



Stock Market Prediction using Text Opinion Mining: A Survey

Arti Buche*, Dr. M. B. Chandak

Department of Computer Science & Engineering, Ramdeobaba College of Engineering & Management,
Nagpur, Maharashtra, India

Abstract— *The behavioural economics has been profoundly affected by the news and stock prices which are closely related and usually influence stock market investment greatly. Text Opinion Mining is a method which holds the historic data for predicting the future directions and offers a broad way to monitor public sentiments. The historic data helps investors to discover hidden patterns to predict the capability in their investment decisions. Here the investigation consists of various techniques and strategies used to predict the ups and downs of stock market using news articles and to correlate them to the stock market fluctuations. We have surveyed and analysed in this paper techniques and key tasks of opinion mining. An overall picture of developing a software system for opinion mining on the basis of survey and analysis is proposed.*

Keywords— *Opinion Mining, stock market prediction, news articles*

I. INTRODUCTION

The stock market prediction using news or personal opinions has attracted much attention of academia and business. Text mining technology has increasingly becoming important for easy and quick availability of news information which can be accessed due to sheer volume of opinion rich web resources. In recent years, our social and political system has a great impact of public sentiments and emotions. Similarly, the postings of social media have helped reshaping businesses.

Text opinion mining is designed to help investors to predict the hidden patterns from the data available in digital form. The main challenge is the prediction of stock market values to extract meaningful characteristics from news articles and decide which stock to buy, hold and sell. The news articles generally contain both positive and negative aspects of stock market values, which make it difficult for financial analysts to predict the underlying truth. Such analysts have to deal with vast information available in newspapers, magazines and other resources in order to investigate the stock trend movement to make a prediction model. [1]

Opinion Mining and sentiment analysis are the techniques used to detect and extract subjective information from text documents. The sentiment classification is the main challenge in opinion mining which depends on mood, judgement or evaluation of any object like film, book etc in the form of a text document, sentence or features which can be labelled as positive or negative. [2]

The paper is organized into following sections: data collection used for generating historic data, sentiment analysis techniques, prediction algorithms used to predict the stock movement and the evaluation measures. Last section concludes the overall survey and discusses some future aspects of research.

II. DATA COLLECTION

There are varied purposes to exploit rich and unique data across disciplines. The major criterion for the improvement of the quality services are the user opinions which can be rendered and can be used for the enhancement of deliverables. A good understanding of products, services and business are provided by blogs, review sites and micro blogs.

A. Blogs

Blogosphere is the name associated to all blog sites. People write on blogs about the topics they want to share with publicly. Because of its ease and simplicity of creating blog posts, blog is of free form and unedited nature. Every topic of interest on blogosphere has large number of posts on blogs. Blogs are used as sources of opinion in many of the studies related to text opinion mining.

B. Review Sites

The decision makers are the opinions for any user in making a purchase. The reviews are generated by the users for products; services and business are largely available on internet. For sentiment classification the reviewers data can be collected from the websites which hosts millions consumers product opinions like Flipkart, Amazon (product reviews), Economicstimes, Moneycontrol (stock market reviews).

C. Micro-blogging

The communication tool very popular among Internet users is micro-blogging. Micro-blogging websites such as Twitter, Tumblr, and Facebook contain millions of messages appearing daily. Twitter messages sometimes express stock market related opinions which are used as data source for prediction of stock market movement. [8]

III. SENTIMENT ANALYSIS TECHNIQUES

Sentiment classification or binary classification or polarity classification is mainly classified into machine learning approach, lexicon based and hybrid approach. Sentiment classification techniques are used for analysing subjective information in a large number of texts, and many studies like stock market prediction.

D. Machine Learning Approach

Machine learning is a system capable of integrating and acquiring the knowledge automatically. The systems that can exhibit efficiency, self-improvement and effectiveness are the systems that learn from experience, analytical observation, training and other means. To test the knowledge acquired interpreted and analysed a corresponding knowledge organization are usually used by a machine learning system. The type of input available is one of the taxonomy based machine learning algorithms which also depends on outcome of the algorithm.

1) Supervised learning methods make use of large number of text documents as input. It generates labels with the help of function which maps inputs to desired outputs as they labelled by human experts and are called as training examples. Supervised learning method can be applied to any text classification problems, e.g., Naïve Bayes classification, Neural Networks, Support Vector Machines, Maximum Entropy and Bayesian Network.

2) Unsupervised learning models are used to find out the labels for the text documents such that the document is classified into predefined categories. When it becomes difficult to find these labels unsupervised learning is used. During system training it uses a set of inputs, like clustering, labels which are not known. To express opinions classification is performed using some fixed syntactic patterns. One such method to compose syntactic patterns is part-of-speech (POS) tagging.

3) Semi-supervised or weak supervision provides supervision at the level of features instead of instances. To combine both labelled and unlabelled examples it generates an appropriate function or classifier. [4]

E. Lexicon Based Approach

The opinion lexicons are used in lexicon based approach for analysing the text. The lexicon based technique is basically classified into two methods dictionary based approach and corpus based approach. The dictionary based approach mainly deals with finding opinion seed word and then performs searching in dictionary for synonyms and antonyms. The corpus based approach includes seed list of opinion words followed by finding other opinion words in turn helping in finding opinion words which are context specific in a larger corpus. Statistical and semantic methods can be used for corpus based approach. [4]

IV. STOCK PREDICTION ALGORITHMS

The analysis of stock market involves investors to predict future of market so that the investments done are successful. Automatic trader agents can be used a component or effective prediction of market can help investors for trade advices. Indirectly sometimes these prediction systems provide supportive information such as direction for future market. The stock market prediction systems can be categorized in different dimensions as follows:

F. NLP based prediction

The prediction of stock price is very complicated task. This prediction based on news articles which uses sentiment analysis one of the techniques in Text Mining. The steps to be followed in NLP based prediction are as follows:

- 1) Collection of reviews based on stock market from news articles
- 2) R.S.S. feed is the main source of news articles. The news articles mainly consist of business and market related news. The top news of stock market is retrieved by using R.S.S feed.
- 3) Sentiment Analysis is applied to the news articles
- 4) Natural Language Processing referred to as sentiment analysis is used to extract from computational linguistics and text documents to extract and identify subjective information. The sentiment analysis tasks determine the polarity of the text at document level, sentence level and feature level as positive, negative and neutral. For calculating the polarity of text POS tagger and SentiWordNet is required. [5]

G. Statistical parameter based prediction

The statistical parameter based prediction method evaluates the stocks by generating and exploring statistics market activity such as volume and past stock prices. To suggest future activity required patterns, charts and other tools are required for identification.

To perform statistical based prediction like past open values or past close values, collection of data of past few years is done on daily basis in terms of open and close which in turn compares prices as up and down by considering sentences in news articles. There are different ways and sources to get the past or historic data like web sources includes business related channels like CNBC, Times Now etc. stock related companies like NSE, BSE etc. and companies which work as mediators like Money Control, Indiabulls etc. Another way is collecting data manually from company's database engine.

In this methodology, firstly fetching of numeric points which are up and down as per news sentences is fetched. Then the occurrence of how many times values went up and down in a specified range in historic data is checked. After checking it next time, it was either positive or negative polarity was assigned. To maintain the count of number of positive and negative a counter is maintained and result is predicted by considering the greater counter value. The combination of NLP based prediction and statistical based prediction is used to generate the final result. [5]

V. EVALUATION MEASURES

The accuracy of sentiment analysis in opinion mining is evaluated by using statistical measures such as precision, recall and F1-score. The performance measures such as precision and recall are used to generate quality results by comparing them.

Recall is the ratio of positive sentiment predicted to total positive sentiments while precision is the ratio of true positive instances predicted to total instances predicted. When recall and precision are combined together it represents F1-score.

H. Analysing Sentiment

The sentiment analysis of news articles was done by domain specific dictionary instead of general dictionary which resulted in greater accuracy prediction. In this case a stock market specific dictionary was developed for the sentiments of the words appearing in the news articles.

Initially several pre-processing was done in order to choose some sentimental words. After choosing, a sentiment scoring formula was applied to finalize the selected words for the dictionary. Depending on the sentiment score i.e. if the score is near 1.0 the word is considered as more positive otherwise if the score is near 0.0 it is considered as negative. If the range is near 0.3 to 0.7 then the word is considered to have equal distribution of both positive and negative words. Finally the news sentiment score was also calculated by applying sentiment scoring formula resulting in sentiment score for the day which was converted into prediction variables to predict stock price fluctuations. [4]

I. Predicting Increasing / Decreasing Stock Prices

The determination of rise and fall of stock market prices correlation with news sentiment and does it results generated differ according to media prediction was the aim of the experiments to be performed. For this experiment, three different groups were generated namely media H, M news and H + M from the historic data.

The prediction of rise and fall of stock prices can be done by using news opinions and the sentiment analysis learned through opinion mining which was used to predict actual stock market movements effectively. The summary for characteristics of news articles was generated depending on media. Firstly the news articles were classified then among them opinions were extracted with an appropriate value and the prediction of stock as up / down was done. [4]

VI. CONCLUSION

This paper investigated on the survey of different methods that gave impact on prediction of stock market using financial news. It also discussed and presented the general procedural flow of the methods.

Many sentiment analysis techniques were discussed in order to perform text opinion mining to generate sentiment score required for the prediction of stock market movement. As a result dictionary was developed which was stock domain specific of sentimental words from the news articles which predicts stock price fluctuations. [6]

In stock prediction algorithms section NLP based prediction and statistical based predictions were discussed which when combined results in assigning polarity in three categories positive, negative, neutral and behaviour of past years historic data. The accurate and reliable prediction results were generated which gave better solution to the consumers regarding investment of their valuable money. [7]

REFERENCES

- [1] Yoosin Kim¹, Seung Ryul Jeong¹, Imran Ghani², Text Opinion Mining to Analyze News for Stock Market Prediction, Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1, March 2014 ISSN 2074-8523; Copyright © SCRG Publication, 2014
- [2] Johan Bollen¹;?,Huina Mao¹;?,Xiao-Jun Zeng²., Twitter mood predicts the stock market., arXiv:1010.3003v1 [cs.CE] 14 Oct 2010
- [3] Azadeh Nikfarjam, Ehsan Emadzadeh and Saravanan Muthaiyah, Text mining approaches for stock market prediction, 978-1-4244-5586-7/10/\$26.00 C 2010 IEEE
- [4] V. S. Jagtap and Karishma Pawar, Analysis of different approaches to Sentence-Level Sentiment Classification, International Journal of Scientific Engineering and Technology (ISSN : 2277-1581) Volume 2 Issue 3, PP : 164-170 1 April 2013
- [5] 1Ruchi Desai, 2Prof.Snehal Gandhi, Stock Market Prediction Using Data Mining, © 2014 IJEDR | Volume 2, Issue 2 | ISSN: 2321-9939
- [6] Walaa Medhat a,* , Ahmed Hassan b, Hoda Korashy b, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal (2014) 5, 1093–1113
- [7] Robert P. Schumaker and Hsinchun Chen, Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System, 12th Americas Conference on Information Systems (AMCIS)
- [8] Adarsh M J and Pushpa Ravikumar, Survey: Twitter data Analysis using Opinion Mining, International Journal of Computer Applications (0975 – 8887) Volume 128 – No.5, October 2015
- [9] Jalaj S. Modha* Prof & Head Gayatri S. Pandi Sandip J. Modha, Automatic Sentiment Analysis for Unstructured Data, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013 ISSN: 2277 128X
- [10] G.Vinodhini and RM.Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

- [11] distributed Fu, K. Lee, D. Sze, F. Chung and C. Ng, "Discovering the Correlation between Stock Time Series and Financial News," Proceedings of the 2008 IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008, pp.880-883.
- [12] M. A. Mittermayer and G. F. Knolmayer, "NewsCATS: A News Categorization And Trading System," Proceedings of the International Conference in Data Mining, 2006.
- [13] S. Ahn and S. B. Cho, "Stock Prediction Using News Text Mining and Time Series Analysis," Proceedings of the KIISE 2010 conference, Vol.37, No.1, 2010.6, pp.364-369.
- [14] V. Sehgal and C. Song, "SOPS: Stock Prediction using Web Sentiment," Seventh IEEE International Conference on Data Mining – Workshops, 2009, pp.21-26.
- [15] W. Paik, M. H. Kyoung, K. S. Min, H. R. Oh, C. Lim and M. S. Shin, "Multi-stage News Classification System for Predicting Stock Price Changes," Journal of the Korea Society for Information Management, Vol.24 No.2, 2007, pp.123-141.
- [16] Yu, Y., Y. Kim, N. Kim, S.R. Jeong, "Predicting the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," Journal of Intelligence and Information Systems, Vol.19, No.1, 2013.3, pp.92-110.
- [17] Yune, H., H. Kim, J.Y. Chang, "An Efficient Search Method of Product Review using Opinion Mining Techniques," Journal of KIISE: Computing Practices and Letters, Vol.16, No.2, 2010.2, pp.222-226.
- [18] Shou-Hsiung Cheng, —Forecasting the Change of Intraday Stock Price by Using Text Mining News of Stockl, IEEE, 2010.
- [19] Mitchell, M.L., Mulherin, J.H., 1994. The Impact of Public Information on the Stock Market. Journal of Finance, 49(3), pp.923-950.
- [20] Fung G.P.C., Yu, J.X., and Lu, H., 2005. —The Predicting Power of Textual Information on Financial Markets. I IEEE Intelligent Informatics Bulletin, 5(1), pp.1-10.
- [21] Mittermayer, M.-A. Forecasting Intraday Stock Price Trends with Text Mining Techniques. Proceedings of the 37th Hawaii International Conference on Social Systems. 2004. Hawaii.
- [22] Johan Bollen, Huina Mao, "Twitter Mood as a Stock Market Predictor," Computer, vol. 44, no. 10, pp. 91-94, Oct. 2011, doi:10.1109/MC.2011.323
- [23] Fung, G.P.C.; Yu, J.X.; Lam, W.: News Sensitive Stock Trend Prediction. In: Proceedings 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Taipei 2002, pp. 481-493.
- [24] Fung, G.P.C.; Yu, J.X.; Lam, W.: Stock Prediction: Integrating Text Mining Approach Using Real-time News. In: Proceedings IEEE Int. Conference on Computational Intelligence for Financial Engineering. Hong Kong 2003, pp. 395-402.
- [25] Gidófalvi, G.: Using News Articles to Predict Stock Price Movements. Project Report, Department of Computer Science and Engineering, University of California, San Diego. <http://www-cse.ucsd.edu/users/elkan/254spring01/gidofalvirep.pdf>, 2001-06-15.
- [26] Ogilvie, P.; Schmill, M.: AAnalyst - Electronic Analyst of Stock Behavior. Project Proposal 791m, Department of Computer Science, University of Massachusetts, Amherst. <http://ciir.cs.umass.edu/~lavrenko/aenanalyst/pitch.pdf>, 1999-10-23.
- [27] Seo, Y.; Giampapa, J.A.; Sycara, K.: Financial News Analysis for Intelligent Portfolio Management. Technical Report CMU-RI-TR-04-04, Robotics Institute, Carnegie Mellon University, Pittsburgh.
- [28] Wüthrich, B.; Cho, V.; Leung, S.; Peramunetilleke, D.; Sankaran, K.; Zhang, J.; Lam, W.: Daily Prediction of Major Stock Indices from Textual WWW Data. In: Proceedings 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. New York 1998, S. 364-368.
- [29] Peramunetilleke, D.; Wong, R.K.: Currency Exchange Rate Forecasting from News Headlines. In: Proceedings 13th Australasian Database Conference. Melbourne 2002, pp.131-139.
- [30] Lavrenko, V.; Schmill, M.; Lawrie, D.; Ogilvie, P.; Jensen, D.; Allan, J.: Language Models for Financial News Recommendation. In: Proceedings 9th Int. Conference on Information and Knowledge Management. Washington 2000, pp. 389-396.
- [31] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, Opinion Mining and Analysis: A Survey, International Journal on Natural Language Computing (IJNLC) Vol. 2, No. 3, June 2013.
- [32] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, An Approach for Opinion Analysis using Expectation Maximization, International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2013.