Web Media and Stock Markets : A Survey and Future Directions from a Big Data Perspective

Qing Li, *Member, IEEE*, Yan Chen, Jun Wang, Yuanzhu Chen, *Member, IEEE*, and Hsinchun Chen, *Fellow, IEEE*

Abstract—Stock market volatility is influenced by information release, dissemination, and public acceptance. With the increasing volume and speed of social media, the effects of Web information on stock markets are becoming increasingly salient. However, studies of the effects of Web media on stock markets lack both depth and breadth due to the challenges in automatically acquiring and analyzing massive amounts of relevant information. In this study, we systematically reviewed 229 research articles on quantifying the interplay between Web media and stock markets from the fields of Finance, Management Information Systems, and Computer Science. In particular, we first categorized the representative works in terms of media type and then summarized the core techniques for converting textual information into machine-friendly forms. Finally, we compared the analysis models used to capture the hidden relationships between Web media and stock movements. Our goal is to clarify current cutting-edge research and its possible future directions to fully understand the mechanisms of Web information percolation and its impact on stock markets from the perspectives of investors cognitive behaviors, corporate governance, and stock market regulation.

Index Terms—Computing methodologies, text mining, financial market, stocks, big data, social media, news

1 Introduction

In traditional finance, the efficient market hypothesis states that a stock price is always driven by "unemotional" investors to equal the firm's rational present value of expected future cash flows [1]. Specifically, stock investors are constantly adjusting their beliefs on the potential market performances of stocks, although they typically disagree on the matter. This disagreement among competing market participants leads to discrepancies between the actual price and the intrinsic value, causing a stock price to fluctuate around a stock's intrinsic value [2], i.e., new information has intricate influences on asset prices. Recent behavioral finance studies have attributed the non-randomness of stock movements, such as overreactions to unfavorable news, to investors' cognitive and emotional biases [3], [4]. Although traditional finance and modern behavioral finance have different views on how information shapes stock movements, both believe that the volatility of the stock market comes from the release, dissemination and absorption of information.

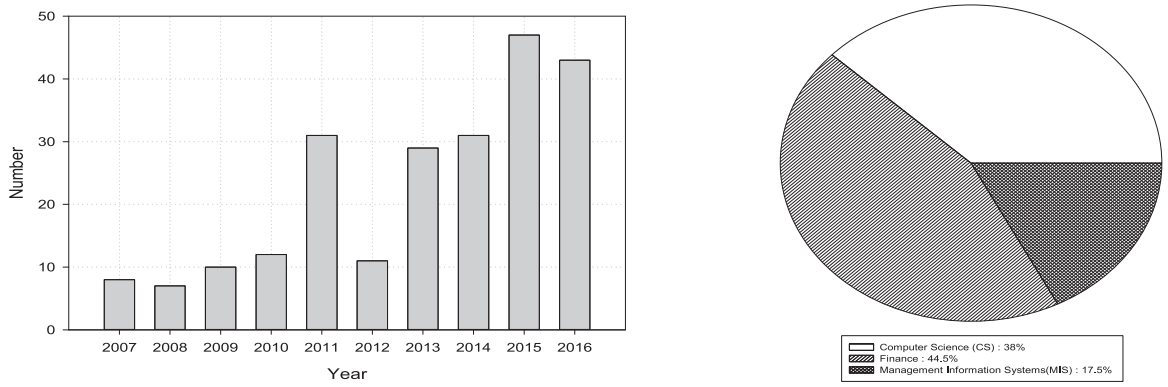
- *Q. Li, Y. Chen, and J. Wang are with the Financial Intelligence and Financial Engineering Sichuan Key Laboratory and the School of Information, Southwestern University of Finance and Economics, Sichuan Sheng 610072, China. E-mail: liq_t@swufe.edu.cn, {504473715, 550599532}@qq.com.*
- *Y. Chen is with the Department of Computer Science, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada. E-mail: yzchen@mun.ca.*
- *H. Chen is with the Department of Management Information System, University of Arizona, Tucson, AZ 85721 USA, and also with the Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Haidian 100084, China. E-mail: hchen@eller.arizona.edu.*

Manuscript received 21 Dec. 2016; revised 3 Oct. 2017; accepted 9 Oct. 2017. Date of publication 16 Oct. 2017; date of current version 9 Jan. 2018. (Corresponding author: Qing Li.) Recommended for acceptance by L. Dong. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2017.2763144

Due to technological advances, the Web has evolved from a technical framework for information dissemination to more of an enabler of social interactions among its users [5]. In particular, traditional news has evolved into various forms of social media, including blogs, tweets/microblogs, discussion boards, and social news. With such broad communication channels, investors can rapidly reach more valuable and timely information. Moreover, the adoption of user engagement in social media effectively magnifies the information contained in the news via comments, votes, and so forth. Such vibrant information creation, sharing, and collaboration among Web users make its impact on stock markets increasingly prominent. A good example is the negative market reaction to a fake tweet about Barack Obama being injured. Specifically, on April 23, 2013, a posting on the Twitter account of the Associated Press reported explosions at the White House that had injured President Obama. This tweet briefly roiled financial markets and caused the Dow Jones Industrial Average to tumble 100 points within 2 minutes.

Web media has become a substantial threat and a critical destabilizing factor that affects the stability of stock markets.

The influence of Web media on stock markets has increased considerably because of its exponentially increased volume and rapid dissemination. It is a substantial challenge to understand the mechanism of Web information percolation and its impact on financial markets. Due to the observation of stock price fluctuations with news feeds, researchers have begun to understand the connections between stock markets and media. The earliest studies relied primarily on empirical research of special cases or linear regression models that simplify the impact of media into the number of news articles instead of their textual content [6], [7]. With technological advancements in natural language processing (NLP) and



(a) Year (b) Category

Fig. 1. Publication Overview.

artificial intelligence (AI), researchers have begun to capture the influence of media on stocks and bridge these connections from a big data perspective. Specifically, to quantify textual influences, NLP techniques, including sentiment analysis and part-of-speech (POS) tagging, are used to extract valuable knowledge from textual media, and AI techniques, including support vector regression (SVR), convolutional neural network (CNN) and tensor-based learning algorithms, are utilized to capture, as precisely as possible, the relationships between high-dimensional market information and stock movements. [8] conducted one of the pilot studies by extracting the sentiment polarity of news articles in Wall Street Daily in terms of the portion of emotional words and found that negative news was statistically associated with downward pressure on the relevant stocks. Meanwhile, [9] discovered that the message sentiment on Yahoo! financial discussion boards was related to stock movements. Subsequently, [10] represented news articles with proper nouns via a POS tagging technique and bridged the connections between breaking news and stock prices using an SVR model. [11] measured public sentiment expressed in tweets and discovered its connection with stock price trends using self-organizing fusion neural networks (SOFNNs). [12] proposed a tensor-based predictive framework to model high-dimensional market information and its intrinsic links to study media-aware stock movements. This series of studies opened up new avenues for understanding the mechanism by which Web information impacts financial markets from a big data perspective.

Studies on media-aware stock movements, which originated in the finance field, have gradually attracted increasing numbers of researchers from management information systems (MIS) and computer science (CS). Fig. 1 shows the number of relevant publications in the past ten years and their distribution in terms of the research field. With the development of information technology, studies quantifying the impacts of Web media on stocks are increasingly common in all of the related research areas. In this work, we investigated 229[1] research articles on this classical problem in the fields of finance, MIS and CS and endeavored to clarify contemporary cutting-edge research and its possible future directions.

1. The full list is available at http://fife.swufe.edu.cn/BIlab/English/achievements_1.htm

The remainder of this article is organized as follows: Section 2 systematically reviews the main available work from the perspectives of media contents, media representation, and a media-to-stock analysis model. Section 3 summarizes the contributions of the representative studies. Section 4 presents speculation on the future research directions. Section 5 concludes with the main findings.

2 Research Methodology and Related Work

Stock movements are essentially driven by various types of information that cover a wide range of topics, including macroeconomics, fundamentals, politics and societies. Studies of information-driven stock movement can be traced back to work on bridging the relationships between annual reports and stock markets [13]. Since the number of financial reports is manageable compared with the huge volume of daily news, the influence of financial reports is generally analyzed via empirical study. By observing the fluctuations of stock markets with news articles, some researchers have begun to investigate the power of the verbal information contained in the news on stock markets. Due to limitations of the techniques available at the time, the influence of the news was simplified using the number of articles as a proxy [14]. With the explosion of information available in the era of social media, some researches have resorted to NLP techniques to convert textual information into a machine-friendly form to precisely and automatically process the influence of Web media [11], [12], [15], [16], [17].

In this section, we first categorize the representative works according to media type and then summarize the core techniques to convert textual information into a machine-friendly form. Finally, we present the cutting-edge analysis models utilized to capture the hidden relationships between Web media and stock movements, along with evaluation metrics. Fig. 2 shows the details of the technique framework to investigate the influence of Web media on stock markets.

2.1 Web Media Content

The research on media-aware stock movements began with financial reports and news articles. With the popularity of Web 2.0, new media sources, such as blogs, tweets/microblogs, discussion boards, and social news, have emerged and played important roles in affecting stock

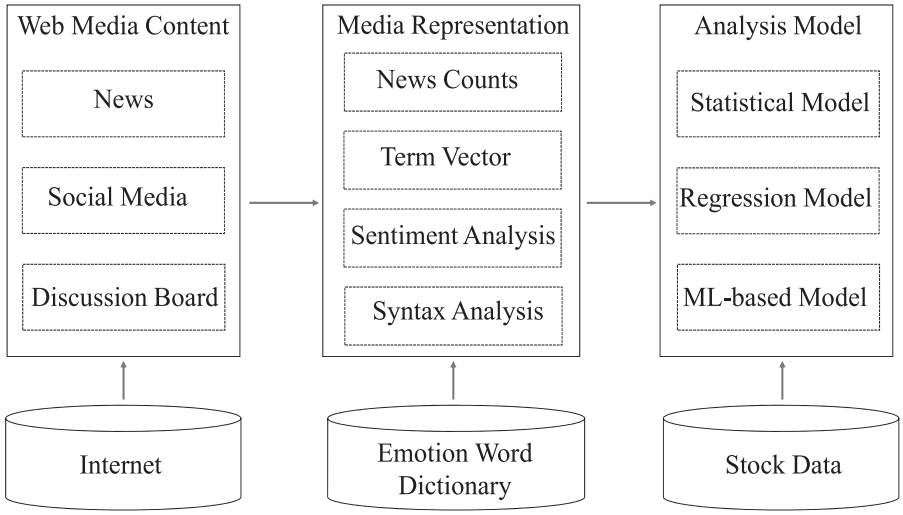


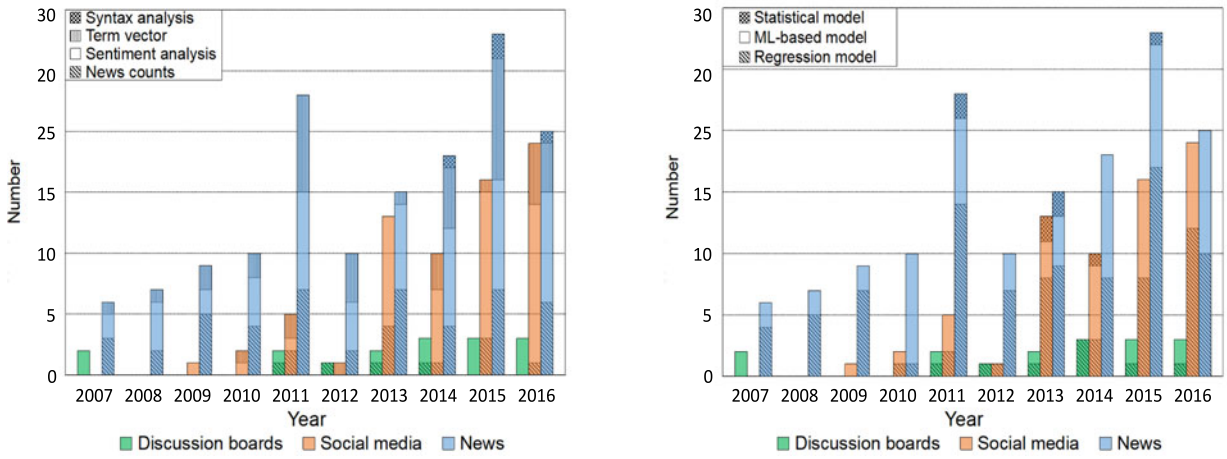
Fig. 2. The framework of the media-based movement analysis.

markets. As a pilot study, [11] found that the emotions of tweets affected stock trends for a brief period after the release of the tweets. In contrast to traditional news, social media allows users to express their opinions and feelings via comments, votes and so forth. Such user engagement efficiently enhances information dissemination and increases the value of the information. In the era of social media, investors' decisions could be influenced by the opinions of others, which may result in a herd behavior in investment. Such technical trends have fostered quantitative studies of media-aware stock movements from a big data perspective. In this section, we review the previous studies on media-aware stock movements in terms of media type, i.e., news articles, discussion boards, and social media, and summarize them in Table 1.

- *News Article:* The study of media-aware stock movements began by observing the influence of breaking news on stock price fluctuations. The issue of whether or how news reports affect stock markets is the most popular research challenge in this area. To address this challenge, researchers have explored the differences between the influences of media and non-media sources and between bad and good news. In addition, they have studied the influences of certain types of news on specific firms. For example, [8] found that high news pessimism indicated downward pressure on market prices and that high or low pessimism tended to increase market trading volume. [18] discovered that the number of news articles was related to the market fluctuations of the Dow Jones Industrial Average (DJIA) and crude oil prices. [19] analyzed the effects of real-time domestic and foreign news about fundamentals on stock returns and found that Portuguese macroeconomic news reduced stock market comovements. [20] used the number of newspaper articles about a stock as a proxy for the stocks overall media exposure and found that mass media could alleviate informational frictions and affect security pricing even if it did not supply genuine news. [21] revealed that local press coverage increased the daily trading volume of local retail investors, from 8 percent to nearly 50 percent depending on the specification. In news-aware stock movement studies, news information is typically obtained from a single news channel or a news archive that contains news articles from several news sources. Based on the statistics from the reviewed article, the most popular data source is Yahoo! Finance, followed by Reuters, the Wall Street Journal, Dow Jones Factiva, PRNewswire, the New York Times, LexisNexis, and Dow Jones Newswire.
- *Discussion board:* The self-publication mechanism of discussion boards make them a perfect channel to collect and reflect the collective wisdom of investors. On financial discussion boards, such as Yahoo! Finance, Sina Finance, and Eastmoney, investors can post their opinions about the future direction of a certain stock, and others can express their support or disagreement via comments or votes. [9] is the earliest work to examine the connection between stock movements and the public sentiment on discussion boards. At nearly the same time, [22] revealed that the performance of a stock and its Web sentiment on Yahoo! Finance discussion boards are closely correlated. [23] further noted that a financial crisis affected Yahoo! Finance forum topics and that different stakeholder groups had distinct effects on stock markets. [15] captured the public sentiment of individual firms from Eastmoney and Sina Finance discussion boards and found that the combination of public sentiment and financial news could be a good indicator of future stock trends. [24] analyzed six major firm-related forums hosted on Yahoo! Finance and revealed statistically significant indicators of firm stock returns in the discussions of the stakeholder groups of each firm. [25] found that the stock prediction task via the Yahoo! Finance forum sentiment analysis achieved better performance than the model using historical prices only.
- *Social media:* At present, social media, including blogs, microblogs, and social news, is burgeoning and one of the most important types of Web media. The interactive user engagement of social media makes market information spread faster than ever. The most cited work on media-aware stock movement is [11], who

TABLE 1
Literature Comparison in Terms of Web Content

Category	Literature	Focus			Analysis Method		
		Market	Scale	Media Source	Response	Features	Model
News	[8]	DJIA, NYSE	Day	WSJ, Dow Jones Newswires	Return	Emotion word number	Linear model
	[18]	DJIA, Nasdaq, S&P 500	Day	CNN News, Fox News, Yahoo News, NY Times	Index	News number	Linear model
	[19]	PSI-20, DJ-30	Day	Bloomberg, MMS, Reuters, IBES	Returns	News/non-News	Linear autoregressive model
	[20]	NYSE, NASDAQ	Month	New York Times, USA Today, Wall Street Journal, Washington Post	Return, Volatility	News coverage	Linear model
	[21]	S&P 500	Day	ProQuest News	Volumes	News coverage	Linear model
Discussion Boards	[9]	Morgan Stanley High-Tech Index	Day	Yahoo! Finance,	Index, volumes, volatility	Sentiment	regression model
	[15]	CSI100	Minute	Web news, Sina Finance, Eastmoney Forum,	Price	Current price, Partial news terms, social sentiment	SVR
	[22]	52 listed firms	Day	Yahoo! Finance,	Price	Sentiment	Statistical model
	[23]	listed firms	Day	Yahoo! Finance,	Price	Sentiment	Linear model
	[24]	6 listed firms	Day	Yahoo! Finance,	Return	message length & numbers, sentiment, transaction records	Linear model
	[25]	18 listed firms	Day	Yahoo! Finance,	Price	Sentiment, words	Graphic model
Social Media	[11]	DJIA	Day	Twitter	Index	Past DJIA, emotions	SOFNN
	[26]	NYSE	Day	Dailies, Twitter, Spinn3r RSS Feeds, LiveJournal blogs	Return, volume	Sentiment	Correlation coefficient
	[27]	NYSE, NASDAQ, AMEX	Day	Twitter	Return, volatility	Sentiment	Multivariate Regression Model
	[28]	S&P 100	Day, Minute	Twitter	Return, Volume, volatility	Sentiment, Message number	Regression Model



(a) Crossing with different approaches to representation

(b) Crossing with different analysis models

Fig. 3. Publication trends by media type.

first revealed the predictability of the effect of tweets on stock movements. [26] extracted sentiment from news, blogs, and microblogs and proposed a sentiment-oriented equity trader. [16] found that Web blogs and consumer ratings are significant leading indicators of firm equity value. [27] established a methodology to extract social sentiment from influential Twitter users within a financial community and found that it provided a more robust predictor of financial markets than the general social sentiment. [28] collected more than 1.2 million tweets related to S&P 100 companies and found that the sentiment of messages positively affects contemporaneous daily abnormal stock returns.

Fig. 3 presents the trend of media type crossing with different media representation approaches (Section 2.2) and analysis models (Section 2.3) during the past ten years. The impact of news on stock markets is a classic research topic and has received consistent attention during past years. Discussion boards provide an accessible channel for researchers to capture collective opinions or wisdom on financial markets. With the popularity of social media, researchers have begun to resort to this new media type to capture social sentiment and explore its potential power for quantifying the influence of Web media on stock markets. Fig. 3 shows that the portion of research studying social media and analyzing its sentiment has increased considerably since 2013, and the adoption of machine learning techniques to analyze the influence of news and social media on markets has continuously increased since 2012.

With advances in NLP and machine learning techniques, researchers have begun to explore the joint influences of different types of Web media [12].

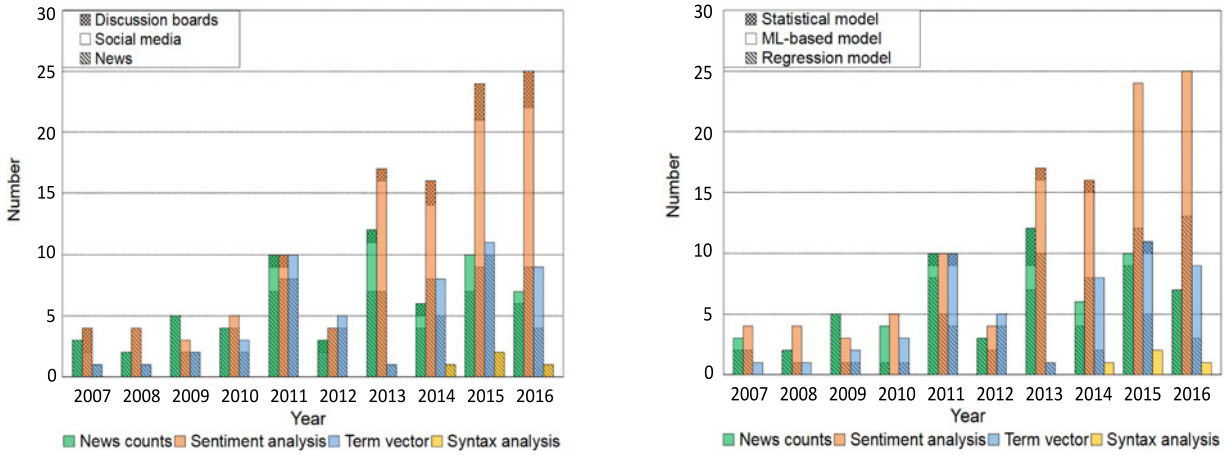
2.2 Media Representation

In traditional finance, the efficient market hypothesis states that "unemotional" investors are constantly updating their beliefs about the directions of markets as they receive new information about national economies or firm fundamentals, which causes stock prices to fluctuate around their intrinsic values [1]. In contrast, modern behavioral finance has discovered that, due to the cognitive biases or emotional impulses of investors, it is common to observe various financial anomalies. For example, the good weather effect states that investors are easily affected by local weather, which leads to a significant correlation between sunshine and stock returns [29], [30]. Therefore, it is critical to extract valuable information that reflects macroeconomics, fundamentals, and investors' emotions from textual Web data. The methods for converting textual data into a machine-friendly form for further media-to-stock analysis can be categorized as follows. These methods are summarized in Table 2:

- *News counts:* Due to limitations of text mining techniques in the early stages of this research, the number of news articles has been widely used as an indicator of the influence of news. This numerical value is generally treated as one of the explanatory (independent) variables in multiple linear regression models to capture the relationship between news and stock market indicators (explained variable), such as the stock price, trading volume, or abnormal return. For example, [31] found that the number of news articles on Dow Jones & Company was directly related to the aggregate measures of stock market activity, including trading volume and market returns. [14] built an econometric model using the number of news articles as the explanatory variable and abnormal returns as the explained variable and found that investors tended to react slowly to bad news. However, quantifying the influence of news using news counts is too simple because the influence of news comes from its content, which includes firm fundamentals, macroeconomic conditions, and professional or peer opinions. Realizing such limitations, researchers have studied various text mining techniques, i.e., term vector, sentiment analysis, and syntax analysis, to extract valuable information from Web media.
- *Term vector:* In natural language processing, the basic approach for representing an article in a machine-friendly form is to transform it into a term vector, where each entry is a weighted term in the article. The weight of a term can be calculated as a boolean value, where the weight of a word is 1 if

TABLE 2
Literature Comparison in Terms of Textual Representation

Category	Literature	Focus			Analysis Method		
		Market	Scale	Media Source	Response	Features	Model
News Counts	[14]	NYSE, AMEX, S&P500	Month	Dow Jones News	Return	News Number	Linear model
	[31]	NYSE, AMEX, OTC	Day	Dow Jones News	Return, Volume	News Number	Linear model
Term Vector	[10]	S&P500	Minute	WSJ	Price	Current Price, Partial news terms	SVR
	[32]	S&P500	Day	Yahoo News	Index	Full news terms	KNN
	[33]	S&P500	Quarter	Yahoo	Stock price	Full news terms	ARIMA, SVR
	[34]	S&P500	Minute	Yahoo! Finance	Stock price	Current price, Proper Nouns	SVR
Sentiment Analysis	[35]	S&P500	Minute	WSJ	Price	Current price, partial terms, news sentiment	SVR
	[36]	CSI100	Minute	Web	Stock price	Current price, partial terms, sentiment words social sentiment	SVR
	[37]	S&P100	Day	Twitter,	Price	Social relation, sentiment	VAR
	[38]	S&P500	Week	Reuters, Bloomberg,	Return, Volatility, Sharp ratio, Draw-down	sentiment	Neural Networks
	[39]	S&P500	Day	Yahoo! Finance,	Return	sentiment, word novelty	SVR
Syntax Analysis	[40]	S&P500	Day, week, month	Reuters, Bloomberg	Index, price	Event tuples	CNN
	[41]	S&P500	Day	Reuters, Bloomberg	Index, price	Knowledge graph	CNN



(a) Crossing with different approaches to content

(b) Crossing with different analysis models

Fig. 4. Methods for representing textual information.

the word exists, 0 otherwise, or by the classical TF/IDF schema in which the word weight, indicating its importance to the topic, is computed in terms of a trade-off between its frequency in the article and the entire corpus [42]. Such a textual representation is called a bag-of-words model. One of the earliest studies can be traced back to the work of [32], which forecasted the daily trends of five major stock market indexes based on Web news, each of which was represented as a term vector. In a similar way, [33] represented a quarterly financial report as a term vector using full words and studied its influence on stock prices. Realizing that some words are irrelevant to the main topic and that using full words may not scale very well, researchers have studied methods to represent articles using "important" words. Both [10] and [34] represented a news article with four alternatives, i.e., full words, noun phrases, proper nous, and name entities, to study the impact of breaking news on stock movements using a support vector regression (SVR) model, and found that proper nouns were the most efficient representation method.

- *Sentiment analysis:* In behavioral finance, abnormal fluctuations of stocks are caused by the emotional impulses of irrational investors [3]. In general, investors may be affected by peer opinions from social media or professional attitudes from news articles. To analyze expert opinions, researchers have resorted to various types of word-based and sentence-based sentiment analysis techniques [43]. For example, [44] measured the negative (positive) sentiment polarity of an article in terms of the proportion of negative (positive) emotional words in the document. Both [39] and [35] utilized OpinionFinder, a document-level sentiment analyzer, to calculate the sentiment index of each news article and found that this index obviously improved predictive precision. [15] proposed a statistical model to detect finance-oriented sentiment words and represented news articles with nouns and financial sentiment words to study their influence on stock markets.

Social media is sufficiently popular that self-publication and free comments on social media can reflect public opinions. Most existing works focus on the public mood of the entire market, and few study the collective opinion on a single stock. [36] modeled a news article as a weighted term vector consisting of a number of nouns and sentiment terms, following the hypothesis that the important concepts of firms' fundamentals in a news article are conveyed by a set of nouns, and irrational investors can be affected by the optimistic or pessimistic sentiment of an article. [37] first constructed a semantic social network in terms of the co-occurrence of two stocks in tweets and calculated the public mood of a stock based on its own and its neighbors' tweets. Due to the noise contained in social media, including advertisements and rumors, precisely capturing the collective mood regarding the market or an individual stock has not yet been realized.

- *Syntax analysis:*In natural language, the full meaning of a sentence is determined by the words and the syntax. The disadvantage of bag-of-words models is the loss of the structural relations between words, which limits these models' potential. For example, the sentence "Samsung phone emits smoke on Indian airline" may cause a downward pressure for both Samsung Corporation and Indian airline if it is represented as a bag of words, as the unstructured terms cannot differentiate the actor ("Samsung") from the place ("Indian airline"). Therefore, some researchers have taken a further step by presenting articles with structured representations of events (e.g., Actor = Samsung phone, Action = emit, Object = smoke, Place = Indian airline), where information extraction (IE) techniques were used to assign the role of each word [40], [45]. For instance, [41] applied a knowledge graph to enrich the structured representations of events for predicting stock volatilities.

Fig. 4 presents the publication trends of these four representation methods crossing with different contents (Section 2.1) and analysis models (Section 2.3) during the past ten years. It can be observed that the syntax analysis

of news has just begun, and progress has been slow due to the considerable difficulty of sophisticated syntax processing. In contrast, the sentimental representation of social media and news has attracted increasing attention from researchers in recent years. As shown in Fig. 4, researchers tend to apply sentiment analysis to extract public emotions from social media since it is an interactive information platform to gather user feedback. However, most existing work has relied on general-domain sentiment analyses. Harvard-IV-4[2] and SenticNet[3] are two popular general-domain emotional word dictionaries used in prior studies [8], [44], [46], [47], [48]. However, general sentiment words may not be emotional in the realm of finance [49]. For example, the general negative sentiment word "tire" is typically used to identify a specific firm in finance. In addition, an emotionless word can be a sentiment in the realm of finance. The word "bear" originally refers to a carnivorous mammal; it also indicates widespread pessimism in the finance domain, such as a "bear market". [49] 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 and proposed a financial sentiment dictionary (Loughran and McDonald Sentiment Word Lists)[4]. To improve the precision of sentiment analysis, it is necessary to determine the document's sentiment in terms of financial sentiment words rather than general sentiment words. [15] proposed an approach to automatically detect financial sentiment words from Web media and used it to sense stock movements, and this approach achieved better performance than one using general emotion words.

2.3 Analysis Model

Once the valuable information is extracted from Web media and represented in a machine-friendly form, various types of analysis models are required to bridge the relationship between the media and stock movements. There are three mainstream methods, i.e., statistical models in statistics, regression models in econometrics and machine learning (ML)-based models in computer science.

- *Statistical model:* Here, statistical model refers to univariate statistical models and bivariate statistical models, which are only able to capture the relationship between stock movements and a single information source without considering other information sources. Univariate analysis models are utilized to test the connections between stock movements and media by examining their statistical significances under different hypothesis tests including the *t*-test, Wilcoxon test, and Kruskal-Wallis test. For example, [50] found that there are statistically significant connections between stock prices and Wikipedia page views via the *t*-test. Bivariate analysis models are used to gauge the relationships between stock movements and media in terms of various types of correlation measures including the Pearson correlation coefficient, Spearman correlation coefficient, and mutual information. For instance, [7] concluded that tweet sentiment could contain statistically significant ex ante information on the future prices of the S&P500 index after analyzing 34 billion postings on Twitters network using mutual information. Both univariate models and bivariate models provide persuasive statistical methods for testing the relationships between stock movements and media from a big data perspective. However, both methods focus on the effect of a single piece of information on stock markets and lack the ability to analyze the joint impacts of various information sources.
- *Econometric regression model:* An econometric regression model specifies the statistical relationships that are believed to exist among the various economic quantities pertaining to a particular economic phenomenon under study. The representative models include the linear regression model, logistic regression model, vector autoregression model and autoregressive integrated moving average model (ARIMA). This approach focuses on the causal relationship between stocks and information sources without considering the interactions among different information sources. For example, [44] analyzed the content of news, especially news sentiment, using linear regression models and found that stock returns were expected to be low in the presence of negative news information. [51] demonstrated the negative relationship between search volume and a stock index using a linear regression model. More relevant studies can be found in the work of [16], [37], [52], as shown in Table 2. In the era of social media, when various types of highly interrelated information produce informational overload, linear regression models tend to fail to discover the complicated nonlinear patterns. For example, [52] were able to capture the impact of social sentiment from Yahoo! finance postings on stock returns in 2000 using linear regression analysis. However, using a similar approach, [53] were unable to link Yahoo! finance to stock movements during the period from 2005 to 2010; during this period, the investing information environment became much more complicated. In fact, [54] found that the linear dependences of stock returns vary over time, but nonlinear dependences are strong throughout. More important, linear regression models only take scalars as independent variables. To fit regression models, each high-dimensionality information source has to be reduced to a scalar, which results in the loss of valuable information.
- *Machine learning (ML) based models:* The advantage of machine learning-based models is their ability to take high-dimensional data as inputs. The common strategy is to concatenate features of different information sources into a super feature vector and apply ML techniques, including neural networks, Bayesian classifiers, and support vector machines (SVMs), to capture the relationship between stocks and information. For instance, [32] applied neural network and K-nearest neighbor (KNN) techniques to predict

2. http://www.wjh.harvard.edu/~inquirer/homecat.htm
3. http://sentic.net
4. https://www3.nd.edu/~mcdonald/Word_Lists.html

future stock indexes based on breaking news. [55] proposed a relevance language model (RLM) to associate stock price trends with news stories. [56] represented each news article as a term vector. The SVM and KNN methods were applied to study the impacts of news on stock movements. [33] studied the relationships between news and stocks using a hybrid predictive model based on both ARIMA and SVR. [11] found that the collective mood states derived from 10 million Twitter feeds were correlated with the value of the Dow Jones Industrial Average (DJIA) over time using self-organizing fusion neural networks (SOFNN). Although vector-based predictive models consider the joint impacts of different information sources on stock trends, their linearization of information weakens or even ignores the intrinsic associations among different information sources. [12] first applied tensor theory to model the complicated market information environment, which is able to capture the interactions of various sources and study their joint impacts on stock market movements.

Table 3 summarizes the relevant studies on media-aware stock movements in terms of different modeling techniques. Fig. 5 presents the publication trends of analysis models crossing with different contents (Section 2.1) and media representations (Section 2.2) during the past ten years. The regression model has received continuous attention because it is a classic method in econometrics, and the ML-based models have gradually become popular since 2010. As shown in Fig. 5, regression models and ML-based models are popularly applied to analyze the textual impacts on markets from news and social media in recent years. In fact, stock markets are strongly affected by various types of highly interrelated information sources. An efficient approach to capture the multi-faceted and multi-relational information held by investors on stock movements has yet to be discovered.

2.4 Data Frequency and Evaluation Metrics

In previous sections, we introduced three core parts, i.e., media source, media representation, and analysis model, which form a complete procedure to analyze media-based stock movements. In addition, there are two more important concerns for this study. One is source data frequency, and the other is evaluation metrics.

Previous studies analyze media impact in different manners, such as years, months, weeks, days, hours, and minutes. The label "Scale" in Tables 1, 2, and 3 indicates the source data frequency of each representative work. In the reviewed literature with experimental evaluations, 169 studies focused on the daily effect of media, 16 articles analyzed the media influence in minutes, 15 studies were on a monthly level, 8 articles analyzed stock trends in weeks, and 3 articles were in hours. In the early stages, economists tended to understand the media effect from a macro perspective by applying regression models in a low-frequency manner, such as days, weeks, months, and quarters. Subsequently, researchers, especially computer scientists, gradually resorted to ML approaches to study media-aware movements in high-frequency manners (minutes or seconds) and over longer experimental periods.

The performance of models in analyzing the media effect on stock markets is evaluated using various metrics according to the model type. Typically, statistical significance is utilized to evaluate the performance of statistical models, which measure the relationship between two data series. Regression models adopt the coefficient of determination (R^2) or its derivatives to determine model fitness [8], [16], [44]. The coefficient of determination is essentially the proportion of the variance in the dependent variable that can be predicted using the independent variables. This provides a measure of how well observed outcomes are replicated by the model.

For ML-based models, most studies focus primarily on predicting stock trends (up or down), which is essentially a binary classification problem. Binary accuracy, the proportion of true results among the total number of cases examined, is a popular approach [11], [32], [55]. In a classification task, precision is the fraction of correctly predicted examples out of all the predictions of a particular class. Recall is the fraction of correctly predicted examples out of all actual members of the class. F-Measure, a harmonic mean of precision and recall, is also used to evaluate performance [22], [56], [59], [60]. Some researchers take a further step by predicting the real value of stock indicators (price, volume, and volatility) and evaluating the model performance using the mean average error (MAE) or its derivatives, such as root-mean-square deviation (RMSE) and mean absolute percentage error (MAPE) [10], [12], [33], [34], [35], [48], [61], [61], [62]. All of these metrics originated the research field of computer science and lack any concern for risk factors. Few researchers have resorted to risk-adjusted return metrics, such as the Sharpe ratio and Sterling ratio, which are prevalent in portfolio management in hedge funds [58]. The label "Metric" in Table 3 indicates the measurement(s) used in each of the representative works.

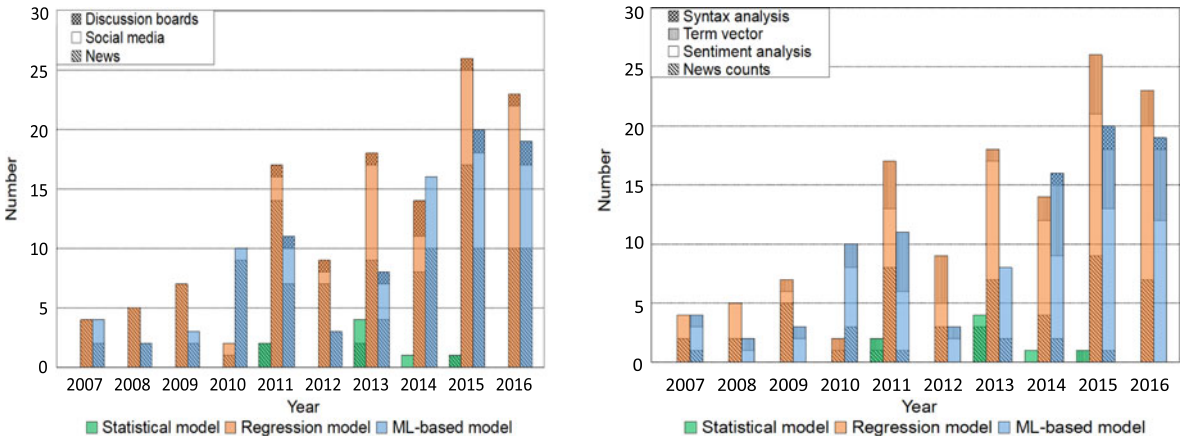
3 Contributions of Representative Works

In this section, we first make a brief introduction of the representative works and then present their unique contributions from several different perspectives.

We select the representative works based on citations and methodological progress. The citation for each reference is based on the citation statistics from Google Scholar as of October 1, 2017. To favor recently published research, we also select papers that were published between 2015 and 2016 in the prestigious journals that have the most cited representative works during the past 10 years. Table 4 presents the most representative studies during the period from 2007 to 2016 in terms of citations. In addition, we rank the journals in terms of the citations of their articles and present the statistical results in Table 5. These journals cover a wide range of topics, from finance and management science to computer science. *Expert Systems with Applications* is ranked first with 7 publications. *Journal of Finance* and *Decision Support Systems* are tied for second place with 6 relevant publications. *Review of Finance Study* is ranked third with 5 articles and followed by *Science Reports* from Nature publishing with 4 publications. These representative studies with high recognition from peer researchers can be roughly classified into three categories according to their foci, i.e., existence, technique, and mechanism. This categorization of relevant work is not mutually exclusive. Some studies may

TABLE 3
Literature Comparison in Terms of the Analysis Model Used

Category	Literature	Focus			Analysis Method			Experiment		
		Market	Scale	Media Source	Response	Predictor	Model	Period	Size	Metric
Statistical Model	[7]	DJIA	hour	Twitter	Price	message volume	Mutual Information	11/12/2012-12/03/2013	10% of all Tweets	statistical significance
	[50]	DJIA	Week	Wikipedia	Index	Page view number	Statistical model	12/10/2007-04/30/2012	30 Wikipedia articles	statistical significance
Regression Model	[16]	NYSE	Day	CNET, Alexa, Google, CRSP, Lexis/Nexis,	Return, risk	Rating volume, Blog emotion, page view, search intensity	VARX	08/01/2007-04/03/2013	9 stocks	determination coefficient
	[44]	S&P500	Day	WSJ	Return	Emotion word number	Linear model	1980-2004	500 stocks	determination coefficient
	[52]	DJIA, XLK	Day	Yahoo! Finance, Raging Bull	Return, Volatility	Emotion Index, Message Number	Linear model	2000	45 stocks	determination coefficient
	[57]	DJIA	Week	Google Trends	Index	Search volume change	Linear model	2004-2011	98 search terms	determination coefficient
	[51]	SPXT	Week	Wikipedia, Google Trends,	Index	Search behavior	Linear model	12/2012-01/2014	55 topics	determination coefficient
ML-based Model	[11]	DJIA	Day	Twitter	Index	Past DJIA, emotions	SOFNN	02/28/2008-12/19/2008	9 million tweets	MAPE, Binary Accuracy
	[12]	CSI100	Minute	Google, Sina, Eastmoney	Stock price	News, firm characteristics, social sentiment	Tensor model	01/01/2011-12/31/2011	89 stocks	Real value, MAE
	[32]	S&P500	Day	Yahoo	Index	News content	KNN	12/06/1997-03/06/1997	5 major Indexes	Binary Accuracy
	[33]	S&P500	Quarter	Yahoo	Stock price	News content	ARIMA, SVR	1994-2010	6 stocks	MAERMSE
	[56]	S&P500	Minute	PRNewswire	Stock trend	News content	KNN, SVM	04/01/2002-12/31/2002	500 stocks	F-Measure
	[55]	S&P500	Hour	Yahoo	Stock trend	Current Price, News content	Language model	10/15/1999-02/10/2000	127 stocks	Binary Accuracy
	[58]	S&P500	Hour	Twitter, Web news	Sterling ratio, sharpe ratio, profit	sentiment, firm characteristics	Genetic programming	08/01/2012-01/30/2015.	index	Finance Metrics



(a) Crossing with different contents (b) Crossing with different representations

Fig. 5. Analysis models.

TABLE 4
Representative Work in Terms of Citations

Reference	Title	Journal	Citations
[11]	Twitter mood predicts the stock market	J COMPUT SCI	2785
[8]	Giving content to investor sentiment the role of media in the stock market	J FINANC	1872
[44]	More than words quantifying language to measure firms fundamentals	J FINANC	1141
[9]	Yahoo! for Amazon: Sentiment extraction from small talk on the Web	MANAGE SCI	908
[20]	Media coverage and the cross-section of stock returns	J FINANC	882
[21]	The causal impact of media in financial markets	J FINANC	493
[57]	Quantifying trading behavior in financial markets using Google Trends	SCI REPORTS	432
[10]	Textual analysis of stock market prediction using breaking financial news: The AZFin text system	ACM T INFORM SYST	412
[63]	Does public financial news resolve asymmetric information	REV FINANC STUD	280
[64]	Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange	EXPERT SYST APPL	254
[16]	Social media and firm equity value	INFORM SYST RES	234
[65]	Wisdom of crowds: The value of stock opinions transmitted through social media	REV FINANC STUD	219
[50]	Quantifying Wikipedia usage patterns before stock market moves	SCI REPORTS	188
[66]	Journalists and the stock market	REV FINANC STUD	186
[17]	The impact of social and conventional media on firm equity value: A sentiment analysis approach	DECIS SUPPORT SYST	168
[67]	When machines read the news: using automated text analytics to quantify high frequency news-implied market reactions	J EMPIR FINANC	163
[68]	Selective publicity and stock prices	J FINANC	159
[69]	How important is the financial media in global markets	REV FINANC STUD	134
[35]	Evaluating sentiment in financial news articles	DECIS SUPPORT SYST	134
[6]	Text mining for market prediction: a systematic review	EXPERT SYST APPL	127
[61]	A quantitative stock prediction system based on financial news	INFORM PROCESS MANAG	103
[70]	Can Internet Search Queries Help to Predict Stock Market Volatility	EUR FINANC MANAG	99
[71]	The effect of macroeconomic news on stock returns: new evidence from newspaper coverage	J BANK & FINANC	94
[72]	Automated news reading: stock price prediction based on financial news using context-capturing features	DECIS SUPPORT SYST	85
[73]	The high-frequency impact of news on long-term yields and forward rates is it real	J MONETARY ECON	79
[74]	Quantifying the relationship between financial news and the stock market	SCI REPORTS	76
[75]	Intraday jumps and us macroeconomic news announcements	J BANK & FINANC	69
[48]	News impact on stock price return via sentiment analysis	KNOWL-BASED SYST	67
[19]	Economic news and international stock market comovement	REV FINANC	65
[53]	Investor sentiment from Internet message postings and the predictability of stock returns	J ECON BEHAV ORGAN	60
[47]	Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction	MIS QUART	57
[60]	Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment	EXPERT SYST APPL	51
[76]	Predicting abnormal returns from news using text classification	QUANT FINANC	49
[46]	The impact of corporate governance press news on stock market returns	EUR FINANC MANAG	48
[15]	The effect of news and public mood on stock movements	INFORM SCIENCES	45
[77]	Information aggregation around macroeconomic announcements: revisions matter	J FINANC ECON	45
[7]	When can social media lead financial markets	SCI REPORTS	44
[25]	Sentiment analysis on social media for stock movement prediction	EXPERT SYST APPL	42
[78]	Underreaction to news in the US stock market	QUART J FINANC	40
[34]	Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies	J AM SOC INF SCI TEC	35

TABLE 5
Journal Influence in Terms of Citations

Journal	Total	Number	Articles
J FINANC	4647	6	[8], [20], [21], [44], [68], [79]
J COMPUT SCI	2785	1	[11]
MANAGE SCI	908	1	[9]
REV FINANC STUD	849	5	[63], [65], [66], [69], [80]
SCI REPORTS	740	4	[7], [50], [57], [74]
EXPERT SYST APPL	512	7	[6], [25], [60], [64], [81], [82], [83]
DECIS SUPPORT SYST	430	6	[17], [35], [36], [72], [84], [85]
ACM T INFORM SYST	416	2	[10], [12]
INFORM SYST RES	234	1	[16]
J BANK & FINANC	168	3	[71], [75], [86]
J EMPIR FINANC	163	1	[67]
EUR FINANC MANAG	147	2	[46], [70]
INFORM PROCESS MANAG	103	1	[61]
QUANT FINANC	83	3	[27], [76], [87]
J MONETARY ECON	79	1	[73]
KNOWL-BASED SYST	75	2	[48], [88]
REV FINANC	65	1	[19]
J ECON BEHAV ORGAN	60	1	[53]
MIS QUART	57	1	[47]
J FINANC ECON	46	1	[77]
INFORM SCIENCES	45	1	[15]
NEUROCOMPUTING	34	3	[33], [38], [58]

fall into multiple categories. Here, we only list the representative articles in single categories according to the main topic of the article. Table 6 summarizes the findings of these representative works for easy review.

The first category is the existence category, which focuses on demonstrating the comovements between stocks and various types of media including news, discussion boards, and microblogs. [9] conducted one of the pilot studies that revealed that there was a direct connection between the sentiment of postings on discussion boards and stock indexes, volumes and volatilities. Although the earliest work on news-aware stock movements can be traced back to [32], Tetlock provided a simple quantitative measure of textual news information in terms of the proportion of negative and positive sentiment words, which has been widely accepted by subsequent studies, especially in the finance field [8], [44]. [11] first discovered that the public mood of tweets was correlated with or even predictive of DJIA values; this work is the most-cited work on media-aware stock movements. In addition to media types, user behaviors related to media release and dissemination have been studied, and these analyses have provided strong support for the identification of media-aware stock movements. [57] detected increases in Google search volumes for queries related to financial markets before stock market declines. [50] found that the number of page views of Wikipedia articles related to financial affairs increased before stock market declines.

The second is the technique category that is devoted to utilizing and enhancing various types of techniques, especially the latest advances in the fields of natural language processing and machine learning, to capture valuable information from textual media and bridge the connections between media and stocks.

- The most popular method for representing text is the term vector method. [32] used full words from articles to form term vectors. [10] found that using some words, especially proper nouns, can achieve better performance. Subsequently, [35] further improved this approach by adding the boolean emotional polarity of each article into the term vector. [15] found that the best result could be obtained by using nouns, adjectives and financial emotional words. The latest achievement is utilizing the syntax structures of sentences to represent news articles [40]. Due to the computational complexity and extraction precision, the successful studies have only parsed the titles of news articles, and no study using full articles has yet been completed. Although various representational approaches have been proposed, financial researchers prefer to use sentiment and ignore the other information in the news articles because it is easier to understand and simpler to implement such an approach [8], [44], [46], [68], [77], [86], [89].
- Regression models are widely used by financial researchers in their analyses to reveal the impacts of media on stock movements, while computer scientists are more interested in advanced predictive models to demonstrate the predictive ability of media. Various techniques have been explored, including k-nearest neighbors, the naïve Bayes classifier [32], decision trees [90], neural networks [11], [40], language models [55], and support vector machines (SVM) [56]. Among them, the support vector machine approach and its derivatives have achieved a series of promising results. Most previous studies have relied on the classifier approach, which is a binary measure, to predict upward or downward stock trends. To predict discrete values of future stock prices, support vector regression (SVR), which is a regression-based variation of SVM, has been used successfully [10], [34], [35], [62]. To study the joint impacts of various information sources, [91] utilized multiple kernel learning to extract the hidden information behind different information sources and integrated them seamlessly for stock prediction. [12] proposed a support tensor regression model that is able to capture the intrinsic relationships among multiple information sources, i.e., social sentiment and firm-specific and financial news. Due to the great successes of deep learning networks in AI, some researchers have begun to explore the hidden relationships between Web media and stocks using deep neural networks such as the deep convolutional neural network (CNN) [40].

The third category is the mechanism category, which investigates the mechanism of Web media percolation and the extent of its impact on stock markets. Rather than proving the existence of media-aware stock movements and seeking out advanced techniques, these studies have focused on understanding the roles of the different types of Web media in affecting stock markets and stock movements. [16] found that social media had greater predictive power for firm equity value than conventional online consumer behavioral metrics. [17] also found that social media had a stronger relationship with firm stock performance

TABLE 6
Contributions of Representative Works

Group	Reference	Contribution
Existence	[44]	This work provides a simple quantitative measure of language and found that the fraction of negative words in firm-specific news stories predicts low firm earnings and that firms stock prices briefly underreact to the information embedded in negative words.
	[9]	This paper proposed a methodology for extracting small investor sentiment from stock message boards and found that stock indexes, volumes and volatilities are related to small investor sentiment.
	[11]	This contribution found that public mood, as measured by tweets, is correlated with or even predictive of DJIA values. The calmness of the public (measured by GPOMS) is predictive of the DJIA, whereas general levels of positive sentiment, as measured by OpinionFinder, are not.
	[57]	This work provides a quantification of the relationship between changes in search volume and changes in stock market prices. It identified increases in Google search volumes for keywords related to financial markets before stock market declines.
	[50]	The number of page views of Wikipedia articles relating to companies or other financial topics increases before the stock market declines.
	[74]	This paper found that movements in financial markets and movements in financial news are intrinsically interlinked.
Technique	[32]	This contribution predicted daily stock indexes using news articles that are represented as weighted term vectors. The k-nearest neighbor, regression analysis, neural network, and rule-based probabilistic model approaches were used as predictive models.
	[10]	This paper modeled financial news using several different textual representations and estimated a discrete stock price twenty minutes after a news article was released using SVR. A proper noun scheme performs better than the standard of bag of words.
	[35]	This work represents financial news articles using proper nouns and overall tone, as evaluated by OpinionFinder. Capturing the sentiment polarity of news articles improves the predictive accuracy of media-aware stock movements.
	[48]	News articles are converted into a sentiment space instead of sentiment polarity. The models with sentiment analysis outperform the bag-of-words model, but simply focusing on positive and negative dimensions did not produce useful predictions.
	[12]	This article introduced a tensor-based information framework to predict stock movements. Market information (news, social sentiment, and transaction records) is represented with tensors. A tensor regression learning algorithm was presented to identify multi-faceted information factors and their intrinsic relationships with stock movements.
	[40]	This work proposed a deep learning method for event-driven stock market prediction. Events were extracted from news texts and represented as dense vectors. A deep convolutional neural network was used to model both the short- and long-term influences of events on stock price movements.
Mechanism	[16]	Social media is a leading indicator of firm equity value and has greater predictive power than conventional online consumer behavioral metrics. Investments in increasing positive blog posts and curtailing negative blog posts would be more effective.
	[17]	Social media has a stronger relationship with firm stock performance than does conventional media, and the impacts of different types of social media vary significantly. Specifically, blog sentiment has a positive impact on stock returns, whereas forum sentiment has a negative impact. Both have positive effects on risk.
	[47]	Wikipedias information aggregation moderates investors negative reactions to bad news and moderates the timings of managers voluntary disclosures of companies earning disappointments.
	[15]	Both news and public mood extracted from discussion boards are utilized to predict media-aware stock movements. Stocks are sensitive to articles on restructuring & earning issues. The firms involved with daily life are more predictable.
	[20]	Stocks with no media coverage earn higher returns than stocks with high media coverage. These results are more pronounced among small stocks and stocks with high individual ownership, low analyst following, and high idiosyncratic volatility.
	[21]	Local media coverage strongly predicts local trading, after controlling for earnings, investors, and newspaper characteristics. Moreover, local trading is strongly related to the timing of local reporting, which is a particular challenge for non-media explanations.

than did conventional media and noted that the impacts of different types of social media varied significantly. For example, blog sentiment had a positive impact on stock returns, whereas forum sentiment had a negative impact. [47] discovered that Wikipedias information aggregation moderated investors negative reactions to bad news and moderated the timings of managers voluntary disclosures of companies earnings disappointments. [15] found that stocks were sensitive to articles on restructuring & earning issues, and firms involved with public interests or daily life, especially in utility supply, real estate, social services, and wholesale and retail trade, tended to be more affected by relevant financial information. [21] discovered that local trading was strongly related to the timing of local reporting. [20] found that stocks with no media coverage earned higher returns than stocks with high media coverage.

4 Directions for Future Work

In the past few decades, studies have been conducted on the impacts of media on stock markets. With the popularity of Web media, especially self-publication and network-based dissemination of social media, the influence of media on stock markets has become increasingly salient. Moreover, with the increasing power of artificial intelligence techniques, especially the ability to handle terabytes of Web information in a short time, extracting more valuable and accurate information from Web media and capturing the hidden relationships between Web media and stock movements with greater sensitivity and precision have become possible. Advances in technology have prompted the birth of the media-aware hedge funds including Derwent Capital, DCM Capital, and Cayman Atlantic. We believe that the study of media-aware stock movements will continue. In this section, we cite areas or aspects in need of future research and advancement.

4.1 Web Contents

Most previous studies have focused on the impact of a single media type on stock movements, such as news, discussion boards, or microblogs. Essentially, market information is multifaceted and interrelated; therefore, it is referred to as a *mosaic* information space [92]. With the increasing media volume and number of dissemination channels, modeling the mosaic-like characteristics of the information and studying the joint effects of multiple information sources are critical tasks. [9] extracted public sentiment from Yahoo! Finance postings for 24 tech-sector firms during the period from July 2001 to August 2001 and concluded that sentiment was positively related to the stock index. However, [53] analyzed 32 million messages on 91 firms posted on the Yahoo! Finance message board in the period from January 2005 to December 2010 and found no evidence that investor sentiment forecasted future stock returns at either the aggregate or individual firm level. One possible explanation for this contradiction is that Web media became more complicated in the five years after the work of [9] and various new information sources and channels appeared. To fully analyze the impacts of Web media on stock markets, studying the joint impacts of different information sources is necessary. Some papers have studied the joint impacts of multiple information sources on stock markets. For example, [26] captured market sentiment from news articles, blogs and microblogs and studied its predictive power for stock prices. [12] extracted public sentiment from finance forums, captured event information and professional opinions from financial news articles, and provided a tensor-based framework to analyze the joint impacts of public sentiment, events, and firm attributes on stock markets. Since Web media is becoming increasingly complicated due to the addition of new information sources, studying the joint impacts of these different sources is important. However, collecting and processing comprehensive information from all Web media sources, including news articles, discussion boards, blogs, and microblogs, remains a considerable challenge in the era of big data.

In addition, rumors, which are one of the most important risk sources in financial markets, have not been studied comprehensively. Most existing studies have focused on "official rumors", which are released via official announcements and in newspapers and magazines. [93] found that positive rumors had positive impacts on stock prices, whereas negative rumors had negative impacts, based on the "rumors" column of the Wall Street Journal. [94] utilized the dashboard column of the Wall Street Journal and found an immediate impact of rumors on abnormal stock returns but no long-term effects. [95] explored the Inside Wall Street column in Business Week and found that positive, significant excess returns were observed the day prior to the publication date, the publication date, and the two days after publication. However, compared to the huge amount of "unofficial rumors" on discussion boards and social media, the effects of mainstream rumors have yet to be thoughtfully explored. Few studies have addressed this issue and have done so with limited experimental data. For example, [96] only analyzed 189 takeover rumors on the Hotcopper discussion board. With the advanced harvesting and analysis techniques available in computer science, it is critical to study powerful automatic rumor-identification techniques and investigate the influence mechanism of "unofficial rumors" in social media.

Finally, it is worth clarifying the effect of rumors on stock movements since the truthfulness of the information contained in rumors is unclear. On one hand, the greater information transparency resulting from clarifying rumors would lead to more stable stock markets. In particular, stock investors are constantly updating their beliefs about future business value using new information. Therefore, transparency enhanced by rumor clarification can increase the stability of stock markets. However, if a market lacks sufficient credibility, investors tend to lose confidence in the companies involved in rumors. Thus, the effectiveness of clarifying the truthfulness of rumors in improving information transparency is inevitably reduced. In reality, irrational investors may underact to clarification announcements by listed companies in the belief that it is better to believe the rumor than not [97]. Such preconceived emotional and irrational behaviors by investors may make efforts to address and clarify rumors ineffective. Few studies have addressed the functionality of rumor clarification. Huberman and Regev [98] found that it is difficult to return the price of a listed firm involved in unfavorable rumors to its real market value even after the rumor has been clarified. Marshall et al. [99] reported that stock prices can rally within 5 days of addressing rumors affecting a firm. Yang and Luo [100] analyzed the impact of rumors on stock returns under different market conditions. Their findings show that the average cumulative abnormal return after clarification is positive in a bull market and negative in a bear market. All of these works still focus on the announcements clarifying "official rumors". However, only a small share of rumors are clarified by these official clarification announcements in mainstream media. It is crucial to understand the role of large-scale rumor clarification in social media from a big data perspective.

4.2 Media Representation

Stock movements are strongly affected by various highly interrelated sources and types of information that cover a wide range of topics including economics, politics, and psychology. The common strategies for media representations in previous studies have relied on either sentiment analysis or term vectors.

In sentiment analysis, general sentiment words may not have an emotional connotation in the realm of finance [49]. For example, the generally negative sentiment word "tire" is typically used to identify a specific firm in finance. In addition, a generally emotionless word can convey sentiment in the realm of finance. The word "bear" originally refers to a carnivorous mammal; it also indicates widespread pessimism in the finance domain, such as a "bear market". To improve the precision of sentiment analysis, it is necessary to determine the sentiment of a document in terms of financial sentiment words rather than general sentiment words. Especially in the era of social media, various types of new words are popularly invented and used, including slang terms. Therefore, it is critical to maintain a financial sentiment dictionary to assist in accurate sentiment analysis for financial documents or web communities.

Term vectors require reduced dimensionality, which inevitably results in the loss of valuable information. To maximally preserve information, the term vector concatenates features of various information sources into a single compound feature vector, but this step inevitably diminishes the interrelations among the various information sources. For example, two news articles released at different times may be textually dissimilar, but both may contain favorable information regarding the same stock. In this scenario, the semantic similarities of different words can be enhanced by the similarities in the corresponding firm-specific data via advanced representation techniques such as tensors [12]. More important, in natural language, the full meaning of a sentence is determined by the words and the syntax. The disadvantage of term vectors is the loss of the structural relationships among words, which limits their potential. Modeling the interactions among different information sources and extracting more valuable information in terms of syntactical analysis are substantial challenges for future research.

4.3 Analysis Model

Recently, a series of breakthrough advances in artificial neural networks has resulted in considerable success in several areas including object recognition in computer vision [101], speech recognition [102], and machine translation [103]. There are two promising techniques to assist us in studying media-aware stock movements. The first is extreme learning machines (ELMs), which is a single-hidden-layer feedforward neural network. It randomly chooses the input weights and hidden-layer biases and analytically determines the optimal output weights instead of tuning them. It avoids difficulties such as local minima and parameters. ELM tends to have better scalability and achieve similar (for regression and binary class cases) or much better (for multi-class cases) generalization performance at much faster learning speeds (up to thousands of times) than traditional SVM [104], [105]. The second is the convolutional neural network (CNN), which is composed of one or more convolutional layers (often with a subsampling step) and then followed by one or more fully connected layers as in a standard multilayer neural network [101]. This deep learning structure allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. A benefit of CNNs is that they are easier to train and have many fewer parameters than fully connected networks with the same number of hidden units. Exploring the power of deep learning for financial markets is very interesting. In addition, recent studies have demonstrated the benefits of tensor-based media representations over vector-based approaches [12], [40], [45]. Studying tensor-based deep neural networks and exploring their potential power for media-aware stock movements are promising avenues for future research.

4.4 Influence Mechanism

Existing studies have focused primarily on the direct impacts of Web media on a certain firm. However, media influence spreads to multiple associated companies rather than being limited to a single firm. In particular, what is the indirect media influence on associated companies when a piece of news affects the trend of a relevant firm? What is the range of the affected companies in terms of media content and firm attributes? If the associated firms of a target company are simultaneously affected by different information, what is the joint media impact on this target company? Due to technical limitations, such media-aware wave effects and superposition effects have never been studied systematically. Understanding how Web media shapes the comovements of relevant firms is critical. This essentially extends the current media influence analysis from one-to-one to one-to-many and many-to-one.

One of the relevant study areas is stock comovement, which identifies homogeneous groups of stocks that have similar movements in terms of returns, trading volumes, and turnover [106]. Previous studies have revealed that stock homogeneity is related to a firms fundamental value [107], investor preferences [108], and uneven information diffusion [82]. In contrast to general stock comovement, media-aware stock comovement emphasizes the short-term impacts of Web media on stock movements. To fully understand media-aware wave effects and superposition effects, a possible solution is to build a media-based enterprise network, as shown in Fig. 6. It first extracts the media features of listed firms from various information sources, including Web news, blogs, microblogs, tweets, and discussion boards. A media-based enterprise network is then constructed according to these media features. Here, each dot represents a listed firm, and the connections between two dots are determined by the media activities of those dots, such as corporate co-exposure in the media, mutual attention from corporate microblogs, and investors' attention paid to both enterprises. It is possible to discover and measure the media-aware wave effect and superposition effect by analyzing the topology of media-based enterprise networks. As a good example of wave effect studies, we can observe the fluctuations of a firm's neighbors in the network once a listed firm is affected by a specific piece of new information. In addition, we can also study a listed firm to understand the superposition effect when it is the neighbor of two other firms that are affected by different news items. However, the efficiency of such an approach has not yet been explored and demands further investigation.

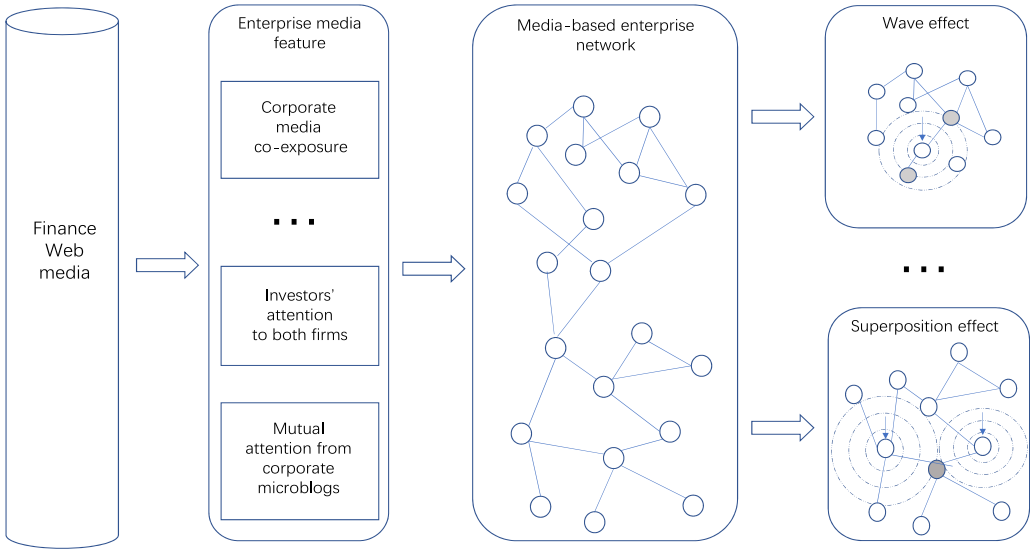


Fig. 6. The framework of media-based enterprise network construction.

5 Conclusion

From the ancient age of transmitting information via smoke signals, to the telegraph era, to the Internet era, information, as an important market factor and price factor, constantly influences and reconstructs financial markets. In this study, we systematically reviewed 229 research articles on media-aware stock movements published between 2007 and 2016 in the fields of finance, MIS, and CS. We endeavored to perform a quantitative analysis to understand, in detail, the mechanisms by which information diffuses through the Web and its impact on stock markets from the perspectives of investors cognitive behaviors, corporate governance, and stock market regulation. The aim of this work was as follows: first, we sought to provide a systematic and comprehensive summary of the contributions and influences of various research activities from different fields on quantifying the influences of Web information on stock markets. Second, we intended to provide a framework for dividing this large challenge into three main issues, i.e., media content, media representation, and analysis model, to clarify the path for further improvements. Third, we sought to provide insightful suggestions for future study based on this comprehensive survey and our hands-on experiences with media-aware stock trading systems [10], [12], [15], [24], [34], [35], [36], [61], [62].

Advancements in quantifying the influence of information on financial markets in the era of social media can be critical for solving several challenging issues from the perspectives of investors cognitive behaviors, corporate governance, and stock market regulation. In particular,

- Investors cognitive behaviors: (a) Will the volume of Web media result in investors being distracted or even ignoring valuable information, causing stock prices to deviate from their intrinsic values? (b) Will public sentiment on social media cause discrepancies between investors? (c) There are various types of Web media, including news articles, blogs, microblogs, and Wikipedia. Does the influence of media vary across different media types? In other words, do investors have selective biases with respect to information sources? Which types of information cause investors to overreact or underreact? (d) What is the influence of rumors in the social media environment?
- Corporate governance: (a) which companies are more vulnerable to breaking news or rumors in the era of social media? (b) Does public opinion on social networks affect a company's business decisions and corporate governance? Will market performance be improved if listed firms make timely and effective disclosures of information via Web media? (c) What are the reasonable measures for listed companies when impacted by rumors? Should they take prompt action to clarify the rumors or remain silent in pursuit of short-term profits?
- Stock market regulation: (a) Is the spread of Web media, especially social media, conducive to reducing information asymmetries among investors, or will it increase illegal, one-sided or false information? (b) Stock markets are always accompanied by rumors. What types of rumors prevail in financial markets? What channels are popularly used to disseminate rumors? (c) To maintain the stability of financial markets, is it possible to monitor the large-scale Web media and perform real-time analysis of its influence on markets to derive early warnings of abnormal returns?

With the recent advancements in computational power, it is possible to quantify the influences of Web media on financial markets, allowing us to better understand the mechanisms of markets, which will be able to protect investors with the most valuable information, assist in corporate management, and provide decision makers with the most reliable inputs for the health of the market.

Acknowledgments

This work has been supported by grants awarded to Dr. Qing Li from National Natural Science Foundation of China (NSFC) (71671141,71401139, 60803106, 61170133),

Fundamental Research Funds for the Central Universities (JBK 171113,JBK 170505,JBK151128,JBK120505) and the Key Lab of Internet Natural Language Processing of Sichuan Province Education Department. It also has been partially funded by grants awarded to Dr. Hsinchun Chen from the US National Science Foundation (ACI-1443019, DUE-1303362, CMMI-1442116, SES-1314631) at the University of Arizona and the China National 1000-Talent Program at the Tsinghua University.

REFERENCES

[1] E. F. Fama, "The behavior of stock-market prices," *J. Bus.*, vol. 38, no. 1, pp. 34–105, 1965.
[2] M. Rechenthin and W. N. Street, "Using conditional probability to identify trends in intra-day high-frequency equity pricing," *Physica A Stat. Mech. Appl.*, vol. 392, no. 24, pp. 6169–6188, 2013.
[3] J. B. D. Long, A. Shleifer, L. H. Summers, and R. J. Waldmann, "Noise trader risk in financial markets," *J. Political Econ.*, vol. 98, no. 4, pp. 703–738, 1990.
[4] A. Shleifer and R. W. Vishny, "The limits of arbitrage," *J. Finance*, vol. 52, no. 1, pp. 35–55, 1997.
[5] T. Berners-Lee, W. Hall, J. Hendler, and D. J. Weitzner, "Creating a science of the web," *Sci.*, vol. 313, no. 5788, pp. 769–771, 2006.
[6] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.
[7] I. Zheludev, R. Smith, and T. Aste, "When can social media lead financial markets?" *Sci. Rep.*, vol. 4, pp. 1–12, 2014.
[8] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," *J. Finance*, vol. 62, no. 3, pp. 1139–1168, 2007.
[9] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," *Manag. Sci.*, vol. 53, no. 9, pp. 1375–1388, 2007.
[10] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, pp. 12:1–12:19, 2009.
[11] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.
[12] Q. Li, Y. Chen, L. L. Jiang, P. Li, and H. Chen, "A tensor-based information framework for predicting the stock market," *ACM Trans. Inf. Syst.*, vol. 34, no. 2, pp. 11:1–11:30, 2016.
[13] S. J. Gray, G. K. Meek, and C. B. Roberts, "International capital market pressures and voluntary annual report disclosures by US and UK multinationals," *J. Int. Financ. Manag. Account.*, vol. 6, no. 1, pp. 43–68, 1995.
[14] W. S. Chan, "Stock price reaction to news and no-news: Drift and reversal after headlines," *J. Financ. Econ.*, vol. 70, no. 2, pp. 223–260, 2003.
[15] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, "The effect of news and public mood on stock movements," *Inf. Sci.*, vol. 278, pp. 826–840, 2014.
[16] X. Luo, J. Zhang, and W. Duan, "Social media and firm equity value," *Inf. Syst. Res.*, vol. 24, no. 1, pp. 146–163, 2013.
[17] Y. Yu, W. Duan, and Q. Cao, "The impact of social and conventional media on firm equity value: A sentiment analysis approach," *Decis. Support Syst.*, vol. 55, pp. 919–926, 2013.
[18] R. Goonatilake and S. Herath, "The volatility of the stock market and news," *Int. Res. J. Finance Econ.*, vol. 3, no. 11, pp. 53–65, 2007.
[19] R. Albuquerque and C. Vega, "Economic news and international stock market co-movement," *Rev. Finance*, vol. 13, no. 3, pp. 401–465, 2009.
[20] L. Fang and J. Peress, "Media coverage and the cross-section of stock returns," *J. Finance*, vol. 64, no. 5, pp. 2023–2052, 2009.
[21] J. E. Engelberg and C. A. Parsons, "The causal impact of media in financial markets," *J. Finance*, vol. 66, no. 1, pp. 67–97, 2011.
[22] V. Sehgal and C. Song, "SOPS: Stock prediction using web sentiment," in *Proc. 7th IEEE Int. Conf. Data Mining Workshops*, 2007, pp. 21–26.
[23] C. Jiang, K. Liang, H. Chen, and Y. Ding, "Analyzing market performance via social media: A case study of a banking industry crisis," *Sci. China Inf. Sci.*, vol. 57, no. 5, pp. 1–18, 2014.
[24] D. Zimbra, H. Chen, and R. F. Lusch, "Stakeholder analyses of firm-related Web forums: Applications in stock return prediction," *ACM Trans. Manag. Inf. Syst.*, vol. 6, no. 1, pp. 2:1–2:38, 2015.
[25] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, 2015.
[26] W. Zhang and S. Skiena, "Trading strategies to exploit blog and news sentiment," in *Proc. 4th Int. AAAI Conf. Weblogs and Social Media*, 2010, pp. 375–378.
[27] S. Y. Yang, S. Y. K. Mo, and A. Liu, "Twitter financial community sentiment and its predictive relationship to stock market movement," *Quant. Finance*, vol. 15, no. 10, pp. 1637–1656, 2015.
[28] T. Li, J. van Dalen, and P. J. van Rees, "More than just noise? Examining the information content of stock microblogs on financial markets," *J. Inf. Technol.*, pp. 1–20, 2017, https://doi.org/10.1057/s41265-016-0034-2
[29] D. Hirshleifer and T. Shumway, "Good day sunshine: Stock returns and the weather," *J. Finance*, vol. 58, no. 3, pp. 1009–1032, 2003.
[30] E. M. Saunders, "Stock prices and Wall Street weather," *Am. Econ. Rev.*, vol. 83, no. 5, pp. 1337–1345, 1993.
[31] M. L. Mitchell and J. H. Mulherin, "The impact of public information on the stock market," *J. Finance*, vol. 49, no. 3, pp. 923–950, 1994.
[32] B. Wüthrich, et al., "Daily stock market forecast from textual Web data," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 1998, pp. 2720–2725.
[33] B. Wang, H. Huang, and X. Wang, "A novel text mining approach to financial time series forecasting," *Neurocomputing*, vol. 83, pp. 136–145, 2011.
[34] R. P. Schumaker and H. Chen, "Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 2, pp. 247–255, 2008.
[35] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decis. Support Syst.*, vol. 53, no. 3, pp. 458–464, 2012.
[36] Q. Li, et al., "Media-aware quantitative trading based on public web information," *Decis. Support Syst.*, vol. 61, pp. 93–105, 2014.
[37] J. Si, A. Mukherjee, B. Liu, S. J. Pan, Q. Li, and H. Li, "Exploiting social relations and sentiment for stock prediction," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1139–1145.
[38] Q. Song, A. Liu, and S. Y. Yang, "Stock portfolio selection using learning-to-rank algorithms with news sentiment," *Neurocomputing*, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.
[39] H. Chen, E. C.-N. Huang, H.-M. Lu, and S.-H. Li, "AZ SmartStock: Stock prediction with targeted sentiment and life support," *IEEE Intell. Syst*, vol. 26, no. 6, pp. 84–88, Nov.-Dec. 2011.
[40] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep learning for event-driven stock prediction," in *Proc. 24th Int. Joint Conf. Artificial Intell.*, 2015, pp. 2327–2333.
[41] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Knowledge-driven event embedding for stock prediction," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 2133–2142.
[42] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY, USA: Addison Wesley Longman Publisher, 1999.
[43] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1/2, pp. 1–135, Jan. 2008.
[44] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," *J. Finance*, vol. 63, no. 3, pp. 1437–1467, 2008.
[45] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Using structured events to predict stock price movement: An empirical investigation." in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1415–1425.
[46] A. Carretta, V. Farina, D. Martelli, F. Fiordelisi, and P. Schwizer, "The impact of corporate governance press news on stock market returns," *Eur. Financ. Manag.*, vol. 17, no. 1, pp. 100–119, 2011.
[47] S. X. Xu and X. M. Zhang, "Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction," *MIS Q.*, vol. 37, no. 4, pp. 1043–1068, 2013.
[48] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowl.-Based Syst.*, vol. 69, pp. 14–23, 2014.
[49] T. Loughran and B. McDonald , "When is a liability is not a liability? Textual analysis, dictionaries, and 10-Ks," *J. Finance*, vol. 66, no. 1, pp. 35–65, 2012.

[50] H. S. Moat, et al.,"Quantifying Wikipedia usage patterns before stock market moves," *Sci. Rep.*, vol. 3, pp. 1–5, 2013.

[51] C. Curme, T. Preis, H. E. Stanley, and H. S. Moat, "Quantifying the semantics of search behavior before stock market moves," *Proc. Natl. Acad. Sci.*, vol. 111, no. 32, pp. 11 600–11 605, 2014.

[52] M. Z. Frank and W. Antweiler, "Is all that talk just noise? The information content of internet stock message boards," *J. Finance*, vol. 59, no. 3, pp. 1259–1294, 2004.

[53] S.-H. Kim and D. Kim, "Investor sentiment from internet message postings and the predictability of stock returns," *J. Econ. Behav. Organ.*, vol. 107, pp. 708–729, 2014.

[54] A. Urquhart and R. Hudson, "Efficient or adaptive markets? Evidence from major stock markets using very long run historic data," *Int. Rev. Financ. Anal.*, vol. 28, pp. 130–142, 2013.

[55] V. Lavrenko, et al., "Language models for financial news recommendation," in *Proc. 9th Int. Conf. Inform. Knowl. Manage.*, 2000, pp. 389–396.

[56] M.-A. Mittermayer and G. F. Knolmayer, "Newscats: A news categorization and trading system," in *Proc. IEEE 6th Int. Conf. Data Mining*, 2006, pp. 1002–1007.

[57] T. Preis, H. S. Moat, and H. E. Stanley, "Quantifying trading behavior in financial markets using Google trends," *Sci. Rep.*, vol. 3, pp. 1–6, 2013.

[58] S. Y. Yang, S. Y. K. Mo, A. Liu, and A. A. Kirilenko, "Genetic programming optimization for a sentiment feedback strength based trading strategy," *Neurocomputing*, vol. 264, pp. 29–41, 2017.

[59] M. Makrehchi, S. Shah, and W. Liao, "Stock prediction using event-based sentiment analysis," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, 2013, pp. 337–342.

[60] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 306–324, 2015.

[61] R. P. Schumaker and H. Chen, "A quantitative stock prediction system based on financial news," *Inf. Process. Manag.*, vol. 45, no. 5, pp. 571–583, 2009.

[62] R. P. Schumaker and H. Chen, "A discrete stock price prediction engine based on financial news," *Comput.*, vol. 43, no. 1, pp. 51–56, 2010.

[63] P. C. Tetlock, "Does public financial news resolve asymmetric information?" *Rev. Financ. Stud.*, vol. 23, no. 9, pp. 3520–3557, 2010.

[64] Y. Kara, M. A. Boyacioglu, and Ö. K. Baykan, "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5311–5319, 2011.

[65] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, "Wisdom of crowds: The value of stock opinions transmitted through social media," *Rev. Financ. Stud.*, vol. 27, no. 5, pp. 1367–1403, 2014.

[66] C. Dougal, J. Engelberg, D. Garcia, and C. A. Parsons, "Journalists and the stock market," *Rev. Financ. Stud.*, vol. 25, no. 3, pp. 639–679, 2012.

[67] A. Gro ß-Klußmann and N. Hautsch, "When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions," *J. Empir. Finance*, vol. 18, no. 2, pp. 321–340, 2011.

[68] D. H. Solomon, "Selective publicity and stock prices," *J. Finance*, vol. 67, no. 2, pp. 599–638, 2012.

[69] J. M. Griffin, N. H. Hirschey, and P. J. Kelly, "How important is the financial media in global markets?" *Rev. Financ. Stud.*, vol. 24, pp. 3941–3992, 2011.

[70] T. Dimpfl and S. Jank, "Can internet search queries help to predict stock market volatility?" *Eur. Financ. Manag.*, vol. 22, no. 2, pp. 171–192, 2016.

[71] G. Birz and J. R. Lott, "The effect of macroeconomic news on stock returns: New evidence from newspaper coverage," *J. Bank. Finance*, vol. 35, no. 11, pp. 2791–2800, 2011.

[72] M. Hagenau, M. Liebmann, and D. Neumann, "Automated news reading: Stock price prediction based on financial news using context-capturing features," *Decis. Support Syst.*, vol. 55, no. 3, pp. 685–697, 2013.

[73] M. J. Beechey and J. H. Wright, "The high-frequency impact of news on long-term yields and forward rates: Is it real?" *J. Monet. Econ.*, vol. 56, no. 4, pp. 535–544, 2009.

[74] M. Alanyali, H. S. Moat, and T. Preis, "Quantifying the relationship between financial news and the stock market," *Sci. Rep.*, vol. 3, pp. 1–6, 2013.

[75] K. P. Evans, "Intraday jumps and US macroeconomic news announcements," *J. Bank. Finance*, vol. 35, no. 10, pp. 2511–2527, 2011.

[76] R. Luss and A. d'Aspremont, "Predicting abnormal returns from news using text classification," *Quant. Finance*, vol. 15, no. 6, pp. 999–1012, 2015.

[77] T. Gilbert, "Information aggregation around macroeconomic announcements: Revisions matter," *J. Financ. Econ.*, vol. 101, no. 1, pp. 114–131, 2011.

[78] N. R. Sinha, "Underreaction to news in the US stock market," *Q. J. Finance*, vol. 6, no. 2, pp. 1 650 005:1–1 650 005:46, 2016.

[79] J. Peress, "The media and the diffusion of information in financial markets: Evidence from newspaper strikes," *J. Finance*, vol. 69, no. 5, pp. 2007–2043, 2014.

[80] K. R. Ahern and D. Sosyura, "Rumor has it: Sensationalism in financial media," *Rev. Financ. Stud.*, vol. 28, pp. 2050–2093, 2015.

[81] P. Nizer and J. C. Nievola, "Predicting published news effect in the Brazilian stock market," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10 674–10 680, 2012.

[82] L. Liu, J. Wu, P. Li, and Q. Li, "A social-media-based approach to predicting stock comovement," *Expert Syst. Appl.*, vol. 42, no. 8, pp. 3893–3901, 2015.

[83] B. Luo, J. Zeng, and J. Duan, "Emotion space model for classifying opinions in stock message board," *Expert Syst. Appl.*, vol. 44, pp. 138–146, 2016.

[84] N. Oliveira, P. Cortez, and N. Areal, "Stock market sentiment lexicon acquisition using microblogging data and statistical measures," *Decis. Support Syst.*, vol. 85, pp. 62–73, 2016.

[85] Y. Shynkevich, T. McGinnity, S. A. Coleman, and A. Belatreche, "Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning," *Decis. Support Syst.*, vol. 85, pp. 74–83, 2016.

[86] I. Medovikov, "When does the stock market listen to economic news? New evidence from copulas and news wires," *J. Bank. Finance*, vol. 65, pp. 27–40, 2016.

[87] F. Lillo, S. Micciché, M. Tumminello, J. Piilo, and R. N. Mantegna, "How news affects the trading behaviour of different categories of investors in a financial market," *Quant. Finance*, vol. 15, no. 2, pp. 213–229, 2015.

[88] B. S. Kumar and V. Ravi, "A survey of the applications of text mining in financial domain," *Knowl.-Based Syst.*, vol. 114, pp. 128–147, 2016.

[89] M. Ammann, R. Frey, and M. Verhofen, "Do newspaper articles predict aggregate stock returns?" *J. Behav. Finance*, vol. 15, no. 3, pp. 195–213, 2014.

[90] T.-T. Vu, S. Chang, Q. T. Ha, and N. Collier, "An experiment in integrating sentiment features for tech stock prediction in twitter," in *Proc. Workshop Inform. Extraction Entity Analytics Social Media Data*, 2012, pp. 23–38.

[91] X. Li, X. Huang, X. Deng, and S. Zhu, "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information," *Neurocomputing*, vol. 142, pp. 228–238, 2014.

[92] J. Francis, J. Douglas Hanna, and D. R. Philbrick, "Management communications with securities analysts," *J. Account. Econ.*, vol. 24, no. 3, pp. 363–394, 1997.

[93] P. L. Davies and M. Canes, "Stock prices and the publication of second-hand information," *J. Bus.*, vol. 51, no. 1, pp. 43–56, 1978.

[94] B. M. Barber and D. Loeffler, "The "Dartboard" column: Second-hand information and price pressure," *J. Financ. Quant. Anal.*, vol. 28, no. 2, pp. 273–284, 1993.

[95] I. Mathur and A. Waheed, "Stock price reactions to securities recommended in business week's "Inside Wall Street"," *Financ. Rev.*, vol. 30, no. 3, pp. 583–604, 1995.

[96] P. Clarkson, D. Joyce, and I. Tutticci, "Market reaction to takeover rumour in internet discussion sites," *Account. Finance*, vol. 46, no. 1, pp. 31–52, 2006.

[97] J. M. Zhao, X. He, and F. Y. Wu, "Study on Chinese stock market rumors and the impact on stock prices," *Manag. World*, vol. 11, pp. 38–51, 2010.

[98] G. Huberman and T. Regev, "Contagious speculation and a cure for cancer: A nonevent that made stock prices soar," *J. Finance*, vol. 56, no. 1, pp. 387–396, 2001.

[99] B. R. Marshall, N. Visaltanachoti, and G. Cooper, "Sell the rumour, buy the fact?" *Account. Finance*, vol. 54, no. 1, pp. 237–249, 2014.

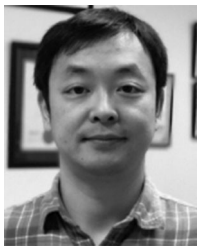
[100] X. Yang and Y. Luo, "Rumor clarification and stock returns: Do bull markets behave differently from bear markets?" *Emerg. Markets Finance Trade*, vol. 50, no. 1, pp. 197–209, 2014.

[101] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.
[102] G. Hinton, et al.,"Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[103] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Int. Conf. Neural Inform. Process. Syst.*, 2014, pp. 3104–3112.
[104] G. B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern., Syst. B (Cybern.)*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[105] X.-K. Wei, Y.-H. Li, and Y. Feng, "Comparative study of extreme learning machine and support vector machine," in *Proc. Int. Symp. Neural Netw.*, 2006, pp. 1089–1095.
[106] R. Morck, B. Yeung, and W. Yu, "The information content of stock markets: Why do emerging markets have synchronous stock price movements?" *J. Financ. Econ.*, vol. 58, no. 1, pp. 215–260, 2000.
[107] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, 1970.
[108] M. S. Rashes, "Massively confused investors making conspicuously ignorant choices (mci–mcic)," *J. Finance*, vol. 56, no. 5, pp. 1911–1927, 2001.

Qing Li received BS and MS degrees from Harbin Engineering University, China and the PhD degree from Kumoh National Institute of Technology, in Feb. 2005. He is a professor with Southwestern University of Finance and Economics, China. Prior to taking that post, he was a postdoctoral researcher with Arizona State University and Information & Communications University of Korea. His research interests lie primarily in intelligent information processing and business intelligence. He served on the editorial board of the *Electronic Commerce Research and Applications*, the *Journal of Database Management*, and the *Journal of Global Information Management*, and the program committees of various international conferences including PACIS, SIGIR, CIKM. He is a member of the IEEE.

Yan Chen received the bachelor's degree from Southwestern University of Finance and Economics, in 2015. He is working toward the PhD degree at Southwestern University of Finance and Economics, China. His research interests lie primarily in data mining and financial intelligence.

Jun Wang is working toward the PhD degree at Southwestern University of Finance and Economics, China. He is also a visiting researcher with Memorial University of Newfoundland in St. John's, Newfoundland. He was awarded the National Scholarship in 2017. His research interests lie primarily in social media, social network analysis, financial analysis and business intelligence.



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

Hsinchun Chen received the BS degree from the National Chiao-Tong University, the MBA degree from SUNY Buffalo, and the MS and PhD drgrees from New York University. He is the University of Arizona Regents' professor and Thomas R. Brown chair professor in management and technology. He is recently served as the lead program director of the Smart and Connected (SCH) Program at the US National Science Foundation (2014-2015), a multi-year multi-agency health IT research program of USA. He is author/editor of 20 books, 290 SCI journal articles, and 160 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. He has served as editor-in-chief of major ACM/IEEE, and Springer journals and conference/program chair of major ACM/IEEE/MIS conferences in digital library, information systems, security informatics, and health informatics. He is also a successful IT entrepreneur. His COPLINK/i2 system for security analytics was commercialized in 2000 and acquired by IBM as its leading government analytics product in 2011. He is internationally renowned for leading the research and development in the health analytics (data and text mining; health big data; DiabeticLink and SilverLink) and security informatics (counter terrorism and cyber security analytics; security big data; COPLINK, Dark Web and Hacker Web) communities. He is a fellow of the ACM, the IEEE and the AAAS. See: http://ai.arizona.edu/hchen.