Text mining approaches for stock market prediction

Conference Paper · March 2010

DOI: 10.1109/ICCAE.2010.5451705 · Source: IEEE Xplore

3 authors:

Azadeh Nikfarjam
Arizona State University
28 PUBLICATIONS **1,792** CITATIONS
SEE PROFILE

Ehsan Emadzadeh
Arizona State University
11 PUBLICATIONS **192** CITATIONS
SEE PROFILE

Saravanan Muthaiyah
Multimedia University
88 PUBLICATIONS **505** CITATIONS
SEE PROFILE

Text mining approaches for stock market prediction

Azadeh Nikfarjam
Faculty of IT, MMU
Cyberjaya, Malaysia
azadeh.nikfarjam@gmail.com

Ehsan Emadzadeh
Faculty of IT, MMU
Cyberjaya, Malaysia
eemadzadeh@gmail.com

Saravanan Muthaiyah
Faculty of Management, MMU
Cyberjaya, Malaysia
saravanan.muthaiyah@mmu.edu.my

Abstract— **Stock market prediction is an attractive research problem to be investigated. 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.**

Keywords: News Mining, knowledge discovery, stock market prediction, data mining

I. INTRODUCTION

Stock market analysis is one of the interesting areas of research. 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. Effective market prediction can help investors with trade advices or can be used as a component inside automatic trader agents. Sometimes prediction systems indirectly help traders by providing supportive information such as the future market direction. E.g. if the direction of a selected stock during 24 hours is predicted to be “up”, buying the stock would be a profitable trading action.

During decades analyzing of stock market was just based on historical market prices. Autoregressive, moving average [1], Genetic Algorithm [2], Neural Networks [3], Support Vector Machines [4], Case-based reasoning [5] and other techniques have examined for analyzing stock market behavior. The problem against precise predictions in these approaches, is modeling the random behavior of the market while there is no justification for it. One of the solutions in justifying the random behavior of the market is considering the impact of un-quantifiable events on the market. Recently some of the researchers have found that news are one the most influential sources that affect stock market and are necessary in achieving to more accurate predictions.

Before further discussion on market prediction methods, it is worthwhile to answer to the question that whether stock market prediction is feasible or not? Some financial specialists believe in Efficient Market Hypothesis (EMH) which states that “stock prices reflect all their relevant information at any given point in time [6][7]”. This means that the stock prices include historical data and general information as well as the private information at any moment which will make it impossible to be predictable. However obtained successes in stock market prediction proved that market forecasting is possible [8][9]. In fact it takes time for the market to adjust itself to the new incoming information. Almost all the successful prediction methods have take advantage of this golden time gap.

Another arising questions while studying news based prediction systems, is about the kind of information that should be looked for in mining the news. To answer this question we shall know the factors that cause special behavior in stock market. There are several factors that influence stock market behavior. In fundamental analysis [10] of the market, some of these factors are considered such as company economic growth, inflation, unemployment, earnings and etc. A successful news analysis would be achieved if the effective information about stock could be extracted from the news content.

To investigate how the news impact can be used in stock market prediction, different previous researches studied and a general flow of the comprising processes is presented. Furthermore some previous researches are compared in terms of implementing the main processes. Basically automatic text classification techniques are used to analyze the incoming news. In addition in some approaches numerical parameters related to stock price are also include to increase prediction accuracy.

First we mention some aspects of the market that is usually considered in designing marker prediction systems. Second we propose the main processes included in a stock prediction system. Finally we will compare existing prediction systems together and introduce the potential future research areas in this field.

II. PREDICTION SYSTEMS

In categorizing the stock market prediction systems different dimensions can be considered:

Input data: Some prediction methods are based on historical market prices and use technical analysis to predict the market. Some other methods are based on analyzing the news content; however combination of historical market data and news can also be used.

978-1-4244-5586-7/10/$26.00 © 2010 IEEE

Prediction goal: The possible market prediction goal can be the future stock price or the volatility of the prices or market trend. Market trend is the general direction of the stock's prices which is upward or downward. Market volatility is defined in [7] as: "the amount of uncertainty or risk about the size of changes in a security's value". A higher volatility means the higher fluctuation of the corresponding stock prices.

Prediction horizon: prediction horizon is the time span in which the prediction would be valid. It can be short-term or long-term prediction. Short-term prediction starts from 5 minutes to 1 hour after the news release and long term starts from 24 hours and can last longer.

A. news based prediction system processes

News based stock market prediction can be considered as a text classification task. Generally the goal is to forecast some aspects of the stock market such as price or volatility based on the news content. Based on prediction goal described in previous section, a set of final classes are defined, such as "Up" (which means this news cause the prices to go up), "Down" (which means this piece of news is probable to causes decrease in prices) and etc. the prediction system is supposed to classify the incoming news into one of these classes.

News based market prediction can be divided into two main phases. "Training phase" and "Operational phase". In operational phase, one of the predefined classes will be assigned to incoming news; however, to make the system ready for the operational phase a classifier should be trained in the training phase. Machine learning techniques are widely used to automate such processes. As a part of the training phase, a set of training data shall be prepared which in our case the train data are the pre-classified news and market information such as market prices. These labeled (pre-classified) news and possibly market numerical data will be processed to be fed into the classifier for training. The trained classifier would be ready to get a piece of news and assign a class to it in operational phase of the system.

Generally, such predictive systems consist of following components which are depicted in Figure 1:

- News labeling
- Classifier input generation
- Classification

1) News labelling

Based on the selected prediction goal, a set of classes are defined for the news and the attempt is to identify the class that each news article belongs to and label them accordingly.

There are two ways to assign labels, manually and automated. In the first, financial market experts will read the news and assign a class based on their opinion. In automated assigning, time stamped numerical market data is analyzed to determine the right class for a piece of news [11][12][13]. Usually a time interval around the news release will be selected and the prices are analyzed in that interval to determine the news impact. For example Fung et al. [14] divided the time series data into independent segments and labeled the segments according to its average slope. Mittermayer and Knolmayer [15] labeled the news based on the percentage change of the price 15 minutes after the news release. If a news article caused at least 3 % increase in the price, it is labeled as "BUY" and labeled "Short" if decreases 3% respectively.

Prediction goal, number of pre-defined classes and prediction horizon are the aspects that affect the method applied in news labeling.

- Prediction goal: If the interest is to detect the impact of news on market volatility, the price changes should be analyzed and the corresponding labels such as "High impact", "Low impact" and etc are assigned to news. On the other hand, if the classification is based on market trends (whether the market will go up or down), prices are analyzed accordingly to discover the news impact on the market direction.
- Number of predefined classes: Number of final classes can affect determining the criteria for assigning a class to a piece of news. E.g. if the number of final classes are 2, then usually the prices in a time range after or before the news release will be compared with the price at the time of news release to decide about the news label [15][16]. On the other hand, if the number of final classes are more than two the percentage of price change in the defined range is used to determine the correct label [14].
- Prediction horizon: Prediction horizon affects selecting time interval to process the market prices e.g. if the goal is to predict the news impact on tomorrow close price, then the horizon will be 24 hours[17], [12] which means during 24 hours of the news release, the prices or other market technical indicators are watched to analyze the impact of the news; On the other hand for short term prediction the time interval can be 5 to 20 minutes after the news release[15],[13].

2) Classifier input generation

The success of any classifier in generating accurate results is highly dependent to the way that the input data are presented. Specific features of the input data are selected to represent the whole document. In classifying the news two important factors should be considered; the news content and the numerical market data such as stock prices. Both of these factors shall be considered for classifier input generation. A summary on applied methods is shown in Table I.

Regarding the comprising elements in the input vector in vector based classifiers, two main approaches are followed.

- News Content
- Combination of news and numerical market data

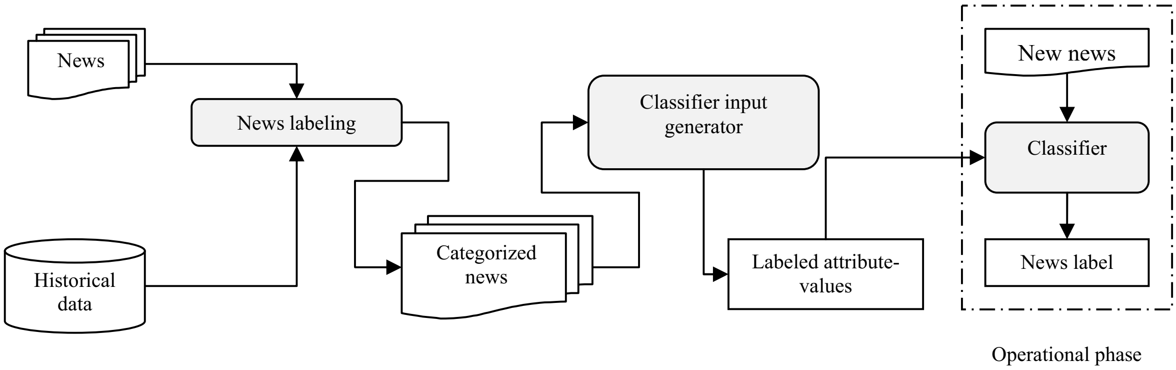

Figure 1. Overall view of a news based stock prediction system

In the first approach the news content is used as input data source, while in the second approach market data such as stock price at the time of news release [13], closing price and change indicator values [18] are included in the classifier input.

Numerical data are the representatives of the real stock market situation around publication of news. By comprising more expressive data about the stock in the input vector, the classification accuracy can be improved. In fact the classifier is expected to find the answer of this question: the current situation of the market is x, and the current news has the content y, so what would be the possible impact of this piece of news on the selected stock?

Representing the news section consists of two main tasks: feature selection and feature weighting. First a set of features will be selected to represent a piece of news and next step is to assign weights to theses selected features. These weighted vectors would be the inputs to classifier.

a) Feature selection

Features are the representatives of a document in a classification problem. Based on the classification goal a set of features from document should be selected which best convey the document content. According to Table I we can see that so far in feature selections two main approaches has been followed.

Generating a term dictionary [8]: A set of terms are gathered and used as the fixed vector elements. Usually a group of financial experts select the representative terms; for each category there exists a set of special vocabulary that if exist in a document the probability of belonging the document to that category would be higher.

Bag of words: in this method all the training news words are extracted. The stop words are removed. Sometimes stemming is done in which the stem of each word is replaced with the original word [19]. Usually by applying tf-idf the terms with highest meaning contribution will be selected as representatives. [16]

Some of the researchers have also tried to use semantic approaches in feature selection such as [17] in which the authors applied concept map or [18] which used Latent Dirichlet Allocation (LDA) based topic extraction mechanism for feature selection.

b) Feature Weighting

Feature weighting is the process of assigning values to the selected terms. In some cases just the presence of a term in a document is important. Therefore the boolean values are used for weighting. Rachlin et al. [11] used the words with higher degree of membership as input features and binary representation for weighting is applied. However by assigning non boolean values, classification can be more accurate. Usually tf-idf is used to calculate the weights [12]. Tf-idf weighting method can be enhanced in different ways. Fung et al.[14] enhanced tf-idf weights by assigning more weights to the terms presented in just one final class. Mittermayer and Knolmayer[20] enhanced tf-idf weighting by multiplying Within Document Frequency (WDF) and tf-idf to weight input features.

3) Classification

Classification is analyzed from two aspects:

- Number of classes
- Classification algorithm

Based on studied approaches so far, 2, 3 or 5 classes are defined for predicting market behaviors. In processing the news generally the goal is to classify the news into two classes either good news or bad news regarding the selected stock [17], [12].Sometimes this classification is extended and another category indicating neutral news is added [14], [13], [15]. If the degree of news influence is important to be identified more final categories will be defined [11].

In most of the methodologies the authors have selected Support Vector Machine (SVM) [21] as their classification algorithm. SVM is a binary classifier which tries to classify the input data by defining a hyperplane or a set of hyperplanes in high-dimensional space. SVM tries to maximize the distance of the hyperplane with the nearest data points of each class. Considering Table II, it can be observed that just SVM and Bayesian classifiers are used in this area. Some methods used combination of binary classifiers to achieve their final classification decision.

TABLE I. SUMMARIZED "CLASSIFIER INPUT GENERATION" METHODS IN STUDIED SYSTEMS

Developed systems	**Feature selection technique**	**Feature representation [weighting]**	**Included elements**	**News Source**
Fung/Lam/Yu [14]	Bag of words Using IBM Intelligent Miner^{TM} for Text	Enhanced TF-IDF: $tf_{t,d} \times CDC \times CSC$	News terms	Reuters Market 3000 Extra
Mittermayer/Knolmayer [15]	Bag of words : defining local dictionaries by selecting most important terms using TF-IDF	Boolean measure of frequency	News terms	press releases provided by PRNewswire
Soni/Eckt/Kaymakt[17]	Concept map	Coordinates of news in the document map	News terms	Financial Times online service, FT Intelligence
Zhai/Hsu/Halgamuge [12]	Bag of words: Term dictionary of documents containing concepts generated(using WordNet) and un important concepts are removed(using TF-IDF)	Binary values	News terms and Technical indicators	Australian Financial Review newspaper
Rachlin/Last/Alberg/Kandel [11]	Pre-defined Term dictionary	TF-IDF weighting	News terms and financial values	Forbes and Reuters websites
Schumaker / Chen [13]	Bag of words, noun phrases, name entities	TF-IDF weighting	News terms and stock price at news release	Yahoo Finance
Mahajan/Dey/Mirajul Haque [18]	Latent Dirichlet Allocation (LDA) based topic extraction mechanism	Correlation between market parameters and extracted topics	News extracted events and Closing value, change percentage and volatility of the stock	Financial news acquired from captialmarket.com website
Lu/Huang/Zhang/Chen [22]	Bag of words modified by part of speech (POS) tags	Not mentioned	News terms	Wall Street Journal

III. Comparison Of Previous Methods

In previous sections various methods applied in news based stock prediction approaches discussed. Summary of the approaches are compared in Table I and Table II. Table I shows the applied methods in generating classifier input generation. In section 2 “input data” was mentioned as one of the stock prediction aspects. Most of the methods have just used the news content as their classifier input data.

However in [12] Zhai et al. trained a classifier dedicated for 7 technical indicators in addition to another classifier for the news classification and finally combined the result of these classifiers. In the method proposed in [13], stock price at time of the news released included in the input vector which increased the prediction accuracy. In addition, [18] used some price indicators in their classifier input vector.

Among proposed methods for feature selection and weighting, just [17] and [18] followed semantic approaches. The rest of the methods used Bag of words as their feature selection in which basically tf-idf technique is applied in term selection and weighting. To improve feature selection [16], [12] and [11] included some pre-defined terms in the classifier input vector and reported that it increased the classification accuracy.

TABLE II. SUMMARIZED CLASSIFICATION AND EVALUATION TECHNIQUES IN STUDIED METHODS

	Classification		**Evaluation**		
Developed systems	***Classifier***	***Number of categories***	***Period***	***Directional accuracy***	***Simulated trader***
Fung/Lam/Yu [14]	SVM	3	7 months		yes
Mittermayer/Knolmayer [15]	SVM	4(3 for training)	9 months	45%	yes
Soni/Eckt/Kaymakt[17]	SVM	2	11 years	56.2%	N/A
Zhai/Hsu/Halgamuge [12]	SVM	2	15 months	70.01%	yes
Rachlin/Last/Alberg/Kandel [11]	Decision tree	5	3 months	82.4%	yes
Schumaker/Chen [13]	SVM	3	1 month	58.2%	yes
Mahajan/Dey/Mirajul Haque [18]	stacked classifier(Decision tree+ SVM)	2	3 years	60%	no
Lu/Huang/Zhang/Chen [22]	maximum entropy (Maxent) model and SVM	N/A	8 years	F-measures between 50.2% and 78.4%	no

Furthermore [14], [12] and [22] modified tf-idf weighting in order to assign higher values to some existing terms in news.

IV. Conclusion And Future Suggestions

In this paper we presented different methods that have been applied to investigate the impact of financial news on stock market prediction. A general flow of processes that is followed by most of the methods is presented and discussed.

As discussed in the "news labeling" section, in determining the news labels, generally the prices are monitored in a time interval around the publication of the news; however technically analyzing the market around the publication of the news can lead to more precise labels. Technical indicators are useful tools that illustrate the real market situation. Using the values of the technical indicators before and after the news release can be more informative than using pure prices. In most of the studied methods, a barrier is defined whereby if the stock price touches that barrier the final decision about assigned class will be made. However sometimes there are some sharp jumps in the prices that are outliers and the prices will be adjusted to their real flow for the next tick, so assigning the class based on just pure prices can be imprecise.

Considering Table I, in generating classifier inputs, it can be observed that most of the developed systems have just used news features and not efficiently used the market data such as prices or technical indicators [23]. Although in [12] authors have exploited seven technical indicators, but indicator values and news are analyzed separately and the analysis results are combined. However if classifier input includes both news and market status at the same time it may end in more accurate results.

Overall, the gap of semantic analysis of the news content can be observed. Some of the researches tried to apply semantics in their classification methods [18] [17]. It seems that by using the novel semantic methods in text classification, a promising result would be achieved for the discussed problem.

References

[1] T. Hellström and K. Holmström, "Predicting the Stock Market," *Technical Report Series IMATOM- 1997-07*, 1998.

[2] K. Kyong-jae and I. Han, "Genetic algorithms approach to feature discretization in aftificial neural networks for the prediction of stock price index," *Expert Systems with Applications*, vol. 19, 2000, pp. 125-132(8).

[3] T.S. Quah and B. Srinivasan, "Improving Returns on Stock Investment through Neural Network Selection," *Expert Syst. Appl.*, vol. 17, 1999, pp. 295-301.

[4] W. Huang, Y. Nakamori, and S. Wang, "Forecasting stock market movement direction with support vector machine," *Computers & Operations Research*, vol. 32, 2005, pp. 2513-2522.

[5] S. Chun and Y. Park, "Dynamic adaptive ensemble case-based reasoning: application to stock market prediction," *Expert Systems with Applications*, vol. 28, 2005, pp. 435-443.

[6] E.F. Fama, "Market Efficiency, Long-Term Returns, and Behavioral Finance," *Journal of Financial Economics*, vol. 49, 1998, pp. 283-306.

[7] E.F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *The Journal of Finance*, vol. 25, 1970, pp. 383-417.

[8] R.A. Haugen, *The new finance : the case against efficient markets*, New Jersey: Prentice Hall, 1999.

[9] M. Kaboudan, "Genetic programming prediction of stock prices," *Computational Economics*, vol. 16, 2000, p. 207–236.

[10] E. Faerber, "Fundamental Analysis," *All about stocks: the easy way to get started*, McGraw-Hill, 2000, pp. 129-168.

[11] G. Rachlin, M. Last, D. Alberg, and A. Kandel, "ADMIRAL: A Data Mining Based Financial Trading System," *2007 IEEE Symposium on Computational Intelligence and Data Mining*, 2007, pp. 720-725.

[12] Y. Zhai, A. Hsu, and S. Halgamuge, "Combining News and Technical Indicators in Daily Stock Price Trends Prediction," *Lecture Notes in Computer Science*, 2007, pp. 1087-1096.

[13] R.P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news:The AZFin Text system," *ACM Transactions on Information Systems*, vol. 27, 2009, pp. 1-19.

[14] G. Fung, J. Yu, and W. Lam, "News sensitive stock trend prediction," *Lecture Notes in Computer Science*, vol. Volume 233, 2002, p. 481–493.

[15] M. Mittermayer and G.F. Knolmayer, "NewsCATS: A News Categorization And Trading System," *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*, 2006, pp. 0-5.

[16] M. Mittermayer, "Forecasting intraday stock price trends with text mining techniques," *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, vol. 00, 2004, pp. 64-73.

[17] A. Soni, N.V. Eck, and U. Kaymak, "Prediction of stock price movements based on concept map information," *IEEE Symposium on Computational Intelligence in Multicriteria Decision Making*, 2007, pp. 205-211.

[18] A. Mahajan, L. Dey, and S.M. Haque, "Mining Financial News for Major Events and Their Impacts on the Market," *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Ieee, 2008, pp. 423-426.

[19] N. Fuhr, "Probabilistic Models in Information Retrieval," *The Computer Journal*, vol. 35, 1992, pp. 243-255.

[20] M. Mittermayer and G. Knolmayer, *Text mining systems for market response to news: A survey*, 2006.

[21] T. Mitchelle, *Machine Learning*, McGraw Hill, 1997.

[22] H. Lu, N.W. Huang, Z. Zhang, T. Chen, and W. District, "Identifying Firm-Specific Risk Statements in News Articles," pp. 42-53.

[23] S.B. Achelis, *Technical Analysis from A to Z*, New York: McGraw Hill, 2001.