Stock Markets Using Web Media Content

Ganesh Babu N\textsuperscript{1}, Selvi R\textsuperscript{2}

\textsuperscript{1}PG Scholar, Department of Computer Science & Engineering, Sri Ramanujar Engineering College, Chennai
\textsuperscript{2}Assistant Professor, Department of Computer Science & Engineering, Sri Ramanujar Engineering College, Chennai

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 23 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.

Keywords—stock market, dissemination, web media, management information system, corporate governance.

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.

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/micro blogs, 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 artificial intelligence (AI), researchers have begun to capture the influence of media on stocks and bridge these connections from a big data perspective.

![Fig.1. Publication Overview](image)

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 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. 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/micro blogs, discussion boards, and social news, have emerged and played important roles in affecting stock 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.

![Fig. 2. The framework of the media-based movement analysis.](image-url)
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 co-movements. [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, micro blogs, 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 first revealed the predictability of the effect of tweets on stock movements. [26] extracted sentiment from news, blogs, and micro blogs 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.
<table>
<thead>
<tr>
<th>Category</th>
<th>Literature</th>
<th>Focus</th>
<th>Analysis Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scale</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Media Source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Response</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Features</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. **DJIA, NYSE**
   - Day
   - WSJ, Dow Jones News wires
   - Return
   - Emotion word number
   - Linear

2. **DJIA, Nasdaq, S&P 500**
   - Day
   - CNN News, Fox News, Yahoo News, NY Times
   - Index
   - News number model
   - Linear

3. **PSI-20, DJ-30**
   - Day
   - Bloomberg, MMS, Reuters, I/B/E/S
   - Returns
   - News/non-News autoregressive model
   - Linear

4. **NYSE, NASDAQ**
   - Month
   - Return, Volatility
   - News coverage model
   - Linear

5. **S&P 500**
   - Day
   - Yahoo! Finance, ProQuest News
   - Volumes
   - News coverage model
   - Linear

6. **Morgan Stanley High-Tech Index CSI100**
   - Day
   - Yahoo! Finance, Index, volumes, volatility
   - Sentiment model
   - Linear regression

7. **Discussion Boards**
   - Day
   - Eastmoney Forum
   - Price
   - Current price, Partial news terms, social
   - SVR

8. **52 listed firms**
   - Day
   - Yahoo! Finance, Price
   - Sentiment model
   - Statistical

9. **listed firms**
   - Day
   - Yahoo! Finance, Price
   - Sentiment model
   - Linear

10. **6 listed firms**
    - Day
    - Yahoo! Finance, Return
    - message length & model numbers, sentiment, transaction records
    - Linear

11. **18 listed firms**
    - Day
    - Yahoo! Finance, Price
    - Sentiment, words
    - Graphic

12. **DJIA**
    - Day
    - Twitter
    - Index
    - Past DJIA, emotions
    - SOFNN

13. **NYSE**
    - Day
    - Dailies, Twitter, Spinn3r
    - Return, volume
    - Sentiment
    - Correlation

14. **NYSE, NASDAQ, AMEX**
    - Day
    - RSS Feeds, LiveJournal Blogs Twitter
    - Return, volatility
    - Sentiment
    - Multivariate Regression

15. **S&P 100**
    - Day
    - Twitter
    - Return, Volume, volatility
    - Sentiment, Message
    - Regression

    - Minute
    - Number

---

Table 1 Literature Comparison in Terms of Web Content

@IJRTER-2019, All Rights Reserved
Fig. 3 Publication trends by media type.

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, macro-economic 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 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 nouns, 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 Opinion-Finder, 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 air-line” 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 Twitter 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 auto regression 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 non-linear 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 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. The most popular method for representing text is the 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.

<table>
<thead>
<tr>
<th>Table 3 Journal Influence in Terms of Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>J FINANC</strong></td>
</tr>
<tr>
<td><strong>J COMPUT SCI</strong></td>
</tr>
<tr>
<td><strong>MANAGE SCI</strong></td>
</tr>
<tr>
<td><strong>REV FINANC STUD</strong></td>
</tr>
<tr>
<td><strong>SCI REPORTS</strong></td>
</tr>
</tbody>
</table>
The first category is the existence category, which focuses on demonstrating the co-movements between stocks and various types of media including news, discussion boards, and micro blogs. [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. 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 exploring 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.

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 micro blogs. 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 micro blogs 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 micro blogs, 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 take-over 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.

4.4 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 feed forward 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 sub sampling 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.

4.5 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 co-movements 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 co-movement, 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 firm's fundamental value [107], investor preferences [108], and uneven information diffusion [82]. In contrast to general stock co-
movement, media-aware stock co-movement 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.

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.

Fig. 3. The framework of media-based enterprise network construction

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, micro-blogs, 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 under react? (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).
Fig. 4. The media-based enterprise working construction

REFERENCES


