# Exploiting Social Relations and Sentiment for Stock Prediction

**Jianfeng Si*** **Arjun Mukherjee†** **Bing Liu†** **Sinno Jialin Pan*** **Qing Li‡** **Huayi Li†**
* Institute for Infocomm Research, Singapore
{ thankjeff@gmail.com, jspan@i2r.a-star.edu.sg}
†Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
{ arjun4787@gmail.com, liub@cs.uic.edu, lhymvp@gmail.com}
‡ Department of Computer Science, City University of Hong Kong, Hong Kong, China
qing.li@cityu.edu.hk

## Abstract

In this paper we first exploit cash-tags ("$" followed by stocks' ticker symbols) in Twitter to build a stock network, where nodes are stocks connected by edges when two stocks co-occur frequently in tweets. We then employ a labeled topic model to jointly model both the tweets and the network structure to assign each node and each edge a topic respectively. This Semantic Stock Network (SSN) summarizes discussion topics about stocks and stock relations. We further show that social sentiment about stock (node) topics and stock relationship (edge) topics are predictive of each stock's market. For prediction, we propose to regress the topic-sentiment time-series and the stock's price time series. Experimental results demonstrate that topic sentiments from close neighbors are able to help improve the prediction of a stock markedly.

## 1 Introduction

Existing research has shown the usefulness of public sentiment in social media across a wide range of applications. Several works showed social media as a promising tool for stock market prediction (Bollen et al., 2011; Ruiz et al., 2012; Si et al., 2013). However, the semantic relationships between stocks have not yet been explored. In this paper, we show that the latent semantic relations among stocks and the associated social sentiment can yield a better prediction model.

On Twitter, cash-tags (e.g., $aapl for Apple Inc.) are used in a tweet to indicate that the tweet talks about the stocks or some other related information about the companies. For example, one tweet containing cash-tags: $aapl and $goog (Google Inc.), is *"$AAPL is loosing customers. everybody is buying android phones! $GOOG"*. Such joint mentions directly reflect some kind of latent relationship between the involved stocks, which motivates us to exploit such information for the stock prediction.

We propose a notion of Semantic Stock Network (SSN) and use it to summarize the latent semantics of stocks from social discussions. To our knowledge, this is the first work that uses cash-tags in Twitter for mining stock semantic relations. Our stock network is constructed based on the co-occurrences of cash-tags in tweets. With the SSN, we employ a labeled topic model to jointly model both the tweets and the network structure to assign each node and each edge a topic respectively. Then, a lexicon-based sentiment analysis method is used to compute a sentiment score for each node and each edge topic. To predict each stock's performance (i.e., the up/down movement of the stock's closing price), we use the sentiment time-series over the SSN and the price time series in a vector autoregression (VAR) framework.

We will show that the neighbor relationships in SSN give very useful insights into the dynamics of the stock market. Our experimental results demonstrate that topic sentiments from close neighbors of a stock can help improve the prediction of the stock market markedly.

## 2 Related work

### 2.1 Social Media & Economic Indices

Many algorithms have been proposed to produce meaningful insights from massive social media data. Related works include detecting and summarizing events (Weng and Lee, 2011; Weng et al., 2011; Baldwin et al., 2012; Gao et al., 2012) and analyzing sentiments about them (Pang and Lee, 2008; Liu, 2012), etc. Some recent literature also used Twitter as a sentiment source for stock market prediction (Bollen et al., 2011; Si et al., 2013). This paper extends beyond the correlation between social media and stock market, but fur-



*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1139–1145,
October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

ther exploits the social relations between stocks from the social media context.

Topic modeling has been widely used in social media. Various extensions of the traditional LDA model (Blei et al., 2003) has been proposed for modeling social media data (Wang et al., 2011, Jo and Oh, 2011; Liu et al., 2007; Mei et al., 2007; Diao et al., 2012). Ramage et al. (2009; 2011) presented a partially supervised learning model called Labeled LDA to utilize supervision signal in topic modeling. Ma et al. (2013) predicted the topic popularity based on hash-tags on Twitter in a classification framework.

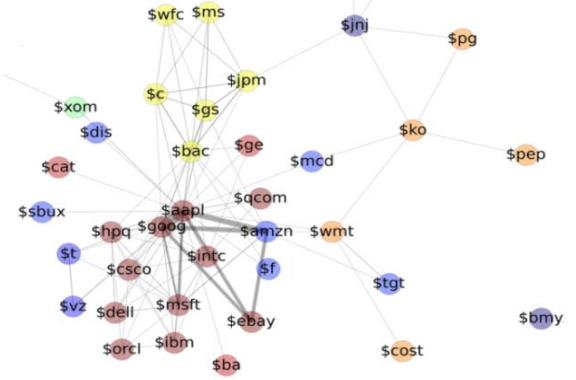



Figure 1. An example stock network.

### 2.2 Financial Networks for Stock

Financial network models study the correlations of stocks in a graph-based view (Tse et al., 2010; Mantegna, 1999; Vandewalle et al., 2001; Onnela et al., 2003; Bonanno et al., 2001). The usual approach is to measure the pairwise correlation of stocks' historical price series and then connect the stocks based on correlation strengths to build a correlation stock network (CSN).

However, our approach leverages social media posts on stock tickers. The rationale behind is that micro-blogging activities have been shown to be highly correlated with the stock market (Ruiz et al., 2012; Mao et al., 2012). It is more informative, granular to incorporate latest developments of the market as reflected in social media instead of relying on stocks' historical price.

## 3 Semantic Stock Network (SSN)

### 3.1 Construction of SSN

We collected five months (Nov. 2 2012 - Apr. 3 2013) of English tweets for a set of stocks in the Standard & Poor's 100 list via Twitter's REST API, using cash-tags as query keywords. For preprocessing, we removed tweets mentioning more than five continuous stock tickers as such tweets usually do not convey much meaning for our task. Finally, we obtained 629,977 tweets in total. Table 1 shows the top five most frequent stocks jointly mentioned with Apple Inc. in our dataset. Formally, we define the stock network as an undirected graph $G = \{V, E\}$. The node set $V$ comprises of stocks, $e_{u,v} \in E$ stands for the edge between stock nodes $u$ and $v$ and the edge weight is the number of co-occurrences. On exploring the co-occurrence statistics in pilot studies, we set a minimum weight threshold of 400 to filter most non-informative edges. Figure 1 demonstrates a segment of the stock network constructed from our dataset.

| $goog | $amzn | $ebay | $msft | $intc |
|---|---|---|---|---|
| 43263 | 23266 | 14437 | 11891 | 2486 |

Table 1. co-occurrence statistics with $aapl.

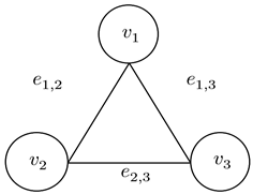



Figure 2. Tweet label design.

### 3.2 Semantic Topics over the Network

Figure 2 illustrates our annotation for each tweet. For a tweet, $d$ with three cash-tags: $\{v_1, v_2, v_3\}$, we annotate $d$ with the label set, $L_d = \{v_1, v_2, v_3, e_{1,2}, e_{1,3}, e_{2,3}\}$. ($e_{1,2}$ is "aapl_goog" if $v_1$is "aapl" and $v_2$ is "goog"). Then, the topic assignments of words in $d$ are constrained to topics indexed by its label set, $L_d$. Given the annotations as labels, we use the Labeled LDA model (Ramage et al., 2009) to jointly learn the topics over nodes and edges. Labeled-LDA assumes that the set of topics are the distinct labels in a labeled set of documents, and each label corresponds to a unique topic. Similar to LDA (Blei et al., 2003), Labeled-LDA models each document as an admixture of latent topics and generates each word from a chosen topic. Moreover, Labeled-LDA incorporates supervision by simply constraining the model to use only those topics that correspond to a document's observed label set (Ramage et al., 2009). For model inference, we use collapsed Gibbs sampling (Bishop, 2006) and the symmetric Dirichlet Priors are set to: $\eta = 0.01, \alpha = 0.01$ as suggested in (Ramage et al., 2010). The Gibbs Sampler is given as:

$$p(z_i = k|z_{-i}) \sim \frac{N(d_i,k)_{-1} + \alpha}{N(d_i,*)_{-1} + |L_{d_i}|*\alpha} * \frac{N(k,w_i)_{-1} + \eta}{N(k,*)_{-1} + |V|*\eta} \quad (1)$$

where $N(d_i, k)$ is the number of words in $d_i$ assigned to topic $k$, while $N(d_i,*)$ is the marginalized sum. $|L_{d_i}|$ is the size of label subset of $d_i$.

$N(k,w)$ is the term frequency of word $w$ in topic $k$. $|V|$ is the vocabulary size. The subscript $_{-1}$ is used to exclude the count assignment of the current word $w_i$. The posterior on the document's topic distribution $\{\theta_{d,k}\}$ and topic's word distribution $\{\beta_{k,w}\}$ can be estimated as follows:

$$\theta_{d,k} = \frac{N(d_i,k)+\alpha}{N(d_i,*)+|L_{d_i}|*\alpha} \quad (2)$$

$$\beta_{k,w} = \frac{N(k,w_i)+\eta}{N(k,*)+|V|*\eta} \quad (3)$$

Later, parameters $\{\beta_{k,w}\}$ will be used to compute the sentiment score for topics.

### 3.3 Leveraging Sentiment over SSN for Stock Prediction

We define a lexicon based sentiment score in the form of opinion polarity for each node-indexed and edge-indexed topic as follows:

$$S(k) = \sum_{w=1}^{|V|} \beta_{k,w}\, l(w),\ S(k) \in [-1,1] \quad (4)$$

where $l(w)$ denotes the opinion polarity of word $w$. $\beta_{k,w}$ is the word probability of $w$ in topic $k$ (Eq.3). Based on an opinion lexicon $O$, $l(w)$ = +1 if $w \in O_{pos}$, $l(w)$ = -1 if $w \in O_{neg}$ and $l(w)$ = 0 otherwise. We use the opinion English lexicon contributed by Hu and Liu (2004).

Considering the inherent dynamics of both the stock markets and social sentiment, we organize the tweets into daily based sub-sets according to their timestamps to construct one $SSN_t$ ($t \in [1,T]$) for each day. Then, we apply a Labeled LDA for each $SSN_t$ and compute the sentiment scores for each $SSN_t$'s nodes and edges. This yields a sentiment time series for the node, $v$, $\{S(v)_1, S(v)_2, \ldots, S(v)_T\}$ and for the edge, $e_{u,v}$, $\{S(e_{u,v})_1, S(e_{u,v})_2, \ldots, S(e_{u,v})_T\}$. We introduce a vector autoregression model (VAR) (Shumway and Stoffer, 2011) by regressing sentiment time series together with the stock price time series to predict the up/down movement of the stock's daily closing price.

As usual in time series analysis, the regression parameters are learned during a training phase and then are used for forecasting under sliding windows, i.e., to train in period $[t, t+w]$ and to predict on time $t+w+1$. Here the window size $w$ refers to the number of days in series used in model training. A VAR model for two variables $\{x_t\}$ and $\{y_t\}$ can be written as:

$$y_t = \sum_{i=1}^{lag} (\vartheta_i^x x_{t-i} + \vartheta_i^y y_{t-i}) + \varepsilon_t \quad (5)$$

where $\{\varepsilon\}$ are white noises, $\{\vartheta\}$ are model parameters, and $lag$ notes the time steps of historical information to use. In our experiment, $\{y_t\}$ is the target stock's price time series, $\{x_t\}$ is the covariate sentiment/price time series, and we will try $lag \in \{2,3\}$. We use the "dse" library in R language to fit our VAR model based on least square regression.

## 4 Experiments

### 4.1 Tweets in Relation to the Stock Market

Micro-blogging activities are well correlated with the stock market. Figure 3 shows us how the Twitter activities response to a report announcement of $aapl (Jan. 23 2013). The report was made public soon after the market closed at 4:00pm, while the tweets volume rose about two hours earlier and reached the peak at the time of announcement, then it arrived the second peak at the time near the market's next opening (9:30am). By further accumulating all days' tweet volume in our dataset as hourly based statistics, we plot the volume distribution in Figure 4. Again, we note that trading activities are well reflected by tweet activities. The volume starts to rise drastically two or three hours before the market opens, and then reaches a peak at 9:00pm. It drops during the lunch time and reaches the second peak around 2:00pm (after lunch). Above observations clearly show that market dynamics are discussed in tweets and the content in tweets' discussion very well reflects the fine-grained aspects of stock market trading, opening and closing.

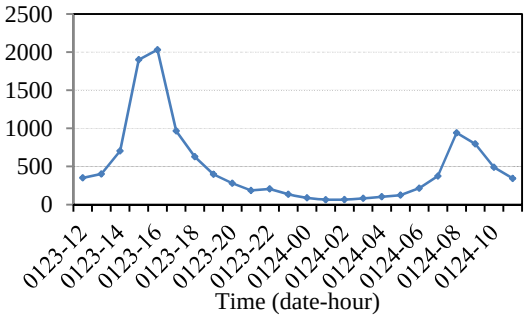



Figure 3. Tweet activity around $aapl's earnings report date on Jan. 23 2013.

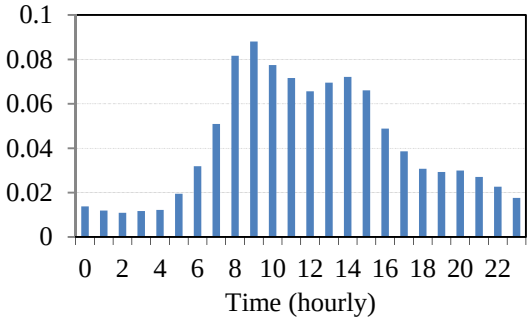



Figure 4. Tweet volume distribution in our data over hours averaged across each day.

### 4.2 Stock Prediction

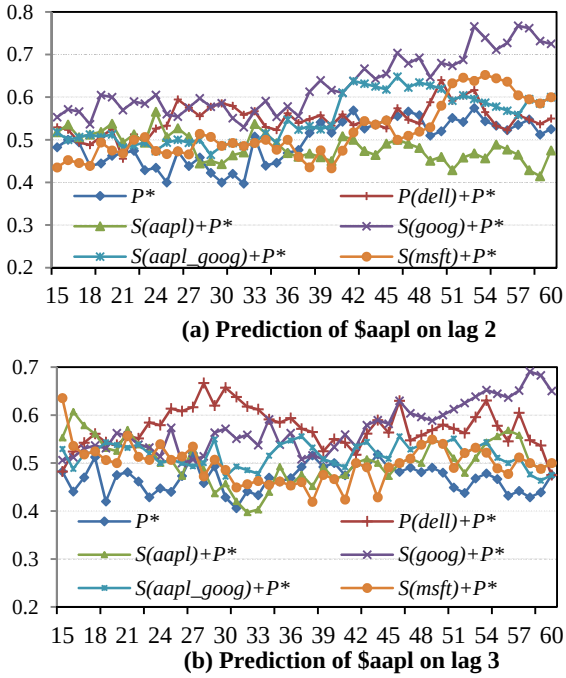
This section demonstrates the effectiveness of our SSN based approach for stock prediction. We leverage the sentiment time-series on two kinds of topics from SSN: 1). Node topic from the target stock itself, 2). Neighbor node/edge topics. We note that the price correlation stock network (CSN) (e.g., Bonanno et al., 2001; Mantegna, 1999) also defines neighbor relationships based on the Pearson's correlation coefficient (Tse et al., 2010) between pair of past price series (We get the stock dataset from Yahoo! Finance, between Nov. 2 2012 and Apr. 3 2013).

We build a two variables VAR model to predict the movement of a stock's daily closing price. One variable is the price time series of the target stock ($\{y_t\}$ in Eq.5); another is the covariate sentiment/price time series ($\{x_t\}$ in Eq.5). We setup two baselines according to the sources of the covariate time series as follows:

1. Covariate price time series from CSN, we try the price time series from the target stock's closest neighbor which takes the maximum historical price correlation in CSN.
2. With no covariate time series, we try the target stock's price only based on the univariate autoregression (AR) model.

To summarize, we try different covariate sentiment ($S(.)$) or price ($P(.)$) time series from SSN or CSN together with the target stock's price time series ($P^*$) to predict the movement of one day ahead price ($P^{**}$). The accuracy is computed based on the correctness of the predicted directions as follows, i.e., if the prediction $P^{**}$ takes the same direction as the actual price value, we increment $\#(posPred)$ by 1, $\#(totalTest)$ is the total number of test.

$$Accuracy = \frac{\#(posPred)}{\#(totalTest)} \quad (6)$$




**(a) Prediction of $aapl on lag 2**




**(b) Prediction of $aapl on lag 3**

Figure 5. Prediction on $aapl. (x-axis is the training window size, y-axis is the prediction accuracy) with different covariate sources.

| | Source | Lag = 2 | Lag = 3 |
|---|---|---|---|
| $P^*$ only | self | 0.49(0.57) | 0.47(0.52) |
| **CSN: $P(.)+P^*$** | dell | 0.55(0.64) | 0.57(0.67) |
| **SSN: $S(.)+P^*$** | aapl | 0.48(0.56) | 0.50(0.61) |
| | **goog** | **0.62(0.78)** | **0.57(0.69)** |
| | aapl_goog | 0.55(0.65) | 0.52(0.56) |
| | msft | 0.52(0.65) | 0.54(0.61) |

Table 2. Performance comparison of the average and best (in parentheses) prediction accuracies over all training window sizes for prediction on $aapl.

Figure 5 details the prediction of $aapl on different training window sizes of [15, 60] and lags. $\{S(aapl), S(goog), S(msft), S(aapl\_goog)\}$ are from SSN, $P(dell)$ is from CSN ($dell (Dell Inc.) takes the maximum price correlation score of 0.92 with $aapl), and $P^* = P(aapl)$ is the univariate AR model, using the target stock's price time series only. Table 2 further summarizes the performance comparison of different approaches reporting the average (and best) prediction accuracies over all time windows and different lag settings. Comparing to the univariate AR model ($P^*$ only), we see that the sentiment based time-series improve performances significantly. Among SSN sentiment based approaches, the $S(goog)$ helps improve the performance mostly and gets the best accuracy of 0.78 on $lag$ 2 and training window size of 53. On average, $S(goog)$ achieves a net gain over $S(aapl)$ in the range of 29% with lag 2 (0.62 = 1.29 x 0.48) and 14% with lag 3 (0.57 = 1.14 x 0.50). Also, $S(aapl\_goog)$ performs better than $S(aapl)$. The result indicates that $aapl's stock performance is highly influenced by its competitor. $P(dell)$ also performs well, but we will see relationships from CSN may not be so reliable.

We further summarize some other prediction cases in Table 3 to show how different covariate sentiment sources ($S(.)$) and price sources ($P(.)$) from their closest neighbor nodes help predict their stocks, which gives consistent conclusions. We compute the $t$-test for SSN based prediction accuracies against that of CSN or price only based approaches among all testing

window sizes ([15, 60]), and find that SSN based approaches are significantly ($p$-value < 0.001) better than others.

We note that tweet volumes of most S&P100 stocks are too small for effective model building, as tweets discuss only popular stocks, other stocks are not included due to their deficient tweet volume.

We make the following observations:

1. CSN may find some correlated stock pairs like $ebay and $amzn, $wmt and $tgt, but sometimes, it also produces pairs without real-world relationships like $tgt and $vz, $qcom and $pfe, etc. In contrast, SSN is built on large statistics of human recognition in social media, which is likely to be more reliable as shown.

2. Sentiment based approaches $\{S(\cdot)\}$ consistently perform better than all price based ones $\{P^*,\ P(\cdot)\}$. For $S(\cdot)$ based predictions, sentiment discovered from the target stock's closest neighbors in SSN performs best in general. This empirical finding dovetails with qualitative results in the financial analysis community (Mizik & Jacobson, 2003; Porter, 2008), where companies' market performances are more likely to be influenced by their competitors. But for Google, its stock market is not so much influenced by other companies (it gets the best prediction accuracy on $S(goog)$, i.e., the internal factor). It can be explained by Google Inc.'s relatively stable revenue structure, which is well supported by its leading position in the search engine market.

3. The business of offline companies like Target Corp. ($tgt) and Wal-Mart Stores Inc. ($wmt) are highly affected by online companies like $amzn. Although competition exists between $tag and $wmt, their performances seem to be affected more by a third-party like $amzn (In Table 3, $S(amzn)$ predicts the best for both). Not surprisingly, these offline companies have already been trying to establish their own online stores and markets.

## 5 Conclusion

This paper proposed to build a stock network from co-occurrences of ticker symbols in tweets. The properties of SSN reveal some close relationships between involved stocks, which provide good information for predicting stocks based on social sentiment. Our experiments show that SSN is more robust than CSN in capturing the neighbor relationships, and topic sentiments from close neighbors of a stock significantly improve the prediction of the stock market.


## Acknowledgments

This work was supported in part by a grant from the National Science Foundation (NSF) under grant no. IIS-1111092).


| Target | *lag* | $P^*$ only | CSN: $P(.)+P^*$ | SSN: $S(.)+P^*$ | | |
|---|---|---|---|---|---|---|
| *goog* | | | *dis*(0.96) | ***goog*** | *aapl* | *amzn* |
| | 2 | 0.48(0.59) | 0.53(0.60) | **0.59(0.65)** | 0.44(0.53) | 0.42(0.49) |
| | 3 | 0.46(0.54) | 0.53(0.62) | **0.56(0.67)** | 0.50(0.59) | 0.43(0.49) |
| *amzn* | | | *csco*(0.90) | *amzn* | *goog* | ***msft*** |
| | 2 | 0.48(0.54) | 0.48(0.55) | 0.47(0.54) | 0.57(0.66) | **0.60(0.68)** |
| | 3 | 0.46(0.53) | 0.49(0.53) | 0.43(0.50) | 0.55(0.63) | **0.57(0.66)** |
| *ebay* | | | *amzn*(0.81) | *ebay* | ***amzn*** | *goog* |
| | 2 | 0.49(0.55) | 0.51(0.57) | 0.44(0.53) | **0.57(0.64)** | 0.56(0.62) |
| | 3 | 0.48(0.58) | 0.49(0.54) | 0.45(0.58) | **0.54(0.64)** | 0.54(0.61) |
| *tgt* | | | *vz*(0.88) | *tgt* | *wmt* | ***amzn*** |
| | 2 | 0.43(0.53) | 0.43(0.54) | 0.46(0.55) | 0.49(0.56) | **0.49(0.59)** |
| | 3 | 0.44(0.50) | 0.40(0.53) | 0.44(0.48) | 0.41(0.48) | **0.48(0.54)** |
| *wmt* | | | *tgt*(0.86) | *wmt* | *tgt* | ***amzn*** |
| | 2 | 0.53(0.59) | 0.53(0.63) | 0.52(0.61) | 0.52(0.60) | **0.60(0.65)** |
| | 3 | 0.53(0.64) | 0.48(0.57) | 0.55(0.66) | 0.48(0.58) | **0.58(0.66)** |
| *qcom* | | | *pfe*(0.88) | *qcom* | *aapl* | ***intc*** |
| | 2 | 0.53(0.6) | 0.55(0.63) | 0.57(0.61) | 0.46(0.54) | **0.63(0.70)** |
| | 3 | 0.54(0.61) | 0.48(0.55) | 0.56(0.65) | 0.51(0.61) | **0.61(0.67)** |

Table 3. Average and best (in parentheses) prediction accuracies (over window sizes of [15, 60]) of some other cases with different covariates, cell of *dis*(0.96) means "$dis" takes the maximum price correlation strength of 0.96 with "$goog" (similar for others in column CSN). The best performances are highlighted in **bold**.

## References


Baldwin T., Cook P., Han B., Harwood A., Karunasekera S., and Moshtaghi M. 2012. A support platform for event detection using social intelligence. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 69-72.

Bishop C.M. 2006. Pattern Recognition and Machine Learning. Springer.

Blei D., NG A., and Jordan M. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3:993-1022.

Bollen J., Mao H., and Zeng X.J. 2011. Twitter mood predicts the stock market. Journal of Computer Science 2(1):1-8.

Bonanno G., Lillo F., and Mantegna R.N. 2001. High-frequency cross-correlation in a set of stocks, Quantitative Finance, Taylor and Francis Journals, vol. 1(1), 96-104.

Cohen J., Cohen P., West S.G., and Aiken L.S. 2003. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, (3rd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates.

Diao Q., Jiang J., Zhu F., and Lim E.P. 2012. Finding bursty topics from microblogs. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 (ACL '12), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 536-544.

Gao W., Li P., and Darwish K. 2012. Joint topic modeling for event summarization across news and social media streams. CIKM 2012: 1173-1182

Hu M. and Liu B. 2004. Mining and summarizing customer reviews. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 22-25. Seattle, Washington (KDD-2004).

Jo Y. and Oh A. 2011. Aspect and sentiment unification model for online review analysis. In ACM Conference in Web Search and Data Mining (WSDM-2011).

Liu B. 2012. Sentiment analysis and opinion mining. Morgan & Claypool Publishers.

Liu Y., Huang X., An A., and Yu X. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 607-614. ACM, New York, NY.

Ma Z., Sun A., and Cong G. 2013. On predicting the popularity of newly emerging hashtags in Twitter. In Journal of the American Society for Information Science and Technology, 64(7): 1399-1410 (2013)

Mantegna R. 1999. Hierarchical structure in financial markets, The European Physical Journal B - Condensed Matter and Complex Systems, Springer, vol. 11(1), pages 193-197, September.

Mao Y., Wei W., Wang B., and Liu B. 2012. Correlating S&P 500 stocks with Twitter data. In Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial '12). ACM, New York, NY, USA, 69-72

Mei Q., Ling X., Wondra M., Su H., and Zhai C. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of International Conference on World Wide Web (WWW-2007).

Mizik N. and Jacobson R. 2003. Trading off between value creation and value appropriation: The financial implications of shifts in strategic emphasis. Journal of Marketing, 63-76.

Onnela J.P., Chakraborti A., and Kaski K. 2003. Dynamics of market correlations: taxonomy and portfolio analysis, Phys. Rev. E 68, 056110.

Pang B. and Lee L. 2008. Opinion Mining and Sentiment Analysis. Now Publishers Inc.

Porter M.E. 2008. The Five Competitive Forces That Shape Strategy.HBR, Harvard Business Review.

Ramage D., Dumais S.T., and Liebling D. 2010. Characterizing microblogging using latent topic models. In Proceedings of ICWSM 2010.

Ramage D., Hall D., Nallapati R., and Manning C.D. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009).

Ramage D., Manning C.D., and Dumais S.T. 2011. Partially labeled topic models for interpretable text mining. In Proceedings of KDD 2011

Ruiz E.J., Hristidis V., Castillo C., Gionis A., and Jaimes A. 2012. Correlating financial time series with micro-blogging activity. In Proceedings of the fifth ACM international conference on Web search and data mining, pp. 513-522. ACM Press, NY (WSDM-2012).

Shumway R.H. and Stoffer D.S. 2011. Time Series Analysis and Its Applications: With R Examples, 3rd ed.

Si J., Mukherjee A., Liu B., Li Q., Li H., and Deng X. 2013. Exploiting Topic based Twitter Sentiment for Stock Prediction. In Proceedings of the 51st

Annual Meeting of the Association for Computational Linguistics. ACL'13, Sofia, Bulgaria, 24-29.

Tse C.K., Liu J., and Lau F.C.M. 2010. A network perspective of the stock market, Journal of Empirical Finance. 17(4): 659-667.

Vandewalle N., Brisbois F., and Tordoir X. 2001. Self-organized critical topology of stock markets, Quantit. Finan., 1, 372–375.

Wang X., Wei F., Liu X., Zhou M., and Zhang M. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. CIKM 2011: 1031-1040

Weng J. and Lee B.S. 2011. Event Detection in Twitter. In Proceedings of the International AAAI Conference on Weblogs and Social Media 2011.

Weng J.Y., Yang C.L., Chen B.N., Wang Y.K., and Lin S.D. 2011. IMASS: An Intelligent Microblog Analysis and Summarization System. ACL (System Demonstrations) 2011: 133-138.