IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Stock Market Prediction Using Text Mining Approaches: A Survey

A.Sahoo[1], J.K.Mantri[2*]

[1, 2] Department of Computer Application, North Orissa University, India

Corresponding Author: jkmantri@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i2.443450 | Available online at: www.ijcseonline.org

Accepted: 14/Feb/2019, Published: 28/Feb/2019

Abstract— Stock market prediction is an attractive research problem to be investigated in the field of computational finance. News contents are one of the most important factors that have influence on market. Considering the news impact in analyzing the stock market behaviour, leads to more precise predictions and as a result more profitable trades. Text mining , or the pragmatic research perspective of computational linguistics, has become increasingly powerful due to data availability and various techniques developed in the past decade. However, no detailed comparison of the systems and their performances is available thus far. This paper tries to describe the main systems developed and presents a survey work for comparing the approaches.

Keywords- Text mining, Natural language processing (NLP), sentimental analysis, stock market prediction

I. INTRODUCTION

Stock market prediction is burning topic in the field of computational finance. Due to its business increment, it has attracted often aid from educator to economics sector. It is impossible to give the prediction of prices of stock market because of stock prices are changed by every second. Hence, the use of data mining techniques to predict financial markets has been extensively studied in numerous publications. Most of the studies use structured data like past prices, historical earnings, or dividends. Text mining approaches are comparatively rare due to the difficulty of extracting relevant information from unstructured data. The first goal of this paper is to briefly describe the prototypes developed so far with a survey of thirty five papers using different text mining techniques on distinct data sets, then a new methodology has been proposed to predict the stock market using Text mining.

II. RELATED WORK

Frederick S.M. Herz, Warrington,Lyle H. Ungar, Philadelphia, Jason M. Eisner, Baltimore, Walter Paul Labys, (2003) presented a method of using natural language processing (NLP) techniques to extract information from online news feeds and then using the information So extracted to predict changes in Stock prices or Volatilities. These predictions can be used to make profitable trading Strategies. More specifically, company names can be recognized and Simple tem plates describing company actions can be automatically filled using parsing or pattern matching on words in or near the Sentence containing the company name. These templates can be clustered into groups which are Statistically correlated with changes in the Stock prices. Their System is composed of two parts: a message understanding component that automatically fills in Simple templates and a Statistical correlation component that tests the correlation of these patterns to increases or decreases in the Stock price. This methods can be applied to a broad range of text, including articles in online news paper such as the Wall Street Journal, financial newsletters, radio & TV transcripts and annual reports. They envision it being used first for newswires such as Bloomberg, or perhaps the AP News WC. In an enhanced embodiment of the System, they suggested that it can be further leverage Statistical patterns in Internet usage data and Internet data such as newly released textual information on Web Pages.

Y.-C. Phung (2005) described a proposed system that extracts key phrases from online news articles for stock movement predictions of Malaysia. Their proposed system was implemented and tested on selected active sectors from Bursa Malaysia, a Malaysian Stock Exchange. The implementation is based on two crucial factors: firstly, stock related news articles are influential in the sense that they can influence a buyer’s attitude and subsequently can cause the stocks to move; and secondly, textual data are considered more superior and contain more information than numeric data because the former not only allows us to predict future stock prices but at the same time provides with reasons as to why it is so. They also reviewed investigations on how the online Malaysian news articles are mined to extract appropriate keyphrases; which are then subsequently used to predict the movements of stock prices in Bursa Malaysia. The main focus of this paper is to implement appropriate text mining techniques to extract the said keyphrases from online news sources. The proposed system is then trained and tested based on a collection of Malaysian news articles from The Star Online: Malaysia Business News . The findings indicate that the implementation is capable of providing valuable keyphrases whether or not original keyphrases are provided.

Robert P. Schumaker and Hsinchun Chen (2007) tried to examine a predictive machine learning approach for financial news articles analysis using several different textual representations: Bag of Words, Noun Phrases, and Named Entities. Through this approach, they investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. They applied their analysis to estimate a discrete stock price twenty minutes after a news article was released. Using a Support Vector Machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables, exhibited that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price (MSE 0.04261), the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return). They also further investigated the different textual representations and found that a Proper Noun scheme performs better than the de facto standard of Bag of Words in all three metrics.

Heeyoung Lee Mihai Surdeanu Bill MacCartney Dan Jurafsky (2008) investigated the importance of text analysis for stock price prediction. In particular, they introduced a system that forecasts companies' stock price changes (UP, DOWN, STAY) in response to financial events reported in 8-K documents. Our results indicate that using text boosts prediction accuracy over 10% (relative) over a strong baseline that incorporates many financially-rooted features. They here concluded this impact is most important in the short term (i.e., the next day after the financial event) but persists for up to five day.

R.P.Schumaker, H.Chen (2009) examined a predictive machine learning approach for financial news articles analysis using several different textual representations: bag of words, noun phrases, and named entities. Through this approach, they investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five week period. They applied their analysis to estimate a discrete stock price twenty minutes after a news article was released. Using a support vector machine (SVM) derivative specially tailored for discrete numeric prediction and models containing different stock-specific variables, it is seen that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price (MSE 0.04261), the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return). They further investigated the different textual representations and found that a Proper Noun scheme performs better than the de facto standard of Bag of Words in all three metrics.

AzadehNikfarjam , EhsanEmadzadeh , Saravanan Muthaiyah (2010) described that Stock market prediction is an attractive research problem to be investigated. They also described news contents are one of the most important factors that have influence on market. Considering the news impact in analyzing the stock market behavior, leads to more precise predictions and as a result more profitable trades. So far various prototypes have been developed which consider the impact of news in stock market prediction. In this paper, the main components of such forecasting systems have been introduced. In addition, different developed prototypes have been introduced and the way whereby the main components are implemented compared. Based on studied attempts, the potential future research activities have been suggested.

Scott Hendry and Alison Madeley (2010) used Latent Semantic Analysis to extract information from Bank of Canada communication statements and investigates what type of information affects returns and volatility in short-term as well as long-term interest rate markets over the 2002-2008 period. Discussions about geopolitical risk and other external shocks, major domestic shocks (SARS and BSE), the balance of risks to the economic projection, and various forward looking statements are found here to significantly affect market returns and volatility, especially for short-term markets. They also described , this effect is over and above that from the information contained in any policy interest rate surprise.

Manisha V. Pinto, Kavita Asnani (2011) provided a framework for predicting stock magnitude and trend for making trading decisions by making use of a combination of Data Mining and Text Mining methods. Their prediction model predicts the stock market closing price for a given trading day 'D', by analysing the information rich unstructured news articles along with the historical stock quotes. In particular, they investigated the immediate impact of the news articles on the time series based on Efficient Market Hypothesis (EMH).

Cambria E, Livingstone A, Hussain A (2012) also described that the adoption of psychological models of emotions has become a common trend among researchers and engineers working in the sphere of affective computing. Because of the elusive nature of emotions and the ambiguity of natural language, however, psychologists have developed many different affect models, which often are not suitable for the design of applications in fields such as affective HCI, social data mining, and sentiment analysis. To this end, they have proposed a novel biologically-inspired and psychologically-motivated emotion categorisation model that goes beyond mere categorical and dimensional approaches. Hence it is concluded that such model represents affective exhibited both through labels and through four independent but concomitant affective dimensions, which can potentially describe the full range of emotional experiences.

Feldman R (2013) tried to explain the main applications and challenges of one of the hottest research areas in computer science i.e Techniques and applications for sentiment analysis. Hence Sentiment analysis (or Opinion mining) is defined as the task of finding the opinions of authors about specific entities. The decision-making process of people is affected by the opinions formed by thought leaders and ordinary people. When a person wants to buy a product online he or she will typically start by searching for reviews and opinions written by other people on the various offerings. Sentiment analysis is one of the hottest research areas in computer science. Over 7,000 articles have been written on the topic. Hundreds of startups are developing sentiment analysis solutions and major statistical packages such as SAS and SPSS include dedicated sentiment analysis modules here.

Enric Junqué de Fortuny, Tom De Smedt, David Martens , Walter Daelemans (2013) Explained that despite the fact that both the Efficient Market Hypothesis and Random Walk Theory postulate , it is impossible to predict future stock prices based on currently available information, recent advances in empirical research have been proving the opposite by achieving what seems to be better than random prediction performance. As a result, they discussed some of the (dis)advantages of the most widely used performance metrics and concluded that is difficult to assess the external validity of performance using some of these measures. Moreover,there remain many questions as to the real-world applicability of these empirical models. In the first part of this study they designed novel stock price prediction models, based on stateof- the-art text-mining techniques to assert whether they can predict the movement of stock prices more accurately by including indicators of irrationality. Along with this, they discussed which metrics are most appropriate for which scenarios in order to evaluate the models. Finally, they discussed how to gain insight into text-mining-based stock price prediction models in order to evaluate, validate and refine the models.

S.Abdullah , Mohammad,Saiedur Rahaman M. S.Rahman (2013) described that Stock market has become one of the major components of economy not only in developed countries but also in third world developing countries. Making decision in stock market is not really easy because a lot of factors are involved with every choice we make. Therefore, a lot of analysis is required to make an optimal move on stock market which may involve price trend, market's nature, company's stability, different news and rumors about stocks etc. The objective of their study is to extract fundamental information from relevant news sources and use them to analyze or sometimes forecast the stock market from the common investor's viewpoint. They surveyed the existing business text mining researches and proposed a framework that uses our text parser and analyzer algorithm with an open source natural language processing tool to analyze (machine learning and text mining), retrieve (natural language processing), forecast (compare with historic data) investment decisions from any text data source on stock market i.e Dhaka Stock Exchange (DSE), capital market of Bangladesh.

Jageshwer Shriwas , Shagufta Farzana (2014) tried to analyze index point of NSE with the help of using classification rules of data mining to find out the pattern and finally this pattern evaluated with the help of graph based analysis based on daily stock price .This pattern shows new indexes and key point which is average value of all the indexes and also analyze the news impact on stock price prediction. Here , they used one year index point of an NSE and analyze it for stock price prediction. This study basically shows the effect of financial news to the prediction of stock market prices as well as daily direction of change in the index.

Yoosin Kim , Seung Ryul Jeong , Imran Ghani (2014) introduced a method of mining text opinions to analyze Korean language news in order to predict rises and falls on the KOSPI (Korea Composite Stock Price Index) which consists of carrying out the NLP (Natural Language Processing) of news, describing its features, categorizing and extracting the sentiments and opinions expressed by the writers. It also identifies the correlation between news and stock market fluctuations. In their experiment, they showed that this method can be used to understand unstructured big-data, and they also reveal ted hat news' sentiment can be used in predicting stock price fluctuations, whether up or down. Hence, they expected that the algorithm extracted experiments can be used to make predictions about stock market movements.

Enric Junqué de Fortuny , Tom De Smedt ,David Martens ,Walter Daelemans (2014) discussed some (dis)advantages of the most widely used performance metrics and concluded that it is difficult to assess the external validity of performance using some of these measures. In the first part of this study they designed novel stock price prediction models, based on stateof-the-art text-mining techniques to assert whether they can predict the movement of stock prices more accurately by including indicators of irrationality. Along with this, they discussed which metrics are most appropriate for which scenarios in order to evaluate the models. Finally, discuss was to how to gain insight into text-mining-based stock price prediction models in order to evaluate, validate and refine the model.

Spandan Ghose Chowdhury, Soham Routh,Satyajit Chakrabarti (2014) proposed a predictive model to predict sentiment around stock price. First the relevant real time news headlines and press-releases have been filtered from the large set of business news sources, and then they have been analyzed to predict the sentiment around companies. In order to find correlation between sentiment predicted from news and original stock price and to test efficient market hypothesis,by ploting the sentiments of 15 odd companies over a period of 4 weeks. Their result shows an average accuracy score for identifying correct sentiment of around 70.1%. They also have plotted the errors of prediction for different companies which have brought out the RMSE and MAE of 30.3% and 30.04% respectively and an enhanced F1 factor of 78.1% by giving the ideas that the comparison between positive sentiment curve and stock price trends reveals 67% co-relation between them, which indicates towards existence of a semi-strong to strong efficient market hypothesis.

M. Jaybhay ,Rajesh V. Argiddi S.S.Apte, (2014) proposed a novel method for the prediction of stock market closing price by considering a multilayered feed-forward neural network which is built by using combination of data and textual mining. The Neural Network is trained on the stock quotes and extracted key phrases using the Backpropagation Algorithm which is used to predict share market closing price. Their paper is an attempt to determine whether the BSE market news in combination with the historical quotes can efficiently help in the calculation of the BSE closing index for a given trading day.

Felix Ming Fai Wong, Zhenming Liu, Mung Chiang (2014) revisited the problem of predicting directional movements of stock prices based on news articles: where their algorithm uses daily articles from The Wall Street Journal to predict the closing stock prices on the same day. They also proposed a unified latent space model to characterize the "co-movements" between stock prices and news articles. Unlike many existing approaches, our new model is able to simultaneously leverage the correlations: (a) among stock prices, (b) among news articles, and (c) between stock prices and news articles. Lastly they concluded that their proposed model is able to make daily predictions on more than 500 stocks (most of which are not even mentioned in any news article) while having low complexity by carrying out extensive backtesting on trading strategies based on our algorithm and the proposed model has substantially better accuracy rate (55.7%) compared to many widely used algorithms. Also ,the return (56%) and Sharpe ratio due to a trading strategy based on proposed model are also much higher then baseline indices.

Joao Guerreiro, Paulo Rita, Duarte Trigueiros (2014) used an advanced Text Mining methodology (a Bayesian contextual analysis algorithm known as Correlated Topic Model, CTM) to conduct a comprehensive analysis of 246 articles published in 40 different journals between 1988 and 2013 on the subject of cause-related marketing. They also described that Text Mining also allows quantitative analyses to be performed on the literature. For instance, it is shown that the most prominent long-term topics discussed since 1988 on the subject are ‘‘brand-cause fit’’, ‘‘law and Ethics’’, and ‘‘corporate and social identification’’, while the most actively discussed topic presently is ‘‘sectors raising social taboos and moral debates’’. Hence, this paper has two goals: first, it introduces the technique of CTM to the Marketing area, illustrating how Text Mining may guide, simplify, and enhance review processes while providing objective building blocks (topics) to be used in a review; second, it applies CTM to the C-RM field, uncovering and summarizing the most discussed topics. But ,Mining text, however, is not aimed at replacing all subjective decisions that must be taken as part of literature review methodologies.

Sadi Seker,Cihan Mert,Khaled Al-Naami,N. Ozalp,U.Ayan (2014)described an information retrieval method for the economy news. The effect of economy news, are researched in the word level and stock market values are considered as the ground proof. The correlation between stock market prices and economy news is an already addressed problem for most of the countries. The most well-known approach is applying the text mining approaches to the news and some time series analysis tech *International Journal Of Social Sciences And Humanity Studies* Vol 6, No 1, 2014 ISSN: 1309-8063 (Online) 70 times over stock market closing values in order to apply classification or clustering algorithms over the features extracted. This study goes further and tries to ask the question what are the available time series analysis techniques for the stock market closing values and which one is the most suitable? In this study, the news and their dates are collected into a database and text mining is applied over the news, the text mining part has been kept simple with only term frequency – inverse document frequency method. For the time series analysis part, they have studied 10 different methods such as random walk, moving average, acceleration, Bollinger band, price rate of change, periodic average, difference, momentum or relative strength index and their variation. In this study they have also explained these techniques in a comparative way and have applied the methods over Turkish Stock Market closing values for more than a 2 year period. Also, they have applied the term frequency – inverse document frequency method on the economy news of one of the high-circulating newspapers in Turkey.

Chanwit Onsumran, Sotarat Thammaboosadee, and Supaporn Kiattisin (2015) focused on the text mining approach of the gold prices volatility prediction model from the textual of economic indicators news articles. Hence a model is designed and developed to analyze how the news articles influence gold price volatility. Here, the selected reliable source of news articles is provided by FXStreet which offers several economic indicators such as Economic Activity, Markit Manufacturing PMI, Bill Auction, Building Permits, ISM Manufacturing Index, Redbook index, Retail Sales, Durable Goods Orders, etc. to build text classifiers and news group affecting volatility price of gold. According to the fundamental of data mining process, each news article is firstly transformed in to feature by TF-IDF method. Then, the comparative experiment is set up to measure the accuracy of combination of two attributes weighting approaches, which are Support Vector Machine (SVM) and Chi-Squared Statistic, and three classification algorithms, which are the k-Nearest Neighbour, SVM and Naive Bayes. Their results show that the SVM method is the most superior to other methods in both attributes weighting and classifier viewpoint.

SS Panigrahi, JK Mantri (2015),In this paper an attempt has been made to implement text based decision tree having all discrete input variables rather than a numerical decision tree where at least one variable is need to be discrete. The available historical data and other technical indicators calculated over the numerical data set of BSE sensex and NSE nifty has been converted and normalized to textual form by certain rules and decision trees are differently constructed for BSE sensex and NSE nifty with application of C4.5 algorithm and compared with the usual decision tree generated directly by applying numerical variables for the same period. The empirical study proves the better efficacy of the proposed model by outperforming the usual decision tree.

Tseng-Chung Tang , Hsu-Tong Deng, Li-Chiu Chi (2015) described about Google search query which is the main feature of symbolic behavior for an investor on his early stage of choosing stocks. Also this mirrors the extent of an investor's concern and preference about specific investment decisions. As such, keyword search volumes can be used as early warning signs of stock market movements. Hence the main purpose and contribution of their study is to employ the Google search volume index as a proxy for investor attention, thereby enabling the observation of price trends as well as the translation into profitable stock investment strategies. Also,this study further investigates the characteristic of time distribution that investors spend on Google search queries under the effects of limited attention to predict market trends.

In the study of Arti Buchek , M. B. Chandak (2016) the behavioral 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. Hence their 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. They have surveyed and analysed techniques and key tasks of opinion mining in this paper. An overall picture of developing a software system for opinion mining on the basis of survey and analysis is proposed here.

Yancong Xie, Hongxun Jiang (2016) used text mining and sentiment analysis on Chinese online financial news, to predict Chinese stock tendency and stock prices based on support vector machine (SVM). Firstly, they collected 2,302,692 news items, which date from 1/1/2008 to 1/1/2015. Secondly, based on this dataset, a specific domain stop-word dictionary and a precise sentiment dictionary are formed. Thirdly, they proposed a forecasting model using SVM. On the algorithm of SVM implementation, by proposing two parameters optimization algorithms to search for best initial parameter setting. Result shows that parameter G (Gaussian Kernel) has the main effect, while parameter C's(cost of misclassification) effect is not obvious. Furthermore, support vector regression (SVR) models for different Chinese stocks are similar whereas in support vector classification (SVC) models best parameters are quite differential their series of contrast experiments show that: a) News has significant influence on stock market; b) Expansion input vector for additional situations when that day has no news data is better than normal input in SVR, yet is worse in SVC; c) SVR shows a fantastic degree of fitting in predicting stock fluctuation while such result has some time lag; d) News effect time lag for stock market is less than two days; e) In SVC, historic stock data has a most efficient time lag which is about 10 days, whereas in SVR this effect is not obvious. In addition, based on the special structure of the input vector, they also designed a method to calculate the financial source impact factor suggesting that the news quality and audience number both have significant effect on the source impact factor. Lastly they calculated that besides, for Chinese investors, traditional media has more influence than digital media

Shri Bharathi, Angelina Geetha (2017) attempts to design and to implement a predictive system for guiding stock market investment. The novelty of their approach is the combination of both Sensex points and Really Simple Syndication (RSS) feeds for effective prediction. Their claim is that the sentiment analysis of RSS news feeds has an impact on stock market values. Hence RSS news feed data are collected along with the stock market investment data for a period of time. Using their algorithm for sentiment analysis, the correlation between the stock market values and sentiments in RSS news feeds are established. This trained model is used for prediction of stock market rates. In their experimental study the stock market prices and RSS news feeds are collected for the company ARBK from Amman Stock Exchange (ASE). Their experimental study has shown an improvement of 14.43% accuracy prediction, when compared with the standard algorithm of ID3, C4.5 and moving average stock level indicator.

Vaishali Ingle; Sachin Deshmukh (2017) collected data for stock market in the form of breaking news from various finance websites. The TF-IDF features extracted from online news data are used for creation of HMM model along with log likelihood values. So that next day's stock price is predicted as either higher or lower than current day's stock price. Results obtained from their proposed model is compared with results from other machine learning predictive techniques such as random forest, KNN, multiple regression, bagging and boosting. Their proposed model produces approximately 70% of accurate prediction.

AdityaBhardwaja,YogendraNarayanb,Vanrajc,Pawana,M.Dutta(2015) described the application of Internet based technologies which had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time. Before investing money, it is very important for investors to predict the stock market. In today's digital world Internet based technologies such as Cloud Computing, Big Data analytics, and Sentiment analysis have changed the way we do business. Sentiment analysis or opinion mining makes use of text mining, natural language processing (NLP), in order to identify and extract the subjective content by analyzing user's opinion, evaluation, sentiments, attitudes and emotions. Here importance of sentiment analysis for Sensex and Nifty has been done to predict the the price of stock.

Chan SW, Chong MW (2017) also described that the growth of financial texts in the wake of big data has challenged most organizations and brought escalating demands for analysis tools. In general, text streams are more challenging to handle than numeric data streams. Text streams are unstructured by nature, but they represent collective expressions that are of value in any financial decision. It can be both daunting and necessary to make sense of unstructured textual data. Hence, they addressed key questions related to the explosion of interest in how to extract insight from unstructured data and how to determine if such insight provides any hints concerning the trends of financial markets. A sentiment analysis engine (SAE) is also proposed here which takes advantage of linguistic analyses based on grammars. This engine extends sentiment analysis not only at the word token level, but also at the phrase level within each sentence. An assessment heuristic is applied to extract the collective expressions shown in the texts. Also here , three evaluations are presented to assess the performance of the engine. First, several standard parsing evaluation metrics are applied on two treebanks. Second, a benchmark evaluation using a dataset of English movie review is conducted. Results show our SAE outperforms the traditional bag of words approach. Third, a financial text stream with twelve million words that aligns with a stock market index is examined. The evaluation results and their statistical significance provide strong evidence of a long persistence in the mood time series generated by the engine. In addition, their approach establishes grounds for belief that the sentiments expressed through text streams are helpful for analyzing the trends in a stock market index, although such sentiments and market indices are normally considered to be completely uncorrelated.

Massimiliano Caporin , and Francesco Poli (2017) retrieved news stories and earnings announcements of the S&P 100 constituents from two professional news providers, along with ten macroeconomic indicators. That also gathered data from Google Trends about these firms' assets as an index of retail investors' attention creating an extensive and innovative database that contains precise information with which to analyze the link between news and asset price dynamics. They also detected the sentiment of news stories using a dictionary of sentiment-related words and negations and propose a set of more than five thousand information-based variables that provide natural proxies for the information used by heterogeneous market players. The first shed light on the impact of information measures on daily realized volatility and select them by penalized regression to perform a forecasting exercise and showing that the model augmented with news-related variables provides superior forecasts.

Chan SW, Chong MW (2017) also described that the growth of financial texts in the wake of big data has challenged most organizations and brought escalating demands for analysis tools. In general, text streams are more challenging to handle than numeric data streams. Text streams are unstructured by nature, but they represent collective expressions that are of value in any financial decision. It can be both daunting and necessary to make sense of unstructured textual data. Hence, they addressed key questions related to the explosion of interest in how to extract insight from unstructured data and how to determine if such insight provides any hints concerning the trends of financial markets. A sentiment analysis engine (SAE) is also proposed here which takes advantage of linguistic

analyses based on grammars. This engine extends sentiment analysis not only at the word token level, but also at the phrase level within each sentence. An assessment heuristic is applied to extract the collective expressions shown in the texts. Also here , three evaluations are presented to assess the performance of the engine. First, several standard parsing evaluation metrics are applied on two treebanks. Second, a benchmark evaluation using a dataset of English movie review is conducted. Results show our SAE outperforms the traditional bag of words approach. Third, a financial text stream with twelve million words that aligns with a stock market index is examined. The evaluation results and their statistical significance provide strong evidence of a long persistence in the mood time series generated by the engine. In addition, their approach establishes grounds for belief that the sentiments expressed through text streams are helpful for analyzing the trends in a stock market index, although such sentiments and market indices are normally considered to be completely uncorrelated.

Frank Z. Xing, Erik Cambria1and Roy E. Welsch (2017) revealed that Natural language processing (NLP), or the pragmatic research perspective of computational linguistics, has become increasingly powerful due to data availability and various techniques developed in the past decade. This increasing capability makes it possible to capture sentiments more accurately and semantics in a more nuanced way. Naturally, many applications are starting to seek improvements by adopting cutting-edge NLP techniques. Financial forecasting is no exception. As a result, articles that leverage NLP techniques to predict financial markets are fast accumulating, gradually establishing the research field of natural language based financial forecasting (NLFF), or from the application perspective, stock market prediction. Hence,their review article clarified the scope of NLFF research by ordering and structuring techniques and applications from related work. This survey also aims to increase the understanding of progress and hotspots in NLFF, and bring about discussions across many different disciplines.

Marcello Backmann (2017) presented a computational framework that aims to predict the changes of stock prices along the day, given the occurrence of news articles related to the companies listed in the Down Jones Index. For this task, an automated process that gathers, cleans, labels, classifies, and simulates investments was developed. This process integrates the existing data mining and text algorithms, with the proposal of new techniques of alignment between news articles and stock prices, pre-processing, and classifier ensemble. The result of experiments in terms of classification measures and the Cumulative Return obtained through investment simulation outperformed the other results found after an extensive review in the related literature. This work also argues that the classification measure of Accuracy and incorrect use of cross validation technique have too few to contribute in terms of investment recommendation for financial market. Altogether, the developed methodology and results contribute with the state of art in this emerging research field, demonstrating that the correct use of text mining techniques is an applicable alternative to predict stock price movements in the financial market.

PavelNetolický,JonášPetrovskýand, František Dařena (2018) used text mining methods to discover if there is a connection between news articles and changes of the S&P 500 stock index. The index values and documents were divided into time windows according to the direction of the index value changes. Lastly achieved a classification accuracy of 65–74%.

III. CONCLUSION AND FUTURE SCOPE

This paper summarizes and compares the different methods developed for predicting the stock market response to news by text mining techniques. Most of the methods forecast price trends, and prediction of volatilities in particular trends of stock prices, exchange rates, and equity indices. Summarizing all the procedures, it is proposed that it is better to implement text based decision tree having all discrete input variables rather than a numerical values using rough-soft set considering the harmonic mean of macro-averaged precision and macro-averaged recall or overall accuracy and the percentage of correct predictions of stock market.

REFERENCES

[1]. Aditya Bhardwaja,Yogendra Narayanb, Vanrajc, Pawana, Maitreyee Dutta, Sentiment Analysis for Indian Stock Market prediction using Sensex and Nifty, Procedia Computer Science 70 (2015) pp 85-91 ,2015.

[2]. Felix Ming Fai Wong, Zhenming Liu, Mung Chiang , Stock Market Prediction from WSJ: Princeton University.

[3]. Heeyoung Lee, Mihai Surdeanu Bill MacCartney, Dan Jurafsky ,On the Importance of Text Analysis for Stock Price Prediction, http://www.sec.gov/edgar.shtml. pp 1-4,2008**.**

[4]. Jageshwer Shriwas1 , Shagufta Farzana ,Using Text Mining and Rule Based Technique for Prediction of Stock Market Price, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459,pp 245-250,2014.

[5]. Robert P. Schumaker and Hsinchun Chen . Textual Analysis of Stock Market Prediction Using Breaking: Artificial Intelligence Lab, The University of Arizona, Arizona 85721, USA, pp 1-29,2007.

[6]. Vaishali Ingle; Sachin Deshmukh , Predictive mining for stock market based on live news TF-IDF features, International Journal of Autonomic Computing, Vol.2 No.4, 10.1504/IJAC.2017.089703,pp 341-345,2017.

[7]. Yancong Xie1, 2, Hongxun Jiang , Stock Market Forecasting Based on Text Mining Technology:A-Support Vector Machine Method. Journal of Computers,doi:10.17706 /jcp.12.6.500-510,2016.

[8]. Arti Buchek , Dr. M. B. Chandak (2016) Stock Market Prediction using Text Opinion Mining: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128x,pp 566-569,2016.

[9]. Azadeh Nikfarjam ; Ehsan Emadzadeh ; Saravanan Muthaiyah .Text mining approaches for stock market prediction,IEEE Xpolre, DOI: 10.1109/ICCAE.2010.5451705,2010.

[10]. Cambria E, Livingstone A, Hussain A ,The hourglass of emotions. In: Lecture notes in computer science, vol 7403. Springer, pp 144–157,2012**.**

[11]. Chan SW,Chong MW, Sentiment Analysis for stock market prediction, Journal of Computers ,pp197-119,2017.

[12]. Chan SW, Chong MW , Sentiment analysis in financial texts. Decis Support Syst. 94:53–64,2017.

[13]. Chanwit Onsumran, Sotarat Thammaboosadee, and Supaporn Kiattisin , Gold Price Volatility Prediction by Text Mining in Economic Indicators News. Journal of Advances in Information Technology Vol. 6, No. 4,243-47,2015.

[14]. Enric Junqué de Fortuny , Tom De Smedt , David Martens , Walter Daelemans Evaluating and understanding text-based stock price prediction models, Information Processing & Management ,Volume 50, Issue 2, pp 426-441,2014.

[15]. Enric Junqué de Fortuny a,⇑, Tom De Smedt b, David Martens a, Walter Daelemans, Evaluating and understanding text-based stock price prediction models, http://dx.doi.org/10.1016/j.ipm.2013.12.002: pp(1- 13),2013

[16]. Feldman R , Techniques and applications for sentiment analysis. Common ACM 56(4):82–89,2013.

[17]. Frank Z. Xing1 · Erik Cambria,Roy E. Welsch , Natural language based financial forecasting: asurvey, https://doi.org/10.1007/s10462-017-9588-9, pp 50-73,2017.

[18]. Frederick S.M. Herz, Warrington,Lyle H. Ungar, Philadelphia; Jason M. Eisner, Baltimore,Stock market prediction using NLP US2003/0135445 AI

[19]. Joa˜o Guerreiro, • Paulo Rita,• Duarte Trigueiros2 , A Text Mining-Based Review of Cause-Related Marketing Literature, J Bus Ethics DOI 10.1007/s10551-015-2622-4, pp (1-17).2015**.**

[20]. Kranti M. Jaybhay , Rajesh V. Argiddi , S.S.Apte, Stock Market Prediction Model by Combining Numeric and News Textual Mining, International Journal of Computer Applications (0975 – 8887) Volume 57– No.19, November 2012, pp (16-22),2014.

[21]. Maecello Backmann, Stock Market Prediction Model, International Journal of Computer Applications pp 121-143 2017.

[22]. Manisha V. Pinto, Kavita Asnani , Stock Price Prediction Using Quotes and Financial News, Volume-1, Issue-5, IJSCE,pp266-269,2011.

[23]. Massimiliano Caporin 1, and Francesco Poli , Building News Measures from Textual Data and an Application to Volatility Forecasting. Econometrics, doi:10.3390/econometrics5030035, pp 1-47,2017.

[24]. Pavel Netolický1, Jonáš Petrovský1, František Dařena1, Text-Mining in Streams of Textual Data Using Time Series Applied to Stock Market. Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis**,** 66(6): 1573 – 1580. 2018

[25]. Robert P. Schumaker and Hsinchun Chen, Textual Analysis of Stock Market Prediction Using Breaking Financial News: pp (1-21),2009

[26]. S.S Abdullah, M. Rahaman and M. Rahman, Analysis of Stock Market using Text Mining and Natural Language Processing.IEEE Xpolre 978-1-4799-0400-6/13,2013**.**

[27]. S.S.Panigrahi, J.K.Mantri, Spinger pp241-268 (2015).

[28]. Sadi evren seker,cihan mert,khaled al-naami,nuri ozalp,ugur ayan time series analysis on stock market for text mining correlation of Economynews, International journal of social sciences and humanity studies, Vol 6, no 1, 2014 ISSN: 1309-8063 (online)pp(69- 91),2014.

[29]. Sheikh SaugatbAbdullah, Mohammad Saudar Rahman : Text Mining-Based Review , International Journal of Emerging Technology and Advanced Engineering ,pp69-81,(2013).

[30]. Bharathi, Angelina Geetha(2017) Sentiment Analysis for Effective Stock Market Prediction, International Journal of Intelligent Engineering and Systems, Vol.10, No.3,DOI: 10.22266/ijies2017.0630.16,pp 146-154.2017.

[31]. Spandan Ghose Chowdhury , Soham Routh , Satyajit Chakrabarti ,News Analytics and Sentiment Analysis to Predict Stock Price Trends, International Journal of Computer Science and Information Technologies, Vol. 5 (3),ISSN 0975-9646,pp 3596-3604.2014.

[32]. Tseng-Chung Tang, Hsu-Tong Deng, Li-Chiu Chi , Textual Analysis of Stock Market Prediction Using Financial News Articles and Google Search Queries International Journal of Innovative Research in Science, Engineering and Technology(An ISO 3297: 2007 Certified Organization)Vol. 4, Issue 8, August 2015, PP 6992-6994,2015.

[33]. Walter Paul Labys, Stock Market Prediction Using Natural Language Processing, Pub. No.: US 2003/0135445 A1, pp 1-10,2003.

[34]. Y.-C. Phung Text mining for stock movement predictions: a Malaysian perspective, www.witpress.com, ISSN 1743-3517,pp 103-111,2005.

[35]. Yoosin Kim , Seung Ryul Jeong , Imran Ghani , 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,pp 1-13,2014.