Chapter 10
Big Data Analytics Techniques in Capital Market Use Cases

10.1 Introduction

Having surveyed the applications of Big Data Analytics in Banking and Financial Services sector in the last chapter, we shall now provide an overview of the possible applications of Big Data Analytics in the Capital Market Use cases.

Application of Big Data Analytics Techniques to Capital Market is identified [1] to be possible in the following functional areas:

(a) Financial data management and reference data management comprising

  (i) historical trading, internal data management challenge and also
  (ii) overall reference data mining to find metadata to deconstruct/reconstruct data models.

(b) Regulation application focusing on fraud mitigation
(c) Risk analytics comprising

  (i) anti-money laundering (AML),
  (ii) know your customer (KYC),
  (iii) rogue trading and
  (iv) on-demand enterprise risk management.

(d) Trading analytics comprising:

  (i) analytics for high-frequency trading and
  (ii) predictive analytics.

(e) Pre-trade decision support analytics including:

  (i) sentiment measurement
  (ii) temporal/bitemporal analytics.

C. S. R. Prabhu et al., *Big Data Analytics: Systems, Algorithms, Applications*,
https://doi.org/10.1007/978-981-15-0094-7_10

(f) Data tagging

In enterprise-level monitoring and reporting, it is hard and difficult to match and reconcile trades from various systems built to different symbology standards, usually resulting in invalid, duplicated and mixed-up trades. Data tagging can easily identify trades and events such as corporate actions and help regulators to detect stress signs easily.

10.2 Capital Market Use Cases of Big Data Technologies [2, 3]

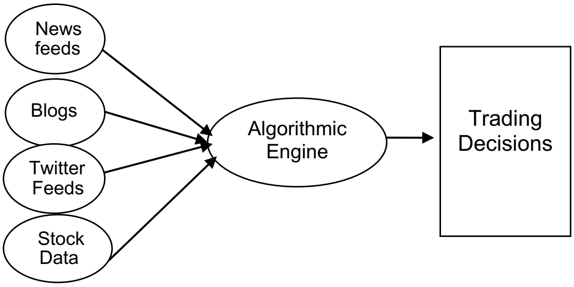
Previously, high-velocity market data was the focus of analytics. But due to changing trading dynamics, unstructured data is now becoming very important. Unstructured data comprises of daily stock feeds, Twitter chats, blogs, etc.

10.2.1 Algorithmic Trading [4–7]

For gaining an edge in Wall Street trading, it is essential to convert unstructured information into machine-readable data. This data is then used by algorithmic traders to produce some alpha (return from market) from the news. These untapped sources can help traders to gain a competitive edge in trading.

As against the past practices of the traders depending on slow incoming news of company specific and industry specific events having occurred over a period of time, today, in the online world, it is the instantaneous and online feedback inputs coming from online news, blogs, Twitter feeds and online stock data that help the trader in making decisions based on Algorithmic Engine. Such Source and Response system to handle responsiveness of important events in the external world influencing trading decisions is shown in Fig. 10.1.

Fig. 10.1 Source and response system in trading



10.2.2 Investors' Faster Access to Securities [3–5, 7]

With today's instantaneous access to the Internet-based stock market information to the investors and the general public as well, information pouring in from quarterly reports or breaking news stories can dramatically affect the share price of a security. As the primary source of event information is provided on the Internet, the impact of such information on trading in stock markets is significant. Therefore, the methodology and technology involved in extracting and using information to support decision-making process become critical.

10.3 Prediction Algorithms

In order to predict the stock market, we have different algorithms and models from the academic community and also industry. We shall now examine and survey the developments in the prediction algorithms and models along with their comparative performance evaluation. For greater accuracy of prediction, we need to examine the impact of various global events on the stock market and the issues they raise.

10.3.1 Stock Market Prediction [3–5, 7]

Due to the huge financial gains at stake, both investors and traders have a great interest in the prediction of stock market which has been identified as an important subject in Finance sector, as also in Engineering, Computer Science and even Mathematics.

As a very huge amount of capital is traded globally every minute, the stock market is an outlet for maximum investments. Therefore, researchers have been striving hard to predict the financial market and hence, stock market prediction generates a great amount of appeal to the researches all over the globe. Inspite of all these research efforts, till today, no specific method has been developed to correctly and accurately predict the movement of the stock prices in the stock markets. Even without a full-fledged and successfully consistent prediction method, some limited successes were noted till now in the prediction process. Auto-regressive and moving average techniques are some of the famous prediction techniques which have dominated the time series prediction techniques for many decades. Inductive learning techniques have also been developed and deployed. K-nearest neighbor (KNN) algorithm and neural network techniques have been applied for prediction. However, their weakness is that they depend heavily on the structural data only while totally neglecting the impact and influence of non-quantifiable information such as news articles.

10.3.2 Efficient Market Hypothesis (EMH)

In this theory, it is stated that financial markets are already and always 'informational efficient'. This means that the current prices at a given instance of time are such that they readily (already) reflect all the known information and automatically, instantaneously change, so as to reflect, at any time, the latest and new information. Therefore, logically, accordingly to this theory, it is impossible to consistently outperform the market by depending on and using the same information which the market already is having with it (except through 'luck', if at all). 'Information' is anything which may affect the prices. There has been a lot of debate on whether financial market is predictable at all or not. Information categories have been proposed as three options: (1) weak (2) semi-strong (3) strong. In weak 'information', only historical information is incorporated and embedded into the current files. In 'semi-strong' case, both historical and current public information are incorporated and embedded into the current files. In 'strong' case the historical information, the current public information and also the current private information such as financial information are incorporated and embedded into the current share prices. The basic tenet of EMH theory is that it is impossible to outperform the market since the market reacts automatically and instantaneously to any given developments such as news or other sudden developments.

10.3.3 Random Walk Theory (RWT)

Random Walk Theory (RWT) maintains that it is impossible to outperform the market and that the prices are determined randomly, even though the historical data and even the current public information has its own impact (same as semi-strong EMH).

10.3.4 Trading Philosophies

Based on the above two theories (EMH and RWT), we have two philosophies: (a) fundamental trading philosophy and (b) technical analysis trading philosophy.

(a) **Fundamental Trading Philosophy**

This philosophy states that the stock price is determined (indirectly) by the vital economic parameters such as indices of inflation, unemployment, Return on Equity (RoE), debt level and individual Price to Earnings Ratio and also especially from the financial performance of the Company itself.

(b) **Technical Trading Philosophy**

This philosophy states that stock price is dependent on historical time series data. This school of philosophy believes that the time of investing in market is the most crucial factor, investment opportunities can be identified by carefully investigating the average price (historically) and value movement in comparison with the current prices. It also believes that the psychological factors of perception (high or low) of price barrier such as support level and the resistance level may also indicate where the opportunities may open up.

10.3.5 Simulation Techniques

Both the Fundamentalists and Technicians have developed certain methodologies and techniques for predicting the price from financial trade articles. Therefore, they adopted simulation techniques where the stock market is analyzed based on simulated (not real) stock markets with simulated (not real) traders by mimicking the real human traders. Being artificial, it is practically possible to perform dissections for identifying key parameters of information. The traders in the simulated situations are programmed to follow a rule hierarchy while trading in response to market changes, especially those based on news articles are updates. The response time of such simulated traders was made to vary, based upon the time elapsed between the point of time of receipt of information and the point of time of reaction. The results were found to be astounding when it was noted in observation that the length of reaction time is dependent on a particular trading philosophy—the traders who showed quick response formed technical strategies while those who waited long formed fundamental strategies. Therefore, we can surmise that technicians may have capitalized on the time lag by acting instantly, before all other traders. This research is able to demonstrate that there exists a week's ability to forecast only for a brief period of time.

10.4 Research Experiments to Determine Threshold Time for Determining Predictability

In a research experiment by Gidafalvi, about 5000 news articles pertaining to 12 stocks were analyzed and it was concluded that in the time interval of 20 minutes before and 20 minutes after some 13 news articles were released, there was a week possibility of predictability of the direction of market movement.

The weakness in predictability is due to the fact that the news articles concerned got repeatedly reprinted across all the news agencies and wire services.

Stronger predictability exists if the first release of the article is isolated, and by using automatic text analysis techniques, it makes possible to capitalize 20 minutes before the human traders start acting.

10.5 Experimental Analysis Using Bag of Words and Support Vector Machine (SVM) Application to News Articles

In an experiment in 2005, Schumaker picked up a large number of news articles (9211) and a very large number of stock quotes (10,259,042) from S&P 500, over a five-week period. Then, the analysis of news articles was performed and the terms which appeared more than three times in the articles were retained. Bag of Words was created with about 4000 terms from 2500 articles with about 5000 noun phrase terms and 2800 named entities, 2800 terms from 2600 articles. The above, when processed by support vector machine (SVM) algorithm derivative, using regression, three metrics M1, M2, M3 were defined for 'closeness' and 'derived accuracy'. M1, M2, M3 are the three models used. M1 uses only extracted articles terms with no baseline price; M2 uses extracted article terms and stock price when the article was released; M3 uses extracted terms at estimated 20 minutes of stock price.

SVM had to perform learning on which terms result in changes in share prices and accordingly adjust their weights according to the severity of price changes. From Closeness results and Directional Accuracy Results, it was found that model M2 gave the closest and the most accurate prediction (for +20 min stock price).

Results showed that M2 which uses the news articles and regression together performs better than pure regression. Therefore, it is essential that impact of news articles be considered for any prediction. This was the conclusion reached in this research.

10.6 Textual Representation and Analysis of News Articles

Many methods are possible to analyze the text contents in a news article. One simple way is tokenizing and using very word in the given text document. This is, however, a human-oriented technique as every word is deployed to indicate syntactic structure of the document. For machine learning algorithms, such structural markings are not required. In order to perform easy and efficient text processing, the normal approach followed is 'Bag of Words'. This is a standard approach for text processing due to its simplicity and ease of use. Over and above, certain parts of speech can be used as features. Noun phrases can be indentified through parsing and a dictionary (or lexicon) is used to identify nouns which may be aggregated, using syntax rules (as nearby words) to form noun phrases.

10.7 Named Entities

Another method called ‘Named Entity Identification’ is based on nouns and noun phrases. By using dictionary and also semantic lexical hierarchy, we can classify nouns and noun phrases into entities such as persons or organizations or locations. By generating a lexical profile across all noun phrases (after analyzing synonyms), it is possible to determine the semantic hierarchy of nouns or entities. Thus, the named entities capture far greater semantics than ordinary ‘Bag of Words’ or even just noun phrases. Even greater semantics is captured by Object Knowledge Model (OKM).

10.8 Object Knowledge Model (OKM) [8]

Objection Knowledge Model (OKM) enables much greater capture of semantics of a text article than other methods (as Bag of Words or Noun Phrases Named Entities) since it captures semantics of activities also, in addition to entities. Thus, not only Named Entities and their attributes are identified but Named Activities and their attributes are also identified in Object Knowledge Model (OKM).

10.9 Application of Machine Learning Algorithms [7]

All machine learning algorithms perform simple linear regression analysis of the old (historical) security trading data for a given time period in recent times in order to determine the price trend of a given stock. In addition, to determine the impact of textual reports or comments, a simple textual analysis technique algorithm by using ‘Bag of Words’ approach is performed in order to determine the keywords in the given text. Finally, all the above inputs are classified into the prediction of stock movement as (1) upwards (2) downwards or (3) unchanged.

Research has been performed on applying (1) genetic algorithms (for classification into two categories), (2) Naïve Bayesian (for classification into three categories) and (3) SVM (for classification into three categories) based on texted news articles or text postings in chat room chats. The outcome of such research indicated that apply to genetic algorithms, the chat room chats were analyzed and stock prices were classified by utilizing the postings and number of words posted on an article daily.

Research on SVM applications for the stock data and articles produced results that this technique is mildly profitable. An attempt was also made to produce an optimum profit trading engine.

10.10 Sources of Data

In all the above experiments and also later experiments, data was collected from secondary data sources such as Yahoo Finance Website which provides intraday 5-min interval data for that day (many primary servers may not provide intraday stock movement data). The news articles are taken from new agencies as Reuters and also from newspapers.

10.11 Summary and Future Work

To summarize, the process of integrating text analysis of news articles with Regression and other machine learning algorithms (as SVM, Naïve Bayesian and genetic algorithms) in an interval from −20 to +20 min is indicated in the following diagram (Fig. 10.2).

Future work can be aimed at improvements and a ‘multimodel regression’ and sentiment analysis on textual data can be integrated to obtain greater accuracy.

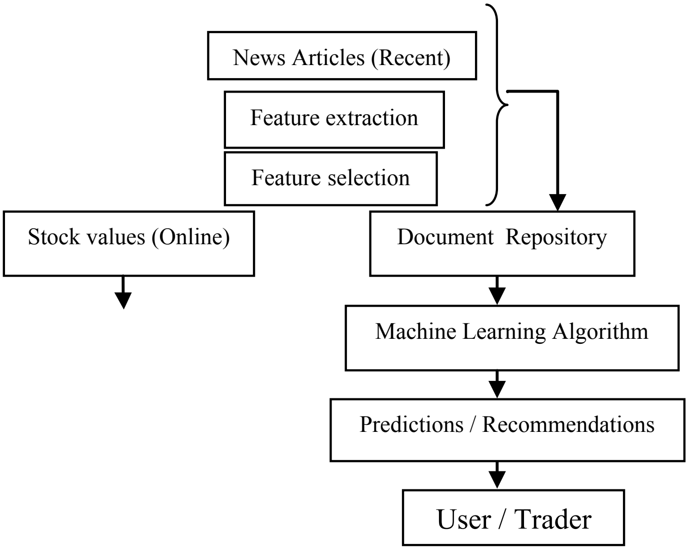


Fig. 10.2 Integrating text analysis with machine learning algorithm

10.12 Conclusion

In this chapter, we have surveyed the use cases, techniques, algorithms for predication and the research performed in determining the predictability of prices in stock markets.

10.13 Review Questions

1. List out where in Capital Market use cases the techniques of Big Data Analytics are applicable.
2. Explain Source and Response System in Trading (with diagram) in Algorithmic Trading.
3. How to provide faster investor access to securities?
4. Explain Prediction Algorithms and Stock Market Prediction.
5. Explain Efficient Market Hypothesis (EMH).
6. Explain Random Walk Theory.
7. What are the two major trading philosophies?
8. Explain simulation techniques deployed in Capital Market Analysis.
9. Explain how Experimental Analysis using Bag of Words and Support Vector Machine (SVM) are applied to new articles.
10. Explain how textual representation and analysis is performed for news articles.
11. Explain the application for machine learning algorithms in Capital Market Analysis.
12. Explain how to integrate text analysis with machine learning algorithms in news article analysis in Capital Market.

References

1. Thomson Reuters, Big data in capital markets: at the start of the journey, White Paper (2014)
2. M. Singh, Big data is capital market. Int. J. Comput. Appl. **107**(5) (2015)
3. T.H.A. Uheug, S.Y.-H. Wu, Trade data service for capital markets, Honors, School of Computer Science and Engineering UNSW (2003)
4. Trading technology survey of exchange technologies (2003). http://www.tradingtechnology.com
5. Capital Market Cooperative Research Centre (CMCRC), http://www.cmcrc.com
6. F.A. Rakhi, B. Benatallah, An integrated architecture for management of capital market system. IEEE Netw. **16**, 15–19 (2002)
7. K.-C. Li, H. Jiang, L.T. Young, Big data algorithms, analytics and applications. https://books.google.co.in/books?isbn=1482240564
8. Object knowledge model definition, Ph.D. thesis, C.S.R. Prabhu Sunrise University, 2015