STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720

www.stmjournals.com

Sentiment Analysis for Financial Predictions: A Review

Puneet Misra[1] and Siddharth Chaurasia[1]*
[1]Department of Computer Science, University of Lucknow, India

Abstract

Sentiment analysis is one of the popular research areas in computer science, as it finds its application in numerous areas like finance, election and product reviews. Behavioral economics suggest that decision of investing in financial markets is driven by feelings like greed, fear. With widespread adoption of technology and digitization of corporate disclosures, news and social media platforms, information expressed through words can be mined and used to gaze into future. Text streams are inherently challenging to interpret the meaning and its effects. Current paper reviews the recent work done for financial market predictions by processing texts from varied sources using machine learning techniques. Support Vector Machine and Naïve Bayes turns out to be extensively used for financial predictions by analyzing text data. Moreover, almost all the studies that used disparate sources of information showed improvement when sentiments were included for predictions thus, highlighting the importance for its inclusion in the prediction process .

Keywords: *Machine Learning; Prediction; Sentiment Analysis; Stock Market; Text Mining*

****Author for Correspondence*** E-mail : siddharth515@gmail.com, Tel : 9410110598

INTRODUCTION

Majority of research that aims at predicting stock market movements uses historical price data. Recently, with growth of digital media like online news, corporate disclosures and announcements are all available to be mined using computational processing techniques. Moreover, as digitization grows so its penetration into our day to day life. Thus, widespread adoptability of social media platforms across the world has made it an interesting stage to share views. Portals like Twitter, Facebook, Telegram all are being used as a platform to express oneself. Therefore, with technological breakthroughs new avenues has been created which can be used to assess and interpret the signals generated by them. With an aim to achieve greater accuracies few studies have even used disparate source of information i.e. both historical and textual data to provide the final predictions.

Rationale behind the application of sentiment and text analysis is taken from behavioral economics which says news, blogs and social media comments may present a picture of sentiments of the larger world for a financial event or stock. The media do not report market status only, but they actively create an impact on market dynamics based on the news they release [1]. Siganos [2] examined the relation between daily sentiment and trading behavior within 20 international markets by exploiting Facebook's Gross National Happiness Index. Interesting correlations were reported in their paper between index and user behavior: First, sentiment has a positive contemporaneous relation to stock returns. Second, sentiment on Sunday affects stock returns on Monday, suggesting causality from sentiment to stock markets. Third, negative sentiments are related to increases in trading volume and return volatility. Thus, all the three observations supported the behavioral finance aspects and their correlation with stock market. Similar to Siganos's work, recently google trends [3] have also found to have predictive properties in the search behavior. Such studies are only limited in number but enforces the support for claims of behavioral economics.

It is quite interesting that, though financial forecasting covers a wide range of ideas from inflation rate prediction to credit scoring, a large proportion of the studies that employed textual data focus on stock market and foreign exchange rate prediction [4]. This appeal can be attributed to transparency of stock

and currency markets. It also indicates amount of focus these two fields receive by general community hence researchers. 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.

Text streams are more challenging to handle than numeric data streams. Text data are unstructured by nature, but they represent collective expressions that are of value in any financial decision [5]. The text mining process consists of two steps: text preprocessing and knowledge extraction. Text preprocessing step converts unstructured data into a document-term matrix, and knowledge extraction step involves mining data for information [6].

This review is divided by the nature of data being dealt with. The correlation is made between type of data being processed and the forecast horizon it may provide. A similarity argument is given on the basis of how it can be mapped to the analysis types done in the domain of finance. Paper first reviews popular approaches being used for predictions by processing texts and then reviews current researches i.e. 2014 onwards as it shows the current state of art techniques and concepts being used in the related field of study.

APPROACHES FOR SENTIMENT CLASSIFICATION

Sentiment classification can be processed by two approaches: Lexicon based, or Machine learning (ML) based. Machine learning yields maximum accuracy while semantic orientation provides better generality [7].

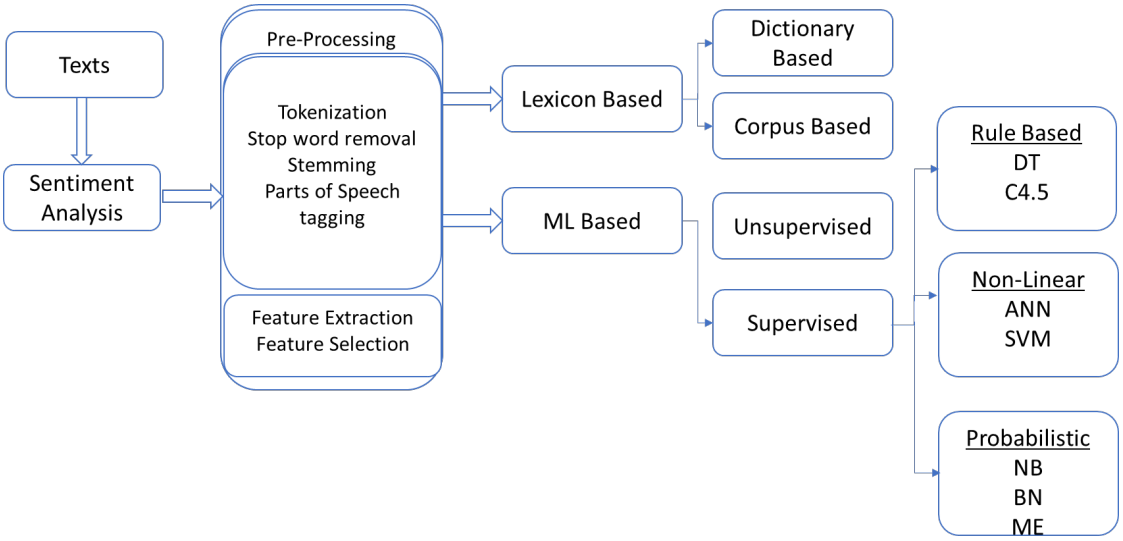
Under lexicon-based approaches, one can use either dictionary or corpus-based approach. Dictionary based approach will use an existing dictionary, which is a collection of opinion words along with their positive or negative sentiment strength. Corpus based approach relies on the probability of occurrence of a sentiment word in conjunction with positive or negative set of words by performing search on very huge amount of texts like Google search [8].

Based on the way learning is performed, machine learning approaches can be divided into two types: Supervised learning or unsupervised learning. Most of the applications discussed in the literature review belong to the supervised learning. For supervised learning, two sets of annotated data both for training and testing is required. A supervised learning setup can be described as: given a text; we try to predict a class based on the real-world continuous data associated with texts meaning. This may be in the form of positive, negative or neutral sentiments with varying degrees.

For machine learning techniques qualitative information available in the text needs to be converted to quantitative which can then be utilized by algorithms. One interpretation is given by Wang [9] with their relative advantages of the two approaches for sentiment analysis. At the very basic level, the word-based n-gram (generative) model and the character-based tagging (discriminative) model are two approaches in the literature. The word-based n-gram approach gives good performance for the in-vocabulary (IV) words; though, it handles out-of-vocabulary (OOV) words poorly. On the other hand, though the character-based approach is more robust for OOV words, it struggles with IV words. The two approaches behave contrarily due to the unit they use, and the model form they adopt, i.e. word vs character (generative vs discriminative).

Popular ML techniques along with important preprocessing steps that are frequently used for the sentiment analysis in financial domain are illustrated in figure1. Support vector machine (SVM), Naïve Bayes (NB), Artificial Neural Network (ANN), Decision Tree (DT) forms the most used techniques.

Fig. 1: Approaches for Sentiment Analysis



ML techniques can be divided into two classes: generative and discriminative. Generative models are built to capture the interaction between all the variables of a system, in order to probabilistically synthesize possible classes. In the context of text mining, it describes how prospective each topic is, and how probable is the word given in the topic. This is how it says documents are generated by the word. A topic is the result of some distribution of words, and words arise because of the topic in the document. Generative models classify the document of words W into topic T by maximizing the joint likelihood: $P(T, W) = P(W|T)P(T)$. Probabilistic methods like Naïve Bayes, Bayesian networks fall in this category. Discriminative model describes how likely a topic is when the set of words are given. The task is to model $P(T|W)$ directly such that T that maximizes this. Discriminative approaches are not concerned with P(T) or P(W) directly [10].

Khadjeh [11] observes SVM as the most used machine learning technique followed by NB for text mining. ANN and K-NN has not drawn much attention with regards to analyzing unstructured text data. In line with Khadjeh, Ravi[6] also reports SVM and NB as two most widely used approaches which are used for stock market prediction by processing texts.

For all the approaches pre-processing is extremely important. The more unstructured it is, greater the importance of pre-processing. Some popular preprocessing steps are: tokenization, stop word removal, stemming, parts of speech (POS) tagging, and feature extraction and representation [8].

Due to sparseness and noise in textual data, feature extraction is also important to get meaningful information. Feature selection approaches can be divided into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used [7]. These techniques can follow two approaches: Bag of Word (BoW) approach or string that retains the sequence of words in the document. BoW approach is easy to use and is widely followed because of its simplicity.

REVIEW OF LITERATURE

Text mining for market prediction is at the intersection of linguistics, machine learning, and behavioral economics [11]. Xing [12] divided the sources of information used for NLFF - Natural Language Financial Forecasting into six classes.

- Corporate disclosures

- Financial reports
- Professional periodicals
- Aggregated News
- Message Boards
- Social Media

Notable reviews since 2014 that have covered text analysis in financial market predictions are [6], [8], [11] and [12]. While, [13] included text analysis as a topic in a broader review of computational methods for financial predictions.

Decisions taken on the basis of corporate disclosures, results and various research reports can be said to be on the basis of fundamentals. The emphasis in these methods is in line with the fundamental analysis, hence, can be said to have a long-term view. Semi-structured information provided by various financial periodicals, financial websites and news aggregators like Yahoo and Google finance may represent the beliefs of notable persons in the industry. Views of subject matter experts can have impacts of different scales depending on the creditability of the publication and author. These may be in the form of analysis of individual stocks or sector or can have more extensive coverage. Such articles can have both technical and fundamental flavors hence can carry the impact from mid-term to long term.

Behavioral economics tells us that emotions can greatly affect individual behavior and decision-making. Bollen [14] investigated whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the DJIA over time. Emotion about financial markets is somewhat contagious and can be discussed and pooled [15]. Hence, most of the studies reviewed in this section share a belief; altogether, social media displays measurable emotion, and the derived sentiment metrics can plausibly predict buying and selling behavior in the short term. There has been significant research over Twitter as sharing platform because of the relatively simple semantics that is expressed in a restricted character length [16]. Thus, we group our review on the basis of the types of inputs used for predictions.

Structured Textual Data

An important part of fundamental analysis is to extract and interpret information embedded in quarterly, annual results and intermediate disclosures. Studies that focus on interpreting corporate information generally have mid to long term views. Though text data is unstructured as compared to historical price information but within the universe of text information this kind of information can be termed as structured or semi-structured. These kinds of declarations have a framework in which declarations are made.

Hajek [17] based on their literature review observed a strong relationship between textual information extracted from annual reports and abnormal stock return. Another prior research by the same author [18] observed existence of non-linear relationships between the sentiment from annual reports and financial performance of the studied stocks. Processing of corporate disclosures is important but have received less weightage from research per se. Reason for this can be attributed to the amount of knowledge that is required to actually understand the reports. Automation in this regard generally provides half-baked information and cannot be relied end to end. Such statements and disclosures still require experts to interpret things though it may ease out the process.

Semi-structured Textual Data

Second important source of information are news channels and views by financial analysts or reports published by independent analysts or brokerage firms. This kind of data can be considered as semi structured as, though, there is no set pattern still language used in these media is a proper. Fair amount of research has been done to utilize information contained in such type of documents.

Uhl [19] analyzed more than 3.6 million Reuters news articles to forecast returns of DJIA using vector autoregressive models. Findings from the research were encouraging as it showed a fair degree of correlation between news and stock movement. Results demonstrated that positive and negative Reuters sentiment have an impact on stock returns. The impact is different with regards to degree and tie. Moreover, adverse sentiment had a higher influence on stock returns than positive sentiment. They turned out to be more obstinate than positive sentiments and may take more time than a few days to absorb the news effect fully. Results indicated Reuters sentiment seems to matter the most after one month, while negative Reuters sentiment appears the most relevant with a lag of three months. Volume is also an indicator of subsequent stock returns, while it is most significant with a delay of two months.

Yoshihara [20] proposed an approach to predicting the trend of stock prices by focusing on news events with long-term effects. Combination of RNN-RBM was employed on the input of news articles represented as word vectors by the bag-of-words representation.

Li [21] captured social sentiment and professional opinions by processing textual information in social media and financial news, respectively, and then represented the whole market information space consisting of these two information sources along with firm characteristics via tensors. A worth mentioning observation made in the study is that the influential power of social sentiment varies with firm characteristics, including trading volumes, turnovers, P/E ratios, and P/B ratios. This is due to the investors bias on industry and firm locality.

Ding [22] applied deep learning to determine event-driven stock price movement. Results from experiments showed that event embeddings based document representations are better than discrete events-based methods, and deep convolutional neural network (CNN) can capture the long-term influence of news event than standard feedforward neural network.

Shynkevich [23] explored the impact of simultaneous analysis of different categories of financial news on the forecasting. News articles are classified as per their relevance to the target stock, its sub-industry, industry, group industry and sector. Chen [24] improved upon the forecasting accuracies by over 4% when financial news was also incorporated in their fundamental analysis based forecast attempts.

In another work where both historical and text information from news channels is used was done by [25]. They proposed an intelligent stock trading system using comprehensive features (ISTSCF) framework to predict future. Stock trading decisions are based on the idea that technical analysis can generate trading indices (TI) based on historical data, while sentiment analysis is used to identify the future trends and generate sentiment indices (SI). Combining the two information extraction techniques improved the model's predictability. The proposed ISTSCF framework consists of stock information extraction, prediction model learning and stock trading decision. Three different methods were united to generate features - First, sentiment analysis is used that provides sensitive market events from stock news articles for sentiment indices, second, TA is utilized that yields effective trading rules based on trading information on the stock exchange for technical indices, Third, trend-based segmentation method (TBSM) that raises trading decisions from stock price for trading signals is employed. Experimental results confirmed that the use of comprehensive features combined with technical indices, sentiment indices and trading signals could improve forecasting performance, which can be further enhanced by SA, TA and TBSM. Additionally, the TBSM also correctly determined trading decisions for obtaining better investment profit.

Ammann [26] analyzed the predictive power of newspaper articles for German stock market returns. Based on newspaper articles published between July 1989 and March 2011 and the main finding is that newspaper stories contained relevant information for predicting future DAX returns.

Unstructured Textual Data

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 [27]. Third area of textual processing has received most attention from the research community. Unstructured data comes in form of social media interactions like twitter feeds, Facebook messaging, WhatsApp chats. Social media engagements are used to express and share the views. Behavioral economics tells us that emotions can greatly affect individual behavior and decision-making. It further states that human decisions are seldom done solely based on intelligence and are driven by emotions like greed and fear. Thus, mining of social media is governed by rational to predict the prevailing sentiments among the public.

Zheng [28] investigated whether measurements of collective emotional states derived from large-scale network feeds are correlated to the stock transaction data over time. Results show that social media sentiments contain statistically significant information on the future prices. Checkley [29] explored the predictability of stock markets on the basis of sentiment metrics extracted from micro-blogging sites Twitter and Stocktwits diverse and high-volume sentiment were found to be predictive of price volatility and traded volume than providing consensus in price direction. The strongest causal links between sentiment and price behavior were found in times of strident and discordant market mood. Similarly, A sentiment analysis engine (SAE) is proposed which takes advantage of linguistic analyses based on grammars in [5]. The engine extends sentiment analysis not only at the word token level but also at the phrase level within each sentence. An apparent co-relation between sentiments and stock indices movement direction was found in this research.

Nofer [30] experimentally presented the results which do not support the view that the simple aggregation of mood states of all individuals in the Twitter blogosphere is sufficient to predict the stock market. Instead, it is necessary to consider the community structure that is a number of followers. They proposed follower-weighted social mood levels which can predict share returns. Experimentally during the six-month testing period, experimental portfolio showed an increase of up to 36 %.

Geva [31] combined the two sources of data for forecasting stock returns in intraday trading. Rational behind integration is, enrichment that the available information can give to the forecasting model which can potentially capture common patterns that may not otherwise be identified when each data source is employed separately. 10 technical indicators were used which were calculated for each 5-min interval. Sliding window approach was used as it is suitable for rapid online calculation. The study shows that integrating market data with textual data contributes to improving the modelling performance and that using more advanced textual data representations further improves predictive accuracy. However, these results strongly depend on the joint selection of both data representation and forecasting algorithm.

Another more recent work where both historical numerical data and textual information from disparate sources was used by [32]. Authors hypothesize that combining disparate online data sources with traditional time-series and technical indicators for a stock can provide a more effective and smart daily trading expert system. In their research work three machine learning models: decision trees, neural networks and support vector machines were used to create an inference engine. Conclusions were in line with the hypothesis that the knowledge base of financial expert systems could benefit from data captured from non-traditional experts like Google and Wikipedia and combining data from different sources can assist to improve the performance of financial systems.

Yang [33] compared two ML techniques NB and SVM to do price prediction using historical prices of 5 stocks from the US market. Textual information from Twitter is showed to improve the prediction accuracy. Dataset taken in the experiments are not exhaustive and cannot be considered as a valid generalization based on the tests conducted. Only 100 tweets for 5 companies were examined which may not represent the proper sample.

Table** Error! No text of specified style in document.. **Primary Studies Aimed at Market Prediction(2014 Onwards)

Article	Applied area	Market	Data Set	Prediction Time Frame	Technique
[5]	Stocks	Hong Kong	Blogs from analysts, Daily news articles from Finet	Long Term	Linguistic based, DT, SVM-RBF
[17]	1402 NYSE Stocks	US	Annual Reports	90 days Price	MLP, NB, SVM, DT, k-NN
[18]	448 Stocks	US	Annual Reports	Long Term	LoR, NN, SVM, DT
[19]	Stock Index	US	Reuters news articles	Few months	VAR
[20]	10 Stocks	Japan	Newspaper articles	Daily trend	SVM, DBN, RNN-RBM
[21]	CSI (100)	China	Social Media, Financial news, Firm Characteristic	20 min	SVM, ANN-BP, Tensor based
[22]	S&P 500 – 15 Stocks	US	News from Reuters and Bloomberg	Daily price prediction	CNN
[23]	S&P 500 - 53 stocks	US	Financial news: Newswire, McClatchy-Tribune Business News and Business Wire	Intraday	SVM, ANN, MKL
[24]	2 Stocks	Taiwan	Financial news, Fundamental and economic data	Long term	AdaBoost (SVM, GA-SVM)
[25]	5 stocks	Taiwan	News on Internet	Daily trading point	SVM with multiple Kernels
[29]	5 Stocks	US	Twitter & StockTwits	Few minutes	LR
[30]	Stock Index	Germany/ Europe	Twitter	One day return	Dictionary based.
[31]	72 S&P500 stocks	US	News from Reuters	One-hour price	FANN, DT+GA, LoR
[32]	Apple Stock	US	Wikipedia pages, Online content produced about a company, technical indicators and company value indicators	Daily direction	DT, SVM
[33]	5 Stocks	US	Historical price and Twitter	Daily direction	NB, SVM
[35]	6 Stocks, 50ETF	China	Baidu search results and Transactional Data	Daily open price	LSTM + Sentiments, MLP & RNN
[36]	Stock Index	US	DJ/Factiva news archive	9,12,25,45 days	Linguistic
[34]	18 stocks	US	Yahoo Message Boards	Daily direction	Aspect based sentiment, Joint Sentiment Topic Model, LDA
[37]	78 - CSI 100 stocks, 13 HKG stocks	China, Hongkong	News Articles, Social media Sentiments	Daily direction	SVM, PCA-SVM
[38]	23 Stocks from HSI	Hong Kong	News Articles and Historical Price Data	Intraday	Multiple Kernel Learning

CNN – Convolutional NN, DJIA – Dow Jones Industrial Average, DBN- Deep Belief Network, GA – Genetic Algorithm, k-NN – K-Nearest Neighbor, LoR- Logistic Regression, LR – Linear Regression, LSTM Long Short Term Memory, LDA – Linear Discriminant Analysis, MKL- Multiple Kernel Learning, PCA -Principal Component Analysis, RNN – Recurrent NN

Nguyen [34] incorporated the sentiments of specific topics of the company in their prediction model. Topics and related sentiments were automatically mined from the texts in a message board by using proposed method and existing topic models. Authors were able to increase the prediction accuracy by around 2% by incorporating sentiments. Stocks that are highly correlated tend to be affected by the same event. Thus, instead of conducting each stock prediction separately and independently, authors

predicted multiple correlated stocks simultaneously through their common properties, which are enabled via sharing the collaboratively factorized low-rank matrices between matrices and the tensor.

Zhang [37] used data from three sources Quantitative price data, Web news data and the data from social media. The events from Web news and the user sentiments from social media are investigated for their joint impacts on the stock price movements through a coupled matrix and tensor factorization framework. A tensor is constructed to fuse heterogeneous data and capture the hidden relations among the events and the investor's sentiments.

Table I presents reviewed primary studies and summarizes different time frames and sources of information that have been widely used along with the ML techniques that have been applied. There is a clear inclination towards ML approaches in recent times. SVM, NB, variants of DT, variants of ANN all are being applied in search of improvement in dealing with textual data.

CONCLUSIONS

Review has been grouped on the basis of different sources of textual information that is used for market predictions. Primary source of information which provides medium to long term views are corporate disclosures like quarterly and annual results. Not many studies have used corporate documents for information extraction as they require fair bit of domain knowledge to be interpreted properly.

Second type of source is news by news channels and reports published by independent experts and brokerage houses. Such documents provide short to medium term of information and are based on events and micro and macro-economic changes. These sources provide independent viewpoints of experts. Fair bit of research has been done in this area and Machine learning techniques have widely been used. Often a dictionary-based approach is used to classify the polarity of document.

Third source is social media platforms like Twitter, or more specific financial discussion platforms like Stocktwits, Moneycontrol. These are most popular to be used to sense short term movements. They represent collective expressions that are of value in any financial decision and may indicate what the herd is thinking. Impact of such expressions are short lived hence response time is also important with regards to media mining. In particular, data provided by twitter has been the most researched one.

Analyzing the recent literature, it can be said that almost all the studies belonging to the three classes contributed towards a predictive capacity of texts in various geographies and with varied timeframes. The predictive capacity of texts when combined with historical approaches also showed a positive impact in the accuracy and hence showing its applicability for financial markets.

Moreover, Junqué [39] though acknowledged the acceptability of deviations from Efficient Market Hypothesis [40] i.e. market's unpredictability in shorter timeframe, but cautioned on usage of judiciously selected performance metrices to actually measure the effectiveness of the predictability.

Machine learning methods are widely being used to analyze the sentiments of the documents. Techniques like SVM, NB and DT are quite popular for text information processing unlike ANN which is the most popular method when dealing with numerical data. Recently, different variants like DBN, LSTM are being studied for analyzing texts related to market. Moreover, from prediction point of view, often sentiment analysis is used to improve upon accuracy provided by historical data. Multiple researches have been done in recent times that have used disparate source of information.

An important point to note is most of the research by analyzing text has been applied in developed economies. US along contributes to more than 50% of such researches. One reason apart from US being biggest economy is such markets are more efficient and is mostly the forerunner in adoption of

new technological advancements. Even common man is more active on social media and freely expresses their views on events that may drive the markets. Results from developed to developing economies cannot be mapped directly as the efficiency level for both are different. Thus, as the technology penetration grows, emerging economies like India, Brazil provide good avenues for further exploration.

REFERENCES

[1] E. A. Gerlein, M. McGinnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," *Expert Syst. Appl.*, vol. 54, pp. 193–207, 2016.

[2] A. Siganos, E. Vagenas-Nanos, and P. Verwijmeren, "Facebook's daily sentiment and international stock markets," *J. Econ. Behav. Organ.*, vol. 107, no. PB, pp. 730–743, 2014.

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

[4] M. J. Schneider and S. Gupta, "Forecasting sales of new and existing products using consumer reviews: A random projections approach," *Int. J. Forecast.*, vol. 32, no. 2, pp. 243–256, 2016.

[5] S. W. K. Chan and M. W. C. Chong, "Sentiment analysis in financial texts," *Decis. Support Syst.*, vol. 94, pp. 53–64, 2017.

[6] B. S. Kumar and V. Ravi, *A survey of the applications of text mining in financial domain*, vol. 114. Elsevier B.V., 2016.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[8] K. Ravi and V. Ravi, *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications*, vol. 89, no. June. Elsevier B.V., 2015.

[9] K. Wang, C. Zong, and K.-Y. Su, "Integrating Generative and Discriminative Character-Based Models for Chinese Word Segmentation," *ACM Trans. Asian Lang. Inf. Process.*, vol. 11, no. 2, pp. 1–41, 2012.

[10] P. Misra and Siddharth, "Financial Market Prediction : Generative Vs Discriminative Methods," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 7, 2018.

[11] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," *Expert Syst. Appl.*, vol. 41, no. 16, pp. 7653–7670, 2014.

[12] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artif. Intell. Rev.*, pp. 1–25, 2017.

[13] R. C. Cavalcante, R. C. Brasileiro, V. L. F. Souza, J. P. Nobrega, and A. L. I. Oliveira, "Computational Intelligence and Financial Markets: A Survey and Future Directions," *Expert Syst. Appl.*, vol. 55, pp. 194–211, 2016.

[14] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[15] N. Oliveira, P. Cortez, and N. Areal, "Some experiments on modeling stock market behavior using investor sentiment analysis and posting volume from Twitter," *Proc. 3rd Int. Conf. Web Intell. Min. Semant. - WIMS '13*, p. 1, 2013.

[16] W. Wei, Y. Mao, and B. Wang, "Twitter volume spikes and stock options pricing," *Comput. Commun.*, vol. 73, pp. 271–281, 2016.

[17] P. Hájek, "Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns," *Neural Comput. Appl.*, vol. 29, no. 7, pp. 343–358, 2018.

[18] P. Hajek, V. Olej, and R. Myskova, "Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making," *Technol. Econ. Dev. Econ.*, vol. 20, no. 4, pp. 721–738, 2014.

[19] M. W. Uhl, "Reuters Sentiment and Stock Returns," *J. Behav. Financ.*, vol. 15, no. 4, pp. 287–298, 2014.

[20] A. Yoshihara, K. Seki, and K. Uehara, "Leveraging temporal properties of news events for stock market prediction," *Artif. Intell. Res.*, vol. 5, no. 1, pp. 103–110, 2015.
[21] Q. Li, J. Wang, F. Wang, P. Li, L. Liu, and Y. Chen, "The role of social sentiment in stock markets : a view from joint effects of multiple information sources," *Multimed. Tools Appl.*, 2016.
[22] X. Ding, Y. Zhang, T. Liu, and J. Duan, "Deep Learning for Event-Driven Stock Prediction," no. Ijcai, pp. 2327–2333, 2015.
[23] Y. Shynkevich, T. M. Mcginnity, S. A. Coleman, and A. Belatreche, "Forecasting movements of health-care stock prices based on different categories of news articles using multiple kernel learning," *Decis. Support Syst. J.*, pp. 1–10, 2016.
[24] Y. J. Chen, Y. M. Chen, and C. L. Lu, "Enhancement of stock market forecasting using an improved fundamental analysis-based approach," *Soft Comput.*, vol. 21, no. 13, pp. 3735–3757, 2017.
[25] J. L. Wu, L. C. Yu, and P. C. Chang, "An intelligent stock trading system using comprehensive features," *Appl. Soft Comput. J.*, vol. 23, pp. 39–50, 2014.
[26] M. Ammann, R. Frey, and M. Verhofen, "Do Newspaper Articles Predict Aggregate Stock Returns?," *J. Behav. Financ.*, vol. 15, no. 3, pp. 195–213, 2014.
[27] E. Cambria, S. Poria, A. Gelbukh, and I. P. Nacional, "Sentiment Analysis Is a Big Suitcase," *IEEE Intell. Syst.*, pp. 74–80, 2017.
[28] J. Zheng, X. Fu, and G. Zhang, "Research on exchange rate forecasting based on deep belief network," *Neural Comput. Appl.*, pp. 1–10, 2017.
[29] M. S. Checkley, D. A. Higón, and H. Alles, "The hasty wisdom of the mob: How market sentiment predicts stock market behavior," *Expert Syst. Appl.*, 2017.
[30] M. Nofer and O. Hinz, "Using Twitter to Predict the Stock Market: Where is the Mood Effect?," *Bus. Inf. Syst. Eng.*, vol. 57, no. 4, pp. 229–242, 2015.
[31] T. Geva and J. Zahavi, "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news," *Decis. Support Syst.*, vol. 57, no. 1, pp. 212–223, 2014.
[32] B. Weng, M. A. Ahmed, and F. M. Megahed, "Stock market one-day ahead movement prediction using disparate data sources," *Expert Syst. Appl.*, vol. 79, pp. 153–163, 2017.
[33] M. Li, C. Yang, J. Zhang, D. Puthal, Y. Luo, and J. Li, "Stock market analysis using social networks," *Proc. Australas. Comput. Sci. Week Multiconference - ACSW '18*, pp. 1–10, 2018.
[34] T. H. Nguyen, K. Shirai, and J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, 2014.
[35] G. Zhang, L. Xu, and Y. Xue, "Model and forecast stock market behavior integrating investor sentiment analysis and transaction data," *Cluster Comput.*, vol. 20, no. 1, pp. 789–803, 2017.
[36] D. Kroujiline, M. Gusev, D. Ushanov, S. V. Sharov, and B. Govorkov, "Forecasting stock market returns over multiple time horizons," *Quant. Financ.*, vol. 16, no. 11, pp. 1695–1712, 2016.
[37] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang, and P. S. Yu, "Improving stock market prediction via heterogeneous information fusion," *Knowledge-Based Syst.*, vol. 143, pp. 236–247, 2018.
[38] X. Li, X. Huang, X. Deng, and S. Zhu, "Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information," *Neurocomputing*, vol. 142, pp. 228–238, 2014.
[39] E. Junqué De Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," *Inf. Process. Manag.*, vol. 50, no. 2, pp. 426–441, 2014.
[40] E. F. Fama, "The Behavior of Stock-Market Prices," *J. Bus.*, vol. 38, no. 1, pp. 34–105, 1965.