DOI 10.29042/2018-3705-3709

Helix Vol. 8(5): 3705- 3709

A Survey on Machine Learning Approach for Stock Market Prediction

[*1]Nischal Puri, [2]Avinash Agarwal, [3]Prakash Prasad
[1]Research Scholar, Priyadarshini Institute of Engg. & Technology, Nagpur
[2]Associate Professor, Ramdeobaba College of Engineering & Management, Nagpur
[3]Professor, Priyadarshini Institute of Engineering & Technology, Nagpur
**Email: nischalspuri@gmail.com*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
Stock market domain is a promising domain in Machine learning approach. Internet technologies helps to gathered various kinds of structured and unstructured data, such as blogs, message threads, an enormous amount of technical data, company meetings, quarterly results of company etc has been accumulated. Currently, many applications have used data mining techniques or machine learning to exploit such data. This study is a comprehensive overview on Machine Learning. Unlike other domains Stock market data has a correlation among different features, and its characteristics change variously with time, political and natural changes. In this paper, we discuss a various methodology for predicting power to manipulate stock prices. However, there is no structured defined frame for solving the problem in stock market domain. We believe the existing ambiguity on the topic is due to its interdisciplinary properties that require both factors affecting the economics as well as artificial intelligence. We review the work related to prediction of stock market based on text-mining and gives an outer picture of the generalized modules. Our comparative analysis expands the theoretical and technical aspects behind each.

Keywords: Machine Learning, Sentiment Analysis, Stock Market Prediction

Introduction

Traders & Investors typically uses tools of two classes of to find stocks to buy-sell; technical and fundamental analysis, both helps to analyze and predict the change in demand-supply (Turner, 2007). This change in demand-supply forms a basis for most forecasting. If buyers are more than seller for a stock, the theory states that the price will increase, and vice versa. The capability to visualize these changes in demand-supply thus gives the ability to trader to book a profitable entry and exit, which is the aim of analysis. Analysis of fundamental requires the study of company basics such as balance sheet, expenses and revenues, annual return, market position etc. While Technical analysis is deals with volume, price data, mainly volume spikes and price patterns (Turner, 2007; Murphy, 1999). TA summaries three areas on which help:

1. Market exploit discounts everything.
2. Prices move as per trends.
3. History repeats itself.

Algorithmic Trading

It is also refer as automated trading by computer program helps to forecasting mechanisms by smart trading agents that are involved in market trades. The decision making has enlarged recently and shaped the frequency trading. Such frequency trading has been popular in stock market and (Evans, Pappas, and Xhafa, 2013) are used Machine Learning, artificial neural networks and genetic algorithms to construct an algorithmic or automated trading.

Sentiment Analysis

It handles by detecting the emotional sentiment in text using specific semantic analysis for a various purpose consider on the way to measure the quality of response in market for a new product and the feedback of customer or to check the product's popularity (Mostafa, 2013 ; Ghiassi, Skinner, & Zimbra, 2013) among people. Thus the research that is focused on study of sentiment are called as ''opinion mining'' (Cambria, Schuller, Yunqing, & Havasi, 2013; Goot, Pouliquen, & Kabadjov, 2009; Balahur, Steinberger, Hsinchun & Zimbra, 2010). It identifies negative and positive terms and handling text by categorizing its emotional stand as negative or positive.

The Generic Overview

At Start text and some market values is fed as input. In next sections discussion are done on mentioned modules, their responsibility and theoretical base.

Input Dataset

Every system consider at least two data source as input, the market data available on stock exchange and the textual data using online resources.

Textual Data

The textual data inputs have various sources and have several types of content. The major sources are bulletin websites and financial websites like Reuters

(Pui Cheong Fung, Xu Yu, & Wai, 2003), The Wall Street Journal (Werner & Myrray, 2004), Financial Times (Wuthrich et al., 1998), Bloomberg (Chatrath et al., 2014; Jin et al., 2013), Dow Jones, Yahoo! Finance (Schumaker et al., 2012) as well as Forbes (Rachlin, Last, Alberg, & Kandel, 2007). The type of the news is either special financial news or general news. The system uses news related to financial domain as it contains noise less compared to general news. We also observe the difference about class of information about companies whereby disclosures and regular reports which have pre scheduled times.

Market Data

The additional source of data required for the systems originates from the values in financial markets in form of indexes or price-points. This values will be used to train the learning algorithms and used for prediction purposes also it forms an input to algorithm of machine learning as feature denoted as independent variable, Earlier study has been focused on prediction of stock market, either in form index of a stock market like the Indian Sensex Index (Mahajan, Dey, & Haque, 2008), Dow Jones Industrial Average (Werner & Myrray, 2004; Bollen & Huina, 2011; Wuthrich et al., 1998), S&P 500 (Schumaker & Chen, 2009), the US NASDAQ Index (Rachlin et al., 2007), Morgan Stanley High-Tech Index (MSH) (Das & Chen, 2007), or the price of stock specific company like Google, Apple, Amazon and Microsoft. The FOREX market has been discussed in many of the works reviewed; more freshly in the works of (Jin et al. 2013; Chatrath et al. 2014).

Pre-Processing

The textual data from the unstructured text should be transformed into a representable format called as structured data and can be treated by the machine. The pre-processing phase plays important role in data mining on the overall outcomes. In aspects of pre-processing there are at least three sub-processes in the reviewed works, namely: selection of feature, dimensionality-reduction and representation of feature.

Feature-Selection

Feature selection is the process of selecting terms appearing in the text for training the set and using only this part as feature in text classification. There are different methods available to study financial articles. Common techniques are to apply a representation of vector in which terms in article are assigned weight and indexed. Using Bag of Words approach, from the article semantically unfilled stop-words list are identified and removed (e.g.; a, the and for). Another way for selecting the feature is to use a subset of terms (J. J. Murphy. 1999), which handles issues regarding scaling of article while still encloses the key text concepts (J. J. Murphy. 1999). Another way is Noun Phrasing. It is done in syntax where identification of parts of speech (i.e., nouns) are done using lexicon and combined on the parts of speech, forming noun phrases using syntactic rules.

Pattern information & Lexicons are used to allot features for machine learning methods or it can be combined in a rules-based approach.

Feature-Representation

Subsequently the determination of features, numeric value is represented to each feature in order to process by machine learning rules. Henceforth, termed as ‘‘feature-representation’’. This numeric value acts like a weight or a score.

The presence or absence of a feature using basic technique is a binary or a Boolean representation where used like 1 and 0 for value representation e.g. a word in the case of a bag-of-words technique as in these works (Schumaker et al.,2012; Mahajan et al., 2008; Wuthrich et al., 1998).

Dimensionality-Reduction

A restricted amount of features is very important because the rise in features can make the clustering or classification problem by decreasing the efficiency of learning algorithms, this situation is known as the curse of dimensionality (Pestov, 2013). Particular dictionaries are specially put together by experts of market like the one used by occurrences (Schumaker & Chen, 2009; Butler & Kešelj, 2009).

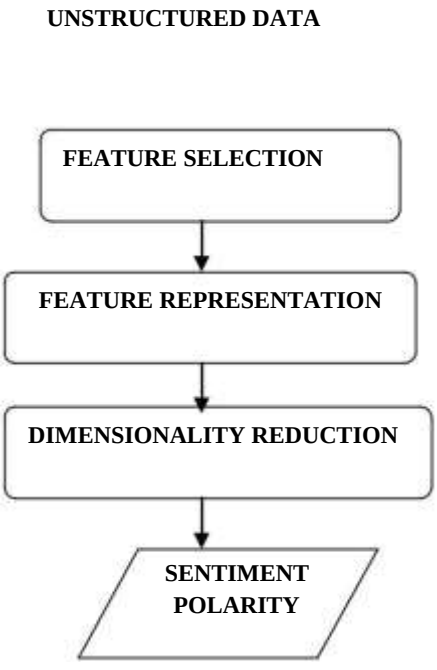

Fig 1 Pre-Processing of Textual Data

S No	Feature Selection	Dimensionality Reduction	Feature Representation	Reference
1	Bag-of-words	Message boards Pre-defined dictionaries	Binary	26
2	Visualization	Thesaurus made by term extraction Tool	Visual coordinates	22
3	OpinionFinder overall tone & polarity	Minimum occurrence per document	Binary	4
4	Structured data	Structured data	Structured data	9
5	Bag-of-words for negative words	Pre-defined dictionary. Havard-IV-4 Psycho-social dictionary	Frequency divided by total number of words	23

Table1: Comparative study of preprocessing of textual data

Machine Learning

This section contains, work to evaluation of the machine learning techniques. It is considered that relating such algorithms is complex and full of strain. Basically the data which is inputted in structure so that structure should learn to categorize an output in terms of market movement such as Up, Down and Steady. Regression Algorithms, Support Vector Machine (SVM), Multi-algorithm experiments & Combinatory Algorithms. Some of the algorithms we have considered are as follows:

Support Vector Machine (SVM)

Another frequently used implementation of SVM is LIBSVM. LIBSVM implements a Sequential Minimal Optimization algorithm recommended in a paper by (Soni et al. 2007; Fan, Chen, and Lin 2005). Using this implementation movement of stock price & prediction is done. The input to the SVM is rather one of a kind in the work of (Soni et al. 2007).

Regression Algorithms

The regression analysis is used to discover the parametric values for a function that helps the function to suits a set of data. The next equation states these relations in symbols. It suggest that regression is the process of measuring the value of a continuous target (y) as a function (F) of one or more forecasters (x1 , x2 , ..., xn), a set of parameters (θ1 , θ2 , ..., θn), & a measure of error (e).

$$y = F(x,\theta) + e$$

The procedure to train a regression model requires outcome the best parametric values for the function that reduce a measure of the error, Eg the sum of squared errors.

Combinatory Algorithms

Here it discusses a class of algorithms which are consists of a number of machine learning procedures combined together. (Das and Chen, 2007) have joined numerous classification algorithms organized by a voting system to find out sentiment of investor. The algorithms are viz., Bayesian Classifier, Discriminant-Based Classifier, Naive Classifier, Vector Distance Classifier, and Adjective-Adverb Phrase Classifier. Accuracy levels are similar to broadly used Bayes classifiers, but false positives are lowered and sentiment accuracy developed. (Mahajan et al. 2008) recognize and categorize a main event that affects the market using a topic extraction mechanism used by Latent Dirichlet Allocation (LDA). Then a trainable classifier which can be used as a stacked classifier is used which is a conglomerate the predictions of various classifiers through a general voting procedure. The voting step is a distinct classification problem. For handling numerical attributes in combination with SVM using sigmoid kernel to project the stacked classifier they practice a decision tree. The average accuracy is 60% of such classification system. A model with Self-Organizing Fuzzy Neural network (SOFNN) deployed by (Bollen and Huina , 2011) to experiment the theory using mood measurements of public can enhance the precision of prediction models of Dow Jones Industrial Average (DJIA) models. A fuzzy neural network which is a learning machine finds the fuzzy system's parameters (i.e. fuzzy sets, fuzzy rules) using approximation techniques. Hence it is considered as a combinatory algorithm

Multi-Algorithm Experiments

(Wuthrich et al, 1998) is one of the initial works of research in this consideration. Although, they carry out their evaluation using various algorithms and measure the results. (Werner and Myrray, 2004) also carry out numerous tests using SVM and Naïve Bayes. Besides, (Groth and Muntermann,2011) engage Naïve Bayes, SVM and Artificial Neural Networks (ANN), k-Nearest Neighbor (k- NN) , in order to find patterns in text data that could explain risk coverage in stock markets. SVM has successfully used in sentiment learning and textual classification , other approaches like Artificial Neural Networks (ANN), k- Nearest

Neighbors (k-NN) , have rarely been reviewed in the context of text mining for market prediction.

Data Training, Testing & Sampling

The unstructured & structured Data were used for the volume used to test & train the system. Around 70 or 80 % of data to train against 30 or 20 % of data to test. The aim of the studied systems is to forecast the movement in a window of future (prediction-window) of the market based on the knowledge in a (training-window) For example, patterns may identify by the a system based on the data (training-window) in order to forecast the movement of stock market on a new day (prediction-window). The timeline of training a window may have two possibilities: sliding or fixed.

Hence, supplementary format helps to solve the problem through which the complete training-window or one side of dynamically sliding to the point where the prediction-window starts. Though, it naturally seems important to apply a sliding window, there are few of the works which really have. This appears to be an aspect that can obtain more consideration in future systems.

Scope for Future Work

The tools of Market forecast based on online mining of text are to be studied thoroughly using the topmost processing power and speed of network in the latest years. This helps to put the role of human responses to the actions in making of stock markets and will help to recognize the market efficiencies. This work recognizes below aspects for future advancement:

Semantics and Syntax

In upcoming years it will require to develop more ontology for particular contexts like product reviews, stock market etc. Semantics can be included in designing of feature-weighting schemes; (Luo, Chen, and Xiong, 2011) suggest a term weighting scheme by operating the semantics for classifying the financial Market and indexing terms.

Syntax deals with their comparative positioning or grouping and ordering. The works are done in the syntactical area was very less. It was categorized into four varieties, namely: verbs, adverbs, nouns, and adjectives. It requires profound syntactic study of phrase level sentiment learning.

Module for Text-mining, textual-source or specialization of application market:

There is enormous potential in imminent aspects for text mining. In the process of mining the text modules like feature-selection, feature-representation and feature-reduction each module to be studied individually for market-prediction.

Machine Learning Algorithms

It has been discussed comprehensively how Naïve Bayes and SVM are preferred, while other machine learning algorithms like K-Nearest Neighbours (k-NN), Artificial Neural Networks (ANN), fuzzy-logic, etc. shows capabilities for text classification and sentiment analysis in works but haven't experimented in context of market data are considerably under-researched.

Experimental Datasets its Availability and Quality

One of the many problems concerning the non-availability of uniform datasets that maps text data onto market data for certain time period that benefit researchers for accepting their experimentation and evaluation efforts. Future scholars are interested to normalize and release datasets for text-mining research in context with stock market-prediction.

Conclusion

The major structures in mining of online text have been studied and some of the key lacunas have been recognized. The assessment was ended on three key aspects, precisely: pre-processing of records, machine learning and the machinery to assess; which breakdowns into many areas. It is a big task to make available a comprehensive review from numerous aspects of data analysis.

Here the papers, we studied the current developments in market forecast models and it has the capability to forecast the movement of market more accurately compare to other techniques. The Machine Learning supports to learn relationships among the data which supports the systems such as stock markets more accurately.

References

1. J. J. Murphy 1999 Technical analysis of the financial markets: a comprehensive guide to trading methods and applications, volume 2. Prentice Hall Press.
2. T. Turner 2007 A Beginner's Guide to Day Trading Online. Adams Media, 2nd edition,
3. Evans , Pappas, Xhafa 2013 Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation Mathematical and Computer Modelling : 1249–1266
4. Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B., & Kabadjov, M. 2009. Opinion mining on newspaper quotations. In Proceedings of the IEEE/WIC/ACM international joint conference on web intelligence and intelligent agent technology. IEEE Computer Society. 523–526
5. Bollen, J., & Huina, M. 2011. Twitter mood as a stock market predictor. Computer, 44: 91–94.
6. 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:39-51

7. Cambria, E., Schuller, B., Yunqing, X., & Havasi, C. 2013. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28: 15–21.

8. 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.

9. Das, S. R., & Chen, M. Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53:1375– 1388.

10. Evans, C., Pappas, K., & Xhafa, F. 2013. Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation. Mathematical and Computer Modelling, 58:1249–1266.

11. Ghiassi, M., Skinner, J., & Zimbra, D. 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40:6266–6282.

12. Groth, S. S., & Muntermann, J. 2011. An intraday market risk management approach based on textual analysis. Decision Support Systems, 50:680–691.

13. Hsinchun, C., & Zimbra, D. 2010. AI and opinion mining. IEEE Intelligent Systems, 25:74–80.

14. 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, WI-IAT '08 Vol. 1: 423–426

15. Majumder, D. 2013. Towards an efficient stock market: Empirical evidence from the Indian market. Journal of Policy Modeling, 35: 572–587.

16. Mostafa, M. M. 2013. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications, 40:4241–4251.

17. Pui Cheong Fung, G., Xu Yu, J., & Wai, L. 2003. Stock prediction: Integrating text mining approach using real-time news. IEEE international conference on computational intelligence for financial engineering, Proceedings: 395–402.

18. Qiming Luo, Enhong Chen, Hui Xiong 2011; A semantic term weighting scheme for text categorization. Expert Systems with Applications, 38 :12708–12716

19. 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, CIDM 2007:720–725.

20. 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.

21. Schumaker, R. P., Zhang, Y., Huang, C.-N., & Chen, H. 2012. Evaluating sentiment in financial news articles. Decision Support Systems.

22. 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: 205–211.

23. Tetlock, P. C. 2007. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62: 1139– 1168.

24. Tomer, J. F. 2007. What is behavioral economics? The Journal of Socio-Economics, 36:463–479.

25. 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.

26. 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).

