$P(e_i=+|w,T=\uparrow)$ represents the contextual relationship between word $w$ and positive word $e_i$ in the paradigm set $S$. This relation can be estimated with the appearance concurrence of the target word, $w$, and the sentiment word, $e_i$, in the article associated with upward trends, and smoothed by the semantic similarity of word $w$ and $e_i$. Specifically,

$$P(e_i=+|w,T=\uparrow)\approx \lambda I^{+}(w,e_i)+(1-\lambda)\times S(w,e_i), \tag{10}$$

where $\lambda$ is the coefficient that adjusts the contribution of the semantic similarity. $I^{+}(w,e_i)$ is the statistical association between word $w$ and positive word $e_i$ in articles with an upward trend which is measured by $\chi^2$. $S(w,e_i)$ is the semantic similarity in terms of semantic relationships in the WordNet.[4] To calculate $S(w,e_i)$, we create $w$ and $e_i$ a feature vector each. It consists of words extracted from three sets: synonyms, class words, and explanation words. These sets are related with $w$ and $e_i$, irrespectively, in WordNet. $S(w,e_i)$ is computed as the normalized vector similarity of the two feature vectors [5].

The estimation of $P(w|T=\downarrow)$ and $P(E=-|w,T=\downarrow)$ are quite similar to the description above with the difference that the positive paradigm words are replaced with negative paradigm words, and the selected documents are associated with a downward trend rather than a upward trend.

*3.2.2. Stock trend*

To calculate the statistical information for sentiment word detection, firm-specific news must be associated with a downward or upward stock trend. Here, the challenge lies in discovering trends from the time series of stock prices. Essentially, this is a curve segmentation problem. The basic idea to address this problem is to plot (or approximate) the discrete data into a curve and then segment it into a series of straight lines [16,20,26,32]. To fully address this problem in the context of stock price, fluctuation and interruption curve patterns must be properly analyzed.

[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.

[4] WordNet is a large lexical database (http://wordnet.princeton.edu/). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

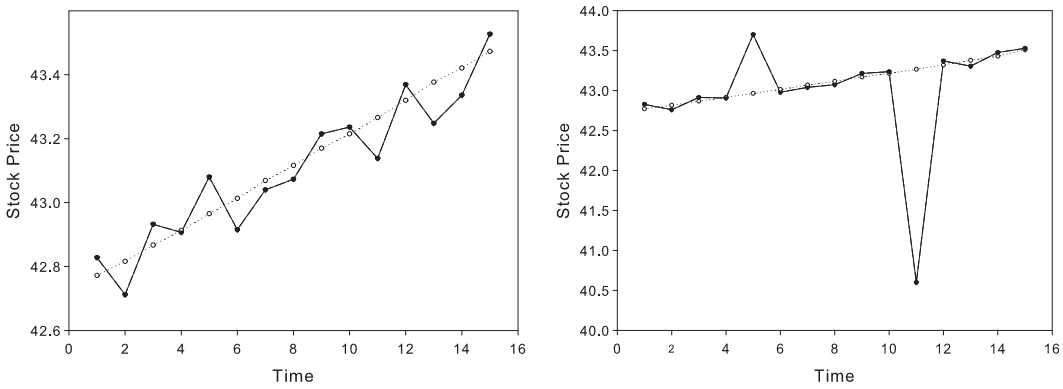



**Fig. 2.** Curve patterns.

- **Fluctuation:** In this study, the purpose of curve segmentation is to determine which articles tend to increase (or decrease) the relevant stock price. Small fluctuations in the curve should not affect the entire trend; small decreases or increases in a roaring (or slipping) trend should still be treated as downward or upward trends, respectively, as a whole (Fig. 2(a)).
- **Interruption:** Similar to fluctuation patterns, significant but temporary decreases or increases in a curve should not affect the entire trend (Fig. 2(b)). Interruptions are generally caused by noise or errors in the original stock transaction data, which can be ignored in our study.

In this article, we propose a segment-and-merge approach for discovering the trends of stock prices. The approach first segments a curve into a number of straight lines and then merges these lines to avoid over segmentation. Specifically, given a time series of a stock price $\{x_i, y_i\}$, where $x_i$ denotes the trading time, $y_i$ represents the corresponding stock price, and $i = 0, 1, 2, \ldots, n$, we first determine the straight line equation, $\alpha y + \beta x + \epsilon = 0$, which would provide a "best" fit for the data points. Here, the best is understood as it is in the least-squares approach: a line that minimizes the squared residuals sum of the linear regression model.

With this equation, we can calculate the distance of point $\{x_i, y_i\}$ to the line with

$$d_i = \frac{|\alpha x_i + \beta y_i + \epsilon|}{\sqrt{\alpha^2 + \beta^2}}. \tag{11}$$

Then, we test this linear regression with the $t$-statistic:

$$t = \frac{\overline{d}}{\frac{s}{\sqrt{n}}}, \tag{12}$$

where $\overline{d}$ denotes the average distance of all the points to the line, $s$ is the standard deviation of these distances, and $n$ represents the total number of points to construct the linear line. If the $t$-statistic is unaccepted, we further split the regression line at the point that has the maximum deviation. Such a segment stage continues on each new split segment recursively until all segments are accepted by the $t$-test.

After segmenting the curve into pieces, we run a merging stage in a bottom-up manner to alleviate the problem from over segmentations and interruption patterns. Over segmentation is generally caused by the greedy nature of the first segment in each split iteration. As shown in Fig. 3(a), segments $c2$ and $c3$ are split by the break-point $b1$; however, their trend divergence

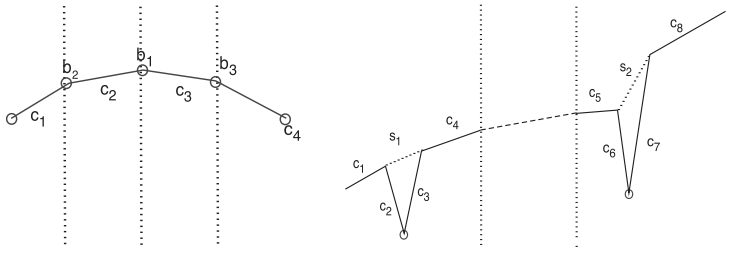



**Fig. 3.** Merging purposes.

is similar. From a trend analysis perspective, the points should belong to the same trend. Therefore, we merge the two adjacent segments if the $t$-test of their combined segment is accepted. Furthermore, to discover interruption patterns, we calculate the slope of each segment and examine whether the significant, but temporary change in the curve interrupts the smooth trend of their adjacent segments. As shown in Fig. 3(b), segment $c_2$ and $c_3$ are interruptions, and $c_1$, $c_2$, $c_3$ and $c_4$ should be combined into one segment because the change of slopes between $c_1$, $c_4$, and $s_1$ is small. However, $c_6$ and $c_7$ cannot be treated as interruptions because of the obvious changes in slopes between $c_5$, $c_8$, and $s_2$. This process is repeated until all interruptions are identified and merged.

### 3.3. Predictive model

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

A variety of machine learning methods for stock market predictions exist, including rough sets theory [8], relevance language model (RLM) [21], support vector machine (SVM) [28], and Naïve Bayesian [37]. However, all of these works focus on directional movements rather than numerical stock prices. In this study, we adopt the extended SVM, i.e., the support vector regression (SVR) model, which applies a regression technique to the SVM to predict numerical values of future stock prices [35].

## 4. Experiments

With an effective methodology to quantitatively analyze the mechanism of information percolation and its impact on stock markets, we are of particular interest to understand whether stock market movements are sensitive to public information. If the stock market movements are sensitive to public information, public information events are subject to different interpretations by investors and present profitable trading opportunities for skilled investors. The trades of informed investors should therefore be more profitable after a news release day. Otherwise, public information reduces asymmetric information, and the trades of informed investors should be less profitable on a news release day [13].

We further investigate the internal mechanism of information percolation on stock markets. In particular, we would like to determine the following:

- Are investors subjective to public mood or news sentiment? If they are, concrete evidence is provided to support a critical hypothesis in behavioral finance: investor sentiments affect stock prices.
- Do news articles with different contents affect the stock market differently? What topic types have strong influence? The answers to these questions are critical for listed firms to enhance their reputation on the Web.
- Does the media influence firms of different characteristics differently? What types of firms are vulnerable to the release of public information? The answers to these questions would provide a good reference for investors to sense stock movements.

Obviously, insights into these issues are essential to protect investors with high quality information, assist in management in the security market, and provide decision makers with the most reliable input regarding the health of the market.

### 4.1. Experimental settings

In this research, we constructed three databases to explore the relationship between the Web media and stock market, i.e., financial news, discussion board posts, and stock transaction data. The details of the experimental data are presented in Table 3. Here, we target the two independent stock exchanges in mainland China: the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE). At the end of 2011, there were 2341 listed stocks on both SSE and SZSE, with a total market capitalization of RMB 20.893 trillion. Prior relevant studies [34,35,40,41] mainly focused on stock exchanges in the United States, particularly the New York Stock Exchange (NYSE). Due to the lack of market makers in Chinese markets[5], this study on SSE and SZSE provides unique insight into the relationship of the media and stock market without interference from the trades of market makers.

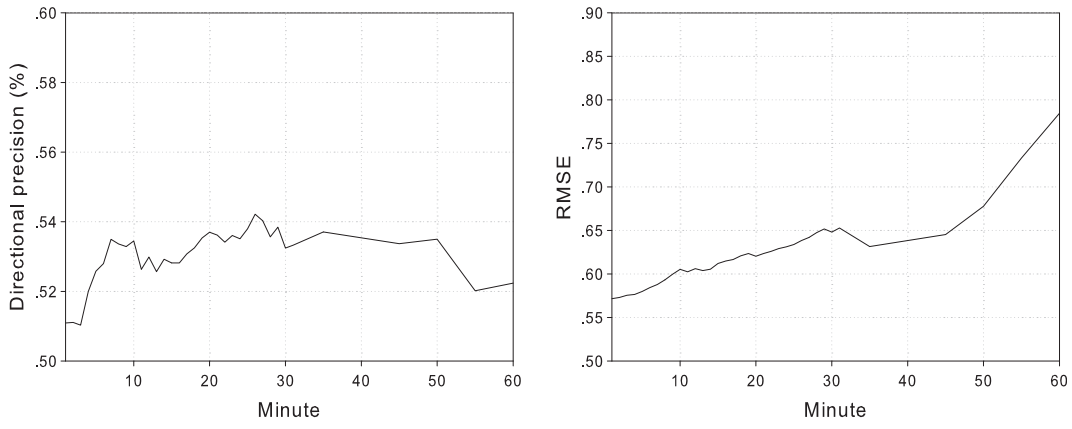
In our experiments, we used data from the first 9 months of 2011 as a training corpus and the last 3 months of 2011 for testing. We removed 11 companies from the available 100 companies due to an inconsistency in the CSI 100 list.[6] In this

[5] The New York Stock Exchange (NYSE) and American Stock Exchange (AMEX) have designated market makers, who act as the official market maker for a given security. The market makers provide a required amount of liquidity to the security's market, and take the other side of trades when there are short-term buy-and-sell-side imbalances in customer orders. In return, the market maker is granted various informational and trade execution advantages – From Wikipedia.

[6] Because the CSI 100 list is adjusted every half a year, we only experimented on the companies listed for the entire year of 2011.

**Table 3**
Three databases for experiment.

| Database | Source | Period | Feature | Description |
|---|---|---|---|---|
| Financial news | 72 Chinese financial websites | 01/01/2011–12/31/2011 | News title, body, publication date | It consists of 124,470 financial news articles related to CSI 100 companies |
| Discussion board | www.sina.com www.eastmoney.com | 01/01/2011–12/31/2011 | Posting, publication date, clicks, | It contains the discussion threads of the CSI 100 companies |
| Stock data | China Stock Market Database (CSMD) | 01/01/2010–12/31/2011 | Stock ID, price, volume, transaction time | It is high-frequency data with intraday transaction information at the granularity of second |




(a) Directional accuracy (b) RMSE

**Fig. 4.** Predictive outlook window.

testing period, the upward trend was 46.12%, the downward trend was 49.53%, and the remaining percentage was maintained. The standard deviation of the stock prices in this testing period was 27.12.

### 4.2. Metrics

Directional accuracy and closeness are two classic evaluation metrics to measure the impact of media on stock movements. As shown in Table 2, previous reports have focused on either or both as evaluation metrics.

In this study, we chose both metrics as our evaluation standards. Specifically, *directional accuracy* measures the upward or downward direction of the predicted stock price compared with the actual movement direction of the stock price. Thus, this metric may be close in prediction but predict a wrong movement direction; the *closeness metric* is used to complement the evaluated difference between the predicted value and the real stock price in terms of root mean squared errors (RMSE).

### 4.3. Time window of prediction

The predictive eMAQT model captures 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 [17] reported that a good outlook time window for a stock forecast is approximately 20 min after the release of relevant information. Subsequent research [34,35] has adopted this window in their stock prediction studies. We next examine the existence of this phenomenon in the Chinese mainland markets, which lack market makers in stock transactions. As shown in Fig. 4, the directional accuracy increases and achieves a best performance of 0.5421 at the 26th minute after news release. A good predictive window of approximately 20 min is observed after news release. This finding indeed agrees with the previous research that reported the existence of lag time between information introduction and stock market correction to equilibrium [22]. That is, the market could be forecasted in short durations after the introduction of new information.

Notably, we determine an optimal predictive outlook window of 26 min for the following experiments in terms of directional accuracy rather than RMSE (Fig. 4). This result was determined because the performance of our investment strategies (Section 4.5) relies on directional accuracy, i.e., the difference between the predicted future price and the price at the point of news release.

**Table 4**
Representation *PN denotes that a news article is represented by a number of weighted proper nouns. *Harvard* represents an article by a number of weighted sentiment words from the Harvard-IV-4 list. *FS* defines an article as a number of weighted sentiment words from the finance-specific sentiment word list.

| Method | RMSE | Directional precision (%) |
|---|---|---|
| PN | 0.6385 | 54.21 |
| Harvard | 0.6291 | 49.38 |
| FS | 0.6157 | 51.72 |
| PN + Harvard | 0.6192 | 54.75 |
| PN + FS | 0.6076 | 55.34 |

### 4.4. Determinants of predictability

In this section, we investigate the predictability determinants of the proposed media-aware trader. Rather than simply show the relationship between the media and stocks, we unveil the hidden factors that determine such connections. In particular, we focus on several classic issues in finance: the role of sentiment in news-driven stock movements; how stock markets respond to news with different content; and how firm characteristics react to media influence.

#### 4.4.1. Does sentiment matter

Previous studies have proven that event information in daily news articles is correlated with stock movements. Because investors are sentimental, as behavioral finance has asserted, several studies have investigated the impact of the sentiment information in response to the news on stock movements [34,35,40,41]. However, general sentiment words developed for psychology and sociology cannot function well in the realm of finance [25]. Determining whether finance-specific sentiment information in articles is useful and how it works with event information would be interesting.

In this study, we extract proper nouns to represent the event information in an article, and sentiment is represented by the finance-specific sentiment words in the article. We adopt FudanNLP, a state-of-the-art lexical analysis system in the Chinese language,[7] as our standard part-of-speech (POS) tagger to extract seven noun categories as proper noun sets from news articles. Sentiment words are detected based on a finance-specific sentiment word list obtained by the suggested algorithm (Section 3.2). Here, we performed a series of preliminary experiments to identify the optimal coefficient ($\lambda$) for smoothing. The best predictive performance is achieved when ($\lambda$) is set to 0.8.

As shown in Table 4, representing news articles with proper nouns can achieve a good directional prediction but attains a poor RMSE. In contrast, the representation based on finance-specific sentiment words can greatly improve the predictive performance, compared with those based on general sentiment words. In addition, we represent news articles with both finance-specific sentiment words and proper nouns. These representations achieve the best performance among the five methods. Therefore, we believe that using the finance-specific sentiment words along with proper nouns is an effective method for representing news articles in the quantitative analyses of media impact on stock markets.

#### 4.4.2. Does public mood matter

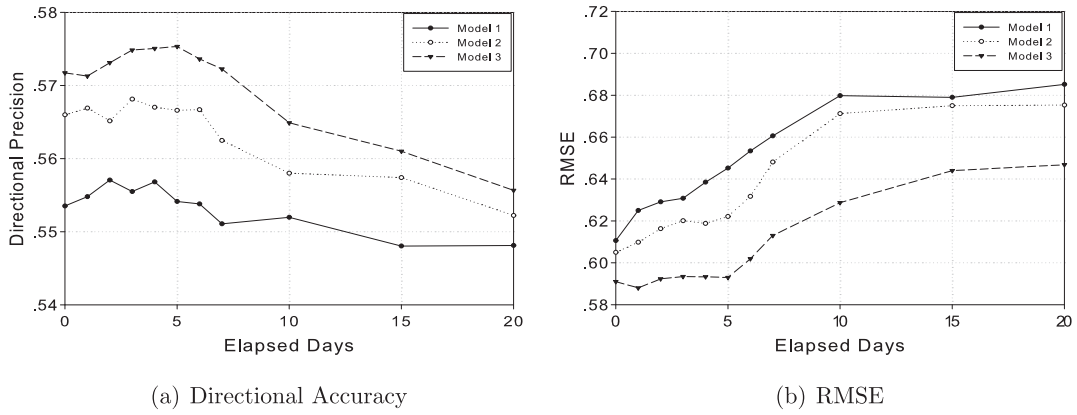
With the widespread adaption of user engagement in the media, investor decisions tend to be influenced by peer emotions, most likely leading to a heard behavior in investment events. To capture the public mood on a specific stock, we analyze firm-specific messages in two premier financial discussion boards from www.sina.com and www.eastmoney.com. Here, we are particularly interested in the pessimistic and optimistic attitudes of the public and their individual and integrated influence on stocks. Thus, we build three models to study these influences. For each stock, Model 1 takes the current stock price, term vectors of relevant news articles, and optimistic public mood ($M_s^+$) as inputs of the predictive model. Model 2 takes the current stock price, term vectors of news articles, and pessimistic public mood ($M_s^-$) as inputs. In Model 3, the inputs are the current stock price,term vectors of news articles, $M_s^+$, and $M_s^-$. The output of these models is the +26 min stock price (i.e., the stock price after a 26-min delay).

From Fig. 5, the pessimistic public mood has a significant contribution in predicting stock movements. Compared with a pessimist outlook, an optimistic public mood has limited power in sensing stock movements. These findings are consistent with a prior study in which the fraction of negative words in firm-specific news stories forecasted low firm earnings [41]. The joint influence of pessimism and optimism in public mood is noticeable, and the impact of public mood lasts for several days. In our experiments, incorporating the public mood from 3 to 6 days can further increase the stock predictive performance. With a time period of less than 3 days or greater than 6 days, the accumulative mood is not sufficient to assist stock predictions.

#### 4.4.3. Does news content matter

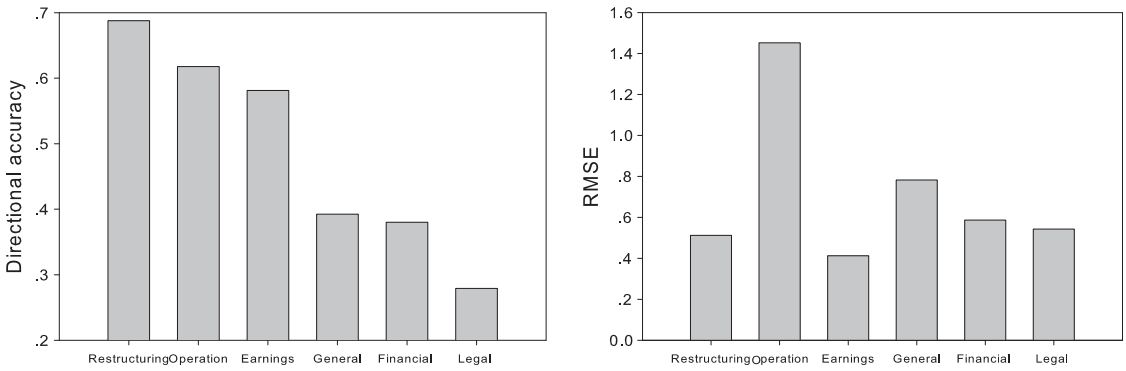
Previous studies have considered all types of news articles equally for predictive model training. However, certain types of news, such as an executive personnel change or a new product release, may be more influential.

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




(a) Directional Accuracy (b) RMSE

**Fig. 5.** Public mood.




(a) Directional Accuracy (b) RMSE

**Fig. 6.** Gist sensitivity.

Here, we classify our training articles into seven categories: corporate governance, earnings reports, and financial, general, legal, operational, and restructuring issues, as suggested by Antweiler and Frank [4]. This classification is achieved after manually labeling 3 months of data as training samples for the SVM classifier and using this trained classifier to automatically categorize other articles in the news corpus with a classification result that has a density of 0.9934 and an F-measure of 0.82. Then, we use the data from the first 9 months of 2011 for training and the data from the last 3 months of 2011 for stock prediction analyses. Because a small number of insufficient corporate governance news exists for classifier training, we only report the performance of the remaining six categories here.

As shown in Fig. 6, news articles related to restructuring issues, including joint venture starts, merges, and bought and sold units, are the most predictable, followed by operational issues (e.g., capacity up, contracts, lay-offs, new products, and supply agreements), earning reports, general issues (e.g., corrections, miscellaneous, and opinions), financial issues, and legal issues in terms of directional accuracy. Notably, although news articles regarding operational issues tend to be predicable based on directional accuracy, the predictive closeness is rather poor compared with that of other news content. We further examine the stocks affected by operational issues news, discovering that the prices of these relevant stocks fluctuate sharply in terms of standard deviation. This fluctuation may lead to a poor closeness while keeping good directional accuracy.

*4.4.4. Does firm matter*

Because articles with different content vary in predictability, understanding how firm characteristics change the impact of Web media is of great interest. In this section, we study the predictive results according to firm characteristics, including trading volume, turnover, price-to-earnings (P/E) ratio, price-to-book (P/B) ratio, risk ($\beta$), and industry sector. In particular, instead of focusing on each single firm, we group 89 firms into three clusters, small, medium, and large, in terms of firm characteristics. This process provides a good birds-eye view to depict the links between firm characteristics and the media influence on stocks. Notably, all of these firm characteristics are calculated according to the overall performance of 2010.

**Table 5**
Predictive performance with trading volume.

| | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Small | 1432 | 0.5605 | 1.2111 |
| Medium | 2258 | 0.5426 | 0.4151 |
| Large | 4081 | 0.5568 | 0.3049 |

**Table 6**
Predictive performance with trading amount.

| | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Small | 1539 | 0.5444 | 0.4990 |
| Medium | 3408 | 0.5465 | 0.8151 |
| Large | 2824 | 0.5666 | 0.2801 |

**Table 7**
Predictive performance with P/E ratio.

| | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Small | 3691 | 0.5515 | 0.4205 |
| Medium | 2617 | 0.5575 | 0.3856 |
| Large | 1463 | 0.5508 | 1.1175 |

**Table 8**
Predictive performance with P/B ratio.

| | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Small | 4002 | 0.5505 | 0.4166 |
| Medium | 2565 | 0.5713 | 0.3183 |
| Large | 1204 | 0.5248 | 1.261 |

Trading volume is the number of shares or contracts that are traded in the capital markets during a given period of time. Higher volume for a stock is an indicator of higher liquidity. As shown in Table 5, high trading volume stocks bring more attention in news reports, and the predictability of these news articles in stock movements is relatively reliable, with a directional accuracy of 0.5542 and RMSE of 0.3624. That is, eMAQT is sensitive to media coverage.

Turnover in the stock market refers to the total value of stocks traded during a specific period of time. As shown in Table 6, stocks with high turnover do not attract as much news coverage as the ones with high volume. However, the predictability of these news articles in stock movements is relatively reliable, compared with the stocks that have small or medium turnover.

The P/E ratio is a valuation ratio of a companys current share price compared to its per-share earnings. Generally, stocks that have higher (or more certain) forecast earnings growth have a higher P/E, and those that are expected to have lower (or riskier) earnings growth have a lower PE. Table 7 shows that stocks with medium P/E are more predicable because some undervalued or overvalued stock data lead to poor predictions.

The P/B ratio is used to compare a stocks market value to its book value and is calculated by dividing the current closing price of the stock by the latest quarters book value per share. A lower or higher P/B ratio could indicate that the stock is undervalued or overvalued. Table 8 shows that stocks with medium P/B are more predicable. A good reason for this finding is that news and properly valued stocks provide good predictability.

Here, we measure the risk of a stock using its beta ($\beta$). The $\beta$ of a stock is a number that describes the correlated volatility of an asset in relation to the volatility of the overall financial market. $\beta$ measures a systematic risk based on how returns co-move with the overall market. A $\beta$ greater than 1 indicates that the movement of the asset is generally in the same direction

**Table 9**
Predictive performance with risk ($\beta$).

| | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Small | 3129 | 0.5410 | 0.8017 |
| Medium | 2604 | 0.5603 | 0.4770 |
| Large | 2038 | 0.5635 | 0.3607 |

**Table 10**
Predictive performance with industry.

| Industry | No. of predictions | Directional accuracy | RMSE |
|---|---|---|---|
| Transport and storage | 328 | 0.5044 | 0.4388 |
| Manufacturing | 1145 | 0.5220 | 1.2600 |
| Construction | 309 | 0.5285 | 0.3653 |
| Mining | 766 | 0.5323 | 0.4054 |
| Finance | 4248 | 0.5618 | 0.4219 |
| Production and utility supply | **61** | **0.5599** | **0.2920** |
| General service | **58** | **0.5705** | **0.2088** |
| Real estate | **355** | **0.5773** | **0.2646** |
| Information technology | **405** | **0.6062** | **0.2593** |
| Social service | **7** | **0.6099** | **0.1637** |
| Wholesale and retail trade | **89** | **0.6476** | **0.2299** |

Bold indicates the consumer-related industries.

as but more than the movement of the market. A higher $\beta$ indicates that a stock is more volatile. For example, most high-tech, Nasdaq-based stocks have a $\beta$ greater than 1, offering the possibility of a higher rate of return but also posing more risk. Table 9 shows that the riskier stocks with high $\beta$ values are more predictable, proving that stocks that are strongly influenced by the daily news are volatile and risky.

Firm performance generally varies by industry. Here, we partitioned the firms into 10 industrial sectors, i.e., mining, real estate construction, transport and storage, finance, wholesale and retail trade, social service, information technology, manufacturing, general service, and production and utility supply. As shown in Table 10, stocks in consumer-related industries are more predicable because consumer news articles are more rigorously considered.

### 4.5. Investment experiments

In this study, we design and implement a quantitative eMedia-aware trader, utilizing media influence on stock markets. To evaluate trader performance, we compare our method with two classic trading strategies, i.e., top-N and simple moving average (SMA)) [19], and one state-of-the-art media-driven trader, AZFinText [35]. We set RMB10,000 (approximately USD1,630) as the investment budget and compare the daily earnings of these approaches in our 3-month evaluation period, during which the CSI Index was down by 5.21% from 2363 to 2240. Notably, we assume zero transaction cost, as previously described [7,21,36,45]. In fact, the transaction costs are effectively absorbed by increasing the volume of each transaction as long as we are generating a profit.[8]

Top-*N* and SMA are two practical and widespread trading approaches in the market [19].

- Top-*N* is a long-term strategy that assumes if a combination of stocks has performed well in the past, the same combination will perform well in the near future. Here, we invest in the top-*N* stocks that performed best over the period from January 1 to September 30, 2011, by buying each stock at the beginning of October 2011 and selling it at the end of the 3-month evaluation period. To determine the optimal *N* combinations, we conduct a series of preliminary experiments with different numbers of stocks, *N*, which we invest according to the performance over the last 9 months in a descending order. We find that the best performance is achieved when the top 30 stocks are selected for investment. As shown in Fig. 7, even with the optimal top-*N* combination, a small loss after the 3-month assessment time is still experienced.
- Different from the long-term strategy Top-*N*, SMA focuses on short-term transactions. The SMA strategy is triggered when an actual market stock price crosses through the daily moving average of the same stock by a certain margin threshold or penetration rate. Specifically, if the change is upward, stocks should be purchased; if it is downward, the stocks should be sold. To determine the optimal penetration rate, we conduct a series of preliminary experiments and determine that the optimal penetration rate for triggering a transaction is over 1% of the invested value. The performance of 3-month daily earnings is shown in Fig. 7. We find that no positive earnings with SMA are observed after the 3-month assessment time.

We also adopt a state-of-art, media-driven trading system, AZFinText [35], as our comparison. AZFinText applies an SVR model to capture the correlation between financial news and stock prices, and the model was improved by incorporating new sentiments to study media influence [36]. Specifically, a sentiment analysis tool, named Opinion Finder, was used to judge the sentiment polarity (good or bad) of a news article. In this study, we use the adjusted AZFinText approach as a benchmark. Due to the lack of Chinese language support for Opinion Finder, we implement a Chinese sentiment analyzer following the sentiment analysis principles of Opinion Finder [47]. Here, we set the optimal outlook prediction window to 23 min and the threshold to trigger transactions at 0.3% of the stock price from our preliminary experiments.

With increasing social media popularity, which cultivates user engagement in information dissemination, the proposed eMAQT further utilizes public emotion to sense stock movements. For this proposed investment trader, both short and long

[8] In Chinese stock markets, the average trading cost is approximately 0.2% and 0.05% of the invested value for long and short selling transactions, respectively.

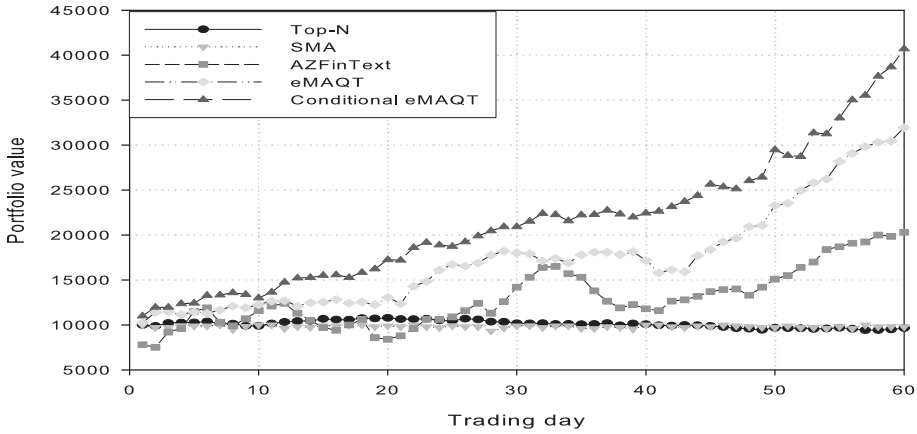



**Fig. 7.** Comparison.

selling strategies are applied to generate profits. Specifically, for long selling, if the predicted future price is greater than the threshold at the time at which an article is released, our trader immediately purchases the stock and then disposes the stock in 26 min. For short selling, if the prediction price is less than the threshold at the time that the article is released, we borrow the stocks and sell it, assuming that the stock price will be cheaper after 26 min for repurchase. To determine the threshold for triggering short and long selling, we perform a series of experiments with different stock price percentages as thresholds. The idealized optimal performance is achieved when we set the threshold to 0.3% of the stock price. In addition, rather than trading on all CSI100 firms, we apply eMAQT to selected firms in terms of firm characteristics (i.e., large trading volume, large amount, large risk and medium P/E, and medium P/B) and top 6 industries, with training and evaluation data that are limited to the best 3 news contents listed in Fig. 6. This conditional eMAQT is powered by the optimized firm combination for stock predictions.

Fig. 7 shows the daily stock returns of these five methods over the assessment period. The daily earnings are calculated based on the earning performance of the previous trading day. Comparing the change of −5.21% in CSI100 and the 103.23% return in AZFinText, the basic eMAQT yielded a remarkable return of 219.50% in three months, and the conditional eMAQT yielded a superior return of 307.10%. With reasonable conditions, the proposed quantitative media-aware trading strategy is quite promising to sense stock movements.

## 5. Conclusion and future work

In this article, we quantitatively investigated the impact of media on stock markets. The profits yielded by eMAQT further prove a previous assertion in finance that public information events are subject to differential interpretations by investors. This result presents profitable trading opportunities for skilled investors to generate a profit post the news release day [13]. That is, stock markets are sensitive to public information in an era of social media.

To identify the essential reasons how Web media influences the stock market, we studied the internal functions of sentiment, firm characteristics, and news content regarding the relation of Web media to stock markets. We found that the sentiment influence originates from two sources: the sentiment in a news article captured by the financial-specific sentiment words and revealed public feeling via postings and comments on financial discussion boards. Such sentiments in Web media cause investors emotions to fluctuate and intervene in their decision making. This result corroborates the basic assertion in behavioral finance that investors are sentimental [11]. With the growing popularity of Web 2.0, various social media outlets exist for readers to express their opinions, including blogs, tweets/microblogs, and social news. An investigation into these new sources to capture public mood is imperative.

Other enlightened findings herein include that Web medias influence on stocks varies by news content and firm characteristics. Specifically, stocks are sensitive to news articles on restructuring and earning issues. The firms strongly involved with public interests or daily life, especially with utility supply, real estate, social service, and wholesale and retail trade, are more predictable in stock markets. However, the article origin, whether official, leaked, or rumored, may have different influences on investors [46,50]. Indeed, further investigations of additional internal Web media functions and their impact on stock markets would be interesting.

## Acknowledgments

This work has been supported by National Natural Science Foundation of China (NSFC) (Nos. 60803106, 61170133 and 91218301), the Fundamental Research Funds for the Central Universities (Nos. JBK120505 and JBK1307200), Science Foundation for Youths of Sichuan Province (Nos. 2013JQ0004, 2011JQ0040, and 2011JTD0028).

## References


[1] A.S. Abrahams, J. Jiao, G.A. Wang, W.G. Fan, Vehicle defect discovery from social media, Decis. Support Syst. 54 (2012) 87–97.
[2] S. Alfarano, E. Camacho, A. Morone, The role of public and private information in a laboratory financial market, 2011. Working paper.
[3] T.G. Andersen, T. Bollerslev, F.X. Diebold, C. Vega, Real-time price discovery in global stock, bond and foreign exchange markets, J. Int. Econ. 73 (2007) 251–277.
[4] W. Antweiler, M.Z. Frank, Do US stock markets typically overreact to corporate news stories?, 2006, Working paper.
[5] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman Publisher, 1999. pp. 41–44.
[6] J. Bollen, A. Pepe, H. Mao, Twitter mood predicts the stock market, J. Comput. Sci. 2 (2011) 1–8.
[7] W.S. Chan, Stock price reaction to news and no-news: drift and reversal after headlines, J. Financ. Econ. 70 (2003) 223–260.
[8] C.H. Cheng, T.L. Chen, L.Y. Wei, A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Inform. Sci. 180 (2010) 1610–1629.
[9] Y.J. Choi, Y.H. Kim, S.H. Myaeng, Domain-specific sentiment analysis using contextual feature generation, in: Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, 2009, pp. 37–44.
[10] D. Cutler, J. Poterba, L. Summers, What moves stock prices?, J Portfolio Manage. 15 (1989) 56–63.
[11] J.B. DeLong, A. Shleifer, L.H. Summers, R.J. Waldmann, Noise trader risk in financial markets, J. Polit. Econ. 98 (1990) 703–738.
[12] A. Duric, F. Song, Feature selection for sentiment analysis based on content and syntax models, Decis. Support Syst. 53 (2012) 704–711.
[13] J.E. Engelberg, A.V. Reed, M.C. Ringgenberg, How are shorts informed?: short sellers, news, and information processing, J. Financ. Econ. 105 (2012) 260–278.
[14] E.F. Fama, Efficient capital markets: a review of theory and empirical work, J. Finance 25 (1970) 383–417.
[15] M.Z. Frank, W. Antweiler, Is all that talk just noise? The information content of internet stock message boards, J. Finance 59 (2004) 1259–1294.
[16] G.P.C. Fung, J.X. Yu, W. Lam, Stock prediction: integrating text mining approach using real-time news, in: Proceedings of IEEE International Conference on Computational Intelligence for Financial Engineering, IEEE, 2003, pp. 395–402.
[17] G. Gidofalvi, Using news articles to predict stock price movements, Department of Computer Science and Engineering, University of California, San Diego, 2001.
[18] E. Gilbert, K. Karahalios, Widespread worry and the stock market, in: Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 2010, pp. 1–8.
[19] F.E. James, Monthly moving averages – an effective investment tool?, J Financ. Quant. Anal. 3 (1968) 315–326.
[20] E. Keogh, S. Chu, D. Hart, M. Pazzani, An online algorithm for segmenting time series, in: Proceedings IEEE International Conference on Data Mining (ICDM), IEEE, 2001, pp. 289–296.
[21] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Language models for financial news recommendation, in: Proceedings of the 9th International conference on Information and Knowledge Management (CIKM), 2000, pp. 389–396.
[22] B. LeBaron, W.B. Arthur, R. Palmer, Time series properties of an artificial stock market, J. Econ. Dynam. Control 23 (1999) 1487–1516.
[23] F. Li, Do stock market investors understand the risk sentiment of corporate annual reports? 54 (2006), Working paper.
[24] Q. Li, J. Wang, Y.P. Chen, Z. Lin, User comments for news recommendation in forum-based social media, Inform. Sci. 180 (2010) 4929–4939.
[25] T. Loughran, B. McDonald, When is a liability is not a liability? Textual analysis, dictionaries, and 10-ks, J. Financ. 66 (2012) 35–65.
[26] D. Lowe, Three-dimensional object recognition from single two-dimensional images, Artif. Intell. 31 (1987) 355–395.
[27] X. Luo, J. Zhang, W. Duan, Social media and firm equity value, Inform. Syst. Res. 24 (2013) 146–163.
[28] M.A. Mittermayer, G.F. Knolmayer, Newscats: a news categorization and trading system, in: Proceedings of the 6th International Conference on Data Mining (ICDM), IEEE, 2006, pp. 1002–1007.
[29] 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, Int. J. Manage. Finance 5 (2009) 179–200.
[30] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval 2 (2008) 1–135.
[31] C. Plott, S. Sunder, Efficiency of experimental security markets with insider information: an application of rational-expectations models, J. Political Econ. 90 (1982) 663–698.
[32] P. Rosin, Techniques for assessing polygonal approximations of curves, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 659–666.
[33] R.P. Schumaker, H. Chen, Evaluating a news-aware quantitative trader: the effect of momentum and contrarian stock selection strategies, J. Am. Soc. Inform. Sci. Technol. 59 (2008) 247–255.
[34] R.P. Schumaker, H. Chen, A quantitative stock prediction system based on financial news, Inform. Process. Manage. 45 (2009) 571–583.
[35] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: the AZFin text system, ACM Trans. Inform. Syst. 27 (2009) 12:1–12:19.
[36] R.P. Schumaker, Y.L. Zhang, C.N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decis. Support Syst. 53 (2012) 458–464.
[37] Y.W. Seo, J.A. Giampapa, K.P. Sycara, Text classification for intelligent agent portfolio management, in: Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems, 2002, pp. 802–803.
[38] R.J. Shiller, Do stock prices move too much to be justified by subsequent changes in dividends?, Am Econ. Rev. 71 (1981) 421–436.
[39] A. Shleifer, R.W. Vishny, The limits of arbitrage, J. Finance 52 (1997) 35–55.
[40] P.C. Tetlock, Giving content to investor sentiment: the role of media in the stock market, J. Finance 62 (2007) 1139–1168.
[41] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: quantifying language to measure firms' fundamentals, J. Finance 63 (2008) 1437–1467.
[42] P.D. Turney, Measuring praise and criticism: inference of semantic orientation from association, ACM Trans. Inform. Syst. 21 (2003) 315–346.
[43] C. Vega, Stock price reaction to public and private information, J. Financ. Econ. 82 (2006) 103–133.
[44] P. Veronesi, Stock market overreactions to bad news in good times: a rational expectations equilibrium model, Rev. Financ. Stud. 12 (1999) 975.
[45] B. Wang, H. Huang, X. Wang, A novel text mining approach to financial time series forecasting, Neurocomputing 83 (2011) 136–145.
[46] 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, Decis. Support Syst. 54 (2013) 1442–1451.
[47] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: a system for subjectivity analysis, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP), 2005, pp. 34–35.
[48] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual Web data, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1998, pp. 2720–2725.
[49] Y. Yu, W. Duan, Q. Cao, The impact of social and conventional media on firm equity value: a sentiment analysis approach, Decis. Support Syst. (2013) Forthcoming.
[50] X.L. Zhu, S. Gauch, Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2000, pp. 288–295.