Stock Market Forecasting Techniques: A Survey

Arti Buche and M.B. Chandak
Department of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

Abstract: This study surveyed and analyzed the prediction of stock price movement using different sentiment analysis techniques which is given as an input to different machine learning methods. Using predictive analysis and technical analysis various technical and economic indicators can be leveraged with the help of opinions extracted from the text in order to identify and optimize the stock price movement decisions to maximize trading profits.

Key words: Stock price movement, sentiment analysis, machine learning, indicators, maximize trading, text

INTRODUCTION

Stock Analysis is basically an evaluation process of a particular trading instrument an investment sector or the market as a whole (Ahn and Cho, 2010). The stock analysts determine the future activities of investment sector, market or instrument (Peramunetilleke and Wong, 2002). It is a method for investors and traders to make stock decisions by studying and evaluating historic and current data. The investors and traders attempts to gain an edge by making informed decisions in the market (Fung *et al.*, 2005).

Predictive analytics is deeply integrated into our current society as it plays a large role in the normal consumer's daily routine. Its starts from protecting email, to what insurance to be paid, stock decisions, thus, the outcome is determined with use of this technology (Fung *et al.*, 2002).

The focus of this study is on the technical analysis of the stock movement. The technical analysts analyze the market and also are more concerned with price and volume as well as demand and supply factor that move the market (Lavrenko *et al.*, 2000). To perform these processes many techniques can be applied which are discussed in this study. The study is organized into following sections: methods of stock market prediction, sentiment analysis techniques, machine learning techniques, evaluation measures and lastly concludes with the future research work.

MATERIALS AND METHODS

Methods for stock movement prediction: The investments to be done successfully, the analysis of stock market are done to predict the future for the investor (Schumaker and Chen, 2009). Trade advices by automatic trader agents can help investors for effective prediction of market. Sometimes these prediction systems indirectly provide supportive information like direction for future market. The stock market prediction methods can be categorized in different as follows:

High dimension data: The efficiently handling of high dimensional data has long been done by support vector machines with good performance. Therefore, SVM is chosen as prediction model for data of high dimension. 6 sets of features are designed to assess the effectiveness of sentiment analysis. The first feature set consists of only historical prices. In other methods the mood information is incorporated in prediction model. The scaling was between [-1, 1] for all the feature values (Fung *et al.*, 2002).

Price only: The historical prices are only used in this method to predict stock market. In this method, the investigation of patterns in the history of stock is done (Fu *et al.*, 2008). In addition to this investigation, this model was used as a paradigm to evaluate whether sentimental integration is effectives as compared with other sentiment models (Nayak and Braak, 2007). The features used in the method for training of machine are $price_{t-1}$ and $price_{t-2}$ which are considered as the movements of price which is either up or down, i.e., (up, down) for the transaction dates t-1 and t-2, respectively (Fung *et al.*, 2003).

Human sentiment: In this method, prediction model uses sentiments annotated by human which are integrated in

Corresponding Author: Arti Buche, Department of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India

addition to historical prices (Ferreira *et al.*, 2005). The users explicitly select a sentiment label posts such as few sentiment labels or stock decisions like "buy", "hold", "sell", "strong buy" and "strong sell". In this method, only annotated sentiments by users are used instead of using all messages and discard other messages. Among them only relevant sentiments are considered whereas remaining information such as message content is removed. The purpose of this human sentiment method is to predict the stock movement, depending on the mood annotated by human. The human sentiments are considered as the strongest features for stock market prediction (Gidofalvi and Elkan, 2001).

Sentiment classification: In this classification, a model was built to extract sentiments from the messages without explicit sentiments (Nikfarjam *et al.*, 2010). Training for the classification model built was done by using textual data containing sentiments (Sehgal and Song, 2007). The remaining messages were classified using the model into five classes buy, sell, hold, strong buy and strong sell (Modha *et al.*, 2013).

Topic based method: People express their views and opinions on a certain topics on social media sites, blogs etc. One such popular topic is stock market prediction or predicting market trend for stocks. People think that future price is based on topic and the related sentiment which predicts whether the stock price goes up or down (Ogilvie and Schmill, 1999). Hence, this method concludes that two models can be used to predict such a pair of topic and sentiment called as new feature of topic-sentiment for the stock prediction model (Mittermayer and Knolmayer, 2006).

Aspect based sentiment: In this method, mixtures of hidden topics and the sentiments are not considered in the dataset (Anonymous, 2019). The messages in the form of textual data are represented as a list of topics and the sentiment value related to it. This method uses consecutive nouns of a sentence from a topic (Bollen *et al.*, 2011).

Sentiment analysis techniques: Opinions and sentiment obtained from human communications which are posted online or social networks play an important role in prediction models (Adarsh and Ravikumar, 2015). Customer sentiment when captured electronically expressions go beyond fact which conveys emotion, mood and opinion which carries enormous business value (Cheng, 2010). Briefly it is the voice of the customer, voter and opinion leader (Mittermayer, 2004). The systematic

rating of human affective states is called as sentiment analysis which detects the polarity according to positive, negative or neutral or mixed or according to emotions, feelings or mood (Buche *et al.*, 2013). Sentiment analysis also has an important part for business purposes. Following are the approaches for pre-processing of text which can be adopted in finding public sentiment (Buche *et al.*, 2013).

Pre-process text: Simplification and cleaning of text is done by preprocessing it. Preprocessing text can create meaningful features from text easily (Jagtap and Paear, 2013). The basic operations that can be performed in preprocessing are as follows:

- Removal of stop words
- To search and replace target strings, regular expressions can be used
- Conversion of multiple related words to a single canonical form which is called as lemmatization
- POS tagging which filters specific part of speech from the text
- Normalization of text
- Special characters, numbers and sequences of repeated characters which belong to certain classes can be removed
- Emails and URLS can be identified and removed (Buche *et al.*, 2013a, b)

Detect languages: The language detection algorithm can be used to identify different languages. This algorithm analyses each row in the text and assigns a probability score, so that, the language with highest score can be detected. For example: stock opinions can be differentiated from different stock exchange markets like English versus non English opinions (Lavrenko *et al.*, 2000).

Extract key phrases from text: A phrase might be a compound noun, a single word or a modifier, the extraction of one or more meaningful phrases is done from a given natural language text (Buche *et al.*, 2013a, b). The extraction of phrases is considered as potentially meaningful context for capturing the topic of sentence or identifying sentiment which can be a combination of modifier and noun (Desai and Snehal, 2014).

Extract N-grams features from text: It featurizes the text and extracts most important piece of information. In this feature extraction, a dictionary of N-grams is generated from the input text (Buche *et al.*, 2013a, b). It helps in

applying various metrics to the N-gram list to reduce dimensionality of the data and identify the most valued information from the N-grams (Seo *et al.*, 2004).

Feature hashing: Feature hashing performs conversion of a stream of text into a set of features which are represented as numeric value. Training a text analysis model can be done using this hashed feature set. Feature hashing represents text documents of variable length as numeric feature vectors of equal length and also achieves dimensionality reduction. If feature hashing is not applied then it would be treated as a categorical feature with many distinct values (Paik *et al.*, 2007).

Latent Dirichlet Allocation (LDA): This approach groups unclassified text into a number of categories. LDA is often used in natural language processing to find similar text. LDA algorithm generates a probabilistic model which identifies groups of topics. It is a generative approach in which strong assumptions between the text and categories can be avoided and words to mathematically model topics can be used (Medhat *et al.*, 2014).

Named entity recognition: It is the most important area of research in natural language processing and machine learning as it can answer to many real world questions like name of person, current location, companies mentioned in news article, products mentioned by company, etc (Buche *et al.*, 2013a, b). It basically provides the list of named entities from the inputted dataset which contains text. In short named entity recognition can provide three types of entity identification: people, locations and organizations (Wuthrich *et al.*, 1998).

Feature selection: Feature selection mainly includes exploring data and developing a new model. It is the process of selecting a subset of relevant features used for building analytical model. Feature selection methods are as follows:

Filter based feature selection: The features with greatest predictive power are selected are identified from the dataset.

Fisher linear discriminant analysis: The linear combination of features variables that can be best grouped into separate classes of data are identified.

Permutation feature importance: The feature variables for a trained model and test dataset are considered and the permutation feature importance computes scores.

Machine learning techniques: Machine learning has been used, since, decades for stock market to outrun human traders. Machine learning is best leveraged as means to optimize prediction and output using historic experimental data. The machine learning algorithms allows analysts to design models for stock market price prediction. These algorithms also allow developing of new models based on historic data. Predictive analytics and machine learning are typically used together under specific circumstances as they are able to go beyond programmatic logic or standard rules of engines developed by mere mortals (Dash and Dash, 2016). Machine learning workflow includes following phases:

- Data
- Create model
- Evaluate model
- Refine and evaluate model
- Deploy the model
- Test and use the model

The machine learning aided methods for analyzing and predicting growth of stock can be described as follows:

Supervised learning: This type of machine learning uses datasets that are already known to create a model which can be used for prediction. The datasets known which includes input data along with known response values are called as training datasets. The training datasets are then used by supervised learning algorithms to build a new model for predictions based on new input values with known outcomes (Anonymous, 2018a, b). Supervised learning is separated into general categories of algorithm which can be used for stock market predictions which are as follows: (Khaidem *et al.*, 2016)

Classification: The determination of type, class of an item or category or row of data classification machine learning method is used. With the help of this method for example following can be classified:

- Sentiment identification as positive or negative
- Email filters as junks, spams or good can be classified

The tasks of classification can be organized as classification is binary, i.e., A or B or multiclass classification which includes prediction of multiple categories by using a single model. The selection of an appropriate algorithm is to be done to generate a classifier or a classification model. The factors to be considered can

be number of classes, different outcomes to be predicted, distribution of data or training time. As classification is a supervised machine learning method, it requires a labeled training dataset. Following are the algorithms of classification machine learning method:

Multiclass decision forest: This algorithm constructs multiple decision trees and voting is performed on the most popular output class.

Multiclass decision jungle: A recent extension to decision forests is decision jungles. After training the model with this method, prediction of target which has multiple values can be carried out.

Multiclass logistic regression: The well-known method in statistics is logistic regression depending on which prediction of the probability of outcome is done and it is also popular for classification of tasks. The trained model after using this algorithm can be used to predict values for new input examples.

Multiclass neural network: In this algorithm, neural network model is built which is used for predicting a target that can be multiple values. For example, letter or digit recognition, pattern recognition or document classification.

One-vs.-all multiclass: Using one-vs.-all approach, classification model is built that can predict multiple classes. It can predict three or more possible outcomes. This approach also uses binary classification methods that require multiple output classes.

Two-class averaged perceptron: It's a simple version of neural network. In this approach, classification of inputs is done to generate several possible outputs based on linear function and are combined with set of weights which are derived from feature vector and hence they are named as 'perceptron'.

Two-class Bayes point machine: This algorithm uses Bayesian approach to linear classification called the 'Bayes point machine'. It creates an untrained binary classification model. It chooses one 'average' classifier which efficiently approximates the optimal Bayesian average of linear classifiers.

Two-class boosted decision tree: Predictions in this algorithm are entirely based on ensemble of trees together that gives prediction. It uses boosted decision tree algorithm in which second tree corrects the errors of first tree, the third tree corrects for the first and second trees and so forth.

Two-class decision forest: In this algorithm, machine learning model is created based on the decision forest algorithm. This algorithm is a good choice if a target with a maximum of two outcomes is to be predicted.

Two-class decision jungle: In this approach, a machine learning model is created that is based on supervised ensemble learning algorithm called decision jungles. A decision jungle consists of decision Directed Acyclic Graphs (DAGs).

Two-class Locally Deep Support Vector Machine: Support Vector Machines (SVMs) are popular and well-researched class of supervised learning methods. They are used in both linear and non-linear classification. This approach creates a two-class, non-linear SVM classifier that optimizes prediction efficiently.

Two-class logistic regression: The popular statistical technique which is mostly used for prediction of probability of an outcome is logistic regression. This approach creates a logistic regression model which can be used to predict two (and only two) outcomes.

Two-class neural network: This algorithm creates a neural network model which can be used in predicting a target value which consists of only two values.

Two-class support vector machine: It creates a model based on support vector machine algorithm. SVM Models are used in many applications, from information retrieval to text and image classification (Anonymous, 2018b).

Regression: It is basically, a methodology which is used widely from engineering to education. For example, it can be used in predicting regional data or future enrollments. In modeling regression algorithm tried to learn the value of a function for a particular instance of data. By determining the contribution of each feature from data, regression algorithms can incorporate input from multiple features to regression function. These algorithms perform prediction of one or more continuous variables such as profit or loss based on columns in dataset (Khaidem *et al.*, 2016). Regression method of machine learning consists of following algorithms for modeling:

Linear regression: It is a common statistical method which attempts to establish a linear relationship between one or two independent variables and dependent variable or numeric outcome. Regression basically predicts a numeric target.

Simple regression: Simple regression involves a single independent variable and a dependent variable. It is classic regression problem.

Multiple linear regression: This involves two or more independent variables which contributes to a single dependent variable. It is also called as multivariate linear regression as in many problems multiple inputs are used to predict single numeric outcome.

Multi-label regression: The task of this regression is of predicting multiple dependent variables with single model. In this regression a sample can be assigned to multiple different labels.

Bayesian linear regression: This approach uses linear regression with additional information of a prior probability distribution. The generation of estimation of parameters is carried out by the combination of prior information about the parameters with likelihood function.

Boosted decision tree regression: It creates regression of trees using boosting. The dependency of each tree on prior trees is called as boosting. The step-wise fashion is used in building of regression tree using a predefined loss function. In each step this function measures the error and in next step corrects it. In this algorithm, using residual of the trees that preceded it, learning is performed.

Decision forest regression: This algorithm is based on decision trees. Each tree in regression decision forest predicts the output as a Gaussian distribution.

Fast forest quantile regression: Quantile regression model predicts values for a specified number of quantiles. Quantile regression helps to understand the distribution of the predicted value instead of getting a single mean prediction value. Applications of this method are predicting stock prices, evaluating student performance, discovering predictive relationship.

Neural network regression: This method builds a regression model using customizable neural network algorithm. Neural network regression model is best suited for problems where traditional regression model cannot fit a solution.

Ordinal regression: This regression method builds a model which is used to predict ranked values. For example, URLs in ranked search results order of finishers in a race, survey response that capture user's preferred brand.

Poisson regression: Poisson regression method is used to build a model which is used to predict numeric values typically counts. The main conditions for this method are response variable uses Poisson distribution, counts cannot be negative, use non-whole numbers only (Anonymous 2018a, b).

Clustering: Clustering is known as a non-supervised machine learning technique. This method can be used for unlabeled data. It is a technique which groups the data points into similar clusters also called as segmentation. For example, in text analysis clustering can be used to group sentences with similar topics or sentiment. Clustering can be performed either on labeled data or unlabeled data. In unlabeled data, temporary label for group of data is assigned to the data points which are closest together and create clusters around a centroid. As far as labeled data is considered the labels can be used to drive the number of clusters or can be used just as another feature.

One of the clustering category is 'k-means' clustering which is one of the simplest and best known unsupervised machine learning algorithm. This algorithm can be used in many machine learning tasks such as detecting abnormal data, clustering of text documents and analysis of text documents prior to classification or regression (Mitchell and Mulherin, 1994).

Evaluation measures: Azure machine learning studio is a tool to build, test and deploy predictive analytics solutions for the data. It can be used as a testing tool for the prediction models designed as it consists of large number of machine learning algorithms. Machine learning is a technique that allows machines to use historic data for the forecasting future behaviors, trends and outcomes.

It also consists of different modules that enable to prepare data as input, develop experiments, visualization and deploy models at cloud scale. Using the components of this studio predictive analytics experiment can be developed, iterated and trained. The azure machine learning studio allows quick experimental runs on predefined models and iterations using different variations. It also enables efficient learning, debugging and result analysis (Anonymous 2018a, b).

Future work: Stock market can be viewed as a particular text mining or opinion mining problem. Most people consider the stock market as unpredictable as the movement in stock exchange depends on capital gains and losses. The prediction of stock market behavior can be considered as an intelligent treatment of past and present financial data. In future research work, prediction of stock market indicators can be done on basis of text analytics which include the following procedure:

- Preprocessing and clean the text dataset
- After preprocessing the text, extract numeric feature vectors
- Train classification or regression or clustering model
- Score and evaluate the model
- Deploy the model

Considering the above procedure three different models can be built to predict stock market trend namely week prediction model, daily prediction model and monthly prediction model. In daily model, predictions of the trend can be done for next day by considering data on daily basis as input similarly second model of week which consists of weekly data as input and third model consists of monthly data as input. In these prediction models, historical data and people opinions along with sentiments are used to build the model. By considering various patterns in the dataset to predict the market trend whether it is up or down. As per the survey conducted, the dataset of historic prices and sentiments decision boosted tree or support vector machine and logistic regression can perform better for the prediction (Nayak *et al.*, 2016).

CONCLUSION

The study investigated on various stock market forecasting techniques which can give impact on stock market prediction decision making. It also discussed various algorithms and methods with general procedural flow. The main focus of the study was to perform opinion mining in order to generate accurate and reliable predictions regarding investment of consumer's valuable money.

REFERENCES

Adarsh, M. and P. Ravikumar, 2015. Survey: Twitter data analysis using opinion mining. Intl. J. Comput. Appl., 128: 34-36.

Ahn, S. and S.B. Cho, 2010. Stock prediction using news text mining and time series analysis. Proc. KIISE. Conf., 37: 364-369.

Anonymous, 2018a. Multiclass neural network. Microsoft Azure, USA. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network

Anonymous, 2018b. Technical Indicator descriptions. Investopedia, New York City, USA. https://www.investopedia.com/terms/w/williamsr.asp

Anonymous, 2019. US S&P 500 index historical data. Yahoo Finance, New York, USA. https://ca.finance.yahoo.com/quote/%5EGSPC/history/?guccounter=1

Bollen, J., H. Mao and X. Zeng, 2011. Twitter mood predicts the stock market. J. Comput. Sci., 2: 1-8.

Buche, A., D. Chandak and A. Zadgaonkar, 2013a. Opinion mining and analysis: A survey. Intl. J. Nat. Lang. Comput., 2: 39-48.

Buche, A., M.B. Chandak and A. Zadgaonkar, 2013b. An approach for online analysis using expectation maximization. Intl. J. Innovative Res. Comput. Commun. Eng., 1: 1064-1073.

Cheng, S.H., 2010. Forecasting the change of intraday stock price by using text mining news of stock. Proceedings of the 2010 International Conference on Machine Learning and Cybernetics Vol. 5, July 11-14, 2010, IEEE, Qingdao, China, ISBN:978-1-4244-6526-2, pp: 2605-2609.

Dash, R. and P.K. Dash, 2016. A hybrid stock trading framework integrating technical analysis with machine learning techniques. J. Finance Data Sci., 2: 42-57.

Desai, R. and G. Snehal, 2014. Stock market prediction using data mining. Intl. J. Eng. Dev. Res., 2: 2780-2784.

Ferreira, T.A., G.C. Vasconcelos and P.J. Adeodato, 2005. A new evolutionary method for time series forecasting. Proceedings of the 7th Annual International Conference on Genetic and Evolutionary Computation (GECCO) Vol. 5, June 25-29, 2005, ACM, Washington, USA., pp: 2221-2222.

Fu, T.C., K.K. Lee, D. Sze, F.L. Chung and C.M. Ng, 2008. Discovering the correlation between stock time series and financial news. Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '08) Vol. 1, December 9-12, 2008, IEEE Computer Society Washington, DC, USA., ISBN:978-0-7695-3496-1, pp: 880-883.

Fung, G.P.C., J.X. Yu and H. Lu, 2005. The predicting power of textual information on financial markets. IEEE. Intell. Inf. Bull., 5: 1-10.

Fung, G.P.C., J.X. Yu and W. Lam, 2002. News sensitive stock trend prediction. Proceedings of the International Conference on Pacific-Asia Knowledge Discovery and Data Mining (PAKDD 2002), May 6-8, 2002, Springer, Berlin, Germany, ISBN:978-3-540-43704-8, pp: 481-493.

Fung, G.P.C., J.X. Yu and W. Lam, 2003. Stock prediction: Integrating text mining approach using real-time news. Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering, March 20-23, 2003, IEEE, Hong Kong, China, ISBN:0-7803-7654-4, pp: 395-402.

Gidofalvi, G. and C. Elkan, 2001. Using news articles to predict stock price movements. MSc Thesis, University of California, San Diego, California, USA.

Jagtap, V.S. and K. Pawar, 2013. Analysis of different approaches to sentence-level sentiment classification. Intl. J. Sci. Eng. Technol., 2: 164-170.

Khaidem, L., S. Saha and S.R. Dey, 2016. Predicting the direction of stock market prices using random forest. Appl. Math. Finance, 1: 1-20.

Kim, Y., S.R. Jeong and I. Ghani, 2014. Text opinion mining to analyze news for stock market prediction. Intl. J. Adv. Soft Comput. Appl., 6: 1-13.

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie and D. Jensen *et al.*, 2000. Language models for financial news recommendation. Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM '00), November 06-11, 2000, ACM New York, USA., pp: 389-396.

Medhat, W., Hassan, A. and H. Korashy, 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J., 5: 1093-1113.

Mitchell, M.L. and J.H. Mulherin, 1994. The impact of public information on the stock market. J. Finance, 49: 923-950.

Mittermayer, M.A. and G.F. Knolmayer, 2006. Newscats: A news categorization and trading system. Proceedings of the 6th International Conference on Data Mining (ICDM'06), December 18-22, 2006, IEEE, Hong Kong, China, pp: 1002-1007.

Mittermayer, M.A., 2004. Forecasting intraday stock price trends with text mining techniques. Proceedings of the 37th Annual Hawaii International Conference on System Sciences, January 5-8, 2004, IEEE, Bern, Switzerland, ISBN: 0-7695-2056-1, pp: 1-10.

Modha, J.S., G.S. Pandi and S.J. Modha, 2013. Automatic sentiment analysis for unstructured data. Intl. J. Adv. Res. Comput. Sci. Software Eng., 3: 91-97.

Nayak, A., M.M. Pai and R.M. Pai, 2016. Prediction models for Indian Stock Market. Procedia Comput. Sci., 89: 441-449.

Nayak, R. and P. Te Braak, 2007. Temporal pattern matching for the prediction of stock prices. Proceedings of the 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM '07) Vol. 84, December 01, 2007, Australian Computer Society, Australian, ISBN:978-1-920682-65-1, pp: 95-103.

Nikfarjam, A., E. Emadzadeh and S. Muthaiyah, 2010. Text mining approaches for stock market prediction. Proceedings of the 2010 2nd International Conference on Computer and Automation Engineering (ICCAE) Vol. 4, February 26-28, 2010, IEEE, Singapore, Singapore, ISBN:978-1-4244-5569-0, pp: 256-260.

Ogilvie, P. and M. Schmill, 1999. Ænalyst-Electronic analyst of stock behavior. MSc Thesis, University of Massachusetts, Massachusetts, USA.

Paik, W.J., M.H. Kyung, K.S. Min, H.R. Oh and C.M. Lim *et al.*, 2007. Multi-stage news classification system for predicting stock price changes. J. Korean Soc. Inf. Manage., 24: 123-141.

Peramunetilleke, D. and R.K. Wong, 2002. Currency exchange rate forecasting from news headlines. Proc. Australas. Database Conf. Melbourne, 24: 131-139.

Schumaker, R.P. and H. Chen, 2009. Textual analysis of stock market prediction using breaking financial news: The AZFin text system. ACM. Trans. Inf. Syst., 27: 1-19.

Sehgal, V. and C. Song, 2007. Sops: Stock prediction using web sentiment. Proceedings of the 7th IEEE International Conference on Data Mining Workshops (ICDMW 2007), October 28-31, 2007, IEEE, Omaha, Nebraska, USA., ISBN:978-0-7695-3019-2, pp: 21-26.

Seo, Y.W., J. Giampapa and K. Sycara, 2004. Financial news analysis for intelligent portfolio management (No. CMU-RI-TR-04-04). Master's Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Wuthrich, B., D. Permunetilleke, S. Leung, W. Lam and V. Cho *et al.*, 1998. Daily prediction of major stock indices from textual www data. HKIE. Trans., 5: 151-156.