The impact of financial news and public mood on stock movements

Qing Li, Tiejun Wang, Qixu Gong
Southwestern University of Finance and Economics, China

Abstract

In an era of internet, investors can rapidly obtain more valuable and timely information. With such rapid information influx, investor decisions tend to be influenced by peer and public emotions. In this article, we propose a quantitative media-aware trading strategy to investigate the media impact on stock markets. Our main findings are (1) fundamental information of firm-specific news articles can enrich the knowledge of investors and affect their trading activities; (2) crowd 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.

Key words: Text Mining, Social Media, Financial news, Stock Market, Trading Strategy

1. **Introduction**

Even though efficient market hypothesis (EMH) [1] that claims investors are rationale enough is occupying the main stage of financial market, recent studies have demonstrated that investors' emotion does influence their decisions [2]. Further evidence is offered by Li [3] and Schumaker [4] who discovered that sentiment elements in financial news articles indicate stock returns. To systematically investigate Web media's impact on 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 related to the companies of 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. Second, we study the sentiment impact on stocks, especially the emotions evoked by public news. 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 ways. 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.

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 [3, 6, 7]. With technological advancements, 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.

[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 Shanghai and the Shenzhen securities market

However, 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 [8, 9, 10, 11, 12]. 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 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 Figure 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.

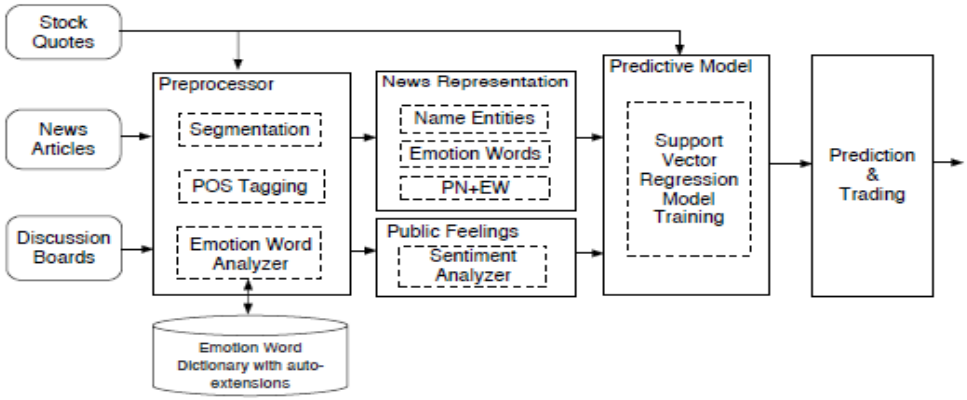


Figure 1: eMAQT: an eMedia-Aware Quantitive Trader.

3.1.1 Representation of news articles

For news items, we model an article as a weighted term vector, V, with a number of nouns and sentiment terms selected from the article. 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.

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. 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 addition, an emotionless word can be a sentiment one in the realm of finance to some degree. The word “bull” originally refers to a male bovine animal; it also indicates an upward trend in financial market, 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.

In this study, we incorporate 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* [14].

• *A firm-specific article with a positive (negative) tone is typically in accordance with an upward (downward) price trend of a relevant stock* [13].

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

$$\begin{aligned} P^{+}(w) &= P(w \mid E=+, T=\uparrow) \\ &\approx P(w \mid T=\uparrow) P(E=+ \mid w, T=\uparrow) \\ &= P(w \mid T=\uparrow) \sum_{i=0}^{M} P(e_i=+ \mid w, T=\uparrow), \end{aligned} \tag{1}$$

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, ↑ and ↓, indicating upward and downward price trends, respectively, M represents the total number of positive words and $\boldsymbol{e_i}$ is the sentiment word in the paradigm set $\boldsymbol{S}$[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 by using standard TF/IDF weighting schema [5][15]. Specifically, we define the weight of term t in document d as shown below:

$$w(t,d) = (0.5 + 0.5 \times \frac{f(t,d)}{\max_{t'} f(t',d)}) \times I(t) \tag{2}$$

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:

$$I(t) = \log \frac{C}{c(t)} \tag{3}$$

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

$$W(t) = \frac{w(t,d_0)}{\max_{t'} w(t',d_0)} \tag{4}$$

---

[2] 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

3.1.2. Public mood of a stock

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, www.sina.com and www.eastmoney.com.

Here, we measure the public mood on stock ***S*** from two aspects: optimism (M+) and pessimism (M −). The optimistic mood on stock ***S*** is measured as

$$M_S^+ = \sum_{i=0}^{\tau} \sum_{j=0}^{K} \frac{P_{i,j} \times W_j}{l_i} \times T_i \tag{5}$$

where Pi,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 the i-th day, Wj represents the weight of posting j, and Ti defines the time factor. The posting weight Wj 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:

$$W_j = \frac{c_j}{\max_{t'} c_{t'}} \tag{6}$$

where cj 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 [16], the time factor *Ti* is defined to tune the influence power with time passed,

$$T_i = e^{-i/\beta} \tag{7}$$

Here, *i* is the number of the passed days that we consider a sentiment to have continued influence, and β represents a constant for 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

$$M_S^- = \sum_{i=o}^{\tau} \sum_{j=0}^{K} \frac{N_{i,j} \times W_j}{l_i} \times T_i \tag{8}$$

3.2. **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 [17], relevance language model (RLM) [9], support vector machine (SVM) [11], and Naive Bayesian [18]. 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 [5].

4. **Experiments**

With an effective methodology to quantitatively analyze the mechanism of information percolation and its impact on stock markets, we are particularly interested in understanding 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 [19].

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.

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. Here, we target the two independent stock exchanges in mainland China: the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE). 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[3]. In this 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

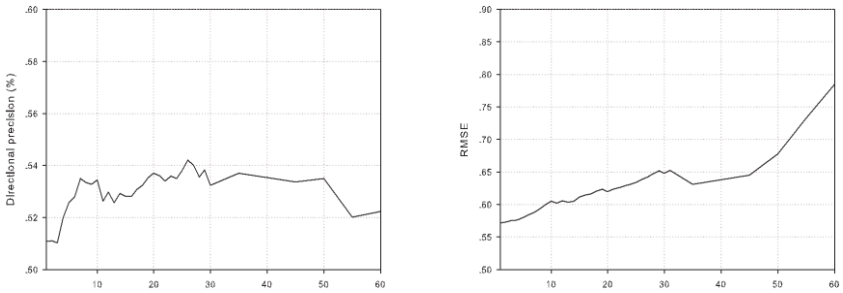
Directional accuracy and closeness are two classic evaluation metrics to measure the impact of media on stock movements. In this study, we chose both metrics as our evaluation standards. Specifically, directional accuracy measures the upward or downward direction of the predicted stock price compared 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

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

are particularly interested in the outlook time window of the predictive model. Gidofalvi [20] reported that a good outlook time window for a stock forecast is approximately 20 minutes after the release of relevant information. In our study, the directional accuracy increases and achieves a best performance of 0.5421 at the 26th minute after news release. Notably, we determine an optimal predictive outlook window of 26 minutes for the following experiments in terms of directional accuracy rather than RMSE (See Figure 1).



(a) Directional accuracy (b) RMSE

Figure 1: Predictive outlook window

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?

As shown in Table 1, representing news articles with proper nouns can achieve a good directional prediction but attains a poor RMSE. 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.

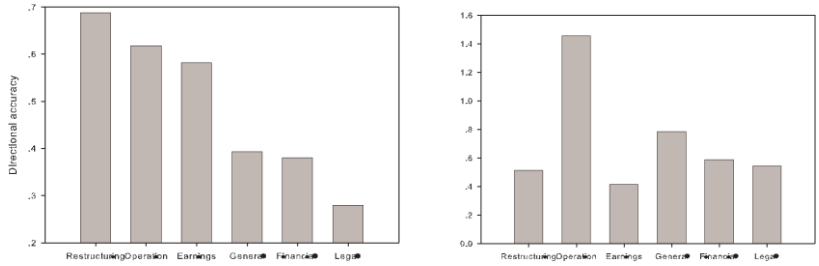
Table 1: *Representation *PN denotes that a news article is represented by a number of weighted proper nouns. Harvard represents an article by a number of weighted sentiment words from the Harvard-IV-4 list. FS defines an article as a number of weighted sentiment words from the finance-specific sentiment 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%

4.4.2. Does news content matter?

As shown in Figure 2, 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), 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.



(a) Directional Accuracy (b) RMSE

Figure 2: Content sensitivity

4.4.3. 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.

Table 2: Predictive performance with different 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

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 2, 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 0.5426 and RMSE of 0.3049. Hence, eMAQT is sensitive to media coverage.

4. Conclusion and future work

In this article, we quantitatively investigated the impact of media on stock markets. The prediction data given by eMAQT further prove a previous assertion in finance that public information events are subject to differential interpretations by investors. Specifically, stocks are sensitive to news articles on restructuring and earning issues. However, the article origin, whether official, leaked, or rumored, may have different influences on investors [21, 22]. Indeed, further investigations of additional internal Web media functions and their impact on stock markets would be interesting.

References

[1] E.F. Fama, Eficient capital markets: A review of theory and empirical work, Journal of Finance 25 (1970) 383-417.

[2] J.B. DeLong, A. Shleifer, L.H. Summers, R.J. Waldmann, Noise trader risk in financial markets, Journal of Political Economy 98 (1990) 703-738.

[3] F. Li, Do stock market investors understand the risk sentiment of corporate annual reports ?

(2006) 54. Working Paper.
[4] R.P. Schumaker, Y.L. Zhang, C.N. Huang, H. Chen, Evaluating sentiment in financial news articles, Decision Support Systems 53 (2012) 458-464.
[5] R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using breaking financial news: The AZFin text system, ACM Transactions on Information Systems 27 (2009) 12:1-12:19.
[6] J. Bollen, A. Pepe, H. Mao, Twitter mood predicts the stock market, Journal of Computational Science 2 (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 (2003) 223-260.
[8] A.S. Abrahams, J. Jiao, G.A. Wang, W.G. Fan, Vehicle defect discovery from social media, Decision Support Systems 54 (2012) 87-97.
[9] 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) (2000) 389-396.
[10] T. Loughran, B. McDonald, When is a liability is not a liability? Textual analysis, dictionaries, and 10-ks, Journal of Financial 66 (2012) 35-65.
[11] M.A. Mittermayer, G.F. Knolmayer, Newscats: A news categorization and trading system, Proceedings of the 6th International Conference on data mining (ICDM), IEEE, pp. 1002-1007.
[12] R.P. Schumaker, H. Chen, Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies, Journal of the American Society for Information Science and technology 59 (2008) 247-255.
[13] P.C. Tetlock, Giving content to investor sentiment: The role of media in the stock market, Journal of Finance 62 (2007) 1139-1168.
[14] P.D. Turney, Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems 21 (2003) 315-346.
[15] 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.
[16] P.C. Tetlock, M. Saar-Tsechansky, S. Macskassy, More than words: Quantifying language to measure firms' fundamentals, Journal of Finance 63 (2008) 1437-1467.
[17] C.H. Cheng, T.L. Chen, L.Y. Wei, A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences 180 (2010) 1610 - 1629.
[18] 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, pp. 802-803.
[19] F.E. James, Monthly moving averages - an efiective investment tool? Journal of Financial and Quantitative Analysis 3 (1968) 315-326.
[20] G. Gidofalvi, Using news articles to predict stock price movements, Department of Computer Science and Engineering, University of California, San Diego (2001).
[21] G.A.Wang, J. Jiao, A.S. Abrahams, W.G. Fan, Z.J. Zhang, Expert rank: A topic-aware expert finding algorithm for online knowledge communities, Decision Support Systems 54 (2013) 1442-1451.
[22] 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 1Research and Development in Information Retrieval (SIGIR), pp. 288-295.