DATA AND TEXT MINING OF FINANCIAL MARKETS USING NEWS AND SOCIAL MEDIA

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2012

By
Zhichao Han
School of Computer Science

Contents

Word Count: 15602

List of Tables

List of Figures

Abstract

Much research has investigated using both data mining, with technical indicators, and text mining, with news and social media. The combination of news features and market data may improve prediction accuracy. Despite of this, existing systems do not appear to have efficiently or effectively integrated news features and market data.

In this dissertation, various of data and text mining techniques are used to identify, investigate and evaluate valuable features and methodologies in stock price performance forecasting on specific securities using technical indicators and textual data such as news, blogs and tweets.

A two-stage architecture utilizing data and text mining technologies is used to predict stock prices. A stock price performance forecasting workflow is designed based on current and past stock prices, tweets, blogs and news. The Latent Dirichlet Allocation (LDA) is utilized to model topics of documents and Principal Component Analysis (PCA) is used to reduce feature dimension.

Ultimately, the tests involving feature combination with numeric and textual data and the proposed technical indicator features with the sentiment score series from tweets yield the best results of all, with classification accuracy for next day stock movement performance (SMP) prediction at 77.50% and next day stock movement direction (SMD) prediction at 80.29%. The SMP is evaluated based on customized criteria and the SMD is assessed based on the comparison of the current closing price and the next *n*th day closing price.

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank my supervisor, Prof John Keane. He gave me much valuable guidance and many suggestions on design approaches and showed much patience in my project.

I would like to thank Dr Xiaojun Zeng and Huina Mao who gave me comments on my dissertation.

I would also like to thank Mr Ian Cottam who instructed me the usage of Condor at Manchester, which helped much in the experiments.

I would like to extend my thanks to Karl and Chris Pearson who helped me with my English and proofread my dissertation.

I would like to acknowledge my friends and my parents for their support and encouragement.

Above all, I would like to thank God.

Chapter 1

Introduction

1.1 Project context

The financial market is recognized as a complicated and non-linear system [2]. Stock market prediction has attracted much attention from academia and business as large amounts of evidence indicate that stock market prices can be at least partially predicted [12, 25, 33, 57]. However, there are so many factors such as politics and natural events affecting stock markets that make forecasting stock prices technically challenging.

Technical analysis is a general methodology to predict price movement and trading volume based on historical data. [35] To address complex and noisy time series of stock prices, many researchers have applied machine learning techniques such as Artificial Neural Networks (ANN) [26, 71], Genetic Algorithms (GA) [17, 26] and Support Vector Machines (SVM) [31, 74] to improve prediction performance. Work by Atsalakis and Valavanis [4] indicates that neural networks and neuron-fuzzy models outperform traditional models in most cases. However, it remains a challenge to tune the model structures of neural networks and neuron-fuzzy models. A two-stage architecture with SVM [29, 68], which decomposes the time series into smaller similar regions, has been shown to be more competitive than a single SVM model where non-stationary factors are considered; in comparison, the prediction results of the two-stage architecture evaluated by Mean Absolute Error (MAE) are 10% more accurate.

According to the Efficient Market Hypothesis [23], all available information is reflected in market prices. However, it is believed that it takes time for the market to respond to the new information. [47, 53] Various research [49, 58, 63] has investigated the prediction of stock prices using text mining of financial news and the directional accuracy of the forecast, which vary from 45% to 60% in terms of accuracy and hence

are not ideal. It is suggested in [52] that the combination of news features and market data may improve prediction accuracy. Despite this, existing systems [49, 58, 63] do not appear to have efficiently or effectively integrated news features and market data such as prices and technical indicators.

Sentiment expressed in financial texts is also relevant to price movement. With the growth of social networks, sentiment analysis of both traders and the public has been adopted to help forecast price movement. A series of work [9, 28, 66, 82] has investigated stock price prediction using Twitter. However, it is suggested in [42] that the general word list, which is manually selected based on the general context, for sentiment analysis may misclassify financial text. The tweets that were used in Bollen et al.'s research [9] appeared to be irrelevant within the financial context; thus indicating that it might be unsuitable to use the general word list in tweet sentiment analysis for forecasting the movement of a specific stock price. Furthermore, in [9] it is suggested that tweets alone would be insufficient as unexpected news and events are often not well reflected in public sentiment until some time has passed.

1.2 Aims and objectives

The aim of this dissertation is to identify, investigate and evaluate valuable features and methodologies in stock price movement performance forecasting specific securities using technical indicators and textual data.

In order to do that, a stock movement forecasting workflow is designed based on current and past stock prices and volumes, news, blogs and tweets. In this system, different components such as sentiment analysis, context analysis and dimensionality reduction are implemented. Experiments are conducted based on a two-stage architecture or using the features from sentiment analysis, dimensionality reduction and feature combination of numeric and textual data. In context analysis, the performance of clustering algorithms is to be compared. The performance of the experiments is evaluated by classification accuracy.

1.3 Research process

In this dissertation, data mining and text mining technologies utilizing price factors are used to predict stock movement performance. Data mining is a process to identify new knowledge from existing large data sets. [14] Text mining refers to the process

of discovering interesting patterns from text documents. [67] Data mining techniques, such as regression and dimension reduction, and text mining techniques, such as bag-of-words and sentiment analysis, are used to predict stock movement performance (SMP).

1.3.1 Data collection

The daily open-high-low-close (OHLC) prices and volumes of the S&P securities have been collected ranges from September 20, 2006 to July 19, 2012.

The news related to the stocks in S&P 100 is obtained from Reuters Site Archive[1]. Only PRNews Wire, Business Wire, Market Wire along with Globe Newswire are used. Around 44000 articles have been collected from the years 2010 and 2011 over 78 companies in the S&P 100.

Blogs to be used for the analysis have been fetched from SeekingAlpha[2], which is an American stock market analysis website. SeekingAlpha ranks the second in the search results when "blog" and "stock" are searched for at Delicious[3], a popular social bookmark service. Note: the first result from Delicious is irrelevant to stock markets. The blog writers on SeekingAlpha do not deliver their blogs with a high frequency. For example, the average blog posts for Google and Apple in the focus article category from January 1, 2012 to June 11, 2012 are 1.35 and 4.54 per day. The experiment results on blogs obtained from other sources in this dissertation may vary. All the 22007 analysis articles on the S&P 100 stocks have been collected up to June 11, 2012.

Twitter tweets have been collected through the Twitter Search API[4], with the keyword $TICKER like $GOOG and $YHOO. 634k tweets related to the stocks in S&P100 have been archived in one week (April 28 - July 19, 2012).

1.3.2 Prediction methods

The prediction target in this dissertation is the performance of securities movement. Customized criteria, classified into *good* and *bad*, are set in order to evaluate the performance. The performance will be regarded *good* if more *good* criteria are met than *bad* criteria, and vice versa. If the numbers of the *good* and *bad* criteria are identical,

[1] http://www.reuters.com/resources/archive/us/
[2] http://seekingalpha.com/
[3] http://delicious.com/search?p=blog+stock
[4] https://dev.twitter.com/docs/api/1/get/search

the performance will be deemed as *uncertain*. The customized criteria are given in 4.1.2.

The stock movement performance (SMP) prediction model used in this dissertation is a Support Vector Regression Machine (SVR) [65]. The performance is mapped into 0˜1 as the prediction target in the SVR. The projection details is given in 4.1.2. The classification *good* / *uncertain* / *bad* is based on regression results. The evidence of the validity of converting regression to classification was given in [73].

As well as prediction of SMP, a stock movement direction (SMD) prediction model is adopted in order to compare the performance of the approaches in this dissertation with the work in [9] where concrete future the DJIA are predicted and directional movement accuracy is also provided.

The SMD prediction is based on the comparison of the current closing price and the closing prices on the next days. The targets *up* (1) or *down* (0) are indicated by the future closing price being greater / less than the current one. If they are equal, the target will be *unchanged* (0.5). The projection from 0˜1 to *up* / *unchanged* / *down* is similar to the SMP projection on *good* / *uncertain* / *bad*, as is illustrated in 4.5.

1.3.3 Features used for prediction

Sentiment features: Sentiment scores are generated from textual data alone. They are used as the input of the prediction model.

Context analysis: Technical indicator features from numeric data are introduced into the prediction model. Clustering is conducted based on technical indicator features before stock performance prediction. There is an SVR model for each cluster. The SVR models are modelled based on bag-of-words (BOW) models and topic features from textual data.

Feature combination: Feature combination is conducted on two groups of experiments. One group is based on the same features used in context analysis, namely, technical indicator features, BOW and topic features. The purpose of this group of experiments is to verify if the clustering in context analysis leads to better performance. The other group of experiments is based on technical indicator features and the sentiment features, which shows the best performance as a single type of features. The purpose of this is to identify a better model from feature combination.

1.3.4 Evaluation

The tuning of SVR parameters is conducted based on a grid search. The SMP and SMD models are evaluated by 10-fold cross-validation. The instances are kept in time order so as to ensure that the instances are independent enough in time span. The result accuracy is the mean accuracy from the 10 experiments.

To assess the SMD prediction model using technical indicator features and tweet sentiment features it is be compared with the Dow Jones Industrial Average (DJIA) movement directional prediction accuracy obtained in [8]. From the literature, the greatest directional movement prediction accuracy is reported in [8] among the related work [3, 9, 13, 19, 28, 37, 46, 48, 60] on stock market prediction using sentiment analysis. However, the prediction model in [8] is different to the SMD as their model predicted the concrete future prices of DJIA while SMD predicts the movement direction of the S&P 100 stocks.

1.4 Contribution

This dissertation investigates the predictive power of technical indicator features and sentiment analysis, context analysis and feature extraction techniques on news, blogs and tweets. The features from tweets give the best performance among the textual data. Sentiment analysis on tweets is found to give the highest prediction accuracy, which may be linked with the fact that tweets are the most intuitive and simple source in emotion among the three textual data.

In the experiments where the technical indicators features and the sentiment score series are combined, the prediction targets of the modelling and evaluation are the SMP and the SMD on the next *n*th day. The prediction accuracy of the next 1st day is 77.50% for SMP and 80.29% for SMD. The prediction accuracy of the next 5th day is 89.23% for SMP and 92.90% for SMD.

In the experiments of context analysis, the greatest accuracy for SMP using tweet features is 67.25% and 71.02% for the next 1st and 5th days respectively. In the experiments of feature extraction, the best results for SMP using tweet features are 62.91% and 65.90% for the next 1st and 5th days respectively.

In summary, this dissertation shows a promising approach using sentiment analysis on tweets. The results of feature combination of tweet sentiment features and technical indicators appear satisfactory as well.

1.5 Dissertation overview

The rest of the document is organized as follows. Chapter 2 presents the literature background; Chapter 3 describes the deign and implementation of the models; the setup of the experiments is given in Chapter 4; analysis of the experiment results is given in Chapter 5; Chapter 6 presents conclusions and recommendations for future work.

Chapter 2

Background and general context

Many researchers have investigated using data mining with technical indicators [22, 29, 34] and using text mining with news [45, 50, 52] and social media [9, 81, 82]. This chapter is organized as follows: firstly, technical background in text mining, sentiment analysis, etc. is discussed in 2.1; secondly, research background in stock market forecast is given in 2.2; finally, a summary is given in 2.3.

2.1 Technical background

2.1.1 Time series similarity analysis

Stock prices, technical indicators and sentiment scores extracted from textual data can be represented in the form of time series. In this subsection, various similarity measures are represented. [72]

Euclidean distance

Euclidean distance is the distance between two points connected by a line. The formula is illustrated in Eq.2.1 where p and q are two points in n-dimension.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2.1}$$

Pearson's correlation coefficient

Pearson's correlation coefficient (PCC) is defined as the covariance of two vectors divided by their standard deviation production, as is illustrated in Eq.2.2 where X and

Y are two vectors.

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.2}$$

Short time series distance

Short time series distance (STS) is proposed by Möller-Levet et al. [51]. The definition STS distance is defined in Eq.2.3. The similarity is compared to the differences of the values in the time series.

$$d_{STS} = \sqrt{\sum_{i=1}^{n-1}(\frac{X_{i+1} - X_i}{t_{i+1} - t_i} - \frac{Y_{i+1} - Y_i}{t_{i+1} - t_i})} \tag{2.3}$$

2.1.2 Learning algorithms

Supervised and unsupervised learning methods are widely used in data mining and text mining. Training instances in supervised learning are labeled and sued to derive a model, whereas all data is used to derive models in unsupervised learning. In this subsection, unsupervised learning algorithms such as K-means [44], GHSOM [59] and LDA [7], and supervised learning algorithms, such as SVM [18] and SOFNN [39] are introduced.

K-means

K-means [44] is a basic clustering technique that aims to minimize the total distance of the data points to the cluster centers. The distance can be defined as either Euclidean distance or other similarity measures given in 2.1.1.

Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) [36] are unsupervised neural networks that order the inputs on a grid in a lower dimension via their similarity. The basic units in the network are called nodes or neurons. Each node is assigned a weight. The closest node is picked as the winner according to each input instance. Finally, all the weights of the winner's neighbor nodes are updated. The procedure is repeated until the network converges.

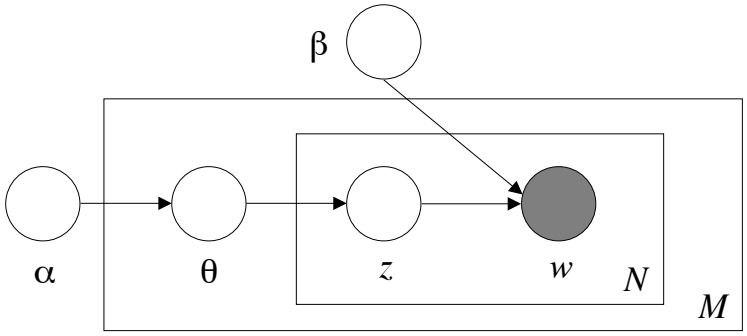


Figure 2.1: Graphic presentation of LDA[7]

It is a major issue for SOM to tune its parameters. To address this issue, the Growing Hierarchical Self-Organizing Map (GHSOM) [59] has been proposed as an extension of SOM where the neighbor size and structure are automatically tuned in a hierarchical and horizontal way.

Latent Dirichlet Allocation (LDA)

A Latent Dirichlet Allocation (LDA) [7] is a popular topic model. An intuitive idea of topic models is that a document consists a collection of topics. For instance, a news article on a new Google product can be categorized into topics such as 'Internet' and 'Business'. However, the names of the topics are unknown as LDA is an unsupervised model.

The basic components of LDA models are *word*, *document* and *corpus*. A *word* is the basic unit, which is denoted as w. A *document* consists of a sequence of words, which is denoted as $\mathbf{w} = \{w_1, w_2, ..., w_N\}$. A *corpus* is made up of documents, which is denoted as $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$.

A graphic presentation of LDA is illustrated in Fig.2.1. In the figure, α and β are corpus-level parameters. θ denotes the joint distribution a topic mixture. $\mathbf{z}$ is a set of topics. N and M are the numbers of words and documents.

The topic distribution of a document in is calculated in Eq.2.4.

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_n p(z_n|\theta) p(w_n|z_n, \beta) \tag{2.4}$$

Support Vector Machine (SVM)

A Support Vector Machine (SVM) [18] is a supervised learning method. The aim of an SVM is to maximize the margin while the constraint function is satisfied. An example

of a linear model is illustrated in Eq.2.5. In the equation, y_i is the target value of the ith instance and $\mathbf{x_i}$ is the input feature vector of the ith instance.

$$\min \frac{1}{2}||\mathbf{w}||^2, s.t. y_i(\mathbf{w}^T\mathbf{x_i} - b) \geq 1 \tag{2.5}$$

The margin of the model is defined in Eq.2.6. The aim to maximize the margin is equivalent to minimizing Eq.2.7. This minimization problem can be solved by the quadratic programming optimization.

$$m = \frac{2}{||\mathbf{w}||} \tag{2.6}$$

$$\frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{2.7}$$

The version of SVM for regression is known as the Support Vector Regression (SVR) [65].

SOFNN

A Self-Organizing Fuzzy Neural Network (SOFNN) [39] is a neural network based on Ellipsoidal Basis Function (EBF) neurons made up of a center vector and a width vector. The five layers are the input layer, the EBF layer, the normalized layer, the weighted layer and the output layer. The SOFNN learning procedure consists of the parameter and structure learning.

The output of SOFNN can be written as Eq.2.8 where $d(t)$ denotes the expected output, $p_i(t)$ are the regressors, θ_i represents the model parameters to be tuned and $\varepsilon(t)$ is the difference between the target output and predicted output.

$$d(t) = \Sigma_{i=1}^{M} p_i(t)\theta_i + \varepsilon(t) \tag{2.8}$$

The structure learning consists of adding and pruning neurons. The system error criterion and if-part criterion are used to decide if there is a need to add an EBF neuron. The overall generalization performance is checked by the system error criterion checks. The if-part criterion considers the performance of existing EBF neurons. Second derivative information is adopted by a neuron pruning process to find excessive neurons. [39]

2.1.3 Text processing

Stemming and lemmatizing

There are many different forms of a simple word, such as tenses and derived words. For instance, 'happy' has a noun form , which is 'happiness' and an adverb form, which is 'happily'. They represent the same meaning. In order to reduce the feature dimension, such words should be filtered out in the process under the same root, 'happy'.

Before words are processed, they have to be stemmed or lemmatized in order to reduce feature dimensions. Words with the same stem are treated as a single feature. The Porter stemming algorithm [56] is a popular stemming method. Lemmatizing is different from stemming as context and dictionary lookups are involved in lemmatizing while stemming is only concerned with suffixes. For example, 'worse' can be recognized as 'bad' by lemmatizing algorithms but not by stemming algorithms.

Textual features

Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF weight is a Natural Language Processing (NLP) technique, which reflects the importance of a word in a document. The term frequency is the number of times a term appears in a document. The higher the frequency of a term, the more informative it is. The inverse document frequency is the measurement for how rare a term is documents.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{2.9}$$

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{2.10}$$

The formula of TF-IDF is given in Eq.2.9 where t represents a term, d represents a document, $|D|$ denotes the total number of the documents in the corpus and $|\{d \in D : t \in d\}|$ is the number of documents which contain the term.

Term Presence: Term frequencies play an important role in term weighting. However, Pang et al. [54] indicates that term presence yields a better performance in sentiment analysis than term frequency. Term presence is represented as a boolean value in a vector. If a term appears in a document, it will be assigned True or 1 in the feature vector.

Parts of speech (POS): POS is important in NLP as it is a simple technique to reduce ambiguity [76]. It is also a necessary procedure for lemmatizing.

Negation: It is essential to consider negation while processing short messages like tweets. A 'not' might change the entire meaning of a sentence. A negation can be encoded into initial features. Das and Chen [20] tried appending 'NOT' to the terms around 'no' or 'do not' to solve the negation problem. For example, in the sentence 'I do not like apples.', the term extracted will be 'like_NOT' instead of 'like'.

Bag-of-words: Bag-of-words is a classical model used in NLP where a piece of document is represented as a term frequency vector. It is assumed that although the term orders and syntax are missing, the major information is contained in the term frequency vector.

2.1.4 Sentiment analysis

Dictionary

A dictionary-based approach is a simple technique to generate the word list for sentiment analysis. First, a small set of seed mood words and an online dictionary, such as WordNet[1], are given. Then new synonyms and antonyms are added into the word list. This will be repeated until no new word is found. However, a major weakness of the dictionary-based approach is that mood words within specific domains are difficult to find. [40]

Besides online dictionaries, synonyms can be identified through co-occurrences of terms. Deerwester et al. [21] argue that the features extracted from Latent Semantic Index (LSI) contain the information of synonymy and polysemy. Latent Dirichlet Allocation (LDA) [7], in the spirit of LSI, is a generative model, which can be used to identify basic linguistic patterns. Topic distribution vectors in LDA models can be regarded as a representation of similar term distributions.

Tetlock [69] uses the General Inquirer's (GI) Harvard IV-4 psychosocial dictionary[2] to convert Wall Street Journal (WSJ) columns into numeric values. The transformation is made by word count in the GI categories. The values are recentered so as to reduce the semantic noise in the columns. An alternative word list[3] made by Loughran

[1] http://wordnet.princeton.edu/wordnet/download/
[2] http://www.wjh.harvard.edu/~inquirer/homecat.htm
[3] http://nd.edu/~mcdonald/Word_Lists.html

and McDonald [42] specially designed for financial contexts is considered. They claim that general negative word lists may not reflect the true sentiment in financial contexts.

Profile of Mood States (POMS)

Bollen et al. [9] investigate stock price forecast using public moods in stock market forecast and obtain an accuracy of 86.7%. POMS [8] is used in this dissertation to conduct sentiment analysis. Twitter tweets are transformed into stemmed normalized terms first, where stop words are removed. Then they are processed as follows:

1. Score the tweets using the POMS-scoring function given in Eq.2.11. Each tweet t is denoted in the term set of w. The POMS emotion adjectives are represented as p_i for mood dimension i.

$$\mathcal{P}(t) \rightarrow m \in \mathbb{R}^6 = [||w \cap p_1||, ||w \cap p_2||, \cdots, ||w \cap p_6||] \tag{2.11}$$

2. Normalize the emotion vector, as is illustrated in Eq.2.12.

$$\hat{m} = \frac{m}{||m||} \tag{2.12}$$

3. Aggregate emotion vectors for particular dates and denote them as m_d, as is given in Eq.2.13 . A period of k-day mood is represented as $\theta_{m_d}[i,k]$.

$$m_d = \frac{\Sigma_{\forall t \in T_d} \hat{m}}{||T_d||} \tag{2.13}$$

$$\theta_{m_d}[i,k] = [m_i, m_{i+1}, \cdots, m_{i+k}] \tag{2.14}$$

4. Normalize mood vectors with z-scores.

$$\tilde{m}_i = \frac{\hat{m}_i - \bar{x}(\theta[i, \pm k])}{\sigma(\theta[i, \pm k])} \tag{2.15}$$

$$\tilde{\theta}_{m_d}[i,k] = [\tilde{m}_i, \tilde{m}_{i+1}, \cdots, \tilde{m}_{i+k}] \tag{2.16}$$

Lydia

Zhang and Skiena [81] applies the same sentiment analysis techniques in the Lydia sentiment analysis system[27]. The Lydia data is made up of time series of the counts of positive and negative words appearing with the corresponding entities.

$$Polarity = \frac{p-n}{p+n} \tag{2.17}$$

$$Subjectivity = \frac{p+n}{N} \tag{2.18}$$

Two important indicators are represented in Eq.2.17 and Eq.2.18. The numbers of positive and negative references are represented as p and n respectively. The total number of references is denoted as N.

2.1.5 Feature selection and extraction

An abundance of features can be extracted by data and text mining techniques from news and the stock market such as investors' sentiments, topics in the news related to the companies and the trends of stock prices. High dimensional data not only causes the curse of dimensionality [5], but also causes high computational time and resources. Hence, feature selection and extraction techniques are necessary to reduce the dimensionality of the data.

Wrappers and Filters

Wrappers and filters are both popular feature selection techniques. The difference between them is that wrappers evaluate each addition of a feature via a specified classifier while filters evaluate the features independently of classifiers. Features have been sorted according to the scores obtained by utility functions. Compared with wrappers, the features obtained by filters usually have higher error with specific classifiers, but at the same time, it saves computational time and resources.

Yang and Pedersen [80] show that document frequency (DF), information gain (IG) and χ^2-test (CHI) yields effective performances using k Nearest Neighbor (kNN) [78] and Linear Least Squares Fit mapping (LLSF)[79]. Using IG thresholding, a kNN classifier obtained better performance (from 87.9% to 89.2%) on Reuters corpus category identification, with a 98% reduction in unique terms. A proper threshold of feature selection is to ensure that a transformation from a document to a word count vector does not lead to a zero vector. CHI and DF shared similar performance, which was around 88% for kNN and 85% for LLSF [80].

DF is the simplest utility function, which counts the number of the documents

where a term appears. The basic assumption of DF is that the rare term are non-informative and less impactive on classifier performances [79].

IG represents the information gained when the candidate attribute is added. It is given in Eq.2.19. A feature is represented as a and all the examples are denoted as Ex. H is the entropy function, illustrated in Eq.2.20.

$$IG(Ex,a) = H(Ex) - \Sigma_{v \in values(a)} \frac{|\{x \in Ex | values(x,a) = v\}|}{|Ex|} \times H(\{x \in Ex | values(x,a) = v\}) \tag{2.19}$$

$$H(X) = -\Sigma_{i=1}^{n} p(x_i) \log p(x_i) \tag{2.20}$$

χ^2 statistic is used to estimate the dependence between two variables. The formula is given in Eq.2.21 where$A$ denotes the co-occurrence of t and c, B represents the number of t appearing alone, C is the number of times that c occurs without t, D denotes the times that neither t nor c appears and N is equal to $(A+B+C+D)$.

$$\chi^2(t,c) = \frac{N \times (AD - BC)}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \tag{2.21}$$

χ^2 statistic is then converted into two scores in [80], which are given in Eq.2.22 and Eq.2.23.

$$\chi^2_{avg}(t) = \Sigma_{i=1}^{m} P_r(c_i) \chi^2(t,c_i) \tag{2.22}$$

$$\chi^2_{max}(t) = \max_{i=1}^{m} \{\chi^2(t,c_i)\} \tag{2.23}$$

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [24] is a popular technique to extract expressive information from high dimensional data. The aim of PCA is to minimize redundancy and maximize the signal of the extracted features. The orthonormal matrix P can be found via the following steps. First, choose a normalized direction of m-dimensional space with the maximized variance. Second, find another direction where its variance is maximized and orthonormal to all the previous chosen directions. Repeat the second step until all the m vectors are found. [64]

2.1.6 Evaluation

Cross-validation

Cross-validation is a technique used to estimate if the results of learning algorithms are generalized. K-fold cross-validation is a common type of cross-validation. A data set is split into K subsamples and K iterations of training and testing are conducted. Each time, one subsample is left for testing and the remainder is used for training.

MAE

The Mean Absolute Error (MAE) is usually adopted for the regression performance evaluation. MAE is defined in 2.24 where $\hat{\theta}$is the predicted value and θ is the real value.

$$MAE = \frac{1}{n}\Sigma_i|\hat{\theta}_i - \theta_i| \tag{2.24}$$

2.2 Stock price movement research background

2.2.1 Numeric data analysis

Technical indicators are usually adopted by investors to analyze stock price movements. Much research has been done on the combining of soft computing technology with technical analysis in stock analysis, and a better prediction result or a higher rate of return is usually achieved. There are many choices for parameters of the indicators. For example, the Relative Strength Index (RSI) shows the strength of price movement trends. The parameter of RSI is the time span, which represents the length of the trends, which can be 10 days, 20 days or any other desired length of time. A more detailed introduction to technical indicators is given in Appendix B.

Enke and Thawornwong [22] applied Evolutionary Algorithms (EA) to achieve ideal parameters for technical indicators. In their work, the Moving Average Convergence / Divergence (MACD) indicator and the RSI oscillator were chosen to generate buying or selling signals. The aim of this work was to maximize the yields and to minimize transaction costs, trend risk and VIX risk. The trend risk evaluates the quality of the trends suggested by indicators. VIX, often referred to as the fear index, is calculated based upon the risk neutral expectation of the S&P 500 variance. The tuning

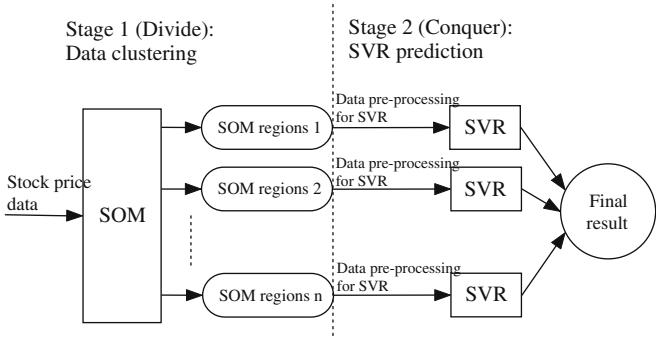


Figure 2.2: The two-stage architecture [29]

of the parameters by EA improved the profit by nearly 5 times that obtained by typical usage of MACD and RSI.

A two-stage architecture, using a Self-Organizing Map (SOM) and Support Vector Regression (SVR), appears to capture the dynamic input-output relationship inherent in financial time series forecasting [29, 68]. In [29], the Exponential Moving Average (EMA) and close prices that were projected into Relative Difference in Percentage of Price (RDP) were used as model inputs. The predicted target was the RDP in the following 5 days. In order to determine the size of SOM, the Growing Hierarchical Self-Organizing Map (GHSOM) [59] is adopted. The results of [29] are estimated by normalized Mean Squared Error (NMSE), Mean Absolute Error (MAE), etc. These showed that the two-stage architecture outperformed a single SVR model. The regression result evaluated by MAE is 10% better than that of a single SVR model. Their two-stage architecture is shown in Fig.2.2. The ICA-SVR method for two-stage model was proposed by Lu et al. [43]. Independent Component Analysis (ICA), which is a dimension reduction technique, was first applied to price series to remove noise. Then an SVR was employed to build the prediction model. A better performance, with an improvement of around 8% evaluated by MSRE, was achieved, compared to the single SVR model.

Kim [34] compared the prediction accuracy of the direction of changes in the daily Korea Composite Stock Price Index (KOSPI) obtained by using an SVM, a Back-Propagation Neural Network (BPNN) and Case-Based Reasoning (CBR). Various technical indicators such as RSI and Commodity channel index (CCI) were chosen as model inputs. The SVM obtained the highest accuracy, which is 57.8% while the accuracy obtained by BPNN and CBR is 54.7% and 52.0% respectively. Huang et al. [31] applied an SVM to forecast weekly movement of NIKKEI 225 index. The

S&P 500 Index and the exchange rate of US Dollars against Japanese Yen (JPY) were the inputs of the models. The accuracy of the SVM on directional prediction was 73%, which was better than that of Random Walk (RW), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Elman Backpropagation Neural Networks (EBNN). The ability of an SVM to minimize the structural risk enables it to be more robust to overfitting. [11]

2.2.2 News analysis

Although some specialists [47, 53] believe that all relevant information is included in stock prices, it still takes time for investors to respond to the new information. In this case, news analysis is likely to assist in price movement predictions.

Text mining techniques such as bag-of-words models and topic models are widely used in news classification tasks. The targets of the instances in stock price prediction models are assigned by future price movements. [52]

Newscats [49] adopted a bag-of-words model with a local dictionary. The prediction was made based on the performance of the stock prices in the next hour. The vector features were represented by the presence of the words but not their frequency. The frequency of movement prediction was 15 seconds. An overall classification (good/no move/bad news) accuracy of 45% was obtained. This relatively disappointing result may be due to the short length of the prediction introducing too much noise into the analysis.

Mahajan et al. [45] used Latent Dirichlet Allocation (LDA) to identify topics of financial news. The stacked classifier adopted was designed based on an SVM and decision tree. The average directional accuracy achieved was 60%. Different temporal and behavior patterns were discovered in different topics and contexts. This work shares a similar idea to the two-stage architecture approach [29, 68].

Schumaker and Chen [63] applied an SVM to S&P 500 stocks with four feature representations: bag of words, noun phrases, proper nouns (a subset of terms from noun phrases) and named entities (essentially specialized proper nouns). The representation of proper nouns was regarded as the hybrid form of noun phrases and named entities and it achieved the best performance among the four textual features (58.2% in directional accuracy and 0.04433 in MSE for closing price results).

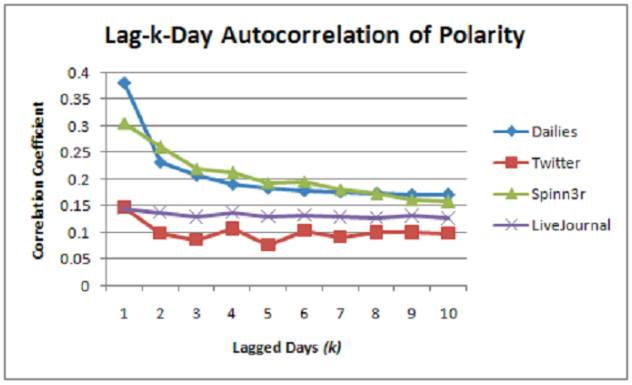


Figure 2.3: Correlation coefficient analysis of Polarity's Lag-k-Day autocorrelation for Dailies (News), Twitter, Spinn3r (blog), and Live-Journal (blog) severally. [81]

2.2.3 Blogs, tweets and other analysis sources

With the growth of social media, traditional news press releases are no longer the only major information sources for investors.

Zhang and Skiena [81] compared different sorts of text sources, namely news, blogs and tweets, and discovered that the sentiments expressed by blogs and tweets had a longer impact duration than those conveyed by news. The correlation coefficient of news dropped dramatically after the first lagged day, as is illustrated in Fig.2.3.

Words that express mood as tags such as #Hope, #Happy and #Fear are investigated by Zhang et al. [82] who discovers that the emotional tweet percentage is positively correlated to VIX but negatively correlated with Dow Jones, NASDAQ and S&P 500. For instance, the correlation coefficients of #Hope to Dow, NASDAQ, S&P 500 and VIX are -0.381, -0.407, -0.373 and 0.337 respectively. It is also implied in [82] that during the period when economic conditions are uncertain, words with stronger emotions such as hope, fear and worry are more likely to be used in tweets.

Bollen et al. [9] estimated public emotions from Twitter feeds via the following two approaches:

1. OpinionFinder[4], which rates positive and negative attitudes

2. The Google-Profile of Mood States (GPOMS) [8] that evaluated the public's sentiments in 6 dimensions, namely, Calm, Alert, Sure, Vital, Kind, and Happy.

The mood states, represented in the form of time series and combined with daily prices, were modeled using a Self-Organizing Fuzzy Neural Network (SOFNN) to predict the

[4]http://www.cs.pitt.edu/mpqa/opinionfinder.html

Dow Jones Industrial Average (DJIA) closing price of the next day. Among positive and negative moods and the 6 dimensions, calm was observed to have the highest Granger causality relation with DJIA, whose p-value vary from 0.013 to 0.065 corresponding to 2 to 6 lagged days. A directional accuracy of 86.7% was obtained.

Work [3, 13, 19, 28, 37, 48, 60] has followed the methodology proposed by Bollen et al.'s, but none has obtained similar accuracy [9]. Instead of word counts in word lists, Hsu et al. [28] has used Singular Value Decomposition (SVD) grouping as high-level features in sentiment analysis and achieved a better prediction accuracy (82% for the training in 6 groups), compared with the word list extended from the Profile of Mood States (POMS) (less than 60%). The SVD, a factorization of a real matrix, was used to identify the biggest components in the word covariance matrix. Their POMS directional accuracy was far lower than GPOMS [8] as they did not build the word list using proximity correlation and the calm feature was not identified in their experiment.

It is indicated in [1] that Bollen et al.'s experiment should be evaluated carefully and Sprenger and Welpe's work [66] is recommended because their approach is more straightforward and detailed. Sprenger and Welpe [66] have investigated the relationships between individual stock prices and tagged tweets (e.g. with $AAPL, $GOOG). Retweets and followership were taken into consideration so each tweet had a different weight. It was observed that the bullishness, the message volume and the agreement are highly associated with the trading volume, whose F-value is 200.5, 287.2 and 201.9 respectively ($p-value < 0.001$). Ruiz et al. [62] have identified two groups of features and stock market events:

1. The features in the first group measure the overall activity in Twitter, such as numbers of re-posts.

2. The second group concerns the properties of an induced interaction graph, such as connected components.

The results showed that the number of connected components of the constrained subgraph was the best feature with regard to correlation[5], especially with regarding to the trading volume.

[5]The average cross-correlation coefficient of the number of connected components to the trading volume is 0.33 on the first lagged day.

2.3 Summary

This chapter has considered a number of research methodologies such as learning algorithms and evaluation methods are introduced in the first section. Various aspects of research on stock price prediction using technical indicators, news and social media have been considered. In the next chapter, the design principle and implementation of each module in the forecasting system are presented.

Chapter 3

Design approach

This chapter presents the design principle and implementation of each module of the forecasting system. In 3.1, the description of the data used in this dissertation is given; 3.2 illustrates the preprocesses for the features such as calculations for technical indicators and the transformation for raw text from news, blogs and tweets; 3.3, 3.4, 3.5 and 3.6 give the details and purposes of the modules and experiments; finally, a summary is given in 3.7.

3.1 Data preparation

Daily prices: Daily prices are fetched via the TTR module[1] in R-project. The implementation of the system is mainly in Python. The module that collects historical data is implemented in R. After the data is processed by R, it is stored in static files for later use by Python modules.

News: The news is obtained from Reuters Site Archive. The news used in this dissertation is published by PRNews Wire, Business Wire, Market Wire and Globe Newswire. As the news are filtered by manually selected keywords like company alias names and their stock tickers, some unrelated news are also retrieved. For instance, news title containing 'PM' which means 'after noon' can be confused with Philip Morris whose ticker symbol is 'PM'. The details of how news are selected are illustrated in Appendix A. The news crawler is implemented in Python.

[1] http://cran.r-project.org/web/packages/TTR/index.html

Blogs: Blog articles are fetched from SeekingAlpha. At SeekingAlpha, the relationship between blog articles and companies is visible to readers, thus the acquirement of blogs is easier than that of news. Matching rules are not necessary for the blog crawler. The crawler for blogs is implemented in Python.

Twitter tweets: Twitter tweets are collected through the Twitter Search API, with the keyword $TICKER such as $GOOG and $YHOO. However, there is a default API rate limit. The unauthentic API calls are allowed 150 requests per hour. In order to tackle the API limit issue, Condor at Manchester[2] is used to run the tweet collectors. The Twitter tweet collector is implemented in Python.

3.2 Preprocess

3.2.1 Technical indicators

Much research [17, 26, 31, 71] on stock forecasting with technical indicators has proved the utility of them. In this dissertation, a bunch of technical indicators[3] are used to perform modelling. The details of these indicators can be found in Appendix B.

However, indicators have been invented to support investors' decisions but not for supporting predictive computer-derived models. Two major transformations are conducted so that the indices are more readable as features.

Converting signals Some indicators contain signal indices, which hint key points to investors related to long or short stocks. Signals usually are generated when signal indices cross the main indices. Suppose T_m is the main index and T_s is the signal index. Signals are converted as Eq.3.1.

$$Sig = \frac{T_m - T_s}{T_m} \tag{3.1}$$

Normalizing As the indicators have different value ranges, it is wise to convert them into the same value range so that modelling will not be affected by dominant features. The normalizing is conducted as Eq.3.2.

[2] http://condor.eps.manchester.ac.uk/

[3] ADX, aroon, Bollinger Bands, CCI, ROC, DPO, EMV, MFI, OBV, RSI, Stochastic Oscillator, SMI, TDI, TRIX, VHF, Williams Accumulation / Distribution and WPR

$$\tilde{s_i} = \frac{s_i - min(s)}{max(s) - min(s)} \tag{3.2}$$

3.2.2 Bag-of-words model

Raw text in news, blogs and tweets are converted into word count vectors. The text of news, blogs and tweets published on the same day are merged into one document. The following steps are conducted on the single document.

1. Apply Porter stemming algorithm to convert words to their stem or root forms.
2. Remove the words with the same roots of the words appearing in the stop-lists[4].
3. Remove the words with a frequency less than 50 times in all documents of news and blog documents in order to save computational resources.

The term preprocess is implemented in Python. The associated library for the Porter stemming algorithm is from nltk[41].

3.2.3 Topic modelling

BOW models ignore the semantic relationships among words. [30] Terms with different stems sometimes share the similar semantic meanings. In order to disambiguate the terms and reduce the feature dimensionality, a topic model is utilized. Topic modelling is a popular solution used to identify synonyms among word stems. Latent Dirichlet Allocation (LDA) [7] is adopted in this dissertation. The number of topics range from 2^5 to 2^{10}. The key words in the results of modelling with fewer topics are likely to be real topics while the key words in the results generated with more topics are likely to be synonyms.

The LDA library used is from gensim [61].

3.3 Sentiment analysis

Previous work [9, 69, 82] has demonstrated the predictive value of sentiments from social media to stock market movement prediction. It is necessary to identify which sentiment analysis methodologies are suitable for specific security price prediction. In this section, several sentiment analysis approaches are illustrated.

[4]Generic, names, dates and numbers, geographic and currencies from http://nd.edu/ mcdonald/-Word_Lists.html

3.3.1 Dictionaries

It is easy to identify the polarity of a word with dictionaries containing positive and negative tags. There are two well-accepted dictionaries, namely, General Inquirer (GI) as is adopted in the work in [69, 70] and Loughran and McDonald Financial Sentiment Dictionaries (LM) as is used in the work in [46].

In this dissertation, both dictionaries are used to extract positive and negative sentiments from text sources. There are many other categories or tags in those two dictionaries such as “Modal Words Strong” and “Uncertainty Words”. Whether those extra tags are useful in prediction are evaluated in the experiments reported here.

GI dictionaries consist of 182 categories, which are for general use. Some of the categories are relatively associated to sentiments such as ”Positiv” (positive outlook) and ”Negativ” (negative outlook), while some seem to be irrelevant such as ”DAV” in which words describe an action, e.g. ”run, walk, write”. ”Positiv” and ”Negativ” are two large valence categories.

There are 6 categories in LM, namely, negative words, positive words, uncertainty words, litigious words, modal words strong and modal words weak. As LM is developed for financial text analysis, there are no irrelevant tags such as ”MALE” and ”Female” which are categories in GI.

3.3.2 Polarity and Subjectivity

The sentiments are presented in the Lydia style. The formula of polarity and subjectivity in the Lydia style are given in Eq.2.17 and Eq.2.18 .

With the GI dictionaries, the positive score and the negative score are the counts of the words in the category ”Positiv” and the category ”Negativ” respectively. As for LM, the scores are from the positive words and negative words.

Groups of topic distribution features are used to extract sentiments as well. They are generated by LDA [7]. The topics modeled by LDA have no manual guidance for polarity. The polarity score of the each LDA topic is evaluated by χ^2 statistics. The equation is given in Eq.2.21.

In this dissertation, χ^2 statistics of the topics are calculated based on the SMP score on the next nth day . Before normalizing, the scores range from -3 to 3. The calculation of the SMP scores is illustrated in 4.1.2. As it is a binary classification, c stands for *good* and $\bar{c}$ stands for *bad*. The label of *uncertain* is omitted in topic polarity calculation. If the SMP score is above 0, the A in Eq.2.21 of the clustered topics whose

distribution is above 0 in the instance will increase. If the SMP score is below 0, the B in Eq.2.21 of the topics whose distribution is above 0 in the instance will increase. Thus, the higher the χ^2 statistics is, the more positive a topic is, and vice versa.

In each group, the number of positive and negative topics are kept no more than 50. For instance, there are 50 positive topics and 1 negative topic for the next day prediction of the security NOV where the topic number of the model is 64. The other 13 positive topics[5] are removed from the list in order to remove noisy topics whose polarity is not obvious.

3.3.3 Smoothed sentiment scores

This dissertation follows the POMS smoothing style to normalize sentiment scores using the z-score. However, in Eq.2.15, future information is included in the equation, which may make the sentiment features be involved with future sentiment information that should not be contained in "current" features. The equation is modified so as to remove the information from the future, as is illustrated in Eq.3.3. $\bar{x}(m_{i-k+1}, m_{i-k+2}, \ldots, m_i)$ represents the average sentiment score from the m_{i-k+1}th day to the m_ith day. σ is the standard deviation function.

$$\tilde{m}_i = \frac{m_i - \bar{x}(m_{i-k+1}, m_{i-k+2}, \ldots, m_i)}{\sigma(m_{i-k+1}, m_{i-k+2}, \ldots, m_i)} \tag{3.3}$$

3.4 Context analysis

In [29], the two-stage architectures showed around 10% improvement evaluated by MAE in prediction using price data. It therefore is a reasonable approach to apply the same architecture in analysis of textual data. The same piece of news of different trends may have a different affect on investors.

Context analysis is applied before training on text sources, as in the two-stage architecture approach used in [29]. In this dissertation, context analysis is based on clustering historical price / volume information. The textual data contains the latest information that might affect investors, but the historical price data represents concrete results of investors' decisions. Thus it is more intuitive and easier to analyze the trends and contexts from price data than from textual data, even with time series of sentiment scores.

[5] Maybe the number of the other positive topics is under 13 as some topics might be neither positive nor negative.

Input variable	Calculation
EMA15	$(p(i) - EMA_{15}(i)/EMA_{15}(i)$
RDP-5	$(p(i) - p(i-5))/p(i-5)$
RDP-10	$(p(i) - p(i-10))/p(i-10)$
RDP-15	$(p(i) - p(i-15))/p(i-15)$
RDP-20	$(p(i) - p(i-20)/p(i-20)$

Table 3.1: Input features of context analysis

The features used in clustering are the same as in [29]. The original closing prices are turned into the percentage of price (RDP) and EMA_{15}. It is claimed in [68] that this transformation makes the data more predictive. The calculation is given in the Tab.3.1. In this table, $p(i) = EMA_3(i)$.

After context analysis, data are split into sub-data. Regression is conducted on each group of sub-data. Thus, the models for the data with different contexts are different.

The context analysis is conducted on features from both technical indicators and textual data. In [29], there is no comparison of the performance among Growing Hierarchical Self-Organizing Map (GHSOM) and the other algorithms discussed. In this dissertation, K-means and GHSOM are both used to cluster historical prices to see which is more suitable for context analysis.

K-means from scikit-learn[55] is used; GHSOM is implemented in Python according to [59].

3.5 Feature extraction

High dimensionality data may lead to the curse of dimensionality [5]. There are many groups of high dimensionality data adopted in this dissertation, such as word occurrences in news, blogs and Twitter tweets with around 8000 features, and topic features generated by LDA with 512 and 1024 attributes.

Hsu et al.'s work [28] shows that high-level features may give a more promising result than sentiment features. The best accuracy result obtained with SVD features was 82% while the result achieved with sentiment features was 60%. Their prediction target is the movement of DJIA, while in this dissertation the prediction targets are price performance of specific securities in the S&P100. It is to be discussed in the

Price/volume related data	Features from textual data
Technical indicator features shown in 3.2.1	Promising bag-of-words models
Technical indicator features shown in Tab.3.1	Promising bag-of-words models
Technical indicator features shown in 3.2.1	Promising sentiment features
Technical indicator features shown in Tab.3.1	Promising sentiment features

Table 3.2: Features used in combination experiments

results of feature extraction experiment to see which performs better in price movement prediction for specific securities.

The PCA is a popular feature extraction technology; The associated library used in this dissertation are from scikit-learn[55].

3.6 Feature combination

In 3.4, the context analysis with historical price information is discussed. In spite of some possible improvements in price movement prediction, especially the prediction with textual data, it is hard to confirm the fact that the clustering before classification is a promising methodology, as it introduces new information, namely, price and volume data besides the textual features. In order to ensure the power of context analysis, the features of technical indicators and textual data are combined.

As is shown in Tab.3.2, several groups of experiments are conducted in this dissertation in order to identify a promising combination.

Thus, there are three purposes in this group of experiments. One is to compare the performance of the technical indicator features proposed in this dissertation and the features used in [29, 68]. Another purpose is to see if the context analysis is better than feature combination as both introduce price and volume related features beyond the features extracted from textual data. The final purpose is to identify if there is any improvement in feature combination of numeric and textual features.

3.7 Summary

The design principle and implementation of each module in the forecasting system have been given in this chapter. The description of the data, the preprocesses used on the data before modelling and the details and purposes of the modules and experiments have been discussed. In the next chapter, the setup of the experiments will be given.

Chapter 4

Experimental framework

In this chapter, the setup of the experiments is illustrated. The baseline experiment is shown in 4.1. The experiments with sentiment features are given in 4.2. The experiments with the two-stage architecture, feature extraction and feature combination are given in 4.3, 4.4 and 4.5.

4.1 Basic experiments

A group of basic experiments are conducted to establish a baseline. These experiments are conducted on features from technical indicators and BOW models for news, blogs and tweets. The following experiments share the same workflow as the baseline experiments.

4.1.1 Features

Technical indicators: Technical indicators illustrated in 3.2.1 and their transforms are adopted as features.

BOW models: Textual data from news, blogs and tweets are converted into vector features. Only the words outside the stoplists and with high frequencies (> 50 in all news and blogs) are remained in the vector features.

4.1.2 Prediction classes

In the experiments, the prediction target is stock movement performance (SMP) on securities in S&P in the next nth day. The classes for SMP are *good*, *uncertain* and

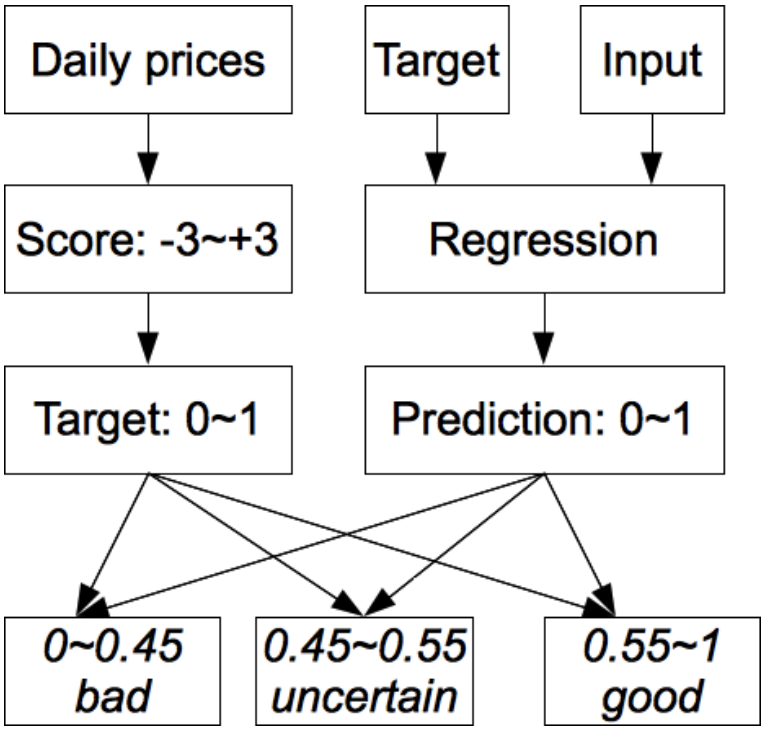


Figure 4.1: Performance prediction

bad.

The following criteria are used to decide if the price change prediction accuracy is *good*.

1. The closing price of the next *n*th day is higher than that of today.
2. The high price and the low price of the next *n*th day are both higher than that of today.
3. The closing price of the *n*th day is higher than the opening price of that day.

There are also three criteria for describing how *bad* a price change is.

1. The closing price of the next *n*th day is lower than that of today.
2. The high price and the low price of the next *n*th day are both lower than that of today.
3. The closing price of the *n*th day is lower than the opening price of that day.

The workflow of SMP scoring is illustrated in Fig.4.1. If any of the good criteria is met, the SMP score is incremented. If any of the bad criteria is met, the SMP score is

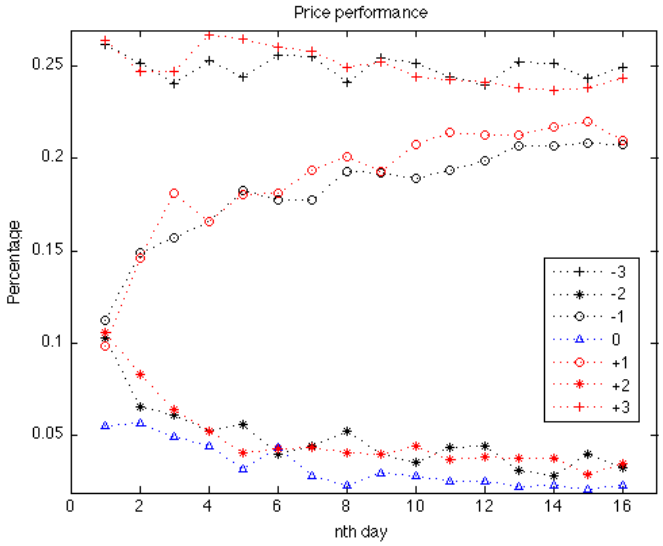


Figure 4.2: SMP score distribution

decremented. Thus, the SMP scores range from -3 to +3. Finally the SMP scores are normalized to 0˜1.

Regression is used in the experiments. The target for the regression model is the SMP score. If the regression result is greater than 0.55, the classification result will be *good*. If the regression result is less than 0.45, the classification result will be *bad*. Otherwise, it will be *uncertain*, where the target score should be 0.5.

In the experiments in this dissertation, the predicted days are the next first working day to the next fifth working day from the current day.

The percentage distribution trend of the SMP scores is illustrated in Fig.4.2. The percentages of SMP score of -3, -2, 0, +2 and +3 show a downward tendency when the prediction cycle increase while the proportions of the SMP score -1 and 1 obviously increase with the days. This indicates that the SMP tends to be very good with score +3 (very bad with score -3) or slightly good with score +1 (slightly bad with score -1) when the prediction day interval increase.

4.1.3 Training and evaluation

An SVR is used for regression in this dissertation. The library of SVR used is from libsvm[16]. The type of SVM is set to epsilon-SVR and the kernel type is set to radial basis function. There is a Python interface provided by libsvm, thus it works well with the experiment system. Both parameters used in tuning SVR, c and γ, range from 2^{-2}

to 2^4.

The results are evaluated by 10-fold cross-validation. The data split into 10 folds is kept in time order so that the independence of the instances in time span is guaranteed.

The modelling is conducted on each security, and the targets are the SMP of nth day ($n = 1...5$). There are more than 70 securities[1] in each data source, which means more than 350 datasets used to train and evaluate the models in each group of experiments. This job is distributed on Condor at Manchester so as to save time .

4.2 Experiments using sentiment features

Two groups of experiments are conducted using sentiment features, namely, ready-made dictionaries and topic features. The prediction target in these experiments is the SMP. The purpose of these groups of experiments is to see if the topic features show better potential than ready-made dictionaries.

4.2.1 Features from ready-made dictionaries

Three experiments are conducted on the General Inquirer(GI) and the Loughran and McDonald Financial Sentiment Dictionaries (LM).

1. Tag counts
2. Tag counts and sentiment scores
3. Sentiment scores

The sentiment scores consist of polarity and subjectivity, which are given in Eq.2.17 and Eq.2.18.

4.2.2 Features from topic distributions

Two experiments are conducted on topic distributions generated by LDA. There are 6 groups of topic distributions where the topic numbers are 2^5, 2^6, 2^7, 2^8, 2^9 and 2^{10}.

1. Topic distributions
2. Sentiment scores

[1]Some securities are missing in Reuters news because of the immature security matching mechanism.

In the experiments on sentiment scores, features with raw scores and smoothed scores in z-score are evaluated. Both of them are presented as features of a single day as is illustrated in Eq.4.1 and 5 days as is shown in Eq.4.1.

$$f = \{polarity, subjectivity\} \tag{4.1}$$

$$f_{series} = \{f_{-4}, f_{-3}, f_{-2}, f_{-1}, f_0\} \tag{4.2}$$

The features of sentiment score series are conducted with the topic numbers, which yield the best results with features of sentiment scores in news, blogs and tweets.

4.3 Experiments using context analysis

Context analysis is to cluster with historical price information before regression. The prediction target in these experiments is the SMP. Context analysis is done both on features of technical indicators and textual data. K-means and GHSOM are two unsupervised learning algorithms used in this group of experiments.

The clustering is based on the data described in Tab.3.1. The modelling with textual data is only conducted on the bag-of-words models.

The parameter for K-means, k, is set to 9, which is close to the average number of clusters obtained by GHSOM. The parameters of GHSOM, $\tau_1$and τ_2, are set to 0.5 and 1.

4.4 Experiments using feature extraction

PCA is conducted on bag-of-words models of news, blogs and Twitter tweets, the promising groups of the LDA topic features, which yield good prediction results in sentiment analysis[2] and the LDA topic features with the largest number of topics. The experiments on the features with the largest number of topics are to see if they provide a greater accuracy than the features that obtain the best results from sentiment analysis.

The prediction target in these experiments is the SMP.

The numbers of the components of PCA are set to meet the lowest amount of the variance that needs to be explained; the amount of variance is set to be 0.9.

[2]LDA#512 for news and LDA#64 for blogs and tweets.

4.5 Experiments using the features combined with technical indicators and textual data

As is illustrated in 3.6, features are combined in Tab.3.2. In order to compare the performance of the groups of feature combination with the basic technical indicator experiments, the experiments using features of technical indicators have been conducted again in the time period January, 2010 to December, 2011 and April 25, 2012 to July 19, 2012 in order make the time period of the data similar to news and tweets respectively.

A further group of experiments has been conducted on an extended prediction cycle up to the performance on the next 16th day on the features from tweets. There are two groups of features on technical indicators whose time spans are around 6 years and 3 months respectively. The data in 3 months matches the time span of the tweet data. The aim of the experiments conducted on 6 years is to check if the data in a much shorter time span gives similar performance.

Two prediction targets are used in this group of experiments. One target is SMP, as illustrated in 4.1.2. The other target is the stock movement direction (SMD), which is the comparison of the current closing price with closing prices the next *n*th days . The purpose of the latter target is to make the approaches in this dissertation more comparable to other work.

An SVR is used for SMD prediction. The SMD targets, *up* (1) or *down* (0), are indicated by the future closing price being greater / lesser than the current one. If they are equal, the target will be *unchanged* (0.5). The projection from 0˜1 to *up* / *unchanged* / *down* is similar to the SMP projection. If the regression result is greater than 0.55, the classification result will be *up*. If the regression result is less than 0.45, the classification result will be *down*. Otherwise, it will be *unchanged*, where the target score should be 0.5.

As can be seen from Fig.4.3, the SMD class percentage distributions of *up*, *unchanged* and *down* are stable in the prediction intervals from day 1 to day 16.

4.6 Summary

In this chapter, the framework of the experiments has been discussed. The baseline experiments and the workflow of all the experiments have been given. The details for the setup and parameters for the algorithms used in the experiments have also been

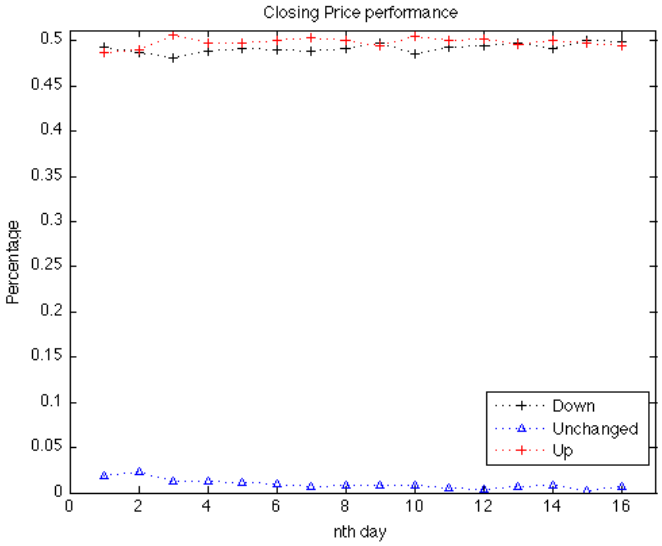


Figure 4.3: SMD score distribution

discussed. In the next chapter, the results and their analysis will be presented.

Chapter 5

Results and analysis

The analysis of the results is given in this chapter. 5.1 shows the results on the baseline experiments; the performance of two ready-made dictionaries and LDA models on sentiment analysis is discussed in 5.2; the analysis on context analysis and feature extraction is illustrated in 5.3 and 5.4; the results of feature combination are discussed in 5.5; finally, a summary is given in 5.6.

5.1 Basic experiments

The results of the basic experiments are shown in Fig.5.1. The details of the results can be found at Tab.D.1. It is clear that the prediction power of technical indicators surpasses those of the bag-of-words models of news, blogs and tweets. The accuracy of SMP prediction for the 2nd day is above 75%, while the results of the experiments on the textual data are all below 70%.

Among the textual data, the bag-of-words models of Twitter tweets yield the best results. The SMP classification accuracy is almost above 60%. Investors' comments might express more confidence or disappointment than news and blogs or the extraction of the polarity is easier with the bag-of-words models of Twitter tweets.

A further issue found in the results is that the longer is the prediction period, the greater is the accuracy. It seems that trends of price movements play a crucial role in trading. The poor performances of the near future predictions might be caused by market fluctuations.

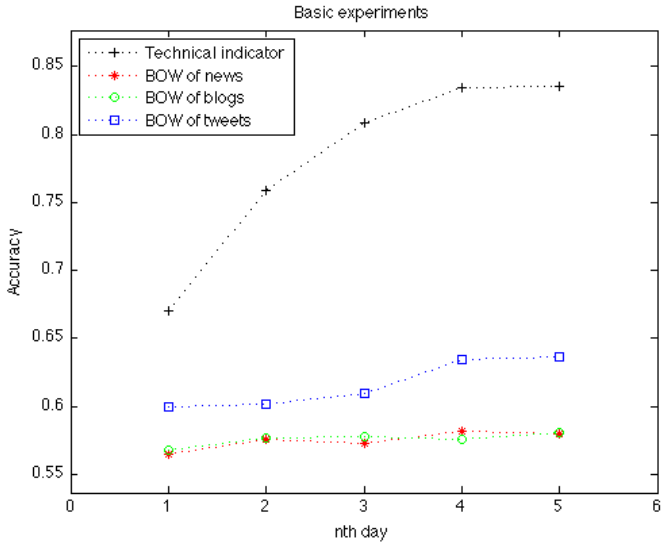


Figure 5.1: Results of basic SMP experiments: "BOW" stands for bag-of-words model.

5.2 Experiments using sentiment features

5.2.1 Sentiment scores from GI and LM

The SMP prediction accuracy using GI and LM sentiment scores are illustrated in Fig.5.2, Fig.5.3 and Fig.5.4. The details of the results are illustrated in Tab.D.2, Tab.D.3 and Tab.D.4. It is obvious that the results of the LM groups are better than the results of GI with the textual data of news while GI outperforms LM with the textual data of tweets. It is not clear which dictionary category is better with regard to the textual data of blogs. This may be linked with the fact that more facts and financial terms are used in news and more words with general emotion semantics appear in tweets while blogs are mixed with facts and judgements so that neither of them outperforms the other.

The increasing accuracy with increasing forecasting days is clearly shown in Fig.5.2 and Fig.5.4, as the news and tweets are instant reactions or reports to current situations, while realtime reflection is not a feature of blogs.

Tab.5.1 shows the average standard deviation of the *good/uncertain/bad* classification results. The results of GI are always more stable than those of LM, especially in the groups of tweets. The text analysis during the design of LM is based on form 10K (annual financial report). The instability may stem from the differences between the contexts of financial reports and the text sources used in this dissertation.

LM appears more suitable for financial articles. If news can be identified that it has

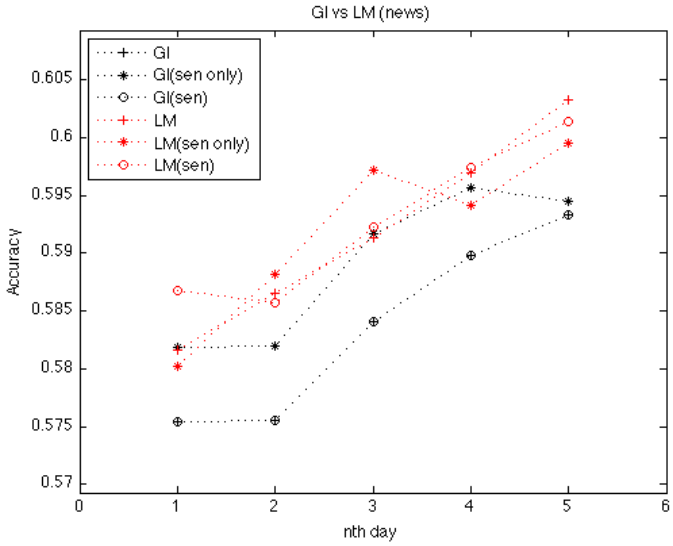


Figure 5.2: Results of SMP prediction with GI and LM (news): In groups with "sen only", the instances only have the polarity and subjectivity scores as features. In groups with "sen", the instances have both dictionary category counts and sentiment scores as features.

Source	GI	LM
News	6.00%	6.21%
Blogs	5.66%	6.1%
Tweets	5.84%	6.76%

Table 5.1: Average standard deviation of SMP prediction accuracy with GI and LM: The standard deviation is averaged over all three methods (tag counting, sen, sen-only) and prediction days (1˜5)

a more financial context such as a quarter report, then the models of LM can be used otherwise models of GI can be used.

5.2.2 Sentiment scores from topics generated by LDA

Topic distributions

The top key words for the topics modeled by LDA with the topic number 32 and 128 are illustrated in Tab.5.2 and Tab.5.3 respectively. The topic key words generated by LDA with other topic numbers are shown in Appendix C. The words appearing in the tables are stemmed. The semantics of the topics in Tab.5.2 are vague. The topics listed are similar to each other. "Market" and "company" appear in many topics. It is easier

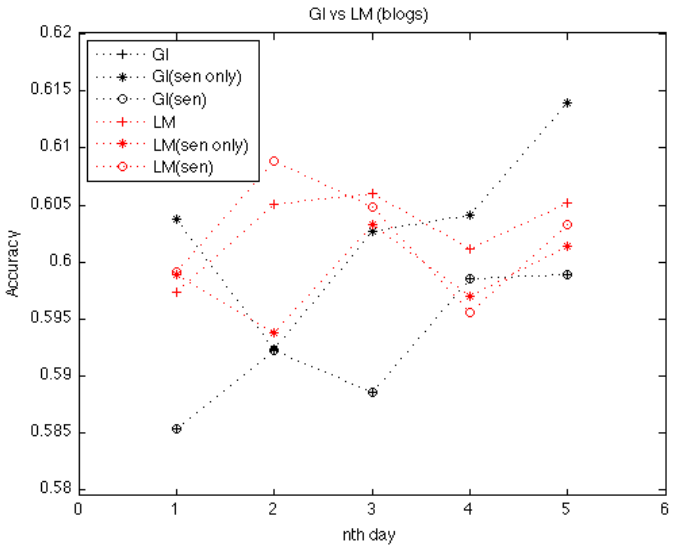


Figure 5.3: Results of SMP prediction with GI and LM (blogs)

to identify the semantics in the topics in Tab.5.3. For instance, topic#1 is a topic of manufacturing industry and topic#8 is a smartphone topic.

The results of SMP prediction accuracy using topic distributions are illustrated in Tab.5.4, Tab.5.5 and Tab.5.6[1]. It can be seen that the groups with fewer topics obtain

[1]In Tab.5.6, the average accuracy of the prediction on the 3rd day with the #topic 512 is 62.688%, while the result with the #topic 128 is 62.693%.

Topic ID	Top 5 key words
0	market, company, research, service, invest
1	verizon, company, oper,wireless, service
2	company, market, data, manag, buss
3	company, research, oracl, market, product
4	market, provide, company, service, time
5	company, buss, rsquo, oper, market
6	company, service, buss, product, manag
7	service, rsquo, market, product, oper
8	invest, market, earn, include, nyse
9	service, buss, provide, company, custom

Table 5.2: Top topics modeled by LDA with topic number 32

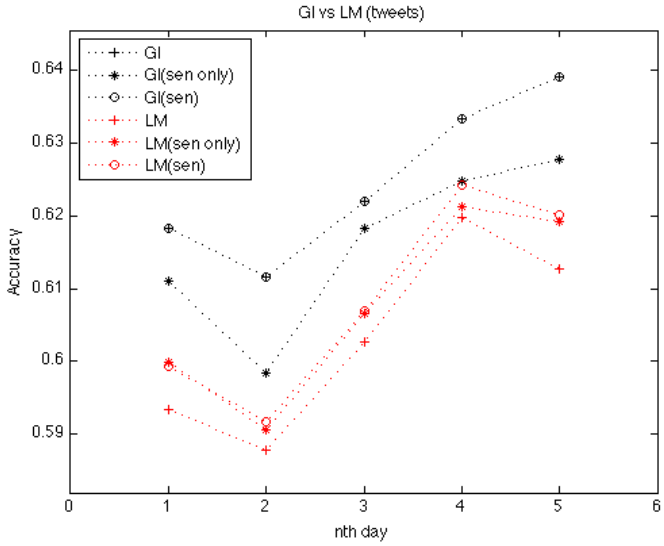


Figure 5.4: Results of SMP prediction with GI and LM (tweets)

slightly better results than the groups with more topics.

Sentiment scores

The sentiment polarity of the LDA topics is estimated by χ^2-statistics. Tab.5.7 and Tab.5.8 show the topics with polarity in blogs and tweets related to CSCO with LDA topics trained with number 64. The topic list in Tab.5.7 is a complete one. The topics appear to be so because the LDA models are trained from news and blogs, but not tweets. The topic #1 and #0 are negative in Tab.5.7, but they are positive with extremely low scores in Tab.5.8. The textual data, blogs and tweets, are generated in

Topic ID	Top 5 key words
0	market, manufacture, group, product, industri
1	service, buss, global, company, provide
2	company, exelon, dividend, sharehold, statement
3	para, comcast, servicio, client, como
4	pharmaci, caremark, plan, health, provide
5	store, mart, retail, pharmaci, custom
6	invest, rank, http, nyse, research
7	service, mobil, commun, buss, verizon
8	mobil, iphon, phone, smartphon, android
9	manag, virtual, service, desktop, client

Table 5.3: Top topics modeled by LDA with topic number 128

#Topic	1	2	3	4	5
32	57.77%	58.37%	**59.75%**	**59.81%**	59.58%
64	**58.26%**	**59.00%**	59.00%	59.31%	**59.66%**
128	57.73%	58.12%	58.98%	59.03%	59.16%
256	57.26%	58.92%	59.07%	59.67%	59.37%
512	57.34%	58.04%	58.89%	59.38%	59.51%
1024	57.45%	58.21%	58.80%	59.16%	59.46%

Table 5.4: Results of SMP prediction with topic distributions (news)

#Topic	1	2	3	4	5
32	**59.58%**	59.69%	**59.91%**	60.13%	**60.77%**
64	58.79%	**59.89%**	59.63%	59.93%	60.53%
128	59.54%	59.93%	59.87%	59.81%	59.91%
256	58.80%	59.08%	60.10%	59.65%	60.15%
512	58.83%	59.63%	59.66%	**60.14%**	60.30%
1024	58.35%	58.59%	58.65%	58.77%	59.89%

Table 5.5: Results of SMP prediction with topic distributions (blogs)

#Topic	1	2	3	4	5
32	**62.71%**	61.59%	62.45%	63.93%	63.85%
64	62.32%	**62.07%**	62.62%	63.83%	**64.05%**
128	61.62%	60.84%	**62.69%**	**63.94%**	63.80%
256	61.59%	61.14%	62.53%	63.46%	63.65%
512	61.63%	61.03%	62.69%	63.67%	63.92%
1024	61.36%	60.55%	61.87%	63.34%	63.13%

Table 5.6: Results of SMP prediction with topic distributions (tweets)

Topic	χ^2 score
#3: company, product, earn, market, develop	+0.015537
#1: company, product, market, buss, service	-0.023575
#0: buss, service, company, market, provide	-0.005821

Table 5.7: Topics with polarity (LDA64-Tweet-1day-CSCO, complete topic list)

Topic	χ^2 score
#24:company, growth, product, include, oracl	+0.005494
#10:market, buss, product, company, service	+0.002405
#26:company, market, invest, amazon, service	+0.001887
#5:store, buss, company, include, game	+0.001132
#3:company, product, earn, market, develop	+0.001099
...	...
#1:company, product, market, buss, service	+0.000667
#0:buss, service, company, market, provide	+0.000030
...	...
#8:company, market, technolog, product, buss	-0.005450
#2:company, invest, market, research, buss	-0.005420
#4:company, research, market, manag, invest	-0.005420
#22:product, provide, service, system, develop	-0.000668
#20:market, company, intel, oper, provide	-0.000551
...	...

Table 5.8: Topics with polarity (LDA64-Blogs-1day-CSCO, partial topic list)

different time periods. It is also possible that these two topics seem to be neutral. An improvement might be achieved if a stricter filter is added to deicide the polarity of a topic to a context. The accuracy results for the SMP prediction on the first future day on CSCO are 53.42% for blogs and 67.96% for tweets.

The accuracy results for the SMP prediction using sentiment scores from LDA topics are shown in Fig.5.5, Fig.5.6 and Fig.5.7. The details of the results are shown in Tab.D.5. In the experiments with news, the group with the number of topics 512 always yields the best average performance. In the experiments with blogs and tweets, the groups with the number of topics 64 always achieve the highest average accuracy.

A large topic number seems to be suitable for sentiment analysis in news while a small topic number is suitable for sentiment analysis in blogs and tweets. The advantage of the large number of topic number in news analysis may be linked with the fact that the news are comprehensive while the source of blogs is a financial analysis blog site, where topics are more concentrated in financial topics.

The best performance of the small topic number in tweets analysis is associated with the training dataset in modelling of the topics. Compared with the features with the topic distributions where the largest accuracy difference is around 1%~2%, the advantage is not so obvious. Topic distributions in tweet analysis are sparse in models with more topics. Some tweets even have no topic distribution in any topics. If the LDA topics are modeled with more sufficient Twitter tweets, the results may be more accurate.

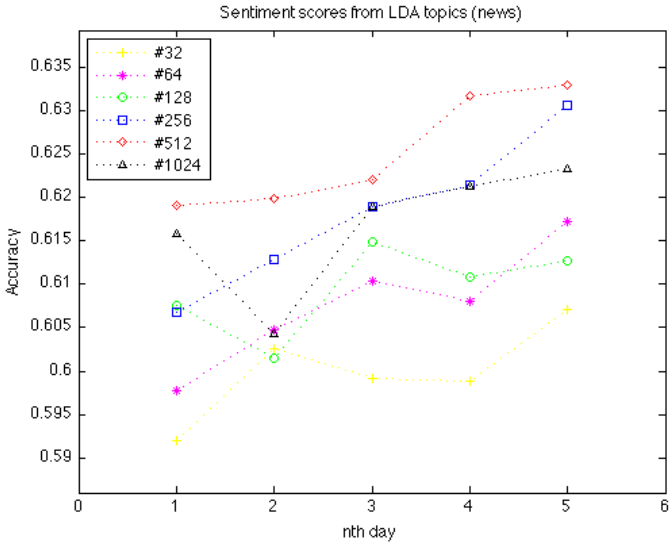


Figure 5.5: Results of SMP prediction with sentiment scores from LDA (news)

The sum and mean χ^2 statistics of topics[2] from day 1 to day 5 are illustrated in Fig.5.8. There is one common feature shared by news, blogs and tweets, which is that the models yield the best performances with the topic number where the mean χ^2 statistics of topics are the lowest or the second lowest. As for the sum of the χ^2 statistics, it grows with the number of topics in news and blogs, while there is a peak in tweet line at topic number 2^6 and the sum of the χ^2 statistics of the rest with the topic number more than 2^6 are extremely low. It indicates that models with higher topic numbers are not suitable for tweet analysis.

In summary, it seems that the higher sum and lower mean of χ^2 statistics lead to better performance. Higher sum and lower mean imply there are a few topics with high χ^2 statistics while the χ^2 statistics of most of them are low.

[2]The absolute values of positive and negative χ^2 statistics.

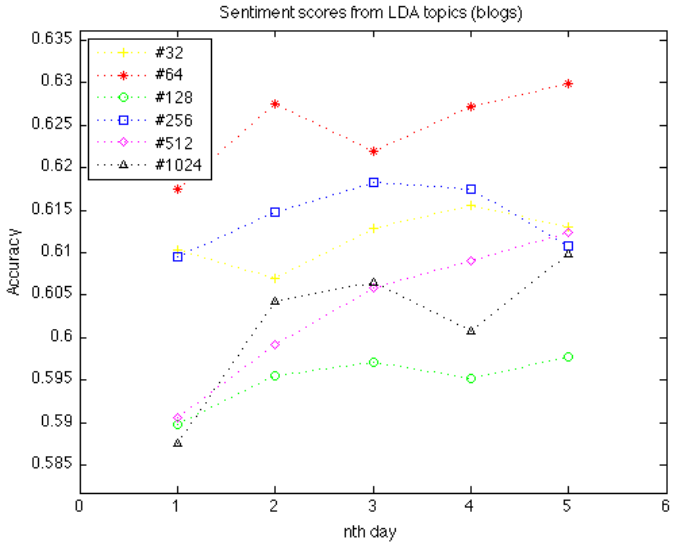


Figure 5.6: Results of SMP prediction with sentiment scores from LDA (blogs)

Features	1	2	3	4	5
Sen-score	**61.90%**	61.98%	62.21%	63.17%	63.29%
Sen-score series	61.75%	**62.23%**	**62.73%**	**63.77%**	**63.71%**
Smoothed score	61.26%	61.95%	61.85%	62.22%	62.13%
Smoothed series	60.95% *	60.98% *	61.54% *	61.39% *	61.84% *

Table 5.9: Results of SMP prediction using sentiment series and smoothed scores (topic#512, news): The best results in each prediction day are bold and the worst results are marked with "*".

Unsmoothed and smoothed sentiment score series

The results of average accuracy of series and smoothed scores for the SMP prediction are given in Tab.5.9, Tab.5.10 and Tab.5.11.

In the experiments with news, as is shown in Tab.5.9, the groups with smoothed scores do not give better performance than the original scores. The groups with smoothed score series are even worse, having the worst average accuracy in all groups. The group with sentiment score series (sentiment scores within 5 days) has the best results in most of the prediction days.

The poor performance of the normalization with the z-score is likely to be associated with the fact that news is made up of facts and events, which have little coherence with the previous news. The improvement made by score series might be associated with the context effects on investors.

In the experiments with blogs and tweets, as is shown in Tab.5.10 and Tab.5.11,

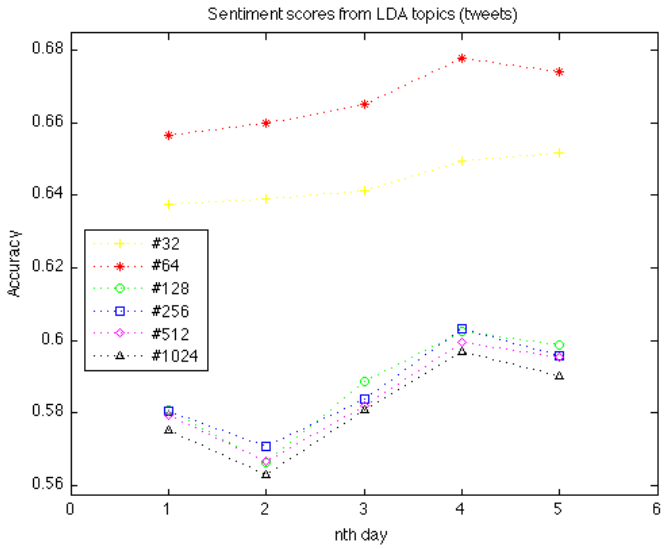


Figure 5.7: Results of SMP prediction with sentiment scores from LDA (tweets)

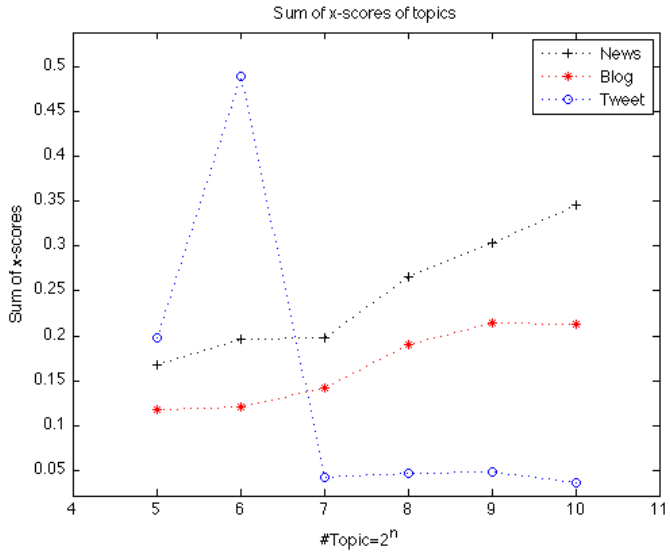
Features	1	2	3	4	5
Sen-score	**61.75%**	**62.74%**	62.19%	62.72%	62.99%
Sen-score series	61.37%	62.41%	**62.88%**	61.94%	63.25%
Smoothed score	61.04%	62.35%	62.77%	**63.10%**	**63.71%**
Smoothed series	59.26% *	61.90% *	61.53% *	60.55% *	61.53% *

Table 5.10: Results of SMP prediction using sentiment series and smoothed scores (topic#64, blogs)

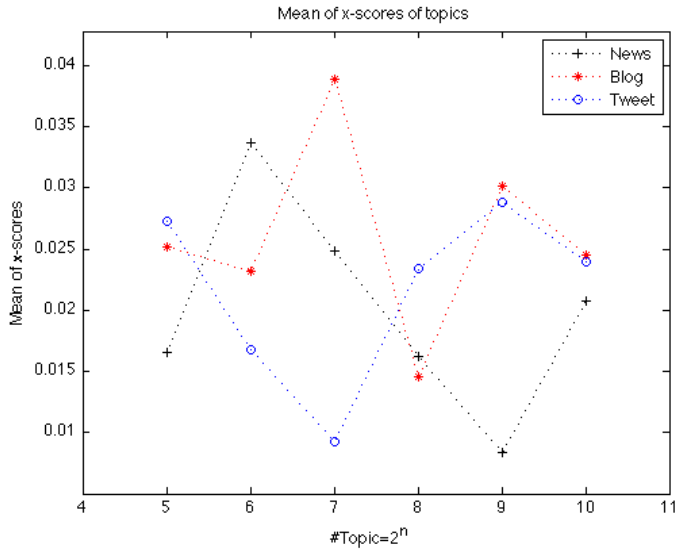
the groups with the features of smoothed scores have a better performance than in the analysis with news. However, the groups with the features of smoothed score series still give the worst performance. The score series in tweet analysis show good improvement, increasing the accuracy around 2% in the third day to fifth day prediction from the original sentiment scores.

The better performance of smoothed score in blogs and tweets, compared with the performance in news analysis, may be linked with the continuity of the investors. The text in blogs and tweets has the sentiment reflection of investors, which makes the normalization meaningful. The smoothed sentiment score series (5-day smoothed sentiment score) yield the worst performance all the time as these features may contain too much historical sentiment and too little current sentiment information. The one-day smoothed score may be better because the historical information does not dominate the instances.

News and tweets are instant reflection of the facts and events, which makes raw



(a) Sum



(b) Mean

Figure 5.8: χ^2 statistics of LDA topics

Features	1	2	3	4	5
Sen-score	**65.67%**	65.98%	66.52%	67.79%	67.40%
Sen-score series	65.63%	**66.93%**	**68.10%**	**70.35%**	**70.19%**
Smoothed score	65.53%	65.19%	66.87%	68.38%	67.89%
Smoothed series	64.65% *	65.31% *	65.22% *	66.38% *	66.35% *

Table 5.11: Results of SMP prediction using sentiment series and smoothed scores (topic#64, tweets)

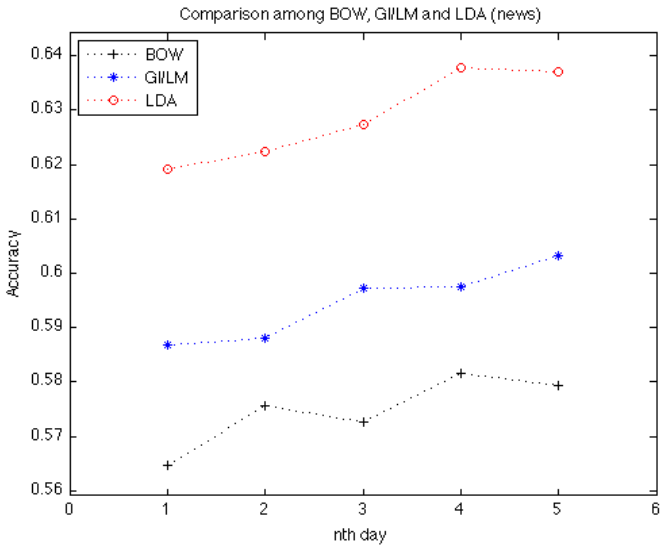


Figure 5.9: The comparison with BOW, GI/LM and LDA in SMP prediction (news)

sentiment scores more meaningful. Blogs are written after a period of thinking. There may be more considered and long-time influential thoughts, reflections and informed conjectures contained in blogs. The effect to writers, readers or others who share the similar ideas conveyed by the blogs may also be longer and have more significance. Thus the features of sentiment scores yield the best results in most of the experiments on news and tweets analysis, but not on blog analysis.

5.2.3 Comparison of BOW models, dictionary-based and topic-based sentiment analysis

The comparisons of BOW models, dictionary-based and topic-based sentiment analysis are illustrated in Fig.5.9, Fig.5.10 and Fig.5.11. The result accuracy is for the SMP prediction. It is apparent that features with sentiment analysis yields better accuracy results and topic-based sentiment analysis outperforms the others.

This indicates that the polarity detection based on modeled topics are more practical than that based on ready-made dictionaries, such as the General Inquirer (GI) and the Loughran and McDonald Financial Sentiment Dictionaries (LM). Contexts of the text source are different from the contexts where the dictionaries are made. Even though the LM is designed for financial applications, the semantics of words in news are still different from those in annual reports.

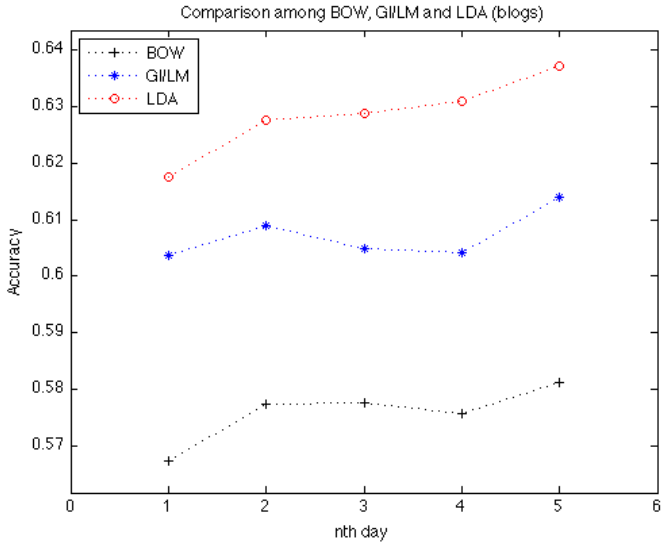


Figure 5.10: The comparison with BOW, GI/LM and LDA in SMP prediction (blogs)

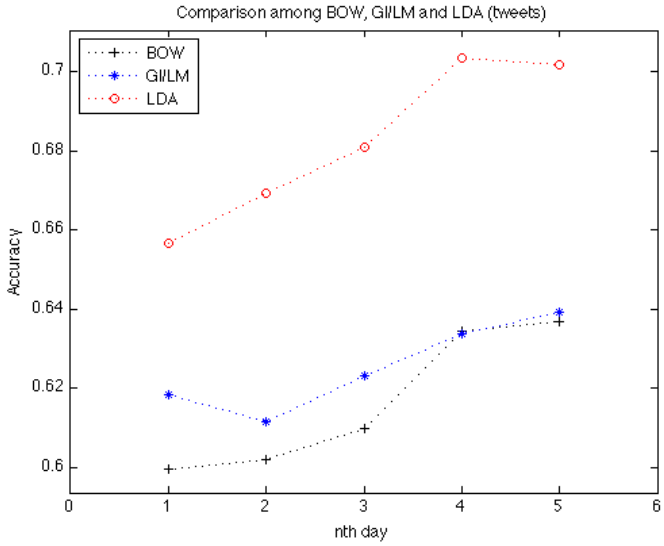


Figure 5.11: The comparison with BOW, GI/LM and LDA in SMP prediction (tweets)

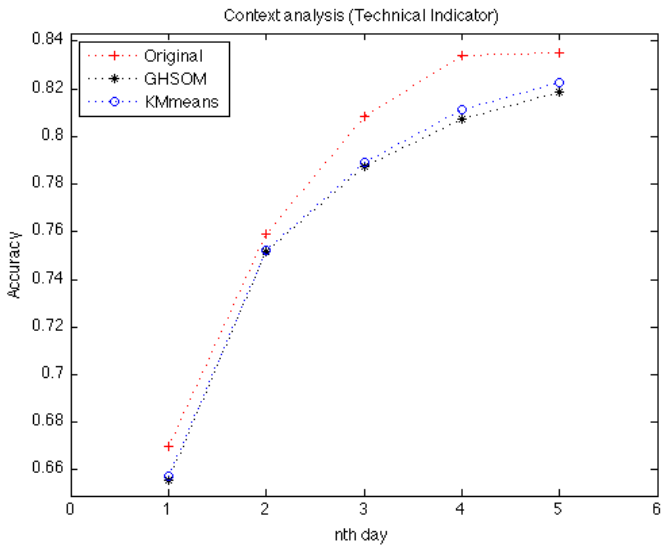


Figure 5.12: Results of SMP prediction using context analysis (technical indicators)

5.3 Experiments using context analysis

The results of SMP prediction accuracy with context analysis and features of technical indicators are shown in Fig.5.12. It is obvious that the original model without context analysis performs better than the ones with context analysis, via either K-means or GHSOM.

There is no improvement made by the two-stage architecture, which is different from the conclusions made in [29, 68]. This outcome may be contingent on the different features and prediction targets. The reason why context analysis weakens the performance of SVR may be that the clustering beforehand weakens the non-linear relationships among features.

The results of average accuracy with context analysis and features of textual data are shown in Fig.5.13 ,Fig.5.14 , and Fig.5.15. The details of the results can be found at Tab.D.6. In these experiments, context analysis improves the accuracy results by around 4%. Here K-means is more likely to yield a greater accuracy than GHSOM in news and blogs, but not in tweets analysis.

The improved results, which take with numeric features into consideration, are not sufficient to prove that the two-stage architecture is a good methodology for text analysis. The combination of numeric and textual features is a simple way in which historical price/volume data can also be considered. The comparison of the two-stage architecture and feature combination is given in 5.5.

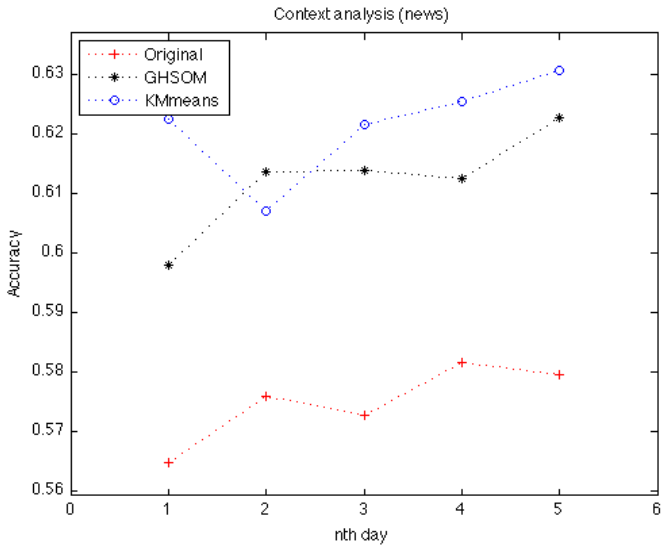


Figure 5.13: Results of SMP prediction using context analysis (news)

The different performances of GHSOM and K-means also make it difficult to conclude which is better for context analysis. More experiments based on specific industry sectors such as the technology and financial sectors may better assess the qualities of the algorithms for this type of application.

5.4 Experiments using feature extraction

The results of average accuracy for the SMP prediction with PCA applied on textual data are shown in Fig.5.16, Fig.5.17 and Fig.5.18. The details of the results are illustrated in Tab.D.7.

The groups with PCA applied on bag-of-words models yield the best results in blog and tweet analysis. The accuracy results of groups with PCA applied on LDA topic distributions are lower as the topic features are obtained through dimension reduction. The improvement made by PCA on the group of topic#1024 (O-LDA-1024 and LDA-1024) is the least, both in terms of blog and tweet analysis. More variance of the groups with topic features is lost even though the variance parameters are set to the same value.

As for the experiments on news, it seems that the information kept in the topic features is still expressive. A further reason might be that the news is more comprehensive than blogs and tweets, which is also a possible reason why topic#512 is the best for sentiment analysis from topic features in news analysis.

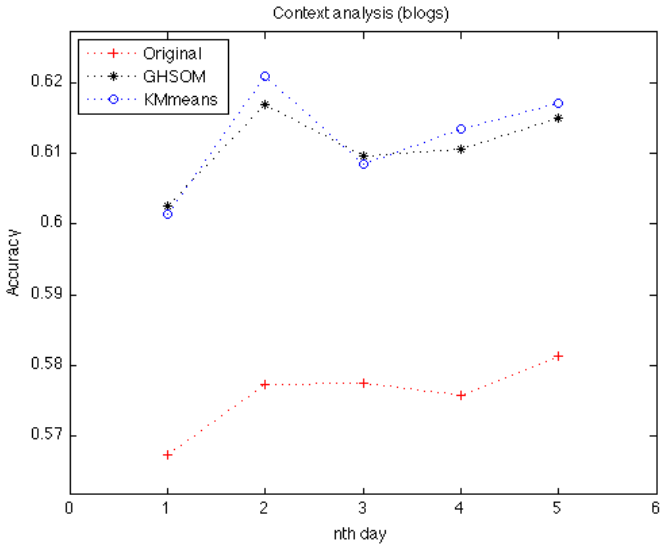


Figure 5.14: Results of SMP prediction using context analysis (blogs)

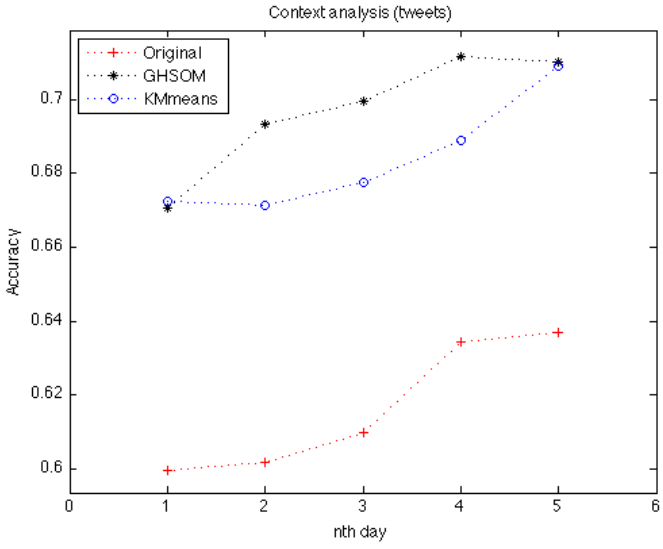


Figure 5.15: Results of SMP prediction using context analysis (tweets)

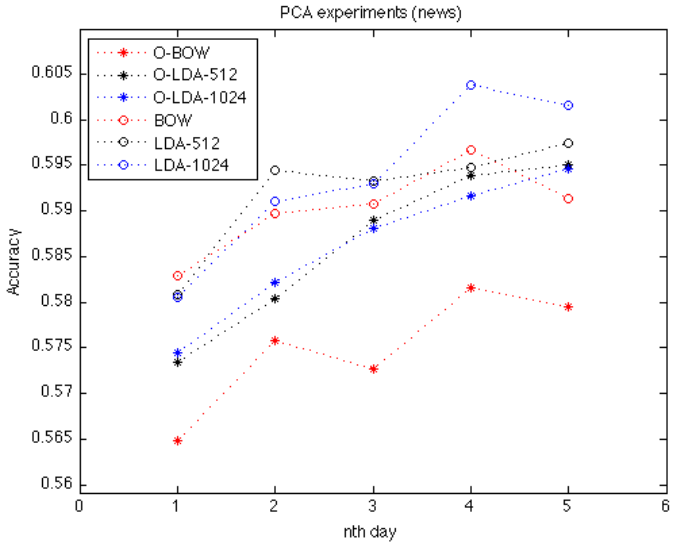


Figure 5.16: Results of SMP prediction with PCA (news): "O-..." stands for the original features before applying PCA.

In terms of dimension reduction, PCA seems to be more powerful than LDA, as the groups of BOW give better prediction results than the groups that starts with the "O-LDA-" groups. However, this conclusion needs to be tentative as PCA is modeled based on specific securities while LDA is modeled based on all the news and blogs.

5.5 Experiments with feature combination

5.5.1 Combination of bag-of-words models and technical indicators

The results for the SMP prediction using feature combination of bag-of-words models (BOW) and technical indicators are shown in Fig.5.19, Fig.5.20 and Fig.5.21. The experiment results on CA are the same in 5.3. They are illustrated in the graphs so that the comparison is clearer. The details of the experiments are illustrated in Tab.D.8.

In the experiments on news and blogs, it is apparent that the context analysis with K-means outperforms the other groups and the application of feature combination shows a slight improvement (no more than 1%). In the experiments on tweets, the combination of BOW and TI-2 achieves the best accuracy and CA ranks the second. Both of the extra features, TI-1 and TI-2, contribute to more obvious improvements in tweet analysis than those in news and blogs.

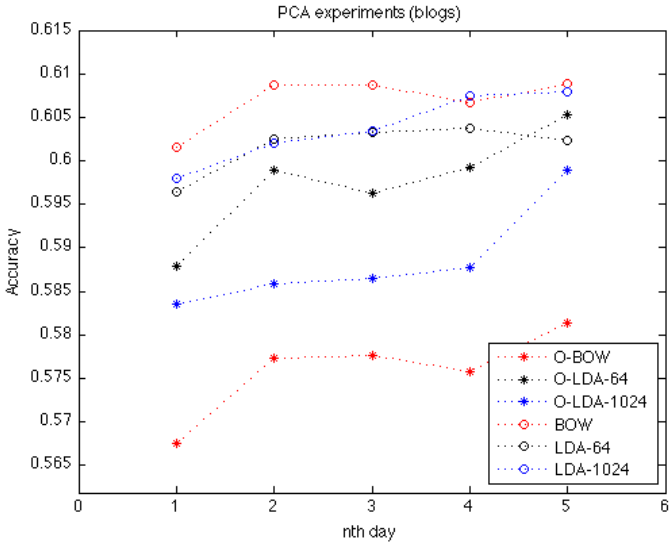


Figure 5.17: Results of SMP prediction with PCA (blogs)

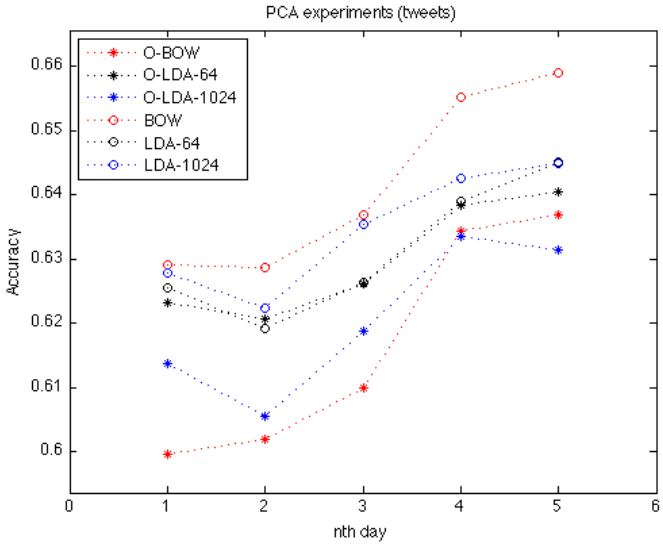


Figure 5.18: Results of SMP prediction with PCA (tweets)

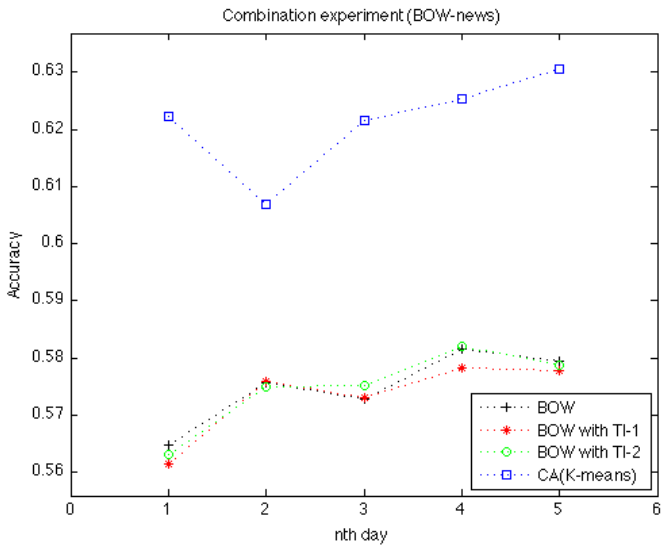


Figure 5.19: Results of SMP prediction using feature combination with BOW and technical indicators (news): TI-1 is the features described in Tab.3.1. TI-2 is the features proposed in this dissertation, as is illustrated in 3.2.1. CA stands for context analysis. The result details of CA can be viewed in D.6.

The poor performances of the combination features in news and blog analysis may be associated with the dominance of the bag-of-words features. There are more than 8000 words represented in the vectors while the number of the technical indicator features is only 34. As for the tweets, the common words, which are counted in the count vectors are much less than those counted in news and blogs, due to Twitter's limit on characters and user habits.

There is no issue with regard to the dominance of word counts in context analysis as clustering and classification are independent, thus it yields a better performance on all three data sources than feature combination. In tweet analysis, CA outperforms BOW with TI-1 where the indicator features used are identical. In a word, context analysis is a promising methodology in the applications where some features are dominant over the others in number. In this experiment, the numbers of BOW features are far more than those of indicator features.

5.5.2 Combination of sentiment scores and technical indicators

So far basic technical indicators have yielded the best performances in this dissertation, thus they are used to examine the predictive power of the combination features.

The results for the SMP prediction using feature combination with sentiment score

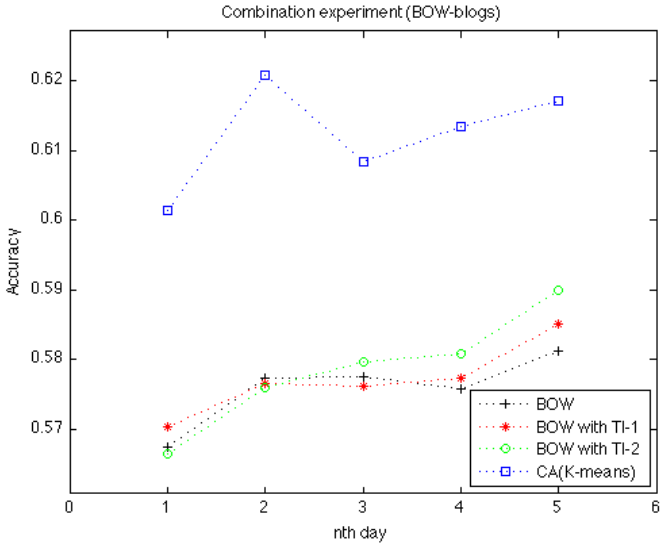


Figure 5.20: Results of SMP prediction using feature combination with BOW and technical indicators (blogs)

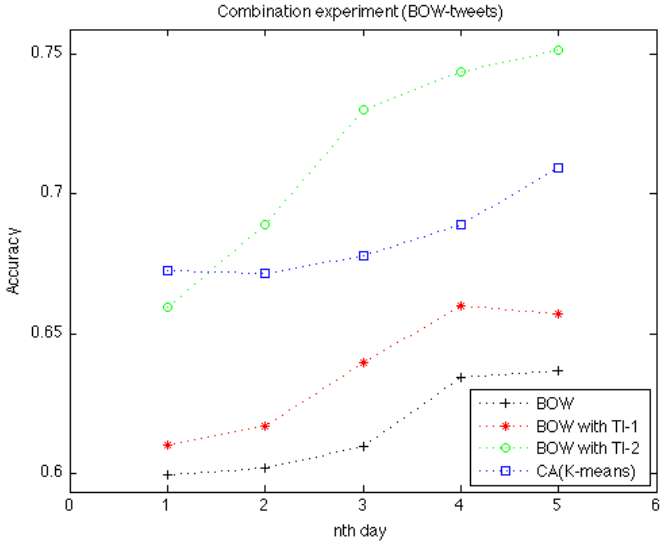


Figure 5.21: Results of SMP prediction using feature combination with BOW and technical indicators (tweets)

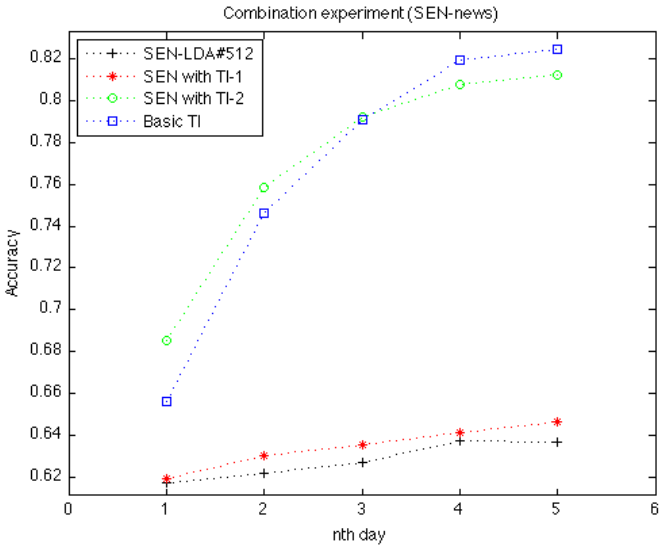
series and technical indicators are given in Fig.5.22, Fig.5.23 and Fig.5.24. The details are shown in Tab.D.8. The accuracy results of technical indicators in the three groups are different as the training periods of the data sets are not identical. The topic sentiment score series generated by the LDA model with topic number 512 is used in the news analysis. The sentiment features used in blog and tweet analysis are generated LDA models with topic number 64.

In Fig.5.22, the performance of Basic TI and SEN with TI-2 is similar to each other and both are around 15% over SEN-LDA#512 and SEN with TI-1, in the prediction for the 5th day. In Fig.5.23, it is obvious that Basic TI gives the best result. SEN with TI-2 ranks the second. SEN with TI-1 is even a little worse than SEN-LDA#64. In Fig.5.24, we note with interest that SEN with TI-2 is more than 10% over Basic TI, obtaining the greatest accuracy for the prediction of all 5 days. The performance of SEN with TI-1 is much better than SEN-LDA#64, showing around 5% improvement in accuracy.

It is further interesting to note that the combination features give totally different outcomes from the three data sources. SEN with TI-2 is similar to Basic TI with news analysis, much worse than Basic TI with blog analysis and much better than Basic TI with tweet analysis. This indicates that sentiment features used in this dissertation for news and blogs analysis do not improve the prediction results with technical indicators, but these features with tweets show promise. The worse performance of sentiment features of blogs may be linked to their non-realtime feature. If the experiments are conducted over a longer prediction time, such as a week or a month, a better prediction accuracy may be obtained.

As for the technical feature selection, there is no doubt that the predictive power of the technical indicators (TI-2) proposed in this dissertation is much better than that of the features (TI-1) used in [29].

The results for the prediction within an extended prediction cycle on feature combination of tweet sentiment features and technical indicator features are demonstrated in Fig.5.25. The lines in red are the prediction results of SMD. The lines in black are the prediction results of SMP. The lines with “plus” (TI-2 in 6 years) drop after day 5 and rises slowly after day 7. But the lines with “circle”(TI-2 in 3 months) show a slight drop after day 5 but rise immediately after day 6, not following the trend of the lines with “plus”. This indicates that the model trained by technical features in 3 months is not a general model due to the short time span of data. Thus it is hard to ensure the improvement made by the combination with tweet sentiment.



,

Figure 5.22: Results of SMP prediction using feature combination with sentiment score series and technical indicators (news)

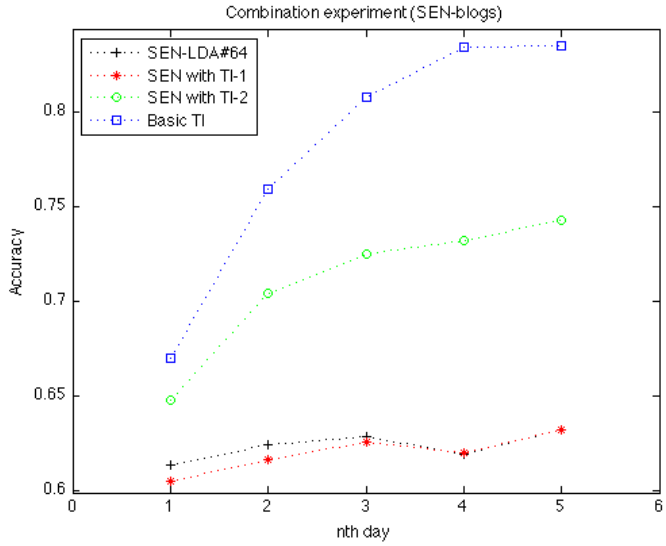


Figure 5.23: Results of SMP prediction using feature combination with sentiment score series and technical indicators (blogs)

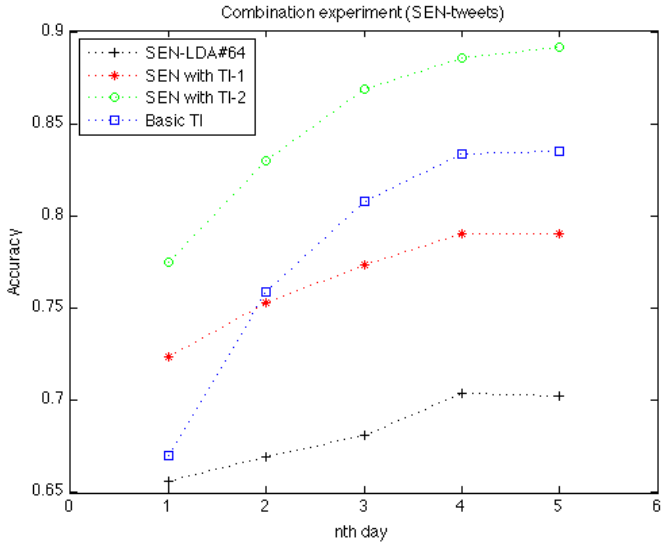


Figure 5.24: Results of SMP prediction using feature combination with sentiment score series and technical indicators (tweets)

Days	1	3	5	10	16
TI-2 SMP	66.99%	80.76%	83.45%	82.13%	84.15%
TI-2 SMD	66.36%	80.99%	84.07%	84.00%	86.38%
TI-2 (short) SMP	65.77%	81.40%	86.04%	88.84%	91.01%
TI-2 (short) SMD	64.78%	82.51%	87.37%	90.37%	92.22%
SEN SMP	65.63%	68.10%	70.19%	77.20%	77.41%
SEN SMD	68.67%	70.32%	74.61%	77.32%	78.27%
SEN+TI-2 SMP	77.50%	86.94%	89.23%	93.91%	94.43%
SEN+TI-2 SMD	80.29%	89.67%	92.90%	95.51%	96.25%

Table 5.12: Partial results of prediction accuracy of the extended experiment

It is also clearly illustrated in Fig.5.25 that red lines (SMD prediction) show better prediction accuracy than black lines (SMP) in most cases. It seems that it obtains a greater accuracy with the prediction target of SMD. The best result for SEN with TI-2 for SMD prediction is 80.29% on the next day prediction, as is demonstrated in Tab.5.12.

5.6 Summary

The results analysis has been given in this chapter. The discussion on results for sentiment analysis has shown that topic features obtained from LDA models are more

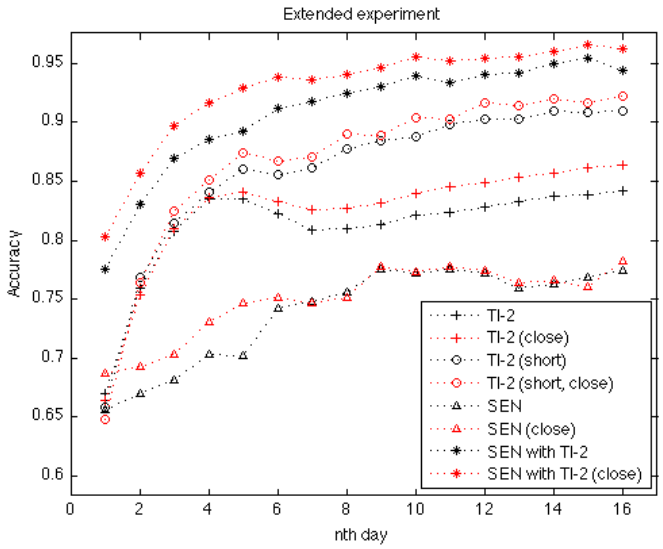


Figure 5.25: Results of the SMD (close) and SMP prediction using feature combination with sentiment score series and technical indicators (tweets)

suitable for dictionary-based approaches. The analysis of the two-stage architectures has indicated that the clustering the numeric data before classification on the textual data improves performance. The results for feature combination of technical indicators and tweet sentiment features have given the greatest performance, which is 77.50% on the next day SMP prediction and 80.29% on the next day SMD prediction. However, the conclusions are tentative due to the relatively short time span of tweet data.

Chapter 6

Conclusions and future work

6.1 Project summary

This dissertation has investigated a variety of text mining and data mining techniques with the aim of identifying predictive patterns from historical price/volume data and textual data. The major findings are listed as follows.

- Overall the selection and transformation of the technical indicators proposed in this dissertation show promise for assisting accurate prediction for SMD and SMP. Compared with the features extracted from textual data, historical data in the markets is more reliable.

- Among the textual data, the features from Twitter tweets yield the best performance in SMP prediction. It indicates that tweets may be the most valuable textual data source for short-term prediction.

- Sentiment score series generated by topic features outperforms the other well-accepted dictionaries with emotion tags. The results for the SMP prediction using the sentiment analysis methods vary from one application to another. As for sentiment normalization, such as the z-score, these do not show any improvements in prediction.

- The two-stage architecture of context analysis improves the performance of the SMP prediction when some features such as word frequency are dominant over others such as technical indicator features in number.

- The PCA applied to the bag-of-words models gives the higher improvements in the SMP prediction, compared with the PCA applied to the topic features. Topic

modelling is a dimension reduction methodology and some of the information in the raw text appears to be lost in the process of topic modelling before the PCA.

- Among the three sources of textual data, the results of feature combination vary from one to another. Tweets showed the best potential in improving the short-term SMP prediction, together with the technical indicator features. The features from news do not improve much. The sentiment scores extracted from blogs worsen the performance with historical price/volume features.

- The best accuracy results in this dissertation, 77.50% on the next day SMP prediction and 80.29% on the next day SMD prediction, are based on feature combination with technical features and sentiment score series on Twitter tweets. The directional prediction (SMD) accuracy is lower than the results obtained in the work in [9] and [28] where 86.7% and 82% in DJIA directional movement prediction are obtained respectively. The results are not so comparable. The results in this dissertation are more tentative than theirs as the time spans of the data used in [9] and [28] are 10 months and 6 months, which are longer than the time span of 3 months of the tweet data used in this dissertation. Furthermore, the prediction target in this dissertation is clearly different from their work [9, 29]. Their work predicts the DJIA while this dissertation predicts the S&P 100 stocks. The classification accuracy of SMD prediction is comparable their directional accuracy. Thus further experiments are necessary to make the results of this dissertation more comparable to other work.

The main weakness of this study is that the quality of the textual data cannot be guaranteed. The mechanism to select news from Reuters archives needs improvements and the quantity of Twitter tweets is limited. If the quality of the textual data is better, the results should be more convincing.

6.2 Future work

Many methodologies have been attempted in this dissertation. There are a number of alternative methods or parameters to each of them. A future study based on the following trials may be interesting.

6.2.1 Two-stage architecture

The performance of K-means and GHSOM in context analysis is difficult to conclude in this dissertation. Further experiments can be conducted based on groups of specific industry sectors such as the technology sector and the financial sector.

The clustering in the two-stage architecture in this dissertation is only applied to price data. It is worth applying it to textual data. As is indicated in 5.2, two ready-made dictionaries, namely, the GI and the LM, give different performance with different data sources. The GI seems to be more suitable for a general context and the LM is likely to be better for financial text. There are many different types of news articles such as product reviews, quarter reports and law issues. It is conjectured that if the types of news articles can be identified before modelling, a better result may be obtained.

6.2.2 Features

Technical indicators: There are hundreds of technical indicators available. More experiments can be conducted over features with other indicators. The presentation forms of indicators are variable as well. Some indicators are presented in a form with prices or volumes, which is not friendly to learning algorithms, such as OBV and William A/D. They can be transformed into percentage changes or the position in a range over a period of time.

The parameters used in this dissertation are default values in the TTR package of R-project. The performance may be improved if the parameters are tuned with genetic algorithms.

Topic features: In this dissertation, LDA topics are trained over all documents of news and blogs collected. It is unknown whether if the topic models are trained from specific industries or securities, the semantics of the topic will be more or less accurate. As the sentiment polarity detection, which is based on specific securities, showed a more powerful potential than the sentiment extracted from available dictionaries, it is conjectured that topic models trained on specific industries or sectors might improve the prediction.

Sentiment analysis: The calculations in sentiment analysis with the dictionary categories are based on word count but not on TF-IDF, which may be more intuitive to express on the importance of a word. The way to calculate the polarity of the topics is

based on topic presence (> 0 or not) but not related to their distributions. There may be a different result if the weight of those terms and topics are represented in another way.

6.2.3 Prediction target

The prediction target in this dissertation is quite different from other work [9, 29], which makes it less comparable to others. The same methodologies can be applied on DJIA prediction and close price movement prediction in future experiments.

The features from blogs do not give satisfactory experiment results. Due to the non-realtime feature of blogs, if the prediction target is the change of prices in a week or a month, more useful patterns might be found using the features obtained from blogs.

A further approach may be the change of price (e.g.$\pm 2\%$) or ATR, which is explained in Appendix B. A more specific prediction target such as how much change in ATR (-1/-0.5/+0.5/+1) may be more helpful to investors than vague movement performance (*good* / *uncertain* / *bad*).

Appendix A

News selection

The news from Reuters Archives is selected using the following procedures.

1. Fetch the links with “article” in their URLs.
2. Drop the articles where there is “C O R R E C T I O N” in their tittles.
3. Keep the articles if there is any desired codes or alias of companies appearing in the titles. The list of security codes and alias are listed in Tab.A.1. The case of the code and alias is not concerned as they are transformed into lowercase before matching. The “No code” column describes whether the codes are used in matching as some of the codes such as “PM” and “USB” are commonly used and thus might lead to mismatching.

Table A.1: Rules of matching securities from news titles.

Code	Alias	No code
AA	Alcoa	True
AAPL	Apple	
ABT	Abbott	
AEP	American Electric Power	
ALL	All state	True
AMGN	Amgen	
AMZN	Amazon	
AVP	Avon	
AXP	American Express	
BA	Boeing	True

BAC	Bank of America	
BAX	Baxter	
BHI	Baker Hughes	
BK	Mellon, Bank of New York Mellon, BNY Mellon	True
BMY	Bristol-Myers	
BRK-B	Berkshire Hathaway, BRK	
CAT	Caterpillar	
C	Citi	True
CL	Colgate:	
CMCSA	Comcast	
COF	Capital One	
COP	Conoco, ConocoPhillips, Conoco Phillips	True
COST	Costco	True
CPB	Campbell	
CSCO	Cisco	
CVS	Caremark	
CVX	Chevron	
DD	DuPont, Du Pont	True
DELL	Dell	
DIS	Disney	True
DOW	Dow Chemical	
DVN	Devon	
EMC		
ETR	Entergy	
EXC	Exelon	
F	Ford	True
FCX	Freeport	
FDX	FedEx	
GD	General Dynamics	
GE	General Electric	
GILD	Gilead	
GOOG	Google	
GS	Goldman Sach	

HAL	Halliburton	
HD	Home Depot	True
HNZ	Heinz	
HON	Honeywell	
HPQ	Hewlett-Packard	HP
IBM	International Business Machines	
INTC	Intel	
JNJ	Johnson & JohnsongJ&J	
JPM	JP Morgan	
KFT	Kraft	
KO	Coca-Cola, Coca Cola	True
LMT	Lockheed Martin	
LOW	Lowe's	True
MA	Mastercard	True
MCD	McDonald's	
MDT	Medtronic	
MET	MetLife	
MMM	3M	
MO	Altria	True
MON	Monsanto	True
MRK	Merck	
MS	Morgan Stanley	True
MSFT	Microsoft	
NKE	Nike	
NOV	Oilwell	True
NSC	Norfolk	
NWSA	News Corp	
NYX	Euronext	
ORCL	Oracle	
OXY	Occidental Petroleum	
PEP	Pepsico	True
PFE	Pfizer	
PG	Procter & Gamble	

PM	Philip Morris	True
QCOM	QUALCOMM	
RF	Regions Financial	
RTN	Raytheon	
S	Sprint Nextel	True
SLB	Schlumberger	
SLE	Sara Lee	
SO	Southern	True
T	AT&T, AT & T	True
TGT	Target Corporation, Target Coprt	
TWX	Time Warner	
TXN	Texas InstrumentsgTI	
UNH	Unitedhealth	
UPS	United Parcel Service	
USB	U.S. Bancorp, U.S. Bank, US Bank, US Bancorp	True
UTX	United Technologies CorpgUTX	
VZ	Verizon	
WAG	Walgreen	
WFC	Wells Fargo	
WMB	Williams Companies	
WMT	Wal-Mart	
WY	Weyerhaeuser	
XOM	Exxon	
XRX	Xerox	

Appendix B

Technical Indicators

In this dissertation, the parameters of the calculations of the technical indicators are the default values in the TTR package in R-project.

ADX ADX consists of four indicators, namely, the positive Directional Index (DI_p), the negative Directional Index (DI_n), the Directional Index (DI) and the Average Directional Index (ADX). [77]

$$HiDiff = HI - HI.1$$

$$LoDiff = LO.1 - LO$$

If $HiDiff < 0$ and $LoDiff < 0$or $HiDiff = LoDiff$ then

$$DI_p = 0$$

$$DI_n = 0$$

If $HiDiff > LoDiff$ then

$$DI_p = HiDiff$$

$$DI_n = 0$$

If $HiDiff < LoDiff$ then

$$DI_p = 0$$

$$DI_n = LoDiff$$

$$DX = \frac{DI_p - DI_n}{DI_p + DI_n} \tag{B.1}$$

$$ADX = \frac{ADX_{-1} \times (n-1) + DX}{n} \tag{B.2}$$

In this dissertation, ADX and DX are used as technical indicators after being normalized. And the ratio of DI_p and DI_n, given in Eq.B.3, is one of the indicator features as well.

$$PN_{ratio} = \frac{DI_p}{DI_n} \tag{B.3}$$

aroon Aroon is an indicator that can discover the beginning of trends. It consists of three indices, as is given in Eq.B.4, EqB.5, and EqB.6. [15]

$$AroonUp = 100 \times (\frac{n - PeriodSinceHighestHigh}{n}) \tag{B.4}$$

$$AroonDown = 100 \times (\frac{n - PeriodSinceLowestLow}{n}) \tag{B.5}$$

$$AroonOscillator = AroonUp - AroonDown \tag{B.6}$$

ATR Average True Range (ATR) is a group of estimates of volatility of price series. [77] The calculations are given in the following equations.

$$TrueHigh = max(high_{[0]}, close_{[-1]}) \tag{B.7}$$

$$TrueLow = min(low_{[0]}, close_{[-1]}) \tag{B.8}$$

$$TR = TrueHigh - TrueLow \tag{B.9}$$

$$ATR = \frac{TR_{[-1]} \times (n-1) + TR}{n} \tag{B.10}$$

In this dissertation, TR and ATR are adopted, as well as Eq.B.11, which is another form of true range.

$$TR2 = \frac{close - TrueLow}{TrueHigh - TrueLow} \tag{B.11}$$

Bollinger Bands Bollinger Bands is a popular indicator to measure security volatility and price level. [10] The calculations are given in the following equations.

$$TP = \frac{high + low + close}{3} \tag{B.12}$$

$$BandWidth = 2 \times F \times \sigma(TP) \tag{B.13}$$

In this dissertation, Eq.B.13 is adopted as one of the technical indicator features. In Eq.B.13, F represents the number of Standard Deviations - usually it is 2.

CCI Commodity Channel Index (CCI) is used to discover the beginning and ending of security trends. [38]

$$MATP = MA(TP, n) \tag{B.14}$$

In Eq.B.14, MA is a Moving Average (simple) function and TP is calculated in Eq.B.12.

$$MDTP_x = \frac{\Sigma_{i=x-n}^{x} abs(TP_i - MATP_x)}{n}$$

$$CCI = \frac{TP - MATP}{MDTP * 0.015} \tag{B.15}$$

The calculation of CCI is given in Eq.B.15.

De-Trended Price Oscillator Detrended Price Oscillator (DPO) eliminates the trend in prices. The calculations are given in Eq.B.16.

$$DPO = close - \frac{\Sigma_{i=1}^{n} Close_i}{n} \tag{B.16}$$

EMA Exponential Moving Averages (EMAs) summarize the average prices over a period of time.

$$EMA_1 = Price_1$$

$$t > 1,\ \ EMA_t = \alpha \times Price_t + (1-\alpha) \times EMA_{t-1}$$

EMV The Arms Ease of Movement indicator (EMV) emphasizes days where the stock prices move easily and minimizes days where the stock prices do not change much. The calculation is given in Eq.B.19.

$$MidpointMove = \frac{high - low}{2} - \frac{high_{-1} - low_{-1}}{2} \tag{B.17}$$

$$BoxRatio = \frac{volume/10000}{high - low} \tag{B.18}$$

$$EMV = MidpointMove/BoxRatio \tag{B.19}$$

In this dissertation, a smoothed EMV, which is averaged by SMA (see B), is adopted as well.

MACD The Moving Average Convergence-Divergence (MACD) is the difference of two EMAs, which can reflect the trends of price movements.

$$MACD = EMA[stockPrices, 12] - EMA[stockPrices, 26]$$

MFI Money Flow Index (MFI) is the ratio of money that flowing into and out of a stock. The calculations are given in the following equations.

$$MoneyFlow = TypicalPrice \times volume \tag{B.20}$$

In Eq.B.20, TypicalPrice can be calculated with Eq.B.12.

If $TypicalPrice > TypicalPrice_{-1}$,

$$PositiveMoneyFlow = PositiveMoneyFlow_{-1} + MoneyFlow \tag{B.21}$$

else

$$NegativeMoneyFlow = NegativeMoneyFlow_{-1} + MoneyFlow \tag{B.22}$$

$$MoneyRatio = \frac{\Sigma_i PositiveMoneyFlow_i}{\Sigma_i NegativeMoneyFlow_i} \tag{B.23}$$

$$MoneyFlowIndex = 100 - \frac{100}{1 + MoneyRatio} \tag{B.24}$$

OBV OBV is similar to MFI, which measures money flow.

If $close > close_{-1}$

$$OBV = OBV_{-1} + volume \tag{B.25}$$

else if $close < close\_1$

$$OBV = OBV_{-1} - volume \tag{B.26}$$

else

$$OBV = OBV_{-1} \tag{B.27}$$

ROC Rate of Change (ROC) measures the change of price series. The calculation is given in Eq.B.28.

$$ROC = \frac{close - close_{-n}}{n} \times 100 \tag{B.28}$$

RSI The Relative Strength Index (RSI) illustrates the strength of a stock in its current movement.

$$RSI = 100 - \frac{100}{1 + RS} \tag{B.29}$$

$$RS = \frac{Average\,gain}{Average\,loss} \tag{B.30}$$

SMA Simple Moving Average (SMA) calculates the mean of prices over a period of time. The calculation is given in Eq.B.31.

$$SMA = \frac{\Sigma_i close}{n} \tag{B.31}$$

Stochastic Oscillator Stochastic Oscillator is an indicator that measures how close the price is to the trading range over the past n periods. The calculations are given in Eq.B.32 and Eq.B.33.

$$\%K = 100 \times \frac{close - LowestLow_{last\ n\ periods}}{HighestHigh_{last\ n\ periods} - LowestLow_{last\ n\ periods}} \tag{B.32}$$

$$\%D = MovingAverage(\%K) \tag{B.33}$$

In this dissertation, Stochastic Fast %K, Stochastic Fast %D and Stochastic Slow %D are adopted as features.

SMI SMI is similar to Stochastic Oscillator. The difference is that SMI is used to measure how close the price is to the middle of the trading range. [6] The calculations are given in the following equations.

$$cm = close - \frac{HighestHigh - LowestLow}{2} \tag{B.34}$$

$$hl = HighestHigh - LowestLow \tag{B.35}$$

$$cmMA = EMA(EMA(cm)) \tag{B.36}$$

$$hlMA = EMA(EMA(hl)) \tag{B.37}$$

$$SMI = 100 \times \frac{cmMA}{hlMA/2} \tag{B.38}$$

$$SMISignal = EMA(SMI) \tag{B.39}$$

In this dissertation, SMI (Eq.B.38) is adopted. *SMISignal* is converted into Eq.B.40.

$$ConvertedSMISignal = \frac{SMI - SMISignal}{SMI} \tag{B.40}$$

TDI Trend Detection Index (TDI) is used to discover the beginning and end of moving trends. The calculations are given in the following equations.

$$Mom = price - price_{period} \tag{B.41}$$

$$MomAbs = abs(Mom) \tag{B.42}$$

$$DI = \Sigma_{i=1}^{n} Mom_i \tag{B.43}$$

$$DIAbs = abs(DI) \tag{B.44}$$

$$DIAbsSum = \Sigma_{i=1}^{n} DIAbs \tag{B.45}$$

$$DIAbsSum2 = \Sigma_{i=1}^{2n} DIAbs \tag{B.46}$$

$$TDI = DIAbs - (DIAbsSum2 - DIAbsSum) \tag{B.47}$$

In this dissertation, DI (Eq.B.43) and TDI (B.47) are used as technical indicator features.

TRIX Triple Smoothed Exponential Oscillator (TRIX) is measure of the rate of change of triple EMA. [32] The calculations are given in the following equations.

$$M = EMA(EMA(EMA(price))) \tag{B.48}$$

$$TRIX = 100 \times \frac{M - M_{-1}}{M} \tag{B.49}$$

$$TRIXSignal = EMA(TRIX) \tag{B.50}$$

In this dissertation, $TRIX$ (Eq.B.49) is used. And $TRIXSignal$ is converted, as is shown in Eq.B.51.

$$ConvertedTRIXSignal = \frac{TRIX - TRIXSignal}{TRIX} \tag{B.51}$$

VHF Vertical Horizontal Filter (VHF) is another indicator that attempts to identify start and end of trends. [75] The calculation is given in Eq.B.52.

$$VHF = \frac{HighestHigh_n - LowestLow_n}{\Sigma_i |close_i - close_{i-1}|} \tag{B.52}$$

Volatility Four volatility indicators are used in this dissertation. The calculations of them are given in the following equations.

$$Vol_{close\ to\ close} = \sqrt{\frac{N}{n-2}\Sigma_i(ROC_i - \overline{ROC}_i)} \tag{B.53}$$

$$Vol_{garman.klass} = \sqrt{\frac{N}{n}\Sigma_i(\frac{1}{2} \times log(high_i/close_i)^2 - (2 \times log(2) - 1) \times log(close_i/open_i)^2)} \tag{B.54}$$

$$Vol_{parkinson} = \sqrt{\frac{N}{4 \times n \times log2} \times \Sigma_i(\frac{high_i}{low_i})^2} \tag{B.55}$$

$$Vol_{rogers.satchell} = \sqrt{\frac{N}{n} \times \Sigma(log\frac{high_i}{open_i} \times log\frac{high_i}{close_i} + log\frac{low_i}{open_i} \times log\frac{low_i}{close_i})} \tag{B.56}$$

Williams Accumulation / Distribution The Williams Accumulation / Distribution (AD) attempts to identify the trend of the market.

If $close > close_{-1}$ then

$$AD = AD_{-1} + close - min(low, close_{-1}) \tag{B.57}$$

If $close < close_{-1}$ then

$$AD = AD_{-1} + max(high, close_{-1} - close) \tag{B.58}$$

If $close = close_{-1}$ then

$$AD = AD_{-1} \tag{B.59}$$

WPR The Williams %R (WPR) is similar to stochastic' fast %K. The calculation is given in Eq.B.60.

$$\%R = 100 \times \frac{HighestHigh_n - close}{HighestHigh_n - LowestLow_n} \tag{B.60}$$

Appendix C

Top topics modeled by LDA

Topic ID	Top 5 key words
1	buss, service, company, market, provide
2	company, product, market, buss, service
3	company, invest, market, research, buss
4	company, product, earn, market, develop
5	company, research, market, manag, invest
6	store, buss, company, include, game
7	company, service, market, product, statement
8	service, company, program, verizon, provide
9	company, market, technolog, product, buss
10	manag, company, research, inform, financi

Table C.1: Top topics modeled by LDA with topic number 64

Topic ID	Top 5 key words
1	Halliburton, walgreen, drill, service, company
2	fund, invest, manag, advantag, incom
3	charg, trimestr, revenu, milliard, total
4	mine, resourc, zone, grade, drill
5	product, medic, company, devic, manufacture
6	voor, autonomi, gilead, niet, zijn
7	verizon, custom, service, news, commun
8	employe, employ, benefit, plan, program
9	option, trade, posit, move, time
10	company, freeport, cost, buss, oper

Table C.2: Top topics modeled by LDA with topic number 256

Topic ID	Top 5 key words
1	market, yahoo, research, equity, company
2	commun, service, provide, oper, solut
3	product, inform, corpor, statement, technolog
4	frontier, renam, blodget, rural, local
5	babi, wrap, boomer, planner, deleg
6	digit, video, content, media, experi
7	video, youtub, onlin, content, applianc
8	gener, dynam, system, contract, inform
9	intel, intc, chip, comput, devic
10	claim, firm, concentr, secur, arbitr

Table C.3: Top topics modeled by LDA with topic number 512

Topic ID	Top 5 key words
1	statement, Honeywell, offer, forward, look
2	quartal, millionen, durch, nicht, milliarden
3	assumpt, fals, probabl, base, scenario
4	atla, bernank, wake, debat, blankfein
5	eastwindresearch, expo, mirag, market, company
6	service, network, provide, connect, commun
7	trajectori, akamai, sandisk, sndk, akam
8	countrywid, bofa, mortgag, loss, financi
9	foundat, verizon, commun, program, nonprofit
10	bellevu, hdmi, company, loos, devic

Table C.4: Top topics modeled by LDA with topic number 1024

Appendix D

Experiment results

Day	1	2	3	4	5
Basic technical indicators	66.99%	75.92%	80.83%	83.42%	83.50%
BOW of news	56.47%	57.57%	57.27%	58.16%	57.94%
BOW of blogs	56.74%	57.73%	57.76%	57.58%	58.13%
BOW of tweets	60.00%	60.19%	61.00%	63.44%	63.69%

Table D.1: Details of average accuracy results of basic experiments

Day	1	2	3	4	5
GI	57.55%	57.55%	58.41%	58.99%	59.33%
GI (sen only)	58.19%	58.19%	59.17%	59.57%	59.46%
GI (sen)	57.64%	57.50%	58.42%	58.93%	59.44%
LM	58.17%	58.65%	59.13%	59.70%	60.33%
LM (sen only)	58.02%	58.81%	59.72%	59.41%	59.95%
LM (sen)	58.68%	58.57%	59.22%	59.74%	60.14%

Table D.2: Details of average accuracy results of experiments with features of GI and LM (news)

Day	1	2	3	4	5
GI	58.53%	59.23%	58.85%	59.85%	59.89%
GI (sen only)	60.38%	59.23%	60.27%	60.41%	61.39%
GI (sen)	58.60%	58.95%	59.00%	59.88%	59.76%
LM	59.73%	60.51%	60.61%	60.12%	60.51%
LM (sen only)	59.89%	59.37%	60.33%	59.70%	60.13%
LM (sen)	59.91%	60.89%	60.48%	59.55%	60.33%

Table D.3: Details of average accuracy results of experiments with features of GI and LM (blogs)

Day	1	2	3	4	5
GI	61.82%	61.17%	62.20%	63.32%	63.91%
GI (sen only)	61.10%	59.83%	61.83%	62.48%	62.78%
GI (sen)	61.85%	61.02%	62.30%	63.39%	63.84%
LM	59.35%	58.79%	60.26%	61.97%	61.27%
LM (sen only)	59.99%	59.06%	60.66%	62.13%	61.93%
LM (sen)	59.93%	59.18%	60.69%	62.43%	62.02%

Table D.4: Details of average accuracy results of experiments with features of GI and LM (tweets)

Data Source	Topics	Day				
		1	2	3	4	5
News	32	59.19%	60.25%	59.91%	59.89%	60.71%
News	64	59.78%	60.47%	61.04%	60.81%	61.72%
News	128	60.76%	60.15%	61.48%	61.08%	61.27%
News	256	60.68%	61.28%	61.89%	62.14%	63.06%
News	512	**61.90%**	**61.98%**	**62.21%**	**63.17%**	**63.29%**
News	1024	61.58%	60.43%	61.89%	62.12%	62.32%
Blogs	32	61.03%	60.70%	61.29%	61.56%	61.30%
Blogs	64	**61.75%**	**62.74%**	**62.19%**	**62.72%**	**62.99%**
Blogs	128	58.98%	59.55%	59.72%	59.52%	59.77%
Blogs	256	60.95%	61.47%	61.83%	61.75%	61.08%
Blogs	512	59.06%	59.92%	60.58%	60.90%	61.23%
Blogs	1024	58.76%	60.43%	60.65%	60.08%	60.99%
Tweets	32	63.75%	63.90%	64.13%	64.94%	65.17%
Tweets	64	**65.67%**	**65.98%**	**66.52%**	**67.79%**	**67.40%**
Tweets	128	58.09%	56.65%	58.88%	60.24%	59.87%
Tweets	256	58.06%	57.09%	58.40%	60.33%	59.57%
Tweets	512	57.94%	56.68%	58.20%	59.94%	59.53%
Tweets	1024	57.53%	56.29%	58.09%	59.68%	59.04%

Table D.5: Details of average accuracy results of experiments with sentiment scores from topic distributions

Data Source	Clustering alg.	Day				
		1	2	3	4	5
News	GHSOM	59.78%	61.35%	61.38%	61.23%	62.26%
News	K-means	62.24%	60.69%	62.15%	62.52%	63.05%
Blogs	GHSOM	60.25%	61.69%	60.97%	61.05%	61.49%
Blogs	K-means	60.13%	62.09%	60.84%	61.34%	61.71%
Tweets	GHSOM	67.06%	69.33%	69.97%	71.16%	71.02%
Tweets	K-means	67.25%	67.15%	67.77%	68.90%	70.92%

Table D.6: Details of average accuracy results of experiments with context analysis

Data source	Feature	Day				
		1	2	3	4	5
News	PCA on BOW	58.30%	58.98%	59.07%	59.67%	59.13%
	PCA on LDA#512	58.08%	59.44%	59.33%	59.47%	59.74%
	PCA on LDA#1024	58.05%	59.11%	59.29%	60.37%	60.16%
Blogs	PCA on BOW	60.16%	60.88%	60.88%	60.67%	60.90%
	PCA on LDA#64	59.65%	60.25%	60.34%	60.38%	60.24%
	PCA on LDA#1024	59.80%	60.21%	60.35%	60.75%	60.80%
Tweets	PCA on BOW	62.91%	62.88%	63.69%	65.51%	65.90%
	PCA on LDA#64	62.56%	61.91%	62.63%	63.91%	64.51%
	PCA on LDA#1024	62.78%	62.24%	63.54%	64.25%	64.49%

Table D.7: Details of average accuracy results of experiments with PCA

Data source	Feature	Day				
		1	2	3	4	5
News	BOW+TI-1	56.15%	57.59%	57.31%	57.82%	57.77%
News	BOW+TI-2	56.32%	57.49%	57.52%	58.21%	57.88%
News	SEN+TI-1	61.93%	63.03%	63.58%	64.17%	64.64%
News	SEN+TI-2	68.54%	75.82%	79.23%	80.78%	81.23%
News	Basic TI[1]	65.61%	74.63%	79.05%	81.89%	82.42%
Blogs	BOW+TI-1	57.03%	57.65%	57.63%	57.73%	58.52%
Blogs	BOW+TI-2	56.65%	57.59%	57.98%	58.09%	58.99%
Blogs	SEN+TI-1	60.51%	61.69%	62.59%	62.00%	63.25%
Blogs	SEN+TI-2	64.78%	70.42%	72.54%	73.22%	74.27%
Blogs	Basic TI	66.99%	75.92%	80.83%	83.42%	83.50%
Tweets	BOW+TI-1	61.02%	61.68%	63.96%	66.00%	65.73%
Tweets	BOW+TI-2	65.96%	68.92%	73.03%	74.38%	75.13%
Tweets	SEN+TI-1	72.37%	75.32%	77.35%	79.08%	79.06%
Tweets	SEN+TI-2	77.50%	83.04%	86.94%	88.60%	89.23%
Tweets	Basic TI [2]	65.77%	76.82%	81.40%	84.12%	86.04%

Table D.8: Details of average accuracy results of experiments with feature combination

Bibliography

[1] Rethinking of sentiment analysis. http://micarum.blogspot.com/2012/02/rethinking-of-sentiment-analysis.html, February 2012.

[2] Y.S. Abu-Mostafa and A.F. Atiya. Introduction to financial forecasting. *Applied Intelligence*, 6(3):205–213, 1996.

[3] Tim Pei Hung Hsieh Alvin Chyan and Chris Lengerich. A stock-purchasing agent from sentiment analysis of twitter. http://cs229.stanford.edu/proj2011/ChyanHsiehLengerich-A_Stock-Purchasing_Agent_from_Sentiment_Analysis_of_Twitter.pdf, 2011. [Online; accessed 04-September-2012].

[4] G.S. Atsalakis and K.P. Valavanis. Surveying stock market forecasting techniques-part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009.

[5] R.E. Bellman, R.E. Bellman, R.E. Bellman, and R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1966.

[6] William Blau. *Technical Analysis of Stocks & Commodities*, January 1993.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR, abs/0911.1583*, pages 1–10, 2009.

[9] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2011.

[10] J. Bollinger. *Bollinger on Bollinger bands*. McGraw-Hill Professional, 2001.

[11] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[12] K.C. Butler and SJ Malaikah. Efficiency and inefficiency in thinly traded stock markets: Kuwait and saudi arabia. *Journal of Banking & Finance*, 16(1):197–210, 1992.

[13] Trusheim S. Chakoumakos, R. and V. Yendluri. Automated market sentiment analysis of twitter for options trading. http://cs229.stanford.edu/proj2011/TrusheimChakoumakosYendluri-Automated_Market_Sentiment_analysis_of_Twitter_for_Options_Trading.pdf, 2011. [Online; accessed 04-September-2012].

[14] S. Chakrabarti, M. Ester, U. Fayyad, J. Gehrke, J. Han, S. Morishita, G. Piatetsky-Shapiro, and W. Wang. Data mining curriculum: A proposal (version 1.0), 2006.

[15] Tushar S. Chande. *Technical Analysis of Stocks & Commodities*, September 1995.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] D.Y. Chiu and P.J. Chen. Dynamically exploring internal mechanism of stock market by fuzzy-based support vector machines with high dimension input space and genetic algorithm. *Expert Systems with Applications*, 36(2):1240–1248, 2009.

[18] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

[19] Maxime Goutagny Daniel Debbini, Philippe Estin. Modeling the stock market using twitter sentiment analysis. http://cs229.stanford.edu/proj2011/DebbiniEstinGoutagny-ModelingTheStockMarketUsingTwitterSentimentAnalysis.pdf, 2011. [Online; accessed 04-September-2012].

[20] S.R. Das and M.Y. Chen. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

[21] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[22] D. Enke and S. Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29(4):927–940, 2005.

[23] E.F. Fama. The behavior of stock-market prices. *The journal of Business*, 38(1): 34–105, 1965.

[24] K.P. FRS. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[25] L.A. Gallagher and M.P. Taylor. Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, pages 345–362, 2002.

[26] A.K. Ghoshal, T. Mukherjee, and S. Dhar. An empirical study in indian share market using neural network and genetic algorithm. *Asian Journal of Research in Banking and Finance*, 1(1):1–19, 2011.

[27] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, pages 219–222, 2007.

[28] E. Hsu, S. Shiu, and D. Torczynski. Predicting dow jones movement with twitter. `http://cs229.stanford.edu/proj2011/HsuShiuTorczynski-PredictingDowJonesMovementWithTwitter.pdf`, 2011. [Online; accessed 04-September-2012].

[29] S.H. Hsu, JJ Hsieh, T.C. Chih, and K.C. Hsu. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, 36(4):7947–7951, 2009.

[30] J. Hu, L. Fang, Y. Cao, H.J. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 179–186. ACM, 2008.

[31] W. Huang, Y. Nakamori, and S.Y. Wang. Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32 (10):2513–2522, 2005.

[32] Jack K. Hutson. *Technical Analysis of Stocks & Commodities*, 1(5).

[33] M.G. Kavussanos and E. Dockery. A multivariate test for stock market efficiency: the case of ase. *Applied Financial Economics*, 11(5):573–579, 2001.

[34] K. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307–319, 2003.

[35] C.D. Kirkpatrick and J.R. Dahlquist. *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.

[36] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[37] Volodymyr Kuleshov. Predicting the dow jones using twitter. `http://cs229.stanford.edu/proj2011/Kuleshov-CanTwitterPredictTheStockMarket.pdf`, 2011. [Online; accessed 04-September-2012].

[38] Donald Lambert. *Commodities(now called Futures)*, 1980.

[39] G. Leng, G. Prasad, and T.M. McGinnity. An on-line algorithm for creating self-organizing fuzzy neural networks. *Neural Networks*, 17(10):1477–1493, 2004.

[40] B. Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, pages 627–666, 2010.

[41] Edward Loper and Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL `http://dx.doi.org/10.3115/1118108.1118117`.

[42] T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[43] C.J. Lu, T.S. Lee, and C.C. Chiu. Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems*, 47(2):115–125, 2009.

[44] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[45] A. Mahajan, L. Dey, and S.M. Haque. Mining financial news for major events and their impacts on the market. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 423–426. IEEE, 2008.

[46] H. Mao, S. Counts, and J. Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *Arxiv preprint arXiv:1112.1051*, 2011.

[47] K. McKinley. Stock market efficiency and insider trading. *Issues in Political Economy*, 8, 1999.

[48] A. Mittal and A. Goel. Stock prediction using twitter sentiment analysis. 2011. [Online; accessed 04-September-2012].

[49] M.A. Mittermayer. Forecasting intraday stock price trends with text mining techniques. In *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2004.

[50] M.A. Mittermayer and G.F. Knolmayer. Newscats: A news categorization and trading system. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1002–1007. Ieee, 2006.

[51] C. Möller-Levet, F. Klawonn, K.H. Cho, and O. Wolkenhauer. Fuzzy clustering of short time-series and unevenly distributed sampling points. *Advances in Intelligent Data Analysis V*, pages 330–340, 2003.

[52] A. Nikfarjam, E. Emadzadeh, and S. Muthaiyah. Text mining approaches for stock market prediction. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 4, pages 256–260. Ieee, 2010.

[53] J.R. Nofsinger. The impact of public information on investors. *Journal of Banking & Finance*, 25(7):1339–1366, 2001.

[54] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[56] MF Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

[57] R.R. Prechter. The wave principle of human social behavior and the new science of socionomics. *Recherche*, 67:02, 1999.

[58] G. Pui Cheong Fung, J. Xu Yu, and W. Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402. IEEE, 2003.

[59] A. Rauber, D. Merkl, and M. Dittenbach. The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data. *Neural Networks, IEEE Transactions on*, 13(6):1331–1341, 2002.

[60] Marius Lazer Ray Chen. Sentiment analysis of twitter feeds for the prediction of stock market movement, 2011.

[61] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[62] E.J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM, 2012.

[63] 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 (TOIS)*, 27(2):12, 2009.

[64] J. Shlens. A tutorial on principal component analysis. *Systems Neurobiology Laboratory, University of California at San Diego*, 2005.

[65] A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[66] T. Sprenger and I. Welpe. Tweets and trades: The information content of stock microblogs. *Working Paper Series Technische Universitt Mnchen (TUM)*, page 89, 2010.

[67] A.H. Tan. Text mining: The state of the art and the challenges. In *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, 1999.

[68] F.E.H. Tay and L.J. Cao. Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent Data Analysis*, 5(4):339–354, 2001.

[69] P.C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.

[70] P.C. Tetlock, M. SAAR-TSECHANSKY, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008.

[71] B. Vanstone, G. Finnie, and C. Tan. Evaluating the application of neural networks and fundamental analysis in the australian stockmarket. *Information Technology papers*, page 15, 2005.

[72] T. Warren Liao. Clustering of time series data – a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[73] S.M. Weiss and N. Indurkhya. Rule-based machine learning methods for functional prediction. *Arxiv preprint cs/9512107*, 1995.

[74] Q. Wen, Z. Yang, Y. Song, and P. Jia. Automatic stock decision support system based on box theory and svm algorithm. *Expert Systems with Applications*, 37 (2):1015–1022, 2010.

[75] Adam White. *Futures Magazine*, August 1991.

[76] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, pages 486–497, 2005.

[77] J.W. Wilder. *New concepts in technical trading systems*. Trend Research Greensboro, North Carolina, 1978.

[78] Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 13–22. Springer-Verlag New York, Inc., 1994.

[79] Y. Yang and C.G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12(3): 252–277, 1994.

[80] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 412–420. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[81] W. Zhang and S. Skiena. Trading strategies to exploit blog and news sentiment. In *Proc. of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 375–378, 2010.

[82] X. Zhang, H. Fuehres, and P. Gloor. Predicting stock market indicators through twitter–"i hope it is not as bad as i fear". In *COIN Collaborative Innovations Networks Conference*, pages 1–8, 2010.