Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 10069 - 10082

Research Article

# Text mining a high accuracy tool for Stock Market Prediction: A Critical Review

Shikha Verma[1], Nikita Singla[2], Punam Rattan[3]

## ABSTRACT

Many researchers are now paying a lot of attention towards the predictability of the online text using text mining techniques and combining it with classical data mining of stock market numeric data to increase the accuracy in the prediction systems. Therefore, the current review have analyzed literature on text mining. The review has analyzed the literature, taking in consideration the generic model of text mining so as to evaluate the gaps and to suggest possible research recommendations on the same. The review systematizes a extensive description of literature that facilitates the compilation of research activities for various steps in the suggested generic model of market prediction based on text mining; provides an insight about various lacunas and problems in this aspect so as to build a pipeline for future researchers and lastly gives interdisciplinary and directional suggestions for amelioration of research in this field. The review is thus believed to be holistic as it provides a unique, comprehensive, versatile and multifaceted point of view about on text mining mediated stock market prediction.

***Keywords:*** Text mining, Stock Market, Data miming, Prediction.

## INTRODUCTION

Market economies are very prominent in the modern societies today. Supply and demand equilibrium of the stock markets depicts the economic strength of the nation. Therefore, learning, mapping and predicting future movement of the stock markets is a crucial need.The ability to manage wealth through predicting unfavorable conditions like financial losses or to gain profits by predicting most favorable period to invest, has always been considered as a valuable point of interest. Despite of all these advantages, the nature of markets is so complex and extremely dynamic which makes it difficult to be predicted. Based on their input data, there are generally two types of predictive measures: technical or fundamental analyses. The former uses historic market data and the later uses other data like news about the company or society, country. Due to the ease of availability of the quantitative historic market data and unstructured nature of fundamental data, the prevalence of the technical analysis approaches is seen in the literature as well as a high interest of technical approach is seen in the real world scenario. Also, the source of fundamental data is

[1]Research Scholar, CT University, Ludhiana, India
[2]Dept of Computer Sc., CT University, Ludhiana, India
[3]Lovely Professional University, Ludhiana, India

not definitive or well researched but few reports in past have shown the predictive ability of fundamental data [11, 20, 4, 37, 21]. It has been witnessed that unstructured text available in the form of social news, media postings, blogs etc. for analyzing market movement is a challenging research aspect but literature has shown many hybrid models that included both the technical and fundamental data which when combined depicts higher prediction accuracy [14, 16, 46, 5].

The current paper therefore gives a systematic review on significant research in the past with respect to stock market prediction using text-mining. The main contributions of this work are summarized below:

1. Summering the fundamentals of stock market prediction using a combination of text and data mining in order to improve the prediction accuracy by framing a generic model.

2. A graphical representation of various sources of data used by various authors in past to predict stock market.

3. The gap in the literature is carefully identified and future research recommendation in this field has been made using both narrative and biblio metric style of review.

4. The unique inter-displinary aspect of machine learning, linguistics and behavioral economics in accurate prediction of market movement has been reviewed to bring a foundational concept from a robust comprehension of the previous literature.

## Generic model of stock prediction using machine learning

There have been quite a prevalent number of studies in the bulk of existing literature for stock prediction using machine learning but despite the presence of multiple systems there is no such bibliometric review analysis for the methods designed or used in previous literature. Some authors have tried to make similar efforts in this context like a paper titled “Text mining approaches for stock market prediction” by Nikfarjam et al. [27] or the literature review by Hagenau et al. [15] that presents a collective view of various techniques. Therefore, this section is dedicated to fill in these gaps by making robust literature review onvarious systems developed in the past in the form of a generic model that is depicted in Fig1.

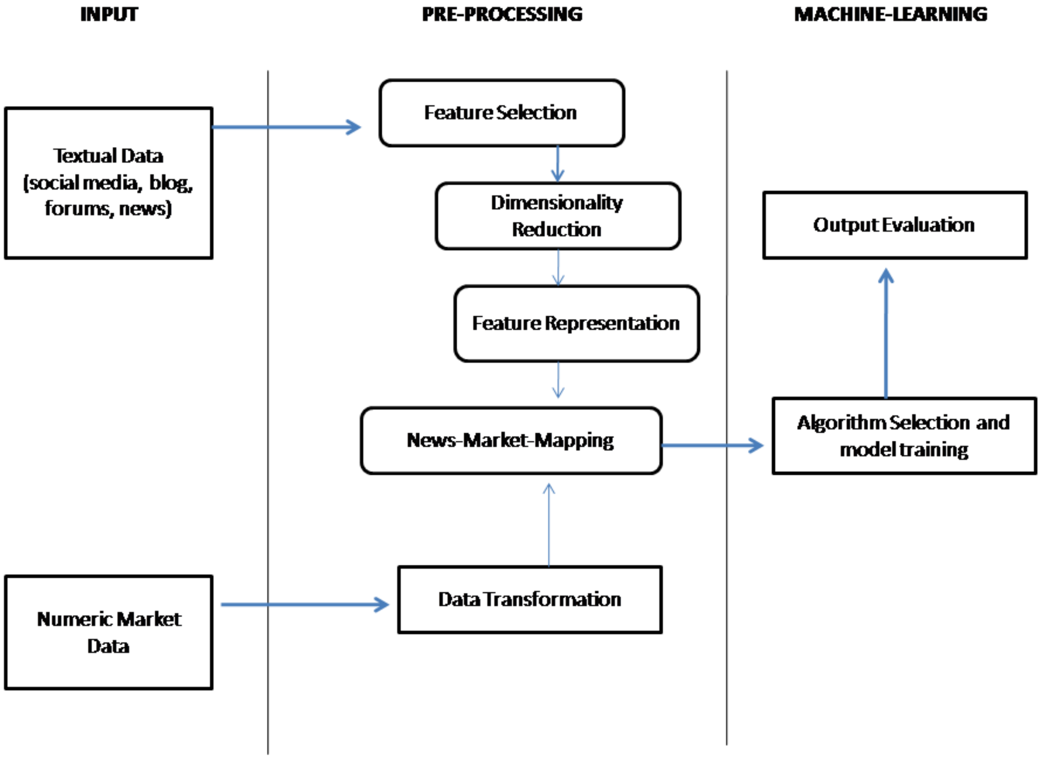

**Fig 1: Generic model of stock prediction using machine learning[21]**

The figure shows that input is fed into system in the form of various types of data (textual/numeric) as per the need of the model which is then preprocessed to remove noise and to improve the output i.e. the market predictive values on the other end.

***Input dataset***

All the systems reviewed during the review analysis depicted two different sources of data viz. textual data from online resources and the numeric market data thus representing the use of both the sources for better accuracy of the predictive model.

*Textual data sources*

Table 1: Comprehensive review on textual market data inputs employed in stock market prediction literature

| **Sources of textual data used** |
|---|
| General news items from text sources like the Wall Street Journal, Dow Jones, Financial Times,Reuters, Bloomberg[45] |
| financial news from sources like OlsenData using HFHF93 data set[28] |
| 600,000 company news from Reuters Market[29] |
| 6602 financial news from various online sources[26] |

| |
|---|
| Yahoo! Finance has also been used as source of textual input. The author s used 1.5 million message postings as data sets [43] |
| 148 direct company news and68 indirect news sets from market-sector retrieved from Australian Financial Review[49] |
| 3493 Financial news from Financial Times [36] |
| 145,110 messages from varying message boards as source of textual input[8] |
| financial news items from Forbes.com, reuters[30] |
| 700 news articles from varying sources from financial news segments of various online sources[25] |
| textual data from sources like Dow Jones News, Factiva news database and wall street journal [38] |
| 2800 financial news sets from Yahoo Finance[33] |
| Annual reports from websites of various companies[3] |
| 12,830 headlines from financial news segment of the leading e-newspaper in Taiwan [17] |
| Corporate filings from management’s discussion and analysissection of 10-K and 10-Q filings from SEC Edgar [22] |
| Adhoc announcements was used as textual data set for corporate disclosure [13] |
| 9,853,498 tweets from twitter.com [2] |
| Broker newsletters from Brokers[24] |
| 5,001,460 tweets from twitter.com [42] |
| Financial news from Yahoo! Finance[34] |
| 361,782 general news items from Bloomberg[19] |
| 52,746 messages from the various sources on internet like Blogs, news and micro blogs, tweets etc.[48] |
| advocated the use of corporateannouncementsand financial news[15] |
| Macroeconomic news from Bloomberg [4] |
| 420 ticker words from micro blogs[37] |
| 1884 stock related news articles from various online financial websites[7] |
| financial news items from both Reuters and Bloomberg[6] |

It has been seen that online textual data have a great influence in the movement of stock markets. This has drawn the attention of many researchers to use textual input from various online sources with varying content types which is described chronologically in this section of the literature review.

The above tabulated depiction of the earlier literature depicts that various news websites like The Reuters, Wall Street Journal, Dow Jones, Financial Times, Forbes, Bloomberg, Yahoo! Finance, were frequently used astext sources. Most of these studies either used general news or financial news as the text type for prediction purpose. However it has also been observed that using financial news is preferred as compared to general news because of lesser noise and lower effort for preprocessing. Further it has been also observed that headlines preferred moreas it is straighterand to the point, thus lacks noise.

*Numeric market data*

Numeric market data in form of price-points or indices are other source of input data that has been prevalently seen in the during the literature review.This helps in achievingaccurate prediction by training the machine learning algorithms. Table 2 gives a chronological insight of the use of numeric market data using different market indices for improved prediction of stock market movement.

Table 2: Comprehensive review on numeric market data inputs and various market index employed in stock market prediction literature

| **Numeric data inputs used** |
|---|
| 33 stock datasets from the “Hang Seng” were recorded for a period of 1 October 2002–30 April2003 on daily basis.[29] |
| Stock prices for a period of 1 January–31 December2002 on daily basis[26] |
| Dow Jones Industrial Average on intraday basis for the entire year of 2000[43] |
| Numerical data aset from Australian StockExchange (BHP Billiton Ltd.) on daily basis from 1 March 2005–31 May 2006[49] |
| Stocks from 11 oil and gas companies for a period from 1 January 1995 to 15 May2006 on daily basis[36] |
| Stocks from “24 tech-sectors in the Morgan Stanley”on daily basis for the months of July and August 2001 on daily basis [8] |
| Daily stock f from US NASDAQ for a period of 7 February to 7 May 2006[30] |
| Stocks from Indian sensex for a period of 5 August 2007 –8 April 2008 on daily basis[25] |
| “S&P and  futurecash flows”for 500 firms on daily basis [38] |
| “S&P of 500 stocks” from 26 October 2004 to 28 November2005 on Intraday basis [33] |
| “Yearly records of 1-Year market drift” for the period of 2003–2008.[3] |
| Daily financial price records from Taiwan Stock Exchange from June–November 2005[17] |
| Quarterly earnings and cash flows as well as records of stock returns on annual basis for a period of 1994–2007[22] |
| Abnormal risk exposure on intraday basis for a period from 1 August 2003 to 31 July 2005[13] |

| |
|---|
| Daily exchange rate for a period from 28 February to 19 December 2008 on daily basis[2] |
| 30 different performance indices on Intraday basis [24] |
| “Daily Stock prices at NASDAQ” in two phases i.e. 1 April 2011 to 31 May 2011& 8 September to 26September 2012[42] |
| “Intraday S&P 500 records” for a period from 26 October to 28 November2005[34] |
| “FOREX exchange rate” from 1 January to 31 December, 2012 on daily basis[19] |
| Two indices for 824 firms viz. “abnormal returns” and “cumulative abnormal Returns”[48] |
| Daily stocks for specific company for a period of 1997–2011 on the daily basis.[15] |
| Korea Stock Price Index recorded on daily basis for a period of January 2000 to December 2016[5] |
| S&P-500 recorded on daily basis for the period of 3 January 2007 to 30 December 2016[35] |

The data in graphical representation of the literature review depicts some crucial findings in this context. stock market indices are the most commonly used textual data type for stock market prediction. However some authors have also resorted to stock price of a specific company for the same.

### *Pre-processing*

The preparation of the collected data has to be made ready for feeding into the machine learning algorithm. Especially with textual data it becomes a critical task because it has to be compulsorily transformed from unstructured to a structured format that is processable by the machine, thus this phase has a key role to play in the any machine learning construct when accurate outcomes are concerned (41). The literature review indicates that there are three sub-processes involved in this phase which includes: feature-selection, dimensionality-reduction, and feature-representation. (Table 3)

Table 3: Comprehensive review on various pre-processing sub processes used in stock market prediction literature used in the previous section of the study

| **Pre-processing method used** |
|---|
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction:**Pre-defined dictionaries<br>**Feature representation:**Binary [45] |
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction:** keyword records |

| |
|---|
| **Feature representation:** Boolean [28] |
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction**: punctuation removal, lowercase conversion, identification and removal of stop-words<br>**Feature representation:** TF-IDF [29] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction**: 1000 terms specific selection<br>**Feature representation:** TF-IDF [26] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction**: top 1000 words selection<br>**Feature representation:**Binary [43] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction**: removing stop words<br>**Feature representation:** TF-IDF and Binary[49] |
| **Feature-selection:**Visualisation,<br>**Dimensionality-reduction**: term extraction using Thesaurus<br>**Feature representation:**Visual coordinates[36] |
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction**: Pre-defined dictionaries<br>**Feature representation:** Discrete values foreach classifier Binary[8] |
| **Feature-selection:** “Bag-of-words model” (financial values),<br>**Dimensionality-reduction**: automatic extraction of influential keywords<br>**Feature representation:**Boolean[30] |
| **Feature-selection:**Latent Dirichlet Allocation<br>**Dimensionality-reduction**: 25 most influential topic extraction<br>**Feature representation:**Binary[25] |
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction**: Pre-defined dictionary<br>**Feature representation:** frequency / totalnumber of words[38] |
| **Feature-selection:** “Bag-of-words model”<br>**Dimensionality-reduction**: Minimum occurrence per document<br>**Feature representation:**Binary [33] |
| **Feature-selection:**Character n-Grams and previous year performance<br>**Dimensionality-reduction**: Minimum occurrence per document<br>**Feature representation:** the frequency of the n-gram in one profile[3] |
| **Feature-selection:** Ordered pairs<br>**Dimensionality-reduction**: Replacement of synonyms<br>**Feature representation:**Weighted based on the rise/ fall ratio of index[17] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction:**Pre-defined dictionaries<br>**Feature representation:**Binary[22] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction:** Feature Scoring via Chi-Squared metrics<br>**Feature representation:** TF-IDF[13] |

| |
|---|
| **Feature-selection:**OpinionFinder<br>**Dimensionality-reduction:**OpinionFinder<br>**Feature representation:**OpinionFinder[2] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction:**Stremming<br>**Feature representation:**Sentiment Value[24] |
| **Feature-selection:**<br>• “Daily aggregate number of positives or negatives on Twitter Sentiment Tool (TST) and an emoticon lexicon.<br>• Daily mean of Pointwise Mutual Information (PMI) for pre-definedbullish-bearish anchor words”<br>**Dimensionality-reduction:**Pre-defined company specific keywords<br>**Feature representation:** “Real number for DailyNeg_Pos&Bullish_Bearish”[42,17] |
| **Feature-selection:**OpinionFinder<br>**Dimensionality-reduction:**Minimum occurrence per document<br>**Feature representation:** Binary[34] |
| **Feature-selection:** “Latent Dirichlet Allocation”<br>**Dimensionality-reduction:** Extraction of topic, manual identification of top topics with special emphasis to currency fluctuations in news articles<br>**Feature representation:** topic distribution of each article[19] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction:** nil<br>**Feature representation:**Binary[48] |
| **Feature-selection:** “Bag-of-words model”,<br>**Dimensionality-reduction:** news frequency<br>**Feature representation:** TF-IDF[15] |
| **Feature-selection:**“Structured Data”<br>**Dimensionality-reduction:**Structured data<br>**Feature representation:**Structured Data [4] |

### *Machine learning*

This section of the literature review focuses on summering machine learning algorithms that has been used by previous authors. The bibliometric analysis points to the fact that machine learning algorithms used in the study are classification algorithms are the most popular ones followed by regression analysis. It has also been noted from the previous literature that comparisons between various algorithms are not an easy task to do as it is full of drawbacks (32). The 6 basic algorithms that could be traced for its prevalence in the literature are mentioned here:

Table 4: Machine leaning algorithms traced from the literature review

| **Algorithms** | **Description** |
|---|---|

| | |
|---|---|
| Support Vector Machine | It is used for supervised learning and is a non-probabilistic binary linear classifier. It finds an appropriate hyperplane that separates two classes with a maximum margin. Thus, training in this algorithm is made in the form of quadratic programming optimization problem. [29, 43, 26, 36, 33, 7] |
| Naïve Bayes | It is the oldest classification algorithm which is based on the Bayes Theorem and is termed “naïve” as it is based on the naïve assumption of complete independence between text features. [45, 43, 22, 48] |
| Decision Rules or Trees: | There has been many efforts in past to create rule-based classification systems[28, 17, 18, 42] |
| Regression Algorithms | There are various approach of using regression algorithm in this context like Support Vector Regression (SVR) that is based on variation of SVM. Other than this another approach of using linear regression models is direct use.[ 9, 38, 34, 15, 19, 4] |
| Combinatory Algorithms | Numerous machine learning algorithms that are stacked or grouped together composes combinatory algorithms which has seen a major relevance in this theme.[ 8, 25, 3, 2] |
| Multi-algorithm experiments | When same experiments are conducted by using a number of different algorithms it raises the accuracy levels and thus multi-algorithm experiment is gaining popularity in this context.[ 43, 13, 21] |

## Inferences and Gaps from the reviewed literature

The review of the above literature gives a generalized inference that research in thisparticular predictive application of text mining is highly scattered and lacks specificity. Many models have been proposed by the researchers but they are specific for specific stock market data (Table 1 & 2). Therefore, a generalized model or approach needs to be developed. Even though many hybrid models have been proposed in the literature that uses a combination of various data mining algorithms and methods to effectively solve the problem of stock market forecasting with high accuracy but they are not universal rather very specific and have been compared only with few existing forecasting tools [14, 5]. Furthermore, after reviewing various authors for different models it is clear that the in maximum cases accuracy ranged from 50 to 70%, thus 50% accuracy has been seen to be a threshold above which results are considered to be acceptably good and report worthy by various authors[49, 3, 22, 12]. Therefore, a continuous dilemma between these models arise when it comes to using it for actual predictions in real world applications as data sets in stock market are very noisy and massive and thus seeks accuracy and precision of prediction.

Apart from the accuracy and specificity, technical indicators or market indices that play a major role in stock market forecasting using data mining also has no specificity in the literature many technical indicators are being used variably by different authors (as mentioned in Table 2) and hence identification of suitable and most influential set of technical indicators needs effort. Such non specificity was also observed in case of performance metrics used by the authors in evaluation of performance and robustness of the models. The literature indicates that there is no specificity in the evaluation matrices as they were selected as per the problem being addressed by the data mining model. Thus, there are no generalized performance matrices or combination of matrices identified that could be definitely used in the data mining process with respect to stock market prediction. Along with this, almost every author has mentioned the importance of pre-processing of data (Table 3) either textual or numeric as data from stock market or related textual data are incomplete, noisy, and massive containing outliers which are to be reduced or removed for better accuracy. Many authors have made efforts to transform data in one scale to another as a preprocessing method but they do not meet the efficiency level required in real-world scenario. Thus, more specific research for selection of accurate pre-processing and feature selection techniques will help in enhancing the proposed prediction models.

Furthermore, many earlier authors have mentioned the importance of examining imbalance of the experimental data while dealing with textual data [10, 39]as it is one of the most important aspect in selection of any model for prediction using data sets from real time. But a peculiar trend has been noticed that very few authors [28, 26, 36]have reported about the status of imbalance of experimental data. Furthermore, it is worth noting that imbalanced dataset with imbalanced classes has to be paid extra attention in terms of devising a suitable feature selection with high dimensionality. Yin et al. [47]have suggested feature-selection as a good way to deal with such data imbalances. In the same notion, Liu et al. [23]used “probability based term weighting scheme” for the same purpose.

Thus, it is vivid from the drawn inferences that earlier literature has a scattered view of various evaluation mechanisms and input data types accompanied by few pre-processing models which causes difficulty in reaching to a concrete level of effectiveness while using text mining along with stock market data for market prediction in real time.

## Future Recommendations

Stock market prediction using text mining along with stock market data is a new and unique field of research to be investigated rigorously making use of the advancements in computational science and digital networking excellence that has tremendously influenced the stock market in last few years and this trend will continue to grow. But text mining for stock market prediction must not be considered as a computer assisted predictive science rather it has wider background that has to be essentially targeted for development of accurate predictive tools. But unfortunately only few researchers have targeted the interdisciplinary prospective which is clear from above literature that it is more focused towards computational modeling and artificial intelligence using textual information. Therefore, to address these issues many authors in past have recommended the background analysis of at least three diverse genera of study namely machine-learning that enables computational modeling and pattern recognition; linguistics which enhances the understanding of the nature of language as well as behavioral-economics that helps in enhancement of the required economic knowledge[31, 40, 44, 1].Such interdisciplinary research is recommended because it helps in understanding the role of human reactions and behavior to various national and global events which ultimately alter market trends. Thus, such interdisciplinary research leads to a better understanding of market performance and will increases its predictability.

## Conclusion

As it is being observed in past that the impact of global financial crisis has been detrimental to the livelihood of almost every person in the world but such impacts could be lowered by having sophisticated, in-depth and extremely pinpoint view of the financial markets. Thus, research in the field of market-prediction using various data that directly is linked with the turmoil caused in the stock market is gaining attention. In the same notion, text-mining using inputs from internet sources is emerging as a highly accurate and comprehensive tool to predict market-movements as it is based on human behavior and its changing psychology at a macro level as a reaction to various contemporary and current events happening in the word.

## REFERENCES

1. Bikas, E., Jurevicˇiene˙ , D., Dubinskas, P., &Novickyte˙ , L. (2013). Behaviouralfinance.The emergence and development trends. *Procedia – Social and BehavioralSciences*, 82, 870–876.
2. Bollen, J., &Huina, M. (2011). Twitter mood as a stock market predictor. *Computer*, 44, 91–94.
3. Butler, M., &Kešelj, V. (2009). Financial forecasting using character n-gram analysis and readability scores of annual reports. In Y. Gao& N. Japkowicz (Eds.),Advances in artificial intelligence. *Berlin Heidelberg: Springer,* pp. 39–51.
4. Chatrath, A., Miao, H., Ramchander, S., &Villupuram, S. (2014). Currency jumps, cojumps and the role of macro news. *Journal of International Money and Finance*, 40, 42–62.
5. Chung, H., Shin, K. S. (2018). Genetic algorithm-optimized long short-term memory network for stock market prediction. *Sustainability*, 10(10):1–18
6. D. Lien Minh, A. Sadeghi-Niaraki, H. D. Huy, K. Min and H. Moon. (2008). Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network. *IEEE Access*, vol. 6, pp. 55392-55404. doi: 10.1109/ACCESS.2018.2868970.
7. Dang, M., Duong, D. (2016). Improvement methods for stock market prediction using financial news articles. In: *IEEE 3rd national foundation for science and technology development conference on information and computer science (NICS),* pp 125–129

8. Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53, 1375–1388.
9. Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., &Vapnik, V. (1997). Support vector regression machines. *In Advances in Neural Information Processing Systems* (pp. 155–161). MIT Press.
10. Duman, E., Ekinci, Y., &Tanrıverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications*, 39, 48–53.
11. Fasanghari, M., &Montazer, G. A. (2010). Design and implementation of fuzzy expert system for Tehran stock exchange portfolio recommendation. *Expert Systems with Applications*, 37, 6138–6147.
12. Garcke, J., Gerstner, T., &Griebel, M. (2013). Intraday foreign exchange rate forecasting using sparse grids. In J. Garcke& M. Griebel (Eds.), Sparse grids and applications. *Berlin Heidelberg: Springer*,pp. 81–105.
13. Groth, S. S., &Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 50, 680–691.
14. Hadavandi, E., Shavandi, H., Ghanbari, A. (2010). Integration of genetic fuzzy systems and artifcial neural networks for stock price forecasting. *Knowl Based Syst* 23:800–808
15. Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features.*Decision Support Systems*, 55, 685–697.
16. Hegazy, Osman &Soliman, Omar, S. & Abdul Salam, Mustafa. (2013). A Machine Learning Model for Stock Market Prediction. International *Journal of Computer Science and Telecommunications*. 4. 17-23.
17. Huang, C.-J., Liao, J.-J., Yang, D.-X., Chang, T.-Y., &Luo, Y.-C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. *Expert Systems with Applications*, 37, 6409–6413.
18. Huang, S.-C., Chuang, P.-J., Wu, C.-F., & Lai, H.-J. (2010). Chaos-based support vector regressions for exchange rate forecasting. *Expert Systems with Applications*, 37, 8590–8598.
19. Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., &Ramakrishnan, N. (2013). Forexforeteller: Currency trend modeling using news articles. *In Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1470–1473). Chicago, Illinois, USA: ACM.
20. KhadjehNassirtoussi, A., Ying Wah, T., & Ngo Chek Ling, D. (2011). A novel FOREX prediction methodology based on fundamental data. *African Journal of Business Management*, 5, 8322–8330.
21. Khan, W., Malik, U., Ghazanfar, M. A., Azam, M. A, Alyoubi, K. H., Alfakeeh, A. S. (2019). Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. *SoftComput.*https://doi.org/10.1007/s00500-019-04347-y
22. Li, F. (2010). The information content of forward-looking statements in corporate filings—a naive Bayesian machine learning approach. *Journal of Accounting Research*, 48, 1049–1102.
23. Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, 36, 690–701.

24. Lugmayr, A., &Gossen, G. (2012). Evaluation of methods and techniques for language based sentiment analysis for DAX 30 stock exchange – a first concept of a ‘‘LUGO’’ sentiment indicator. In A. Lugmayr, T. Risse, B. Stockleben, J. Kaario, B. Pogorelc& E. SerralAsensio (Eds.), *SAME 2012–5th international workshop on semantic ambient media experience*.
25. Mahajan, A., Dey, L., &Haque, S. M. (2008). Mining financial news for major events and their impacts on the market. *In IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, 2008. WI-IAT ‘08 (Vol. 1, pp. 423–426).
26. Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. *In Proceedings of the 37th annual Hawaii international conference on system sciences*, 2004 (pp. 10).
27. Nikfarjam, A., Emadzadeh, E., &Muthaiyah, S. (2010). Text mining approaches for stock market prediction. *In The 2nd international conference on computer and automation engineering (ICCAE),* 2010 (Vol. 4, pp. 256–260).
28. Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, 24, 131–139.
29. Pui Cheong Fung, G., Xu Yu, J., &Wai, L. (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).
30. 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, 2007. *CIDM 2007* (pp. 720–725).
31. Robertson, C., Geva, S., & Wolff, R. (2006). What types of events provide the strongest evidence that the stock market is affected by company specific news? *Proceedings of the fifth Australasian conference on data mining and analystics*, Vol. 61, pp. 145–153. Sydney, Australia: Australian Computer Society Inc.
32. Salzberg, S. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery, 1, 317–328.
33. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions of Information Systems*, 27, 1–19.
34. Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems*.
35. Seo, M., Lee, S., Kim, G. (2019). Forecasting the volatility of stock market index using the hybrid models with google domestic trends. *Fluct Noise Lett* 18(01):1950006, 1–17
36. Soni, A., van Eck, N. J., &Kaymak, U. (2007). Prediction of stock price movements based on concept map information. *In IEEE symposium on computational intelligence in multicriteria decision making* (pp. 205–211).
37. Sun, A., Lachanski, M., &Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, 48, 272–281. doi:10.1016/j.irfa.2016.10.009
38. Tetlock, P. C., Saar-Tsechansky, M., &Macskassy, S. (2008). More than words: Quantifying language to measure firms’ fundamentals. *The Journal of Finance*, 63, 1437–1467.

39. Thammasiri, D., Delen, D., Meesad, P., &Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41, 321–330.
40. Tomer, J. F. (2007). What is behavioral economics? *The Journal of Socio-Economics*, 36, 463–479.
41. Uysal, A. K., &Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, 50, 104–112.
42. Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012). An experiment in integrating sentiment features for tech stock prediction in twitter. *In Proceedings of the workshop on information extraction and entity analytics on social media data* (pp. 23–38). Mumbai, India: The COLING 2012 Organizing Committee.
43. Werner, A., &Myrray, Z. F. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 10, 1259–1294.
44. Wisniewski, T. P., &Lambe, B. (2013). The role of media in the credit crunch: Thecase of the banking sector. *Journal of Economic Behavior and Organization*, 85,163–175.
45. 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*, 1998 (Vol. 3, pp. 2720–2725, Vol.2723).
46. Xiong T, Bao Y, Hu Z, Chiong R (2015) Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. *InfSci* 305:77–92
47. Yin, L., Ge, Y., Xiao, K., Wang, X., &Quan, X. (2013). Feature selection for highdimensional imbalanced data. *Neurocomputing*, 105, 3–11.
48. Yu, Y., Duan, W., & Cao, Q. (2013c). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*.
49. Zhai, Y., Hsu, A., &Halgamuge, S. K. (2007). Combining news and technical indicators in daily stock price trends prediction. *In Proceedings of the 4th international symposium on neural networks: advances in neural networks*, Part III (pp. 1087–1096). Nanjing, China: Springer-Verlag.