TEXT MINING FOR MARKET PREDITION A SYSTEMATIC REIVEW

STAT 517: STATISTICAL PREDICTIVE MODELLING – LITERATURE REVIEW

NAME: VIGNESH J MURALIDHARAN

Contents

1. INTRODUCTION

Stock market prediction is an attractive research problem to be investigated. News contents are one of the most important factors that have influence on the market. Considering the news impact in analyzing the stock market behavior, leads to more precise predictions and as a result more profitable trade. So far various prototypes have been developed which consider the impact of news in stock market predictions. Lots of investors are involved in stock market and they are all interested to know more about the future of market to be able to have more successful investments (Nikfarjam 2010). Recently some researchers have found that news are one the most influential sources that affect stock market and are necessary in achieving to more accurate predictions (Hellstrom and Holmstrom 1998). Other researchers say that ability to predict in a market economy is equal to being able to generate wealth by avoiding financial losses and making financial gains (Bollen and Mao 2011). In general, the predictive measures are divided into technical or fundamental analyses. They are differentiated based on their input data. Most of the researchers in the past have been using technical analysis approaches mainly dude to availability of quantifiable historic market data. Researchers believe that all the content of a specific asset is reflected in the price trends which makes them to use the chars and mathematical chart indicators to guide them through the investment decisions. However, for the fundamental analysis, can be applied to evaluate certain factors such as performance of company, news and environmental conditions (Hellstrom and Holmstrom 1998). When using technical analysis, we usually have structed data which can be trained easily to predict the market, rather when using fundamental analysis data from online in the textural information in social media, news, blogs and forums will be unstructured (Bollen and Mao 2011). Nevertheless, fundamental data available is unstructured text which is most challenging research aspect and therefore the focus of work.

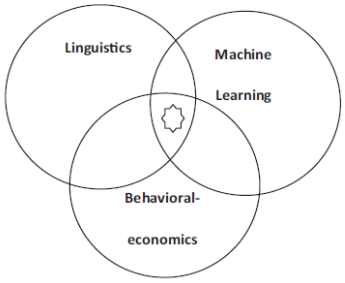

Figure 1 Interdisciplinary between linguistics, machine learning and behavioral economics

In order to address this research problem adequately some fields of study must be included namely linguistics (to understand the nature of the language), machine learning (to enable computational modelling and pattern recognition) and behavioral economics (to establish economic sense). Its considered that the main premise is aligned with recent findings of the market conditions with products of human behaviors involved (Bikas et al. 2013). The other related background topics involve as follows.

1.1 Efficient Market Hypothesis (EMH):

States that the current market price reflects the assimilation of all the information available. This means that given the information, no prediction of future changes in the price can be made. As new information enters the system the unbalanced state is immediately discovered and quickly eliminated by a correct change in market price (Hellstrom and Holmstrom 1998).

1.2 Behavioral economics:

Most technology manufactures will show a base model with options that can be changed according to the buyer's preferences. The way in which these product choices are presented

to buyers will influence the final purchases made and illustrates several concepts from behavioral economic theories ("An Introduction to Behavioral Economics" n.d.).

1.3 Adaptive Market Hypothesis (AMH):

Is based on an evolutionary approach to economic interactions, as well as some recent research in the cognitive neurosciences that has been transforming and revitalizing the interactions of psychology and economics. Although some of these ideas have not yet been fully articulated within a rigorous quantitative framework, long time investment professionals will no doubt recognize immediately the possibilities generated by this new perspective. Only time will tell whether its potential will be fulfilled (Lo 2004).

1.4 Fundamental vs Technical Analysis:

Fundamental analysis calculates future price movements by looking at a business's economic factors, known as fundamentals. It includes economic analysis, industry analysis and company analysis. This type of investing assumes that the short-term market is wrong, but that stock price will correct itself in long run. Profits can be made by purchasing a mispriced security and then waiting for the market to recognize its mistake. It is used to buy and hold investors and value investors, among others. Fundamental analysis looks at financial statements, including balance sheets, cash flow statements and income statements, to determine a company's intrinsic value.

Technical analysis uses a security's past price movements to predict its future price movements. It focuses on the market prices themselves, rather than other factors that might affect them. It ignores the value of the stock and instead considers trends and patterns created by investors emotional responses to price movements. Technical analyses look only at charts, as it believes that all a company's fundamentals are reflected in the stock price. It looks at models and trading rules based on price and volume transformations, such as the relative strength index, moving averages, regressions, inter-market and intra market price correlations, business cycles, stock market cycles and chart patterns (Petrusheva and Jordanoski 2016).

1.5 Sentiment and Emotional Analysis:

It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining and text mining. The growing importance of sentiment analysis coincides with the growth of social media such as reviews forum, microblogs, twitter and social networks (Liu 2012). Sentiment analysis systems are being applied in most every business and social domain because opinions are central to all human activities and key influencers of our behaviors. There are lots of development in the sentiment analysis in present situations the emotions are based on the Minsky's conception of emotions that consist of four affective dimensions (Pleasantness, Attention, Sensitivity and Aptitude). Each dimension has six levels of activation, called sentic levels. Each level represents an emotional state of mind and can be more or less intense, depending on the position on the corresponding dimensions (Li and Xu 2014).

2. RESEARCH GOALS

The main contributions of this work will be summarizing the relevant fundamental economic and computer/information science concepts and to clarify how it is tied up with present situations. Observations on areas with lack of research which can constitute possible opportunities for future work were also discussed in this paper.

3. MAIN WORK

Despite having multiple systems in this area or research the author couldn't find any dedicated and comprehensive comparative analysis of the available systems. (Khadjeh Nassirtoussi et al. 2014) has

tried with outlining the major systems that have been developed around the world. The roles and theoretical foundations of each process in the text mining with an outline is as follows.

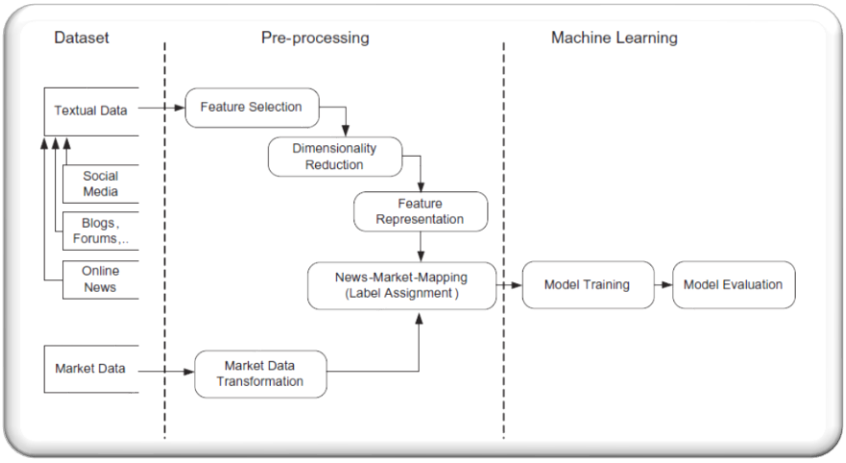

Figure 2: Generic common system components of the process of text mining and prediction system

3.1 Data Input (Text Data & Market Data): Any research using textual data begins with the researcher identifying the corpus of text relevant to the research question of interest. From this corpus, text will be sampled or selected for analysis. Text are generally distinguished from one another by attributes relating to the author and separated by time, topic, act. Many users of textual data find it useful to process raw text prior to analysis, with an aim to identifying more appropriate units of analysis than is offered by the raw text. The most common types of preprocessing are reduction of words to their word stems or lemmas or elimination of words using stop list or based on relative frequency. Other methods of preprocessing textual data aimed at generating units of analysis from sampling units include converting text to "n-grams" defined as sequences of n consecutive items, usually words in political science applications thereby distinguishing command economy from market economy (Benoit n.d.). The authors have taken the data in the form of text data and from the market data form. Some of the data are from Online message boards, twitter post and mood analysis for market predictions, companies' annual reports and frequency of recent news. Some of the market data are taken from the foreign exchange market (FOREX), NASDAQ and stock price of companies like google, yahoo and Microsoft. Almost all the forecast types on any market measures are categorical and discrete values. There are very few researches which have been explored based on linear regression. The concept of high frequency trading which is explored is very similar to term in low latency trading which his amplified in the work of some authors. Some of the researcher's experiments are based on time periods and the annual timeframes of market efficiency convergence time.

3.2 Pre-Processing: Once we have the input data available it is prepared to analyze into the machine learning algorithm. There are at least three sub processes or aspects of pre-processing which have been reviewed in the paper namely feature selection, dimensionality reduction and feature representation. Feature selection aids us to create an accurate predictive model. This can be used to identify and remove unneeded, irrelevant and reductant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. It tries to find a subset of the original variables. Having a limited number of features is extremely important as the increase in the number of features which can easily happen in feature selection in text can make the classification or clustering problem extremely hard to solve by decreasing the efficiency of most of the learning algorithm this

situation is widely known as the curse of dimensionality (Nikfarjam 2010). After the minimum number of features is determined each feature needs to be represented by a numeric value so that it can be processed by machine learning algorithms. These assigned numeric value acts like a score or a weight. The most basic techniques have been used when dealing with text mining-based market predictions are listed in the table below. Generally, in text mining enhanced feature reduction and feature weighting can have significant impact on the eventual text classification efficiency (SHI et al. 2011).

Table 1: Techniques representation of Feature selection, reduction and representations

Reference / Author	Feature Selection	Dimensionality reduction	Feature representation
Wuthrich et al. (1998)	Bag of words	Predefined dictionaries	Binary
Soni et a.(2007)	Visualization	Thesaurus made using term extraction tool	Visul Coordinates
Tetlock et al. (2008)	Negative words usuage	Psychology dictionary predefined dictionaries	Frequency divided by total number of words
Butler et al (2009)	Charcter n-grams method	Minimum occurrence per document	Frequency of the n-gram in one profile
Schumaker and Chen (2009)	Noun Phraces	Minimum occurrence per document	Binary
Jin et al (2013)	Latent Dirichlet Allocation (LDA)	Topic extraction, news articles fluctuations	Topic distribution

3.3 Machine Learning: After the preprocessing is completed and text is transformed into several features with a numeric representation, machine learning algorithms can be engaged. At its most basic, machine learning uses programmed algorithms that receive and analyses input data to predict output values within an acceptable range. In this paper it is attempted to provide a summary of the algorithms used in reviewed works. The main objective is to report what issued so that it helps understand what lacks and may b possible for future research. Basically, the systems are using the input data to learn to classify an output usually in terms of the movement of the market in classes such as up, down and steady. Some of the most famous reviewed works-based algorithms are as follows. Some of the famous works referred by the authors are listed in the table below with the techniques used by each reference also listed.

3.3.1 Support Vector Machine (SVM): Is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side but, in a high dimensional (Kernel map) data its easily transformed in the space which is easy to distinguish.

3.3.2 Regression Algorithms: Is perhaps one of the most well known and well understood algorithms in statistics and machine learning. More specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible (Bollen and Mao 2011). This is studied as a model for understanding the relationship between input and output numerical variables but has been borrowed by machine learning.

3.3.3 Naïve Bayes: Is a family of probabilistic algorithms that take advantage of probability theory and Bayes theorem to predict the tag of text. The probabilities are by using Bayes theorem which describes the probability of a feature, based on prior knowledge of conditions that might be related to that of future.

3.3.4 Decision Rules and Trees: It is a map of the possible outcomes of a series of related choices. It allows an individual or organization to weigh possible actions against one another based on their cost, probabilities and benefits. They can be used either to drive informal discussion or to map out an algorithm that predicts the best choice mathematically (Petrusheva and Jordanoski 2016). It typically starts with single node which branches into several like a tree shape. Induced from training data in a bottom up

specific to general style or a top down approach. The initial sate of a decision rule solution is indeed the collection of all individual instances. This is a non-parametric supervised learning method used for both classification and regression. This requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. The cost of using the tree is logarithmic in the number of data points used to train and test.

3.3.5 Combinatory Algorithms: These are computational procedures which are designed to help solve combinatorial problems involving arrangements of elements from a finite set and selections from a finite set. In many such problems, exhaustive search is not feasible, it operates on the domain of those optimization problems in which a set of feasible solutions is discrete or can be reduced to discrete and in which the goal is to find the best solutions ("Combinatorial Problems" n.d.). Some common problems to solve is through this method. Sometimes multi-algorithm experiments are taken care whereby the same experiments are conducted using several different algorithms.

Table 2: Classification algorithms for Machine learning Techniques

Reference / Author	Algorithm Type	Algorithm Details / Used for
Pui Cheong Fung et al. (2003)	SVM	SVM-Light
Soni et al. (2007)		SVM with Stnadard linear kernel
Hagenau et al. (2013)	Regression Algorithms	SVM with linear kernel, SVR Used to predict the discrete value of the stock return
Chatrath et al.(2014)		Stepwise Sequential Minimal Optimization Used to allow discrete mumeric prediction instead of classification
Yu, Duan, et al. (2013)	Naïve Bayes	Naïve Bayes, Used for sentiment analysis of social media along with effect of conventional media
Li (2010)		Naïve Bayes and dictionary based Used to examine the information content of the forward looking statements in management discussion of the company
Peramunetilleke (2002)	Decision Rules or Trees	Rule classifier, Used to expressing correlation between keywords
Rachlin et al. (2007)		C 4.5, Used to predicting rules with combination of numerical and textual data
Das and Chen (2007)	Combinatory Algorithms	Different classifiers, predict Annual company reports
Butler and Huina (2011)		Self Organizing fuzzy Neural Network, Detect patterns in textual data

3.3.6 Training vs Testing Volume and Sampling: Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing we can minimize the effects of data discrepancies and better understand the characteristics of the model. Typically, most of the data will go towards your training data, while only 10-25% of the data for testing. Training informs a model on how it should work and make its decisions. Testing just gives us an idea of how well the classifier in performing. Two aspects are summarized if the information were available. If the sampling for training and testing was of a special kind, it is specially of interest here is to know if a linear sampling has been followed as in essence the samples are on time series. Some works have clearly mentioned the sampling type like stratified.

3.3.7 Sliding Window Market Analysis: Time series data analysis, such as temporal pattern recognition and trend forecasting, plays an increasingly significant part in temporal data statistics and analysis. Yet challenges still exist in the efficiency of pattern extracting and trend prediction for large multivariate time series ("Sliding window analysis | Coleopterists Corner" n.d.). The paper explains a multi stage clustering approach towards by using dynamic sliding time windows. The segmented series are clustered separately in each time window to product first stage clustering centers, and which are used to generate second stage clustering results involving all time windows. This method can simplify large scale time series mining problems through multistage thus achieve improved efficiency (de Brito and Oliveira 2014). Although it intuitively seems necessary to implement a sliding window, there are very few of the reviewed works which have. This seems to be an aspect that can receive more attention in the future systems of text mining (Borovikov and Sadovsky^ 2014).

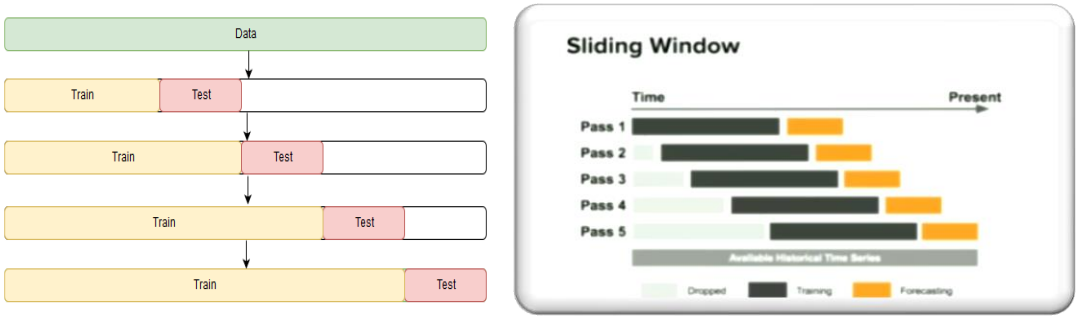

Figure 3 : Difference between Fixed and Sliding window Market Analysis for Time Series Data

3.3.8 Semantics and Syntax: The problem of finding the semantic mappings between two given ontologies lies at the heart of numerous information processing applications. Virtually any application that involves multiple ontologies must establish semantics mappings among them, to ensure interoperability. Examples of such applications arise in myriad domains, including e-commerce, knowledge management, eLearning and tourism. Tackling semantics in an important issue and research efforts are occupied with it in several fronts. A corpus is the task of building structures that approximate concepts from a large set of documents with Latent semantic analysis which has class of techniques with documents represented in term space.

Syntax is also very important and proper observation and utilization of it along with semantics which can improve textual classification accuracy. It is interesting to note that there are several approaches to such syntax-based pattern recognition methods. In the word occurrence method, a pattern is considered as a word that appears in a document. Whereas in the word sequence method, it consists of a set of consecutive words that appear in a document. Topic models such as latent Dirichlet allocation are generative models that allow documents to be explained by unobserved latent topics. The hidden markov model is topic model that simultaneously models topics and syntactic structures in a collection of documents. The idea behind the model is that typical word can play different roles. It can either be part of the content and serve in a semantic or as syntactic structure. Each word gets assigned to a syntactic class, but one class is reserved for the semantic words. Words in this class behave as they would in a LDA topic model, participating in different topics having certain probabilities of appearing document. Some works include part of speech tagging and achieve phrase level sentiment analysis.

3.3.9 Ways of Evaluation: Most of the works are presenting a confusion matrix or parts thereof to present their results. Calculating accuracy, recall or precision ration and sometimes the F-measure being the most common. The accuracy in majority of the cases is reported in the range of 50-70% while arguing for the better than chance results which is estimated at 50%. In the common evaluation approach above 55% have been considered a report-worthy in other parts of literature as well. However, what makes most of the results in this paper a questionable is majority of them surprisingly have not examined or reported if the experimental data is imbalanced or not. The real problem arises, when the cost of misclassification of the minor class samples are very high. Usually when using the trading period we need to simulate the profits and should be measured to evaluate the viability of the system.

$$Accuracy = \frac{\#\ of\ correct\ prediction}{(total\ \#\ of\ predictions\ made)}$$

4. CONCLUSION

Emotion classification has a broad range of applications. An accurate and efficient classification system is of great interest in this study. Behavioral finance with respect to human and social recognition and emotional tolerance studies were identified and understood incoming economic decisions. The review was conducted on three major aspects, namely: pre-processing, machine learning and evaluation mechanism with each breaking down into multiple sub discussions. It is believed to be the first effort of a comprehensive review made with this kind of research. This way of text mining can be used in several platforms like investment banks and financial institutions which can use for predicting specialized market. A focus on market predictive text mining in this research helps the formation of this emerging field as a recognizable and independent field that can be delved into vigorously and not only in shadow of general text mining research. The formation of such independent field of research for market predictive text mining, distinct from product review sentiment analysis or such is a hopeful implication of this work. This work is hoped to help other researchers put the various ideas in this field into perspective more conveniently and become able to make strategic decisions upfront in the design of the future systems.

5. FUTURE WORK

For future work, more linguistics patterns need to be investigated. Market predictions mechanism based on online text mining are just emerging to be investigated rigorously utilizing the radical peak of computational processing power and network speed in the recent times. This research helps put into perspective the role of human reactions to vents in the making of markets and can lead to a better understanding of market efficiencies and convergence via information absorption. In summary, this work identifies the above areas or aspects in need of future research and advancement.

6. REFERENCES

"An Introduction to Behavioral Economics" (n.d.). Available athttps://www.behavioraleconomics.com/resources/introduction-behavioral-economics/.

Benoit, K. (n.d.). "Data, Textual," in *International Encyclopedia of Political Science*, 2455 Teller Road, Thousand Oaks California 91320 United States : SAGE Publications, Inc. https://doi.org/10.4135/9781412959636.n127.

Bikas, E., Jurevičienė, D., Dubinskas, P., and Novickytė, L. (2013), "Behavioural Finance: The Emergence and Development Trends," *Procedia - Social and Behavioral Sciences*, Elsevier, 82, 870–876. https://doi.org/10.1016/J.SBSPRO.2013.06.363.

Bollen, J., and Mao, H. (2011), "Twitter Mood as a Stock Market Predictor," *Computer*, 44, 91–94. https://doi.org/10.1109/MC.2011.323.

Borovikov, I., and Sadovsky^, M. G. (2014), *Sliding Window Analysis of Binary n-Grams Relative Information for Financial Time Series*.

de Brito, R. F. B., and Oliveira, A. L. I. (2014), "Sliding window-based analysis of multiple foreign exchange trading systems by using soft computing techniques," in *2014 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 4251–4258. https://doi.org/10.1109/IJCNN.2014.6889874.

"Combinatorial Problems" (n.d.). Available athttps://www8.cs.umu.se/kurser/TDBA77/VT06/algorithms/BOOK/BOOK4/NODE147.HTM.

Hellstrom, T., and Holmstrom, K. (1998), "7_Predicting the stock market," *Research and Reports Opuscula ISRN HEV-BIB-OP $\$26-SE, Malardalen University, Vasteras, Sweden, 1998*, 37. https://doi.org/10.1.1.57.4327.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. (2014), "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, Elsevier Ltd, 41, 7653–7670. https://doi.org/10.1016/j.eswa.2014.06.009.

Li, W., and Xu, H. (2014), "Text-based emotion classification using emotion cause extraction," *Expert Systems with Applications*, Pergamon, 41, 1742–1749. https://doi.org/10.1016/J.ESWA.2013.08.073.

Liu, B. (2012), "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, 5, 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Lo, A. W. (2004), *The Adaptive Markets Hypothesis: Market Efficiency from an Evolutionary Perspective **

Nikfarjam, A. (2010), "Text mining approaches for stock market prediction - Nikfarjam - 2010.pdf," 4, 256–260.

Petrusheva, N., and Jordanoski, I. (2016), *COMPARATIVE ANALYSIS BETWEEN THE FUNDAMENTAL AND TECHNICAL ANALYSIS OF STOCKS*, *JPMNT) Journal of Process Management-New Technologies, International*.

SHI, K., HE, J., LIU, H., ZHANG, N., and SONG, W. (2011), "Efficient text classification method based on improved term reduction and term weighting," *The Journal of China Universities of Posts and Telecommunications*, No longer published by Elsevier, 18, 131–135. https://doi.org/10.1016/S1005-8885(10)60196-3.

"Sliding window analysis | Coleopterists Corner" (n.d.). Available athttp://coleoguy.blogspot.com/2014/04/sliding-window-analysis.html.