Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 6, June 2016, pg.500 – 506

STOCK MARKET FORECASTING TECHNIQUES: LITERATURE SURVEY

Vivek Rajput[1], Sarika Bobde[2]

[1,2]Department of Computer Engineering, Maharashtra Institute of Technology, Pune, India

rajputvive@gmail.com, sarika.bobde@mitpune.edu.in

Abstract - The goal of this paper is to study different techniques to predict stock price movement using the sentiment analysis from social media, data mining. In this paper we will find efficient method which can predict stock movement more accurately. Social media offers a powerful outlet for people's thoughts and feelings it is an enormous ever-growing source of texts ranging from everyday observations to involved discussions. This paper contributes to the field of sentiment analysis, which aims to extract emotions and opinions from text. A basic goal is to classify text as expressing either positive or negative emotion. Sentiment classifiers have been built for social media text such as product reviews, blog posts, and even twitter messages. With increasing complexity of text sources and topics, it is time to re-examine the standard sentiment extraction approaches, and possibly to re-define and enrich the definition of sentiment. Next, unlike sentiment analysis research to date, we examine sentiment expression and polarity classification within and across various social media streams by building topical datasets within each stream. Different data mining methods are used to predict market more efficiently along with various hybrid approaches. We conclude that stock prediction is very complex task and various factors should be considered for forecasting the market more accurately and efficiently.

Keywords: Sentiment Analysis, Stock market

I. INTRODUCTION

Stock Market prediction has always had a certain appeal for researchers. While numerous scientific attempts have been made, no method has been discovered to accurately predict stock price movement. The difficulty of prediction lies in the complexities of modeling market dynamics. Even with a lack of consistent prediction methods, there have been some mild successes. Stock Market research encapsulates two elemental trading philosophies; Fundamental and Technical approaches. In Fundamental analysis, Stock Market price movements are believed to derive from a security's relative data. Fundamentalists use numeric information such as earnings, ratios, and management effectiveness to determine future forecasts. In Technical analysis, it is believed that market timing is key. Technicians utilize charts and modeling techniques to identify trends in price and volume. These later individuals rely on historical data in order to predict future outcomes. One area of limited success in Stock Market prediction

comes from textual data. Information from quarterly reports or breaking news stories can dramatically affect the share price of a security. Most existing literature on financial text mining relies on identifying a predefined set of keywords and machine learning techniques. These methods typically assign weights to keywords in proportion to the movement of a share price. These types of analyses have shown a definite, but weak ability to forecast the direction of share prices.

A. Sentiment Analysis

Sentiment analysis also known as opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state or the intended emotional communication. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored .Lexical affinity not only detects obvious affect words, it also assigns arbitrary words a probable affinity to particular emotions. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation —Point-wise Mutual Information. Concept-level approaches leverage on elements from knowledge representation such as ontologies and semantic networks and, hence, are also able to detect semantics that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

B. Stock Market

A **stock market** or **equity market** is the aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares); these may include securities listed on a stock exchange as well as those only traded privately. Stocks can be categorized in various ways. One common way is by the country where the company is domiciled. For example, Nestlé and Novartis are domiciled in Switzerland, so they may be considered as part of the Swiss stock market, although their stock may also be traded at exchanges in other countries.

At the close of 2012, the size of the world stock market (total market capitalization) was about US$55 trillion. By country, the largest market was the United States (about 34%), followed by Japan (about 6%) and the United Kingdom (about 6%). This went up more in 2013.

A stock exchange is a place or organization by which stock traders (people and companies) can trade stocks. Companies may want to get their stock listed on a stock exchange. Other stocks may be traded "over the counter", that is, through a dealer. A large company will usually have its stock listed on many exchanges across the world. Exchanges may also cover other types of security such as fixed interest securities or interest derivatives.

Trade in stock markets means the transfer for money of a stock or security from a seller to a buyer. This requires these two parties to agree on a price. Equities (stocks or shares) confer an ownership interest in a particular company. Participants in the stock market range from small individual stock investors to larger traders investors, who can be based anywhere in the world, and may include banks, insurance companies or pension funds, and hedge funds. Their buy or sell orders may be executed on their behalf by a stock exchange trader.

An example of such an exchange is the New York Stock Exchange. The other type of stock exchange is a virtual kind, composed of a network of computers where trades are made electronically by traders. An example of such an exchange is the NASDAQ.

C. Opinion Mining

With the growth of the web over the last decade, opinions can now be found almost everywhere - blogs, social networking sites like Facebook and Twitter, news portals, ecommerce sites, etc. While these opinions are meant to be helpful, the vast availability of such opinions becomes overwhelming to users when there is just too much to digest. Over the last few years, this special task of summarizing opinions has stirred tremendous interest amongst the Natural Language Processing (NLP) and Text Mining communities. 'Opinions' mainly include opinionated text data such as blog/review articles, and associated numerical data like aspect rating is also included. While different groups have different notions of what an opinion summary should be, we consider any study that attempts to generate a concise and digestible summary of a large number of opinions as the study of Opinion Summarization.

The simplest form of an opinion summary is the result of sentiment prediction (by aggregating the sentiment scores). The task of sentiment prediction or classification itself has been studied for many years. Beyond such summaries, the newer generation of opinion summaries includes structured summaries that provide a well-organized breakdown by aspects/topics, various formats of textual summaries and temporal visualization. The different formats of summaries complement one another by providing a different level of understanding. For example, sentiment prediction on reviews of a product can give a very general notion of what the users feel about the product. If the user needs more specifics, then the topic-based summaries or textual summaries may be more useful. Regardless of the summary formats, the goal of opinion summarization is to help users digest the vast availability of opinions in an easy manner. The approaches utilized to address this summarization task vary greatly and touch different areas of research including text clustering, sentiment prediction, text mining, NLP analysis, and so on. Some of these approaches rely on simple heuristics, while others use robust statistical models.

Currently, there are three distinct papers that are related to the study of opinion summarization.
Liu's book [Liu 2006] covers various techniques in opinion mining and summarization. Liu first defines the notion of 'opinion' and 'opinion mining' and introduces basic concepts related to these definitions. Then he describes techniques in opinion mining covering sentiment classification, opinion summarization, and opinion spam detection. While Liu summarizes this area with a novel framework, as this book was published in 2006, his survey does not cover some of the more recent work in opinion summarization. A big portion of Liu's book is dedicated to explaining definitions and techniques in sentiment classification (the simplest form of an opinion summary), and only a small portion of his book discusses the task of summary generation beyond sentiment classification. In addition, most of the opinion summarization works discussed by Liu are rule-based and heuristics-oriented techniques, missing out on some of the probabilistic methods that were published during the same time period.

In 2010, Liu wrote another book chapter about 'Sentiment Analysis and Subjectivity' [Liu 2010]. Although the new book chapter covers some recent articles, the content in general is very similar to the previous book chapter. The focus of the new book chapter is purely sentiment classification techniques, not covering some of the state-of-the-art opinion summarization methods. As there are already multiple surveys touching the sentiment classification task, in our survey, we focus purely on the recent techniques used in opinion summarization that goes beyond sentiment classification or uses sentiment classification as one of the components in summarization.

Pang and Lee's survey [Pang and Lee 2008] on Opinion Mining and Sentiment Analysis provides a better coverage of works related to opinion summarization. Although this survey covers a lot of recent works, it is focused on

'opinion mining' broadly rather than opinion 'summarization'. In Pang's survey, the methods are explained at a very high level, and the classification of related works is different from the view we will take. In Pang's survey, works in opinion summarization are categorized as single document, multi-documents, textual and visual approaches. In our survey, we will provide a breakdown of the techniques used into distinct steps (e.g. step1: aspect/feature extraction, step 2: sentiment prediction, and step 3: summary generation) and attempt to classify the techniques used in each step to provide both a broad perspective and detailed understanding of those techniques. By focusing on the smaller scope of study, we are able to use more sophisticated categorization for opinion summarization. This will allow readers to compare and contrast the approaches with ease.

In this survey, we cover a comprehensive list of state-of-the-art techniques and paradigms used for the task of opinion summarization that goes beyond sentiment classification. We will classify the approaches in various ways and describe the techniques used in an intuitive manner. We will also provide various aspects of evaluation in opinion summarization, which was not covered by other previous surveys. Finally, we will provide insights into the weaknesses of the approaches and describe the challenges that remain to be solved in this area.

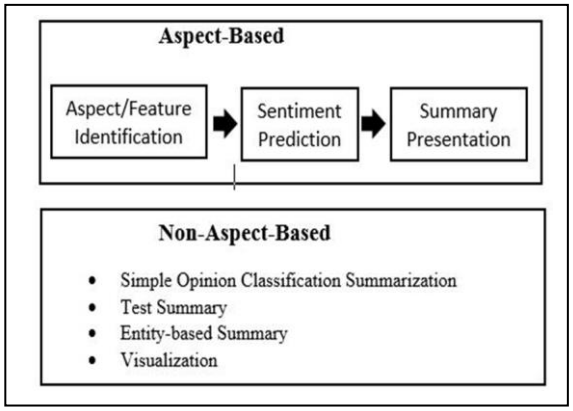

Figure 1: Overview of opinion summarization techniques

II. LITERATURE SURVEY

Kannan, Sekar, Sathik and P. Arumugam in [13] used data mining technology to discover the hidden patterns from the historic data that have probable predictive capability in their investment decisions. The prediction of stock market is challenging task of financial time series predictions. There are five Methods namely Typical price(TP), Bollinger bands, Relative strength index (RSI), CMI and MA used to analyzed the stock index. In this paper the author got the profitable signal is 84.24% using Bollinger Bands rather than MA, RSI and CMI.
Jing Tao Yao and chew Lim tan in [14] used artificial neural networks for classification, prediction and recognition. Neural network training is an art. Trading based on neural network outputs, or trading strategy is also an art. Authors discuss a seven-step neural network prediction model building approach in this article. Pre and post data processing/analysis skills, data sampling, training criteria and model recommendation will also be covered in this article.

Tiffany Hui-Kuang and Kun-Huang Huarng in [15] used neural network because of their capabilities in handling nonlinear relationship and also implement a new fuzzy time series model to improve forecasting. The fuzzy relationship is used to forecast the Taiwan stock index. In the neural network fuzzy time series model where as in-sample observations are used for training and out-sample observations are used for forecasting. The drawback of taking all the degree of membership for training and forecasting may affect the performance of the neural network. To avoid this take the difference between observations. These reduce the range of the universe of discourse.

Md. Rafiul Hassan and Baikunthu Nath in [16] used Hidden Markov Models (HMM) approach to forecasting stock price for interrelated markets. HMM was used for pattern recognition and classification problems because of its proven suitability for modeling dynamic system. The author summarized the advantage of the HMM was strong statistical foundation. It's able to handle new data robustly and computationally efficient to develop and evaluate

similar patterns. The author decides to develop hybrid system using AI paradigms with HMM improve the accuracy and efficiency of forecast the stock market.

Ching-Hseue cheng, Tai-Liang chen, Liang-Ying Wei in [17] this paper proposed a hybrid forecasting model using multi-technical indicators to predict stock price trends. There are four procedures described such as select the essential technical indicators, the popular indicators based on a correlation matrix and use CDPA to minimize the entropy principle approach. Then use RST algorithm to extract linguistic rules and utilize genetic algorithm to refine the extracted rules to get better forecasting accuracy and stock return. The advantage was discovered that produce more reliable and understandable rules and forecasting rules based on objective stock data rather than subjective human judgments.

Fazel Zarandi M.H, Rezaee B, Turksen I.B and Neshat E in [18] used a type-2 fuzzy rule based expert system is developed for stock price analysis. The purposed type-2 fuzzy model applies the technical and fundamental indexes as the input variables. The model used for stock price prediction of an automotive manufactory in Asia. The output membership values were projected onto the input spaces to generate the next membership values of input variables and tuned by genetic algorithm. The type-1 method was used for inference and to increasing the robustness of the system. This method was used to robustness, flexibility and error minimization. It is used to forecast more profitable trading in stock markets.

III. METHODS FOR STOCK MOVEMENT PREDICTION

A. Methods for Stock Movement Prediction

The Support Vector Machine (SVM) has long been recognized as being able to efficiently handle high dimensional data and has been shown to perform well on. Therefore, we chose the SVM with the linear kernel as the prediction model. To assess the effectiveness of sentiment analysis on the message boards, six sets of features are designed. The first one used only the historical prices. The other methods incorporated the mood information into the prediction model. All the feature values were scaled into [−1, 1].

B. Price only

In this method, only historical prices are used to predict the stock movement. The purpose of this method is to investigate whether there are patterns in the history of the stock or not. In addition, this model was used as a baseline to evaluate whether integration of the sentiments is effective by comparing with other sentiment models. Features used for the training of SVM are $price_{t-1}$ and $price_{t-2}$ which are the price movements (up, down) at the transaction dates $t - 1$ and $t - 2$, respectively.

C. Human sentiment

In addition to historical prices, this model integrated the sentiments annotated by human into the prediction model. The MessageBoard dataset, the users explicitly select a sentiment label with their posts. These sentiment labels are “strong buy”, “buy”, “hold”, “sell” and “strong sell”. Instead of using all the messages, we tried to use only the messages with annotated sentiments by the users, and discard the other messages. From these messages, we used only the explicit sentiment and remove other in-formation such as message content. The purpose of this method is that how mood annotated by human can be used to predict the stock. Because the sentiments are annotated by human, this feature is one of the strongest features for stock prediction.

D. Sentiment classification

To utilize the remaining 84.4% of the messages without the explicit sentiments, we tried to build a model to extract the sentiments for those messages. A classification model was trained from the messages with annotated sentiments on the training dataset. Then it was used to classify the remaining messages into five classes (Strong Buy, Buy, Hold, Sell, and Strong Sell).

E. LDA-based method

In this model, we consider each message as a mixture of hidden topics. Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Therefore, we choose the LDA as a simple topic model to discover these hidden topics.

F. JST-based method

The opinion is often expressed on a topic or aspect. When people post the message on the social media to express their opinion for a given stock, they tend to talk their opinions for a certain topic such as profit and dividend. Based on pairs of topic-sentiment, they would think that the future price of that stock goes up or down. From that intuition, we propose a new feature topic-sentiment for the stock prediction model. To extract pairs of topic-sentiment, we tried to use two kinds of models.

G. Aspect-based sentiment

Instead of considering the mixtures of hidden topics and sentiments as in the previous model, in this model the mixtures are not hidden. Each message is represented as a list of topics and their corresponding sentiment values. In our proposed method, the topic is the consecutive nouns in the sentence.

IV. CONCLUSION

Thus we have studied various methods for stock market prediction using sentiment analysis and data mining. Because of usefulness and needs from the people, opinion mining became an active research area. As the volume of the opinionated data increases, analyzing and summarizing opinionated data is becoming more important. To satisfy these needs, many kinds of opinion summarization techniques are proposed. Probabilistic approaches using statistics of terms and heuristic approaches using predefined rules are representative works. Despite of a lot of research efforts, current stock prediction studies still have many limitations and margins for improvement. We finally conclude that stock prediction is very complex task and various factors should be considered for forecasting the market more accurately and efficiently.

REFERENCES

[1] Wenping Zhang, Chunping Li, Yunming Ye, Wenjie Li and Eric W.T. Ngai "Dynamic Business Network Analysis for Correlated Stock Price Movement Prediction", Published by the IEEE Computer Society 2015

[2] Chia-Hsuan Yeh and Chun-Yi Yang "Social Networks and Asset Price Dynamics", IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, , JUNE 2015

[3] Li-Xin Wang "Dynamical Models of Stock Prices Based on Technical Trading Rules Part I: The Models", IEEE TRANSACTIONS ON FUZZY SYSTEMS, AUGUST 2015

[4] Xiaodong Li a, Haoran Xie a,⇑, Li Chen b, Jianping Wanga, Xiaotie Deng "News impact on stock price return via sentiment analysis", ScienceDirect 2014

[5] Hong Keel Sul,Alan R. Dennis,Lingyao (Ivy) "Trading on Twitter: The Financial Information Content of Emotion in Social Media", 2014 47th Hawaii International Conference on System Science
[6] Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inform. Syst.(TOIS) 26 (2008) 12.
[7] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu, Mining social emotions from affective text, IEEE Trans. Knowl. Data Eng. 24 (2012) 1658–1670.
[8] M. Bautin, L. Vijayarenu, S. Skiena, International sentiment analysis for news and blogs, in: Proceedings of the International Conference on Weblogs and Social Media, 2008.
[9] F. Bießmann, J.M. Papaioannou, M. Braun, A. Harth, Canonical trends: detecting trend setters in web data, in: International Conference on Machine Learning, 2012.
[10] E. Cambria, C. Havasi, A. Hussain, Senticnet 2: a semantic and affective resource for opinion mining and sentiment analysis, in: FLAIRS Conference, 2012, pp. 202–207.
[11] E. Cambria, T. Mazzocco, A. Hussain, Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining, Biol. Inspired Cogn. Architec. 4 (2013) 41–53.
[12] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, IEEE Intell. Syst. (2013).
[13] K. Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam, "Financial stock market forecast using data mining Techniques", 2010, Proceedings of the international multiconference of engineers and computer scientists.
[14] JingTao YAO and Chew Lim TAN , "Guidelines for Financial Prediction with Artificial neural networks".
[15] Tiffany Hui-Kuang yu and Kun-Huang Huarng, "A Neural network-based fuzzy time series model to improve forecasting", *Elsevier*, 2010, pp: 3366-3372.
[16] Md. Rafiul Hassan and Baikunth Nath, "Stock Market forecasting using Hidden Markov Model: A New Approach", Proceeding of the 2005 5th international conference on intelligent Systems Design and Application 0-7695-2286-06/05, IEEE 2005.
[17] Ching-Hsue cheng, Tai-Liang Chen, Liang-Ying Wei, " A hybrid model based on rough set theory and genetic algorithms for stock price forecasting", 2010, pp. 1610-1629.
[18] M.H. Fazel Zarandi, B. Rezaee, I.B. Turksen and E.Neshat, "A type-2 fuzzy rule-based experts system model for stock price analysis", Expert systems with Applications, 2009, pp. 139-154.

