Artif Intell Rev (2018) 50:49–73
https://doi.org/10.1007/s10462-017-9588-9

Natural language based financial forecasting: a survey

Frank Z. Xing[1] · **Erik Cambria**[1] · **Roy E. Welsch**[2]

Published online: 27 October 2017

Abstract Natural language processing (NLP), or the pragmatic research perspective of computational linguistics, has become increasingly powerful due to data availability and various techniques developed in the past decade. This increasing capability makes it possible to capture sentiments more accurately and semantics in a more nuanced way. Naturally, many applications are starting to seek improvements by adopting cutting-edge NLP techniques. Financial forecasting is no exception. As a result, articles that leverage NLP techniques to predict financial markets are fast accumulating, gradually establishing the research field of natural language based financial forecasting (NLFF), or from the application perspective, stock market prediction. This review article clarifies the scope of NLFF research by ordering and structuring techniques and applications from related work. The survey also aims to increase the understanding of progress and hotspots in NLFF, and bring about discussions across many different disciplines.

Keywords Financial forecasting · Natural language processing · Text mining · Predictive analytics · Knowledge engineering · Computational finance

1 Introduction

Utilizing textual data to improve modeling of the financial market dynamics has long been the tradition of trading practice. The growing volume of financial reports, press releases, and

✉ Erik Cambria
cambria@ntu.edu.sg

Frank Z. Xing
zxing001@e.ntu.edu.sg

Roy E. Welsch
rwelsch@mit.edu

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[2] MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA

news articles also galvanizes the wish to run this analysis automatically to keep a competitive business advantage, which at least dates back to the 1980's. Interestingly, this is the time that solely exploring historical data became more difficult. According to the analysis of Qian and Rasheed (2004) using the Hurst exponent, the correlation between Dow Jones daily returns and its historical data receded from the 1990's. Apart from econometricians' increasingly complicated pattern mining models, the earliest attempts to import other predictors employed discourse analysis techniques developed from linguistics (Frazier et al. 1984) and Naïve statistical methods such as word spotting (Brachman and Khabaza 1996). However, the idea of automatically analyzing textual information has made little progress for years for many reasons from different aspects. For example, the most popular language model earlier, was bag-of-words, which may not be adequate to the task of comprehensive or deep understanding; the paradigm of knowledge engineering research also bounds the focus on a small portion of highly structured texts.

The construction of ontologies or semantic networks relies on very reliable and noise-free materials, while information about corporations from Internet Stock Message Boards (SMB) and forum discussions (Antweiler and Frank 2004) are seldom considered. In the first decade of this century, the standard financial news analyzing system usually involved a mixed collection of news articles and stock quotes, as described in Schumaker and Chen (2009). News articles are represented with concatenated vectors, for instance, word frequencies together with a one-hot representation of key noun phrases and name entities. Popular machine learning algorithms at that time, usually support vector machines (SVM) (Fung et al. 2003) or evolutionary heuristics (Brabazon and O'Neill 2008), are applied to blend the vector feature with numerical data, to predict stock movements.

From 2010 onward, social media websites such as Twitter, Facebook, etc., have generated an exponentially increasing amount of user content, the news analytics community once developed a special interest in mining this real-time information (Cambria et al. 2014). Numerous papers especially pore over Twitter contents because of the relatively simple semantics conveyed in a restricted character length (Bollen et al. 2011; Si et al. 2013; Wei et al. 2016). Besides of the enrichment in different types of text sources, in this stage, more sophisticated NLP techniques are proposed. Sentiment analysis resources, such as Opinion Lexicon (Hu and Liu 2004), are proposed; topic model (Blei 2012) is used to discover both aspect and the related sentiment (Nguyen and Shirai 2015). Machine Learning methods and knowledge-based techniques are simultaneously used for sentiment analysis as a core component. Neural networks, including a myriad of deep learning variants like convolutional neural networks (CNN) (Ding et al. 2015), restricted Boltzmann machines (RBM) (Yoshihara et al. 2016), long short-term memory (LSTM) networks (Li et al. 2016), etc., are experimented with prediction algorithms. Sometimes these models are also applied together with classic time series models such as autoregressive integrated moving average (ARIMA) (Zhang 2003; Liu et al. 2016). Stepping back for a holistic view, we are at the dawn of the semantics curve of NLP technologies (Cambria and White 2014). NLP systems start to approach human understanding accuracy at the sentence level. Therefore, it is reasonable to expect a long period to witness different approaches to compete before we could reach the next narrative curve within the framework of NLFF.

To provide a landscape of the hotspots, methods, and findings of NLFF research, we survey the most important studies by ordering and structuring them from many different perspectives. We use the following query to search for the relevant literature included in Scopus database: *(TITLE-ABS-KEY("text mining") OR TITLE-ABS-KEY("textual") OR TITLE-ABS-KEY("sentiment analysis")) AND ((TITLE-ABS-KEY("financial") OR*

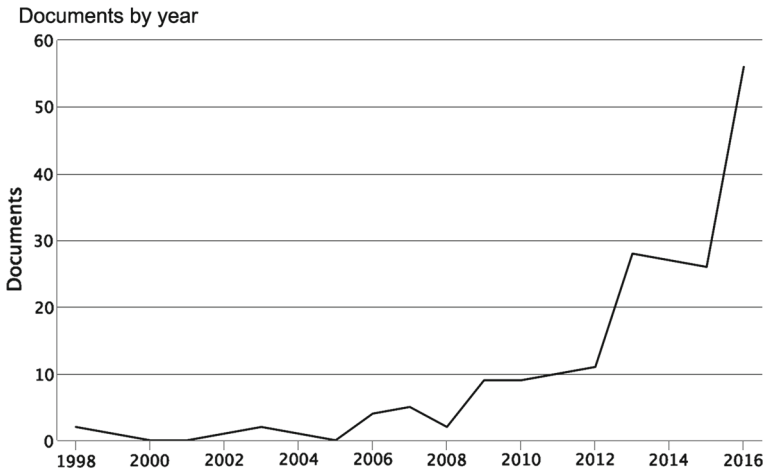

Fig. 1 Research articles published by year (1998–2016)

TITLE-ABS-KEY("stock market")) AND (TITLE-ABS-KEY("prediction") OR TITLE-ABS-KEY("forecasting"))). Figure 1 shows the recent exponential increase of papers in this field.

It is quite interesting that, though financial forecasting covers a wide range of ideas from inflation rate prediction to credit scoring (Tkác and Verner 2016; Schneider and Gupta 2016), a large proportion of the studies that employed textual data focus on stock market and foreign exchange rate (FOREX) prediction. We owe this special appeal of stock and currency markets to three main reasons.

- *Lack of accessibility for many assets* Corporate financial statements are usually internally archived or from scattered sources. For the current stage, it's difficult to agglomerate information from these materials.
- *The nature of other financial products* Treasury securities have simple and policy driven term structure of interest rates. As a result, the correlation between mass textual information and interest rates movement is weak; On the contrary, derivatives have complicated pricing mechanisms and constrained information transparency. These characteristics make the market a gray, chaotic system, which is very sensitive to perturbation. Therefore, any delayed estimation of public mood or topic is not really useful for prediction.
- *Transparency of stock and currency markets* These markets usually have a large capitalization and many participants, which gives weight to the massive opinions of the investors or participants. Public information on the stock market is much more available.

Given the long history and the above properties of stock markets, it is a good venue for discovering and testing our knowledge distilled from the financial markets. Despite the fact that few of the forecasting systems reported in the literature have been shown to make a profit in the long run with transaction cost deducted, many meaningful hypotheses and significant observations have been drawn from stock market data. Figure 2 provides an intuitive grasp of the scope of NLFF. At the intersection of NLP and financial forecasting, NLFF brings together topics that would interest both fields, such as sentic computing, natural language understanding, time series analysis and more.

The rest of this survey is organized as follows: Sect. 2 provides a historical view of how currently approved NLP techniques are derived, plus some basic knowledge of time series modeling; Sect. 3 enumerates and discusses several mainstream philosophies and the

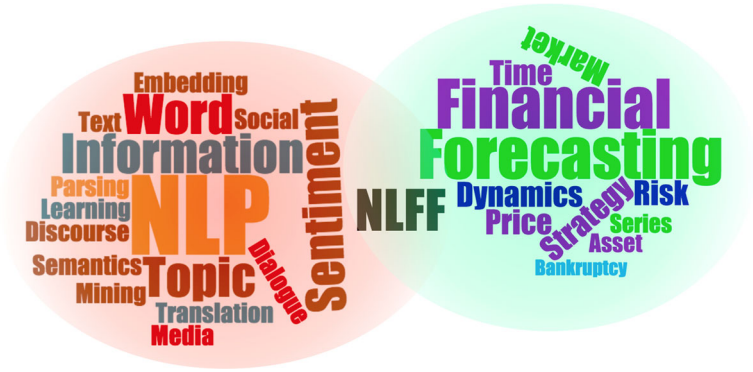

Fig. 2 A word cloud illustration of NLFF bridging the research scope of NLP and financial forecasting

motivation behind different forecasting frameworks; Sect. 4 reviews existing studies from three angles: text source and processing techniques, algorithms for predictive models, and result evaluation; finally, Sect. 5 concludes this survey and proposes future research directions.

2 Background

2.1 Semantic modeling

The idea that language is a set of lexicons and, at the same time, a syntactic system (Bühler 1934) has been proposed even before the inception of NLP. Aligned with this tradition, the early popular approaches of NLP research as well take a view that emphasizes either the expressiveness (Sowa 1987) or language rules (Chomsky 1956). Most of the diversified NLP techniques developed and applied on NLFF these days can still fit into these two categories, or a mix of them.

To represent textual financial data as features that can be easily processed by a computer, most of the early NLFF papers have employed bag-of-words, which represent the semantics of a piece of text by the set of words and the frequency of their appearance. Stopword lists are often used to filter out function words such as "a", "the" etc. An obvious drawback of this technique is that word order is not taken into consideration. This problem can be serious in certain cases. For example, the financial news "Samsung now is gaining advantages on Apple" and "Apple now is gaining advantages on Samsung" lead to opposite reaction in the market, though they share the same bag-of-words representation. Another drawback is, when one meaning is phrased by different words, such as in "Brexit caused a drop in the pound" and "Leaving the EU accelerates pound's slump", this semantic similarity will not be captured.

These problems are well addressed by considering a word with its context. A family of neural network models can be leveraged to generate distributed and compact representation of words (Bengio et al. 2003). With the recent advances in deep learning, this vector representation, or word embedding (Mikolov et al. 2013; Chaturvedi et al. 2016) is better formed. This representation makes it possible to compute semantic similarities. Beside word representations, topic models (Blei 2012) capture the semantics of a collection of documents on a grand scale. At the document level, semantics is discomposed to multiple topics and corresponding relevance coefficients. These techniques enable the analysis of a large volume of financial articles as a whole.

2.2 Sentiment analysis

Sentiment analysis (Cambria 2016) is a 'suitcase research problem' (Cambria et al. 2017) that requires tackling many NLP sub-tasks, including aspect extraction (Poria et al. 2016a), subjectivity detection (Chaturvedi et al. 2017), named entity recognition (Ma et al. 2016), and sarcasm detection (Poria et al. 2016b), but also complementary tasks such as personality recognition (Majumder et al. 2017), user profiling (Mihalcea and Garimella 2016) and multimodal fusion (Poria et al. 2016c). Sentiment analysis is yet another important perspective for NLFF due to the interactive nature of financial activities. According to the five-eras vision of the future web (Owyang 2009), market sentiment will become a prominent factor that influences trading and information flow as well as shaping products and services. This research area of sentiment analysis flourishes along with the trend of Web 2.0. Existing approaches to affective computing fall into three categories: knowledge-based techniques, statistical methods, and hybrid approaches (Poria et al. 2017). Knowledge-based techniques derive from and leverage early age large scale resource-building projects, such as Cyc (Guha and Lenat 1990), Open Mind Common Sense (OMCS) from which ConceptNet (Liu and Singh 2004) was built, and WordNet (Fellbaum 1998). Along with different psychological theories of emotion, computational models of the representation of sentiment were proposed (Marsella and Gratch 2014). Models that take discrete theories of emotion assign core emotion labels to words, for example, WordNet-Affect (Valitutti 2004). Further generalization can categorize words into positive and negative ones according to the primary core emotion, for example, Opinion Lexicons (Hu and Liu 2004). Models that consider dimensional or appraisal theories of emotion add more factors such as subjectivity and intensity to the knowledge base. SentiWordNet (Baccianella et al. 2010) is a good representative. Other popular open domain sources include SenticNet (Cambria et al. 2016), which contains entries at the concept level to tackle the problem of phrases and multiword expressions (Sag et al. 2002; Cambria 2013).

In the financial domain, there are several widely used hand-crafted public resources developed by economists, such as the General Inquirer (Kelly 1975), the Henry Word List (Henry 2008), and the Loughran & McDonald Word List (Loughran and McDonald 2011). Wuthrich et al. (1998) in their pioneer work have also used around 400 expert crafted keyword tuples as influential factors of market movements. Recently, there have been other attempts to automatically build lexicons for the financial domain (Tai and Kao 2013; Hamilton et al. 2016). Both papers used a label propagation framework from some seed words. However, the financial lexicons produced by Tai and Kao (2013) have not been made public. Instead of the sentiment polarity value, there are different fine-grained sentiment spaces that can be applied to financial forecasting. For instance, SenticNet stores four-dimensional values of the hourglass model (Cambria et al. 2012), which is derived from Plutchik's wheel of emotions model. On the other hand, a rather different sentiment space empirically proposed to scale mood aptitude, or subjectivity, by some psychologists called Profile of Mood States (POMS), is quite popular among researchers of finance. The original form of POMS (Shacham 1983) consists of six factors: tension–anxiety, depression–dejection, anger–hostility, fatigue–inertia, vigor–activity, and confusion–bewilderment. Different modified versions of POMS and tools that adopted this idea, such as OpinionFinder (Wilson et al. 2005), are crucial components in the NLFF framework of many studies (Bollen et al. 2011; Nofer and Hinz 2015). These factors are not necessarily independent because redundant representations of sentiment states can be useful. Furthermore, applications of sentiment analysis in pragmatic systems can also be carried out at different levels. The Stock Sonar (Feldman 2013) used to conduct sentiment analysis at both the word level and phrase level. At the end, the system will do polarity classification at a document level.

2.3 Event extraction

Statistical methods extract conjunctions between words, usually depending on a large annotated corpus. For example, Hatzivassiloglou and McKeown (1997) uses a 21 million word Wall Street Journal corpus to mine the relations between adjectives such as "and", "or", "but". As a result, much knowledge about financial phenomena and descriptions can be obtained. Also, these meaningful narratives can be fed into deep neural networks to produce vector representations. For example, Ding et al. (2015) introduced the idea of using deep learning to embed events, which are Actor-Action-Object-Time tuples such as "Google acquires Nest on Jan 13, 2014".

Apart from the above-mentioned context-aware and sentiment analysis approaches, more fundamental NLP techniques that help to analyze text structure, for example, parse trees, POS tagging, named entity recognition, and event modeling (Malik et al. 2011; Ding et al. 2015) are applied as infrastructure for NLFF as well (Schumaker and Chen 2009). Some recent research indicates that a combination of subjective sentiment and objective event facts would take advantage of each other and produce a better forecasting result (Ding 2016).

Based on the capability to extract semantics and sentiments from natural language, the problem of financial forecasting can be modeled at a more abstract level. Time series analysis is a classic technique which gives more weight to endogenous factors. Some research adopts time series model like ARIMA (Liu et al. 2016), generalized autoregressive conditional heteroskedasticity (GARCH) (Moniz and de Jong 2014), and combine it with machine learning techniques. To contemplate introducing more external impact, a monitor of happening activities is required. This can be achieved either by following the framework of, for example, Open Information Extraction (Banko et al. 2007), or leveraging on existing event databases, such as GDELT (Leetaru and Schrodt 2013) or ICEWS (Weidmann and Ward 2010).

3 Philosophy behind financial forecasting

The scope of the financial forecasting task is categorized into a two layer taxonomy according to Kumar and Ravi (2016). In a narrow sense, financial forecasting should cover prediction of key indicators, such as price, volatility, volume and so forth, in FOREX and the stock market. In a broader sense, cyber security affairs like fraud detection and in service, supply chain management, are discussed as well. For market prediction, most studies justify their effectiveness by the goodness of approximation for the realized time series to their prediction. With this capability, in trading simulations they will provide excess return compared to average market participants. For tasks such as credit scoring and customer relationship management, those companies that adopt good forecasting techniques will outperform ignorant competitors in the rapidly changing business environment. A fundamental question to address here is "Where does the excess return come from?"

The most acknowledged answer lays on the negation of the efficient market hypothesis (EMH) (Fama 1970) in a real world case. Actually, if all the participants in a market are "informationally efficient", all deals would be conducted at a fair value. The excess return should come from the passionate or noise traders, which further offends the hypothesis of rational man. To reconcile this problem, behavioral economics has come up with theories that are compatible with the interactive nature of the market and participants, such as the adaptive market hypothesis (AMH). Then excess return can be ascribed to information asymmetry. More recently, as the concept of information overload comes into vision, we realized that

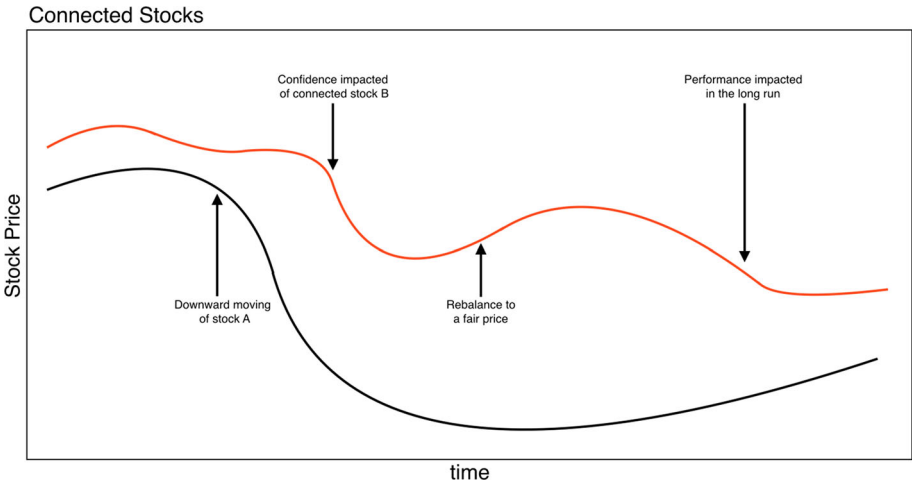

Fig. 3 An illustration of co-movement of connected stocks

even for a market that is "informationally efficient", the ability to quickly utilize and mine the information can be very different among participants. As in heterogeneous agent models (HAM) (Hommes 2006), stock cycles still appear in efficient market.

Traditionally, there are two schools of thought regarding what information to resort to. Technical analysts (Park and Irwin 2004) believe there exist patterns or motifs that would repeat in the future. Consequently, many data mining techniques are applied to historical data to find these patterns (Hajizadeh et al. 2010). Sometimes, these computational methods can be used together with existing technical indicators, such as moving average convergence divergence (MACD). In this case, there seems to be no difficulty in locating the information, but the speed and mining power is crucial. While for fundamental analysis, what information to look at is of more significance. Since many macroeconomic factors are unstructured and scattered from different sources, it is a field where text mining and NLP techniques are frequently employed.

From the perspective of Artificial Intelligence, three sources of information have been most heavily exploited (Ding et al. 2015). Historical time series data such that used by technical analysis, semantic features, and sentiment information extracted from the financial news as valued by fundamental analysis. The latter two sources often involve NLFF techniques.

3.1 A spectrum of perspectives

According to our observations, the intriguing task of financial forecasting attracts researchers with both computer science and finance backgrounds. The ways they formalize this task are diversified. However, these thoughts are compatible with each other. They form a spectrum of perspectives together. We list four typical perspectives as follows.

The connectionist perspective Economists hold the belief that assets in similar sectors will have similar behavior due to the fundamental environment. Corporations that are involved in the same manufacturing chain are also connected in some sense (Cohen and Frazzini 2008), as illustrated in Fig. 3. Market participants are not able to pay full attention to all the assets. This limited attention will induce stock price to under-react to firm-specific information that would potentially influence unmentioned firms as well. Discovering these related firms will generate

return predictability across assets. Moreover, based on the analysis of a natural experiment of the 2003 mutual fund trading scandal, the co-movement of stocks can further be caused by their shared ownership (Anton and Polk 2014). Intuitively, the aggregational behavior of these agents will be reflected in the movement of price. This observation lays another layer under the relation between stocks, hence gives birth to the connectionist perspective. The trading strategy of using abnormally connected ratios derives from it. Prior to this, the practice of finding connected stocks has been explored in real life stock trading. Despite the fact that the underlying mechanism is often uninvestigated, a stock trader will always be interested in finding out the inter-relationships among stocks, such that the movement of one stock could trigger the movements of some other stocks (Fung et al. 2003). The mainstream way to dig into connections is data mining techniques. However, textual data can be used to drive the connection discovery as well.

The portfolio management perspective When constructing trading strategies, certain constraints may significantly change the effectiveness in practice. As market participants allocate their capital into different assets, the portfolio management, or portfolio selection problem is described as simultaneously achieving two goals: maximizing the return and minimizing the risk in the classic Markowitz theory. A standard Markowitz mean–variance model for portfolio selection can be formalized as:

$$\begin{aligned} &\text{minimize} && \overbrace{\lambda\left[\sum_{i=1}^{N}\sum_{j=1}^{N} x_i\sigma_{ij}x_j\right]}^{\text{risk item}} + \overbrace{(1-\lambda)\left[-\sum_{i=1}^{N}\mu_i x_i\right]}^{\text{return item}} \\ &\text{subject to} && \sum_{i=1}^{N} x_i = 1,\ i = 1, 2, \ldots, N \end{aligned}$$

where μ_i is the mean return of asset i; σ_{ij} is the covariance between returns from assets i and j; $0 < \lambda < 1$ is the risk aversion parameter. The proportion of asset i in the portfolio x_i can be negative if it is possible to short this asset.

Therefore, portfolio management can be formalized as an optimization problem. Machine learning techniques can be actively used in solving this optimization problem when asset prices are fast-changing (Li et al. 2015) or weights are allocated across assets sparsely (Shen et al. 2014). More probabilistic modeling of how the rebalance actions can be taken will result in more complicated (hence more general) portfolio representations, such as Bayesian Portfolio Analysis (Avramov and Zhou 2010) and Stochastic Portfolio Theory (Samo and Vervuurt 2016). However, the mutual idea is that the excess return, which is often referred to as "alpha" in portfolio management theory, comes from volatility harvesting (Bouchey et al. 2015; Witte 2015). In practice, the task of "seeking the alpha" depends on risk modeling. Different rebalancing trigger methods have been reported for developed markets and emerging markets (Stein et al. 2013). Sometimes manipulating portfolios can have surprising effects. For instance, two investment portfolios with negative profit expectation can generate positive return expectation when the two investments are not independent (Harmer and Abbott 1999).

The energy system perspective A rather physical way of considering the market is to take it as an energy system. The fundamental analysis assumes the movement in the market is a reflection of real world operation of companies. These companies can either be collaborative or competitive, hence form a dynamic business network. The energy cascading model (ECM) assumes there are two types of business influence that can propagate via links in the network: positive energy that brings up the price and negative energy that drags down the price. The

internal energy of nodes in the business network in the current state can be estimated by sentiment scores deduced from financial news, hence the energy flow and the future states of the network can be calculated (Zhang et al. 2015). For one specific company, energy can also be calculated for various technical indices. The effects of these hidden energy terms on the visible stock price energy can be modeled and fused as a Bayesian network (Ticknor 2013).

The social network perspective The social network perspective derives from the early work in mathematics and was later confirmed by evidence from experimental finance. With plenty of heterogeneous market participants, the simulation suggests that bubbles may easily triple the fundamental price (Bao et al. 2015). This puts into serious question about to what extent the market price depends on real world economic scenarios, or market fluctuation is just a reflection of mass sentiment. As an evidence, some keyword queries data from search engines, such as Google Trends, is proved to be useful to forecast near-term economic indicators (Choi and Varian 2012). Bollen et al. (2011) reported stock market prediction with Twitter mood as well. Generally, they support their claim by illustrating better approximation (drop in mean absolute error) when indicators from social media are taken into consideration. In this perspective, the excess return comes from the correct reaction to the sticky nature of market fluctuation.

4 Walking through the literature

4.1 A review of reviews

The application of NLP techniques to financial forecasting is an emerging research field, the techniques used are also fast developing. As a result, the number of previous reviews is limited. Most of them have been published recently. To the best of our knowledge, one earliest review in the sense of NLFF is Nassirtoussi et al. (2014). Prior to it, some relevant discussions about news impact on stock markets can be spotted within papers, such as Li et al. (2014b) and Hagenau et al. (2013). Other similar topics reviewed either manually conduct text processing Chen et al. (2016) or rely solely on numerical data (Vui et al. 2013), which is not exactly what we discuss here.

According to Nassirtoussi et al. (2014), text mining for market prediction is positioned at the intersection of linguistics, machine learning, and behavioral economics. This review article covers different types of input datasets, pre-processing methods and machine learning techniques employed. Many of the machine learning algorithms presented, such as SVM, Naïve Bayes, and decision rules, are slightly outdated considering the recent research advances yet remain popular in the industry. Limited to the systematic point of view, issues on sentiment analysis are not well addressed too. In comparison, our survey makes two more contributions. The first one is to compare and elaborate on why these systems use diversified set-ups; and the second one is to include recent attention to sentiment analysis, event extraction and deep learning.

Review papers by researchers with a finance background, such as Loughran and McDonald (2016), takes a less engineering view. In Loughran and McDonald (2016), they do not attempt to evaluate the performance of a built system, but focus more on introducing resources used and interpretability. This survey also includes indicators that are seldom considered by computer scientists, such as the concept of readability. Rajput and Bobde (2016) roughly surveyed methods such as SVM, latent Dirichlet allocation (LDA) and aspect-based sentiment analysis as a whole. Another survey of better quality is Kumar and Ravi (2016). In this article,

a two-layer taxonomy of text mining in financial applications is provided. Different studies are grouped according to that taxonomy. Distributional analysis of publication venue, year, and datasets used are reported as well. This article concludes that due to different datasets and evaluation metrics used, it is still an open question about a suitable feature selection method. It also suggested constructing an ontology for each domain, and exploring some potential algorithms such as evolutionary methods, fuzzy-logic based techniques, deep learning, and spiking neural networks.

4.2 Text source and processing

We do not plan to enumerate all the papers that process financial texts for forecasting from Sects. 4.2 to 4.4. However, we try to meet two principles. The articles we include into our discussion here (1) are deemed to be a significant work (received high citation level) and (2) have good coverage of the corresponding categories.

Previous studies leverage a very diversified set of text sources. Both the form and content can be systematically different. We categorize them into six main groups according to length, subjectivity, and the frequency of updates as shown in Table 1. Corporate disclosures are primary sources directly distributed by the company. The motivation to exploit this source derives from the empirically reinforced belief of a relation between price movement and corporate releases. Because of the length and the relatively complicated structure, only a few studies automate exploiting this kind of source with mixed news data, for example, Groth and Muntermann (2011) investigates a collection of disclosures published to fulfill the German security regulations. Financial reports are produced by research institutions. These materials can be similar in form to corporate disclosures, but the content is re-organized and examined by the third party. Though it is considered hard to maintain a balanced source of financial reports, some research still leverages on the highly logical feature of financial reports (Chan and Franklin 2011). Professional periodicals refer to the regular press of media companies that have special authority in finance, like The Wall Street Journal (WSJ), Financial Times (Wuthrich et al. 1998), Dow Jones News Services (DJNS), Thomson Reuters (Fung et al. 2003), Bloomberg (Ding et al. 2015), Forbes (Rachlin et al. 2007), to name a few. Most studies use a mixture of several of the above-mentioned sources. Aggregated news, however, is a service that does not produce its own, but gathers the information from various professional periodicals. News Wire Services or news feeds (RSS) also belong here. Dominant sources are Yahoo! Finance (Lavrenko et al. 2000; Schumaker and Chen 2009; Nguyen et al. 2015), Google Finance and Thomson Reuter Eikon (formerly TR3000 Extra) (Fung et al. 2003). Message boards take the form of a forum. Market participants express their opinion under a directory of different topics. Raging Bull (Antweiler and Frank 2004), Yahoo's message board, Amazon's message board (Das and Chen 2007) are discussed in the literature. Social media is a new and fast-growing source from which financial information can be extracted. Most studies have cast their attention on Twitter (Bollen et al. 2011; Si et al. 2013; Nofer and Hinz 2015). Google Trend is yet another form, for which further processing of natural language is not required with the help of a search engine (Choi and Varian 2012). Generally, social media contains much noise that needs to be filtered by a list of financially related keywords. Corporate disclosures and financial reports are better-structured and more reliable sources. Though less studied in the past, these sources are gaining increasing attention.

We believe that the volume of data for analysis, which varies enormously among different studies, is less important than the frequency they come up. As a result, the volume is not listed as a character in our categorization in Table 1. Information with different propagation speed actually has an effect on a different time scale of market cycles. Texts with low data

Table 1 Financial texts from different sources and examples

Type	Characters	Example
Corporate disclosures	Long length, Subjective tone, Low frequency	Apple Quarter Reports: . . .We are pleased to report third quarter results that reflect stronger customer demand and business performance than we anticipated at the start of the quarter, said Tim Cook, Apple's CEO. . .
Financial reports	Long length, Objective tone, Low frequency	Quamnet Portal: Gold prices went through a week of uncertainty due to mixed economic data. First there were weak retail sales data, which led gold prices to surge, yet investors remained uncertain how the data will affect the upcoming decision of the Federal Reserve. . .
Professional periodicals	Variable length, Objective tone, Mid frequency	Financial Times: The US Consumer Product Safety Commission issued a formal recall notice for 1 million Samsung Galaxy Note 7 smartphones on Thursday, after nearly a hundred reports of overheating batteries. . .
Aggregated news	Mid length, Variable tone, Variable frequency	Yahoo! Finance: Indonesians Declare $8.9 Billion of Singapore Assets for Tax. . . A positive ruling, should remove the uncertainty that may be hampering more participation, said Euben Paracuelles, a Singapore-based economist with Nomura Holdings Inc., in a report Friday. . .
Message boards	Short length, Objective tone, High frequency	Amazon's Board: The fact is. . . The value of the company increases because the leader (Bezos) is identified as a commodity with a version for what the future may hold. He will now be a public figure until the day he dies. That is value
Social media	Short length, Subjective tone, High frequency	Twitter: $AAPL is loosing customers. everybody is buying android phones! $GOOG

frequency and high authority tend to have a profound and long-lasting impact, while high-frequency data reflects short-term volatility and can generate different patterns depending on market microstructure. Because of its continuous effect, the market reaction can attenuate very fast after rounds of adaptation. As a good example, the tweets of US newly elected president Trump have observable effects on the stock price of the company he mentioned at the beginning. However, within a month his tweets no longer have positive relevance with the inter-day price change.

Table 2 includes the concrete information on what kind of sources are investigated as well as the way they are processed for previous studies. It is shown that from the very inception of this research field, professional periodicals are always a crucial text source. When processing this information, filtering text source with a list of keywords or hashtags to a domain specific, or even company-specific materials, rather than taking the noisy data collection as a whole, is common. Only in the five recent years, a large proportion of research papers has cast their attention to social media. Consequently, most studies dealing with social media texts have a very condensed timestamp at the second level. In this situation, machine learning techniques are more actively considered.

There is no clear standard on how long we should watch the market before we start to theorize and implement our model. Some studies have speculated on a very short data span, for instance, 5 weeks (Schumaker and Chen 2009), while some make an effort to trace back to 1980 (Tetlock et al. 2008). The majority takes a span of several months into consideration. Empirically, we suggest investigating into a longer time span with less frequent data, such as corporate disclosures and professional periodicals. While for data from social media, the data span can be shorter as the effects are often intraday. Another consideration is that the time span should not either be too long or too short. Otherwise, the data observed will often be accompanied by deterministic trends. When having a trend, the metrics reported will not be comparable. In this case, the raw data should be differenced before further processing.

Text data processing is the procedure that prepares a well-formatted input. This input will be used for later forecasting by feeding it to the algorithms implemented in a predictive model. Popular formatting techniques can be roughly divided into three groups. The first group is a one-hot representation of keyword, keyword tuples, sentiment word, or more advanced statistics of them. For example, the share of positive mood on all target word occurrences (sum of positive and negative mood states) can be defined as "Social Mood Index (SMI)" (Nofer and Hinz 2015). A time series of weighted mood word density in postings for each day, is defined as optimism–pessimism mood scores (M_s^+ and M_s^-) in Li et al. (2014b). The second group contains specific input formats for certain algorithms such as word embeddings (Ding et al. 2015), or distributional probabilities of the price moving up, down or steady conditional on different words (Antweiler and Frank 2004). Yoshihara et al. (2016) used a standard bag-of-words model to represents the news articles. However, the temporal properties of the articles are emphasized by employing a combination of recurrent neural network and RBM. The trained article representation was later incorporated to tune deep belief networks (DBN) that output an uptrend or downtrend. The third group actively gathers the alignments from texts to different trend motifs (Lavrenko et al. 2000), triggers for related stocks, or simply the directional categories without further semantic or sentiment analysis of these alignments. In other words, this third group representation is similar to association rules.

Additionally, there were many XML-format text sources delivered by the main financial information companies such as Dow Jones Elementized News Feed, Thomson Reuters News Feed Direct, Bloomberg Event-Driven Trading Feed, and NASDAQ OMX Event-Driven Analytics. Perhaps due to some commercial reason, these services are no longer available.

Table 2 Type of financial texts leveraged and how are they processed

References	Text type	Coverage	Frequency level	Data span	Processing
Wuthrich et al. (1998)	Professional periodicals	Stock, Currency, Bond market	est. Hours	6/12/1997–6/3/1998	Manually crafted keyword tuples spotting
Lavrenko et al. (2000)	Aggregated news	Stock market	Minutes	15/10/1999–10/2/2000	Alignment with trends
Fung et al. (2003)	Professional periodical	Stock market	Minutes	1/10/2002–30/4/2003	Alignment with other stocks
Antweiler and Frank (2004)	Message board	Stock market	Minutes	3/1/2000–29/12/2000	Naïve Bayes classifier
Das and Chen (2007)	Message boards	Stock market	Minutes	6/2001–8/2001	Manually crafted sentiment lexicon
Tetlock et al. (2008)	Professional periodical	Stock market	Hours	1980–2004	Bag-of-negative-words
Schumaker and Chen (2009)	Aggregated news	Stock market	Minutes	26/10/2005–28/11/2005	Bag-of-words, Name entities, Noun phrases
Bollen et al. (2011)	Social media	Stock market	Seconds	28/2/2008–19/12/2008	Sentiment classification tool
Chan and Franklin (2011)	Financial reports	Comprehensive	–	Not mentioned	Semantic class, Instance-attribute pair
Groth and Muntermann (2011)	Corporate disclosures	Comprehensive	Days	1/8/2003–31/7/2005	Risk modeling
Ruiz et al. (2012)	Social media	Stock market	est. Seconds	1/2010–6/2010	Graph representation
Schumaker et al. (2012)	Aggregated news	Comprehensive	Minutes	26/10/2005–28/11/2005	Pos/Neg & Sub/Obj classification
Si et al. (2013)	Social media	Stock market	Seconds	2/11/2012–7/2/2013	Dirichlet processes mixture model
Si et al. (2014)	Social media	Stock market	Seconds	2/11/2012–3/4/2013	Semantic stock network
Li et al. (2014b)	Mixed type	Stock market	–	1/1/2011–31/12/2011	Emotion word dictionary
Ding et al. (2015)	Professional periodicals	Comprehensive	Minutes	10/2006–11/2013	Neural tensor network
Nofer and Hinz (2015)	Social media	Comprehensive	Seconds	1/2011–11/2013	Sentiment classification tool
Nguyen et al. (2015)	Message board	Stock market	est. Hours	23/7/2012–19/7/2013	Latent Dirichlet allocation
Yoshihara et al. (2016)	Aggregated news	Stock market	Minutes	1/1/1999–31/12/2008	Recurrent neural network, RBMs

Instead, there are some commercial sources, mostly from content vendors, that directly provide the processed sentiment data. The correlation between Thomson Reuters Datastream and stock returns is examined and believed to exist according to Uhl (2014). Latest released products include TR MarketPsych Indices (TRMI), RavenPack News Analytics (RPNA) and so forth. TRMI covers a wide range of text sources from blogs to main social media sites. While the detailed source list and how they process the texts are not revealed. According to historical testing using the moving average of TRMI to indicate buy/sell pressure, the index has proved to be a significant predictor for Apple's stock price and JPY/USD exchange rate (Reuters 2016).

4.3 Algorithms

Linear regressions and SVM are classic methods that dominate prediction models in the past decades. Regression models are particularly preferred since we can explicitly observe the impact of each factor included and analyze the importance of variables by dropping them out. SVM has a sound mathematical foundation and all the support vectors can be computed. According to Kumar and Ravi (2016), 70% of previous studies have adopted regular methods (decision trees, SVMs, etc) and regression analysis. For articles we discuss here, the proportion is roughly the same. Considering the volume and quality of data available, overly complicated models generally have a poor performance. However, one drawback of linear models is that they rely on strong hypotheses, for example, a Gaussian distribution of dependent variables, which does not always stand up in real world cases. In spite of this there are efforts to estimate some singular distributions (Taleb 2008), the result is often specific to problems and cannot be popularized to various financial indicators. Therefore, neural network and other statistical learning methods, such as Bayesian networks are also widely experimented with. In many studies, the features generated from the texts are combined with numerical data to form a robust input data stream for prediction, in which case an ensemble method can be used to manipulate the combination either on a feature level or a decision level.

It is still an open question as to what category of algorithms is especially appropriate for NLFF (Kumar and Ravi 2016). From Table 3, mainstream algorithms can be placed into four categories: regressions, probabilistic inferences, and neural networks, or a hybrid of them. Our analysis of Table 3 comes to the similar observation as (Kumar and Ravi 2016) that evolutionary computing has been applied for numerical analysis (Brabazon and O'Neill 2008), but seldom discussed in the literature to deal with financial texts.

Regression models are especially suitable for impact analysis. Sometimes, a primary linear regression is directly used with ordinary least square (OLS) to estimate coefficients (Nassirtoussi et al. 2014). For example, Tetlock et al. (2008) uses this method to illustrate that, negatives words in firm-specific news stories robustly predict slightly lower returns on the following trading day. If we want to include more complicated time lags or multiple factors simultaneously, an MR or VAR model is required. Multivariate regression (MR) (Antweiler and Frank 2004) is conducted in two steps. First, a dummy variable is introduced to examine whether different lags of the corresponding factor are predictive. Then, logistic regression is used to adopt all factors with t-statistics to show significance. This approach is good at drawing pairwise conclusions such as "Does factor A have an effect, and to what extent, on factor B". However, we should be cautious that the MR method evades the problem of collinearity and leaves the interaction between predictors untouched. Vector autoregression (VAR) (Si et al. 2014; Nofer and Hinz 2015) can be used to model the time series of sentiment and stock price as a vector together, based on their past values. This is due to the observation

Table 3 Algorithms involved and the implementation details

Reference	Feature formatting	Model type	Implementation
Wuthrich et al. (1998)	Number of tuples occurrences	Naïve Bayes & Association rules	Experimentally tuned k-NN
Lavrenko et al. (2000)	Trend possibility distribution	n-gram language model	Conditional probability maximization
Fung et al. (2003)	Tf-idf weighted key words	Support vector machine	Split-and-merge segmentation
Antweiler and Frank (2004)	Text classification	Regressions	Variable & Lag tuning
Das and Chen (2007)	Lexicon occurrences	Classifiers voting	Discriminant values
Tetlock et al. (2008)	Lexicon based sentiment score	Regressions	Ordinary least square & Dependent variables
Schumaker and Chen (2009)	Binary representation	Support vector regression	Sequential minimal optimization
Bollen et al. (2011)	Temporal mood indicator	Self-organized fuzzy neural network	Online learning
Chan and Franklin (2011)	Textual information database	Inference engine	Multiple decision tree classifiers
Groth and Muntermann (2011)	Labelled lexicon occurrences	Ensemble learning	NB, k-NN, NN, SVM with tuning
Ruiz et al. (2012)	Graph features	Vector autoregression	Least square regression
Schumaker et al. (2012)	Proper Nouns	Support vector regression	Sequential minimal optimization
Si et al. (2013)	Topic based sentiment score	Vector autoregression	Least square regression
Si et al. (2014)	Lexicon based sentiment score	Vector autoregression	Least square regression
Li et al. (2014b)	Tf-idf weighted key/senti words	Support vector regression	Sequential minimal optimization
Ding et al. (2015)	Sequence of event embeddings	Convolutional neural network	Margin loss minimization
Nofer and Hinz (2015)	Weighted Social Mood Index	Vector autoregression	Minimum information criterion
Nguyen et al. (2015)	Topic model parameters	Support vector machine	Linear kernel soft margin
Yoshihara et al. (2016)	Temporal news embeddings	Deep belief network	Greedy layer-wise training

that, not only the public sentiment will cause volatility of the market, the market will also induce fluctuation on social moods. This observation is addressed in Sehgal and Song (2007) by modeling the sentiment score as a probability conditional on the past information released from text sources. Although there are other models, such as copula-based regression (Koleva and Paiva 2009) or structural equation modeling (SEM), that are capable of capturing this correlation, VAR is still currently the most popular model. However, since no theory suggests the interdependencies should be linear, doubts exist (Peters 1991; Xing et al. 2017) about the appropriateness of VAR.

If we solely care about predicting the direction, not the intensity of market movement, SVM can naturally serve as a binary classifier. Many previous studies indeed formalize stock market prediction or more broadly, financial forecasting as a classification problem. Inspired by the idea that the empirical risk minimization principle can also be used to build a regression model, support vector regression (SVR) (Schumaker and Chen 2009; Li et al. 2014b) is proposed to make discrete forecasting. The hyperplane for SVR is also determined by a portion of training data with a sensitivity threshold. Unlike SVM paying attention only to classification accuracy, SVR gives more weight to data far away from the classification hyperplane due to the fact that this type of error would cause a huge loss in practice. The shortcoming of SVR is the necessity of introducing a kernel to map training data into a linear separable higher dimension and an extra threshold parameter. These hyper-parameters are picked manually without much sound reasoning, for instance as in Schumaker and Chen (2009).

The original task discussed in Lavrenko et al. (2000) is financial news recommendation. However, assuming this recommendation is "accurate", a user should be able to make a profit based on it. Therefore, financial news recommendation plus some text analyzing techniques would be equivalent to the task of financial forecasting. Lavrenko et al. (2000) attempts to maximize the probability of a model with trends (M_{trends}) conditional on a set of documents as recommendations $P(M_{trends}|D_1, D_2, \ldots, D_m)$. Using Bayes' theorem, the problem can be converted to maximizing $P(D_1, D_2, \ldots, D_m|M_{trends})$, since $P(M)$ is considered as a uniform prior and $P(D_1, D_2, \ldots, D_m)$ can be estimated from generating these documents from normal English. Assuming independence of documents, the problem is maximizing $\prod P(D_i|M)$, or further decomposes the formula to word level, maximizing $\prod\prod P(w_{ij}|M)$. Chan and Franklin (2011) provides yet another perspective of event sequence extraction. Strictly speaking, it is not a predictive algorithm, while it would be useful to extract structured information from text sources. With the help of a trained inference engine, a trading strategy can be further built on the predicted event sequence.

From many popular self-organizing neural network architectures, Bollen et al. (2011) chose self-organizing fuzzy neural network (SOFNN) which is developed specially for regression problems and is faster than other fuzzy neural network models, such as adaptive neuro-fuzzy inference system (ANFIS). The structure of SOFNN is not different from common fuzzy neural networks. However, the learning process is bifold. In the early phase of self-organizing learning, the number of rules is determined. After the network structure is established, weight parameters are adjusted in the optimization-learning phase. In Bollen et al. (2011), lagged Dow Jones industrial average (DJIA) value and generalized POMS (GPOMS) are simultaneously set as the input of an SOFNN model. The output will be the current value of DJIA. Ding et al. (2015) chose a neural tensor network (NTN) to train event embeddings. Later a sequence of event embeddings with different term span are fed into a CNN for a binary output.

4.4 Results

Previous studies report their results in various forms. Even though some studies argue that their text processing output is a statistically significant predictor (Antweiler and Frank 2004; Nofer and Hinz 2015), three kinds of measurement are commonly acknowledged (see Table 4). The first measurement is directional accuracy, where the forecasting is simply represented in a binary up/down form. Accuracy is the percentage of correct forecasts of the total number of forecasting attempts. Reported accuracy rates are from 40% to around 80%. Theoretically, any accuracy rate that significantly differs from 50% can prove the effectiveness of forecasting results. Though, in fact, accuracy improvements on a benchmark method would be more convincing. To analyze false positive and false negative errors, in addition to accuracy rate, precision and recall may be considered as well, such as in Groth and Muntermann (2011).

The second measurement is the closeness between the forecasted time series and the corresponding real world time series, usually in the form of stock price. Closeness measurement is commonly used for function approximation tasks. Several metrics can be taken in this measurement, such as mean squared error (MSE) (Schumaker and Chen 2009), root mean squared error (RMSE) which is simply the square root of MSE (Li et al. 2014b), mean absolute percentage error (MAPE) (Bollen et al. 2011), mean absolute scaled error (MASE) (Hyndman and Koehler 2006) and more. Reduction in these errors generally means a more precise forecasting result.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{x_i}\right)^2$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x_i}}{x_i}\right|$$

The third measurement is trading simulation results. Specific metrics include average percentage gain per transaction (AGPT) (Lavrenko et al. 2000), accumulated profit for a certain period, profit ratio, or portfolio performance. It is very hard to compare trading simulation results among previous studies because the configurations are quite arbitrarily set. Many studies report simulation result without deducting transaction costs. Simulation results for financial indices and for specific big companies such as Apple (Si et al. 2014) or Amazon are more comparable than results from different sectors. Trading strategies are usually reported if the evaluation part consists of trading simulation. Though more sophisticated trading strategies are developed by financial practitioners, two simple strategies remain the most popular in the experiments. The "buy up/sell down" strategy suggests buying stocks when the forecasted price is rising, and selling stocks when the forecasted price is going down. The "short-term reversal" strategy arbitrages on the overreaction and correction of the market. Both strategies can be equipped with a trigger mechanism, which aligns with the idea of passive management. The traditional rebalance frequency is daily. However, hourly rebalancing (Lavrenko et al. 2000) or 20-minute rebalancing (Schumaker et al. 2012) are also reported. Current NLP techniques are not fast enough to facilitate low delay trading at below the second level. As social media accelerate the fluctuations of the market, there might be pressure to shorten the rebalance frequency. However, daily rebalancing seems a good trade-off between arbitrage efficiency and transaction cost.

Figure 4 categorizes many choices of specific metrics into the taxonomy of three measurements. These research designs are also common for other computational social science

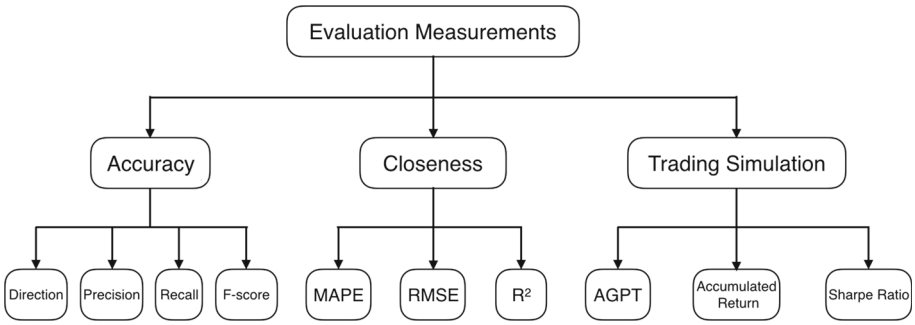

Fig. 4 Taxonomy of measurements reported

problems (Hofman et al. 2017). However, it is worth mentioning that the three measurements are not necessarily correlated. For example, a forecasting method may have very high directional accuracy and work well in most cases, but at the same time being extremely fragile to black swan events. The method that suffers a huge loss in a single transaction can illustrate no profitability in trading simulation. Consequently, we suggest evaluating the forecasting result using all three measurements and make robust comparisons comprehensively. We also observe researchers' preference to analyze the market in their home countries (Njølstad 2014; Nofer and Hinz 2015; Dong et al. 2017), which is often referred to as "home bias" by investors. Despite this effect, most efforts have been made on the New York Stock Exchange (NYSE) and NASDAQ.

5 Conclusion

Our survey presents various NLP techniques used for financial forecasting tasks today, as well as how these techniques are developed. As shown by Fig. 5, NLFF is related to many groups of concepts. The artificial intelligence community tends to consider three major types of representation of textual financial data: semantic, sentiment, and event representation extracted from information sources. Utilizing these data, many studies attempted to build financial forecasting systems and took the underlying financial principles for granted. We explicitly construct a spectrum of philosophies for reference. As one more step, we analyze previous studies from three angles: different type of text sources employed, algorithms, and reported results. Some recent updates, such as the use of deep learning methods for forecasting, are included. In addition, we make an effort to categorize and standardize the measurements used for evaluation. We suggest future research following and covering these three measurements. This would partially solve the difficulty of making comparisons between research results in the scope of NLFF.

We conclude our survey by summarizing some main findings as well as interesting facts from previous studies. Some future directions are provided at the same time.

5.1 Main findings

The illusion of Growth The way growth rate is calculated for each period brings up the illusion of growth when the price of an asset is actually stagnant. Regardless of the movement trajectory of price, the average growth rate is always positive. This mathematical rule alerts us

Table 4 Results reported using different measurements

References	Measurement	Performance	Trading Strategy
Wuthrich et al. (1998)	Direction accuracy of 5 main indices	Ftse 42%, Nky 47%, Dow 40%, Hsi 53%, Sti 40%	Buy up/sell down, rebalancing daily
Lavrenko et al. (2000)	Trading simulation of 127 stocks	Average gain per transaction 0.23%	Short-term reversal, rebalancing hourly
Fung et al. (2003)	Trading simulation of 33 stocks from Hsi	Cumulative profit 6.55%	Buy up/sell down, rebalancing daily
Antweiler and Frank (2004)	Statistic testing for correlation with DJIA & DJII	Significant predictor	No
Das and Chen (2007)	Statistic testing for correlation with MSH-35	Correlation is weak	No
Schumaker and Chen (2009)	Closeness, direction accuracy, trading simulation	MSE 0.04261, Acc 57%, Return 2.06%	Not mentioned
Bollen et al. (2011)	Closeness, direction accuracy for DJIA	MAPE reduction by 6%	No
Chan and Franklin (2011)	Even sequence correct accuracy	Significant improvement (> 7%)	No
Groth and Muntermann (2011)	Accuracy, precision, recall, option simulation	Acc 70%, p 47%, r 70%, significant false positive	No
Ruiz et al. (2012)	Trading simulation on a 10-company portfolio	Return 0.32%	Buy up/sell down, rebalancing daily
Schumaker et al. (2012)	Direction accuracy & trading simulation	Acc 59%, Return 3.30% (sub. news only)	Triggered short-term reversal, rebalancing every 20 min
Si et al. (2013)	Direction accuracy of S&P100 index	Best tuning 68.0%	No
Si et al. (2014)	Direction accuracy on $AAPL	Best tuning 78.0%	No
Li et al. (2014b)	Closeness, direction accuracy, trading simulation	RMSE 0.63, Acc 54.21%, est. Return 4%	Short-term reversal, rebalancing every 26 min
Ding et al. (2015)	Direction accuracy, trading simulation	Acc 65.08%, Avg. Profit Ratio 1.679	Short-term reversal, rebalancing daily
Nofer and Hinz (2015)	Statistic testing for correlation with DAX, trading simulation	Significant predictor, AROR 84.96%	Buy up/sell down of ETF, rebalancing daily
Nguyen et al. (2015)	Direction accuracy	Acc 54.41%	No
Yoshihara et al. (2016)	Direction accuracy, trading simulation	Improved error rates and profit gain than SVM	Buy/sell at MACD turning point

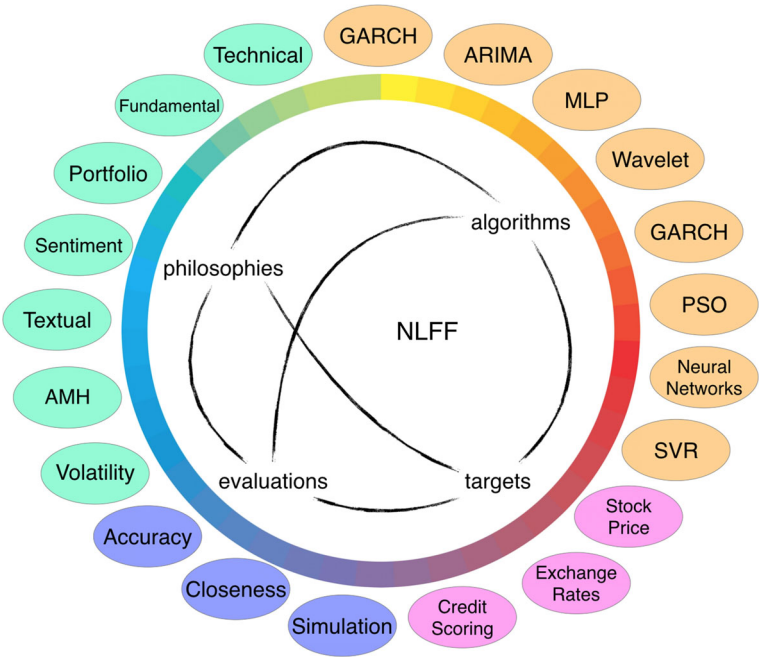

Fig. 5 Topics concerning NLFF, inspired and adapted from the concept wheel of financial markets (Cavalcante et al. 2016)

it is important to reduce volatility with regard to trading strategy. In other words, compounded wealth is reduced dramatically by the square of volatility (Stein et al. 2013). In trading simulations, the gains are not the only indicator that is worth reporting. Realized volatility is a crucial factor to the quality of a trading strategy.

The predictability of Financial News It seems that most previous studies have confirmed the correlation between public mood and the movement of the market, for instance, (Li et al. 2014c; Heston and Sinha 2016). The literature Kittrell (2011) argues that the reversal of sentiment will be slightly ahead of price reversal. As a result, sentiment reversals can serve as buy/sell signals in constructing trading strategies. Though Brown and Cliff (2004) claimed that sentiment levels and changes are strongly correlated with contemporaneous market returns, but have little predictive power for the near-term (weekly) stock market. It refers to the critical problem of time window selection, as elaborated in The "20-minute" Theory. While for the market return itself, long-term memory may exist.

The "20-minute" Theory There exists an optimum time window to foresee the impact of new information released and the market correction to equilibrium. This theory was proposed by LeBaron et al. (1999), and supported by empirical evidence from Schumaker and Chen (2009) and Li et al. (2014a, 2015).

The Monday Effect The effect of less trading volume by institutional investors at the start of a week was first found by Lakonishok and Maberly (1990). Furthermore, the market also tends to be bearish at the start of a new week. Perhaps because people are busy doing other things, observation shows that the number of messages posted and the length of them drop dramatically on the first trading day of a new week (Antweiler and Frank 2004).

The Reversal Effect An increasingly optimistic mood from message boards usually leads to negative return for the next trading day; Disagreement among the posted messages is

associated with increased trading volume for the day, but will decrease trading volume for the next trading day, though this may only apply to developed markets (Antweiler and Frank 2004).

5.2 Future directions

We believe three future directions are very promising in the near term.

Domain Specific Resources Building Previous surveys have pointed out the importance of resource building. For instance, Kumar and Ravi (2016) suggests constructing domain specific ontologies. In fact, the form of knowledge representation is not limited to ontologies, but can also be wordlists, concept databases, manually annotated datasets, etc. Due to the lack of ground truth in the financial domain, Chan and Chong (2017) can only evaluate model accuracy on a popular movie review dataset. Embarrassingly for financial text streams, the paper used the Granger causality test to prove the sentiment index is not random. Some recent attempts have been made to automatically identify sentiment lexicons (Oliveira et al. 2016, 2017) or more straightforwardly, identify the sentiment polarity of information contents (Chang et al. 2016). However, there is a lot to be done before we have a rich and authoritative resource in the financial domain.

Online Predictive Model Online, or real-time algorithms will modify the key variables stored with the model each time a new batch of data comes in. For this reason, online models have very good adaptability, which is necessary for monitoring fast-changing markets. In addition, the short optimum time window requires a quick response in time as well. For numerical data, many studies have discussed online methods to reduce algorithm complexity, such as motif matching (Mueen and Keogh 2010) and online portfolio selection (Li and Hoi 2014). However, at the unstructured textual data stage, little work has been done for exploring online NLP techniques. New approaches absent from the domain are worth trying as well (Kumar and Ravi 2016). For example, directly generate fuzzy rules from textual data, which provides better interpretability.

Comprehensive Evaluation Measurements A main current difficulty to summarize and compare the existing studies is various and incomplete measurements used by researchers. We suggest experiments to at least cover all three kinds of measurements we discussed in Sect. 4.4. Some metrics are not always measuring independent effects. For example, volatility measured by the stability of trading simulation is closely correlated to Sharpe Ratio (Uhl 2014), if we consider volatility as a risk factor. Many efforts are needed to unify these measures.

References

Anton M, Polk C (2014) Connected stocks. J Finance 69(3):1099–1127

Antweiler W, Frank MZ (2004) Is all that talk just noise? The information content of internet stock message boards. J Finance 59(3):1259–1294

Avramov D, Zhou G (2010) Bayesian portfolio analysis. Annu Rev Financ Econ 2:25–47

Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: 7th language resources and evaluation conference, pp 2200–2204

Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O (2007) Open information extraction from the web. In: International joint conference on artificial intelligence, pp 2670–2676

Bao T, Hommes C, Makarewicz T (2015) Bubble formation and (in)efficient markets in learning-to-forecast and -optimise experiments. Tinbergen Institute Discussion Paper TI 2015-107/II. https://www.econstor.eu/bitstream/10419/125108/1/15107.pdf

Bengio Y, Ducharme R, Vincent P (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
Blei DM (2012) Probabilistic topic models. Commun ACM 55(4):77–84
Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
Bouchey P, Nemtchinov V, Wong TKL (2015) Volatility harvesting in theory and practice. J Wealth Manage 18(3):89–100
Brabazon A, O'Neill M (2008) An introduction to evolutionary computation in finance. IEEE Comput Intell Mag 3(4):42–55
Brachman RJ, Khabaza T et al (1996) Mining business databases. Commun ACM 39(11):42–48
Brown GW, Cliff MT (2004) Investor sentiment and the near-term stock market. J Empir Finance 11:1–27
Bühler K (1934) Sprachtheorie. Fischer, Jena
Cambria E (2013) An introduction to concept-level sentiment analysis. In: Lecture notes in computer science (LNCS), vol 8266. Springer, pp 478–483
Cambria E (2016) Affective computing and sentiment analysis. IEEE Intell Syst 31(2):102–107
Cambria E, White B (2014) Jumping NLP curves: a review of natural language processing research. IEEE Comput Intell Mag 9(2):48–57
Cambria E, Livingstone A, Hussain A (2012) The hourglass of emotions. In: Lecture notes in computer science, vol 7403. Springer, pp 144–157
Cambria E, Wang H, White B (2014) Guest editorial: big social data analysis. Knowl-Based Syst 69:1–2
Cambria E, Poria S, Bajpai R, Schuller B (2016) SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives. In: International conference on computational linguistics (COLING), pp 2666–2677
Cambria E, Poria S, Gelbukh A, Thelwall M (2017) Sentiment analysis is a big suitcase. IEEE Intell Syst 32(6):74–80
Cavalcante RC, Brasileiro RC, Souza VL, Nobrega JP, Oliveira AL (2016) Computational intelligence and financial markets: a survey and future directions. Expert Syst Appl 55:194–211
Chan SW, Chong MW (2017) Sentiment analysis in financial texts. Decis Support Syst 94:53–64
Chan S, Franklin J (2011) A text-based decision support system for financial sequence prediction. Decis Support Syst 52(1):189–198
Chang CY, Zhang Y, Teng Z, Bozanic Z, Ke B (2016) Measuring the information content of financial news. In: Proceedings of the the 26th international conference on computational linguistics
Chaturvedi I, Ong YS, Tsang I, Welsch R, Cambria E (2016) Learning word dependencies in text by means of a deep recurrent belief network. Knowl-Based Syst 108:144–154
Chaturvedi I, Ragusa E, Gastaldo P, Zunino R, Cambria E (2017) Bayesian network based extreme learning machine for subjectivity detection. J Frankl Inst. https://doi.org/10.1016/j.jfranklin.2017.06.007
Chen N, Ribeiro B, Chen A (2016) Financial credit risk assessment: a recent review. Artif Intell Rev 45:1–23
Choi H, Varian H (2012) Predicting the present with google trends. Econ Rec 88(1):2–9
Chomsky N (1956) Three models for the description of language. IRE Trans Inf Theory 2(3):113–124. https://doi.org/10.1109/TIT.1956.1056813
Cohen L, Frazzini A (2008) Economic links and predictable returns. J Finance 63(4):1977–2011
Das SR, Chen MY (2007) Yahoo! for amazon: sentiment extraction from small talk on the web. Manage Sci 53(9):1375–1388
Ding X (2016) Research on methodology of market trends prediction based on social media. Ph.D. thesis, Harbin Institute of Technology
Ding X, Zhang Y, Liu T, Duan J (2015) Deep learning for event-driven stock prediction. In: International joint conference on artificial intelligence
Dong L, Wang Z, Xiong D (2017) Stock market prediction based on text information. Acta Scientiarum Naturalium Universitatis Pekinesis. https://doi.org/10.13209/j.0479-8023.2017.037
Fama EF (1970) Efficient capital markets: a review of theory and empirical work. J Finance 25:383–417
Feldman R (2013) Techniques and applications for sentiment analysis. Commun ACM 56(4):82–89
Fellbaum C (1998) WordNet: an electronic lexical database. MIT Press, Cambridge
Frazier KB, Ingram RW, Tennyson BM (1984) A methodology for the analysis of narrative accounting disclosures. J Account Res 22(1):318–331
Fung GPC, Yu JX, Lam W (2003) Stock prediction: integrating text mining approach using real-time news. In: 2003 IEEE international conference on computational intelligence for financial engineering, 2003. Proceedings, pp 395–402. https://doi.org/10.1109/CIFER.2003.1196287
Groth SS, Muntermann J (2011) An intraday market risk management approach based on textual analysis. Decis Support Syst 50(4):680–691
Guha RV, Lenat DB (1990) Cyc: a midterm report. AI Mag 11(3):32–59

Hagenau M, Liebmann M, Neumann D (2013) Automated news reading: stock price prediction based on financial news using context-capturing features. Decis Support Syst 55(3):685–697. https://doi.org/10.1016/j.dss.2013.02.006
Hajizadeh E, Ardakani HD, Shahrabi J (2010) Application of data mining techniques in stock markets: a survey. J Econ Int Finance 2(7):109–118
Hamilton WL, Clark K, Leskovec J, Jurafsky D (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. In: Empirical methods in natural language processing (EMNLP), pp 595–605
Harmer GP, Abbott D (1999) Parrondo's paradox. Stat Sci 14(2):206–213
Hatzivassiloglou V, McKeown KR (1997) Predicting the semantic orientation of adjectives. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp 174–181
Henry E (2008) Are investors influenced by how earnings press releases are written? Int J Bus Commun 45:363–407
Heston SL, Sinha NR (2016) News versus sentiment: predicting stock returns from news stories. Technical Report 2016-048: Board of Governors of the Federal Reserve System, Washington
Hofman JM, Sharma A, Watts DJ (2017) Prediction and explanation in social systems. Science 355(6324):486–488
Hommes CH (2006) Heterogeneous agent models in economics and finance. In: Tesfatsion L, Judd K (eds) Handbook of computational economics II: agent-based economics. Elsevier, pp 1109–86
Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 168–177
Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
Kelly EF (1975) Computer recognition of English word senses. Elsevier, Amsterdam
Kittrell J (2011) Sentiment reversals as buy signals. Wiley, Hoboken, pp 231–244. https://doi.org/10.1002/9781118467411.ch9
Koleva N, Paiva D (2009) Copula-based regression models: a survey. J Stat Plan Inference 139(11):3847–3856. https://doi.org/10.1016/j.jspi.2009.05.023
Kumar BS, Ravi V (2016) A survey of the applications of text mining in financial domain. Knowl-Based Syst 114:128–147
Lakonishok J, Maberly E (1990) The weekend effect: trading patterns of individual and institutional investors. J Finance 40:231–243
Lavrenko V, Schmill M, Lawrie D, Ogilvie P, Jensen D, Allan J (2000) Language models for financial news recommendation. In: Proceedings of the ninth international conference on information and knowledge management, pp 389–396
LeBaron B, Arthur W, Palmer R (1999) Time series properties of an artificial stock market. J Econ Dyn Control 23:1487–1516
Leetaru K, Schrodt PA (2013) Gdelt: global data on events, location, and tone, 1979–2012. In: ISA annual convention, vol 2. Citeseer
Li B, Hoi SCH (2014) Online portfolio selection: a survey. ACM Comput Surv 46(3). https://doi.org/10.1145/2512962
Li Q, Wang T, Gong Q, Chen Y, Lin Z, Song SK (2014a) Media-aware quantitative trading based on public web information. Decis Support Syst 61:93–105
Li Q, Wang T, Li P, Liu L, Gong Q, Chen Y (2014b) The effect of news and public mood on stock movements. Inf Sci 278:826–840
Li X, Xie H, Chen L, Wang J, Deng X (2014c) News impact on stock price return via sentiment analysis. Knowl-Based Syst 69:14–23
Li B, Hoi SCH, Sahoo D, Liu ZY (2015) Moving average reversion strategy for on-line portfolio selection. Artif Intell 222:104–123
Li Q, Jiang L, Li P, Chen H (2015) Tensor-based learning for predicting stock movements. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, pp 1784–1790
Li L, Qin B, Ren W, Liu T (2016) Truth discovery with memory network. CoRR arXiv:1611.01868
Liu H, Singh P (2004) ConceptNet—a practical commonsense reasoning tool-kit. BT Technol J 22(4):211–226
Liu C, Hoi SCH, Zhao P, Sun J (2016) Online arima algorithms for time series prediction. In: Thirtieth AAAI conference on artificial intelligence
Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J Finance 66:67–97
Loughran T, McDonald B (2016) Textual analysis in accounting and finance: a survey. J Account Res 54(4):1187–1230
Ma Y, Cambria E, Gao S (2016) Label embedding for zero-shot fine-grained named entity typing. In: COLING, pp 171–180

Majumder N, Poria S, Gelbukh A, Cambria E (2017) Deep learning based document modeling for personality detection from text. IEEE Intell Syst 32(2):74–79
Malik HH, Bhardwaj VS, Fiorletta H (2011) Accurate information extraction for quantitative financial events. In: Proceedings of the 20th ACM international conference on information and knowledge management
Marsella S, Gratch J (2014) Computationally modeling human emotion. Commun ACM 57(12):56–67
Mihalcea R, Garimella A (2016) What men say, what women hear: finding gender-specific meaning shades. IEEE Intell Syst 31(4):62–67
Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. CoRR arXiv:1310.4546
Moniz A, de Jong F (2014) Classifying the influence of negative affect expressed by the financial media on investor behavior. In: Fifth information interaction in context symposium, IIiX '14, Regensburg, Germany, 26–29 Aug 2014, pp 275–278
Mueen A, Keogh E (2010) Online discovery and maintenance of time series motifs. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '10. ACM, New York,, pp 1089–1098. https://doi.org/10.1145/1835804.1835941
Nassirtoussi AK, Aghabozorgi S, Waha TY, Ngo DCL (2014) Text mining for market prediction: a systematic review. Expert Syst Appl 41:7653–7670
Nguyen TH, Shirai K (2015) Topic modeling based sentiment analysis on social media for stock market prediction. In: The 53rd annual meeting of the association for computational linguistics (ACL), pp 1354–1364
Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42:9603–9611
Njølstad LSH (2014) Sentiment analysis for financial applications. Master's thesis, Norwegian University of Science and Technology
Nofer M, Hinz O (2015) Using twitter to predict the stock market: where is the mood effect? Bus Inf Syst Eng 57(4):229–242
Oliveira N, Cortez P, Areal N (2016) Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis Support Syst 85:62–73
Oliveira N, Cortez P, Areal N (2017) The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Syst Appl 73:125–144
Owyang J (2009) The future of the social web. Forrester Research Inc, Cambridge
Park CH, Irwin SH (2004) The profitability of technical analysis: a review. AgMAS project research report 2004-04, University of Illinois at Urbana-Champaign
Peters EE (1991) A chaotic attractor for the S&P 500. Financ Anal J 47(2):55–62+81. http://www.jstor.org/stable/4479416
Poria S, Cambria E, Gelbukh A (2016a) Aspect extraction for opinion mining with a deep convolutional neural network. Knowl-Based Syst 108:42–49
Poria S, Cambria E, Hazarika D, Vij P (2016b) A deeper look into sarcastic tweets using deep convolutional neural networks. In: COLING, pp 1601–1612
Poria S, Chaturvedi I, Cambria E, Hussain A (2016c) Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM, Barcelona, pp 439–448
Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion 37:98–125
Qian B, Rasheed K (2004) Hurst exponent and financial market predictability. In: Proceedings of the 2nd IASTED international conference on financial engineering and applications, pp 203–209
Rachlin G, Last M, Alberg D, Kandel A (2007) Admiral: a data mining based financial trading system. In: IEEE symposium on computational intelligence and data mining
Rajput V, Bobde S (2016) Stock market forecasting techniques: literature survey. Int J Comput Sci Mob Comput 5(6):500–506
Reuters T (2016) OptiRisk: Marketpsych indices and sentiment analysis toolkit. Products Leaflets Thomson Reuters
Ruiz EJ, Hristidis V, Castillo C, Gionis A, Jaimes A (2012) Correlating financial time series with micro-blogging activity. In: Proceedings of the fifth ACM international conference on web search and data mining, pp 513–522
Sag IA, Baldwin T, Bond F, Copestake A, Flickinger D (2002) Multiword expressions: a pain in the neck for NLP. In: Lecture notes in computer science, vol 2276, pp 1–15
Samo YLK, Vervuurt A (2016) Stochastic portfolio theory: a machine learning approach. In: Proceedings of the thirty-second conference on uncertainty in artificial intelligence (UAI)

Schneider MJ, Gupta S (2016) Forecasting sales of new and existing products using consumer reviews: a random projections approach. Int J Forecast 32:243–256
Schumaker RP, Chen H (2009) Textual analysis of stock market prediction using breaking financial news: the AZFin text system. ACM Trans Inf Syst 27(2):1–19. https://doi.org/10.1145/1462198.1462204
Schumaker RP, Zhang Y, Huang CN, Chen H (2012) Financial fraud detection using vocal, linguistic and financial cues. Decis Support Syst 53:458–464
Sehgal V, Song C (2007) Sops: stock prediction using web sentiment. In: Proceedings of the seventh IEEE international conference on data mining workshops, pp 21–26
Shacham S (1983) A shortened version of the profile of mood states. J Personal Assess 47(3):305–306
Shen W, Wang J, Ma S (2014) Doubly regularized portfolio with risk minimization. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence. AAAI Press, pp 1286–1292
Si J, Mukherjee A, Liu B, Li Q, Li H, Deng X (2013) Exploiting topic based twitter sentiment for stock prediction. In: The 51st annual meeting of the association for computational linguistics (ACL)
Si J, Mukherjee A, Liu B, Pan SJ, Li Q, Li H (2014) Exploiting social relations and sentiment for stock prediction. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 1139–1145
Sowa JF (1987) Semantic networks. In: Shapiro SC (eds) Encyclopedia of artificial intelligence. Wiley, pp 1011–1024
Stein D, Bouchey P, Atwill T, Nemtchinov V (2013) Why does diversifying and rebalancing create alpha? White paper, Parametric
Tai Y, Kao H (2013) Automatic domain-specific sentiment lexicon generation with label propagation. In: The 15th international conference on information integration and web-based applications and services, Vienna, Austria
Taleb NN (2008) Finiteness of variance is irrelevant in the practice of quantitative finance. Complexity 14(3):66–76. https://doi.org/10.1002/cplx.20263
Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: quantifying language to measure firms' fundamentals. J Finance 63(3):1437–1467
Ticknor JL (2013) A bayesian regularized artificial neural network for stock market forecasting. Expert Syst Appl 40(14):5501–5506
Tkác M, Verner R (2016) Artificial neural networks in business: two decades of research. Appl Soft Comput 38:788–804
Uhl M (2014) Reuters sentiment and stock returns. J Behav Finance 15(4):287–298
Valitutti R (2004) WordNet-affect: an affective extension of WordNet. In: Proceedings of the 4th international conference on language resources and evaluation, pp 1083–1086
Vui CS et al (2013) A review of stock market prediction with artificial neural network. In: IEEE international conference on control system, computing and engineering, pp 477–482
Wei W, Mao Y, Wang B (2016) Twitter volume spikes and stock options pricing. Comput Commun 73:271–281
Weidmann NB, Ward MD (2010) Predicting conflict in space and time. J Confl Resolut 54(6):883–901
Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: a system for subjectivity analysis. In: Empirical methods in natural language processing (EMNLP)
Witte JH (2015) Volatility harvesting: extracting return from randomness. CoRR arXiv:1508.05241
Wuthrich B, Cho V, Leung S, Permunetilleke D, Sankaran K, Zhang J (1998) Daily stock market forecast from textual web data. In: IEEE international conference on systems, man, and cybernetics, vol 3, pp 2720–2725
Xing FZ, Cambria E, Zou X (2017) Predicting evolving chaotic time series with fuzzy neural networks. In: International joint conference on neural networks (IJCNN), pp 3176–3183
Yoshihara A, Seki K, Uehara K (2016) Leveraging temporal properties of news events for stock market prediction. Artif Intell Res 5(1):103–110
Zhang GP (2003) Time series forecasting using a hybrid arima and neural network model. Neurocomputing 50:159–175
Zhang W, Li C, Ye Y, Li W, Ngai EW (2015) Dynamic business network analysis for correlated stock price movement prediction. IEEE Intell Syst 30(2):26–33