공학박사학위논문

Prediction of stock price, base rate, and interest rate spread with text data

텍스트 데이터를 이용한 주식 가격, 기준 금리 및 스프레드 예측

2018 년 2 월

서울대학교 대학원

산업공학과

김 미 숙

Prediction of stock price, base rate, and interest rate spread with text data

텍스트 데이터를 이용한 주식 가격, 기준 금리 및 스프레드 예측

지도교수 조 성 준

이 논문을 공학박사학위논문으로 제출함

2017 년 12 월

서울대학교 대학원

산업공학과

김 미 숙

김미숙의 박사학위논문을 인준함

2017년 12 월

위 원 장	이 재 욱	(인)
부위원장	조 성 준	(인)
위 원	장 우 진	(인)
위 원	조 성 배	(인)
위 원	강 필 성	(인)

Abstract

Prediction of stock price, base rate, and interest rate spread with text data

Misuk Kim
Department of Industrial Engineering
The Graduate School
Seoul National University

Methodologies in financial research based on a variety of predictions models have been actively developed for the analysis of market behaviors. The significance of prediction modeling in the financial market cannot be emphasized better especially given that it leads directly to large transaction profit. In terms of applicability for the active agents in the market requires, these research results require both predictability and interpretability. In this study, we propose methodologies suitable for incorporating distinct characteristics across different financial data in the analysis for the purpose of effective prediction modeling. Firstly, we propose a methodology that quantitatively and qualitatively predicts the stock price movements through sentiment analysis of corporate disclosures in the stock market. The proposed method predicts stock price movements by embedding the documents, and the class of documents defined to fit the purpose of our study, to the same projection space based on the distributed

representations learned, and compares the predictive performance against various existing models. The results provide prime evidence of effectiveness of our prediction results through visualization of document sentiments. In addition, we propose a methodology specifically designed for predicting the vote results of the base interest rate, which is the most important factor in the bond market, developed within the premise of the Korean bond market. Our methodology allows computation of sentence sentiments using the monetary policy decision recorded as text data, which is released before the announcement of the vote result, which are then aggregated to the document level to express the document sentiment of monetary policy decision into values. Using these sentiments, we predict the vote results of the base rate. Finally, we define the framework for predicting the spread, the difference between two bond rates with different maturities. The framework mainly considers the following three aspects as the standards for the effectiveness of research: interpretability, proper prediction metrics, and the reporting methods. The framework use wrapper approaches for the practical interpretation of important variables, while using PARE, in combination with MAE, as prediction metrics, for taking into account the tolerance of the spread. Lately, we suggest various visualizations and hierarchical illustration of significant variables as more applicable and effective reporting methods. This dissertation defines a variety of financial problems, proposes analytical methodologies, compares quantitative prediction power, and provide the qualitative evidence. The proposed methodologies prove to serve as a quick and accurate data-driven decision making support tool to active agents in the

real-site.

Keywords: Data Mining, Machine Learning, Word Embedding, Distributed Representation, Bag-of-Words, Sentiment Analysis, Stock Price Prediction, Vote Result Prediction of Monetary Policy Committee, Bond Spreads Prediction, Corporate Disclosures, Monetary Policy Documents, Economic Indicators

Student Number: 2015-30235

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Financial Markets

Financial markets include organized markets where funds are traded within the premise of concrete forms such as stock exchanges, as well as those with abstract meaning such as systematic or repetitive over-the-counter (OTC) transactions. The structure of financial market ensures that suppliers and consumers collectively determine the market prices and fulfill fair transactions assumptions guided by established rules. Institution-wise, financial markets can be broadly divided into indirect financial markets and direct financial markets. Indirect financial markets consist of financial institutions, banks and insurance companies in particular, which connect consumers with suppliers of funds. In the meantime, direct financial market includes stock and bond markets. In this dissertation, we will study direct financial markets.

The stock market is an aggregation of buyers and sellers who trade equities, stocks of publicly held companies, bonds and other securities. There exist two main branches: the stock market and the distribution market. In the issue market, stocks are issued. Then, in the distribution market, the issued securi-

ties are circulated. The distribution market can further be subdivided into: (1) an exchange market, in which transactions occur through the stock exchange, and; (2) the over-the-counter market, where the dealers act as market makers by quoting prices at which they will buy and sell securities. In the stock market, a myriad of indicators serve as means to approximate the performance of the stock market at once, such as the Dow Jones Industrial Average (DJIA, henceforth), the NASDAQ Index, Russell 2000, and Standard and Poor's 500 (S&P500, henceforth). DJIA, comprised of stocks of the largest 30 companies in the United States, is the leading indicator of the U.S. stock market. On the other hand, the S&P500 is the highly reputable measure of the U.S. stock market which considers the 500 largest capitalization stocks traded in the U.S. The stock price that constitutes the index represents the value of the company, and it is calculated simply by dividing the market capitalization by the number of shares outstanding. It is not easy to measure or quantify the value of a company by the products or services it currently produces or offers. On top of the current values, the value of all future cash flows, such as borrowed or invested capital costs, business risk, opportunity cost, and risk-free interest rate, should also be considered in computation. At the same time, stock prices are influenced by the opinions of various investors, both optimistic and pessimistic traders and spectators, every trading day. Each investor judges the entire market based on various indicators of choice such as GDP, unemployment rate, and/or corporate profit. For example, as economy grows, a company invests in projects while expanding its business and hiring more staff. If the stock price is expected to be higher upon the completion (or during the process of) expansion than the

current stock price, a given investor will try to buy the stock and the stock will enter the bull market. In contrast, if selected indicators evoke fear or panic in the market, the investor may act irrationally and keep on selling. A well-known historical example would be the Great Recession of 2008 in the U.S., where the market value declined frantically.

On the contrary, the bond market is generally the secondary market where issued bonds are traded among market participants. A given investor may cash and secure the necessary funding by selling bonds obtained from the bond market while investing the residual funds in the bond market to efficiently manage the assets. The bond market performs two essential economic functions. Firstly, it facilitates the procurement of long-term investment funds by offering liquidity to investors who purchase bonds in the issue market. Secondly, it provides a basis for determining the capital cost of the corporation by forming the fair price of bonds issued. Currently, the bond market is divided into the exchange market and the OTC market. In order to smooth financing for corporate restructuring after the 1997 financial crisis, new bond-related policies were introduced while improving the market infrastructure, through which the Korean bond market has come to develop around the government bond market. Ever since, Korean bond market has expanded both in terms of variety and size. Nowadays, being actively traced are a wide range of items including government bonds, monetary stabilization securities issued by the Bank of Korea, municipal bonds issued by local governments, corporate bonds, and financial bonds. The size of the market increased as well, thanks to the opening of the foreign bond market, the introduction of a primary dealer, the appearance of a bond brokerage company,

and the issuance of various maturity bonds.

In 2015, the market capitalization of the Korean bond market reached 1,780 trillion won, which was greater than the gross domestic product (1,559 trillion won) and the stock market (1,404 trillion won) at that time. The bond market has high predictive power in terms of bond prices because it has low noise, the cause of which is the prevention of reckless speculation by individuals. Moreover, the market is affected by data such as macroeconomic indicators (Cheng 1996). Despite this, however, there is relatively little pertinent research; thus, it is meaningful to analyze and forecast the base rate, which is closely related to bond prices. For example, the coupon rates of bonds are fixed when bonds are issued; consequently, if the base rate is raised, the demand for bonds decreases and their prices decline.

Through many different channels via a variety of means, financial markets play an important role in the well-being of enterprises and the overall economy. Does the financial market not only transfer real economic resources and offers dividends or interest to the market participants, but it also creates liquidity, enabling trades among the investors in the market. Therefore, it is important to analyze various factors that are linked organically to the financial market, and to derive new information from massive amounts of data.

1.2 Data-driven Decision Making

The financial market provides a very rich variety and volume of data. A given company's performance is reported via financial statements, historical data

(Ballings et al., 2015; Qian & Rasheed, 2007), and real-time data. A set of methods have been developed to analyze and predict target values in order to extract the information that suits research objectives. Research on financial markets has mainly focused on the study of stock market behavior (Ariyo et al., 2014; Huang et al., 2014). Above all, fundamental analysis centers around developing a methodology to evaluate company performance and credibility. Fundamental analysis is the most rational, objective, and widely used method because it uses publicly available information such as financial statement analysis. If a company operates well, additional capital will be created and the target value will rise accordingly. Technical analysis, on the other hand, concentrates on predicting the future target values through observing trends in the past target values rather than company fundamentals, based on the assumption that certain patterns will repeat and deliver consistent results. Technical analysis involves capturing patterns from time-series data by creating charts and/or figures with statistics. Well-known examples of the results of such analysis include candle stick patterns, the head and shoulders reversal patterns, and the cup and handle continuation patterns. Finally, a branch of target value prediction studies utilize selected data mining techniques fit for the research objectives, such as artificial neural network or support vector machines. For example, Preis et al. (2013) identified online precursors of stock prices based on search volume data provided by google trends in order to pin down optimal trading strategies. Moat et al. (2013) studied the relationship between Wikipedia's views change and total stock market movements. More recent studies began to incorporate textual data in the analysis of market behaviors, ranging from the comprehensive re-

ports on the company's performance submitted to the Securities and Exchange Commission (SEC) such as 10-Ks and 10-Qs (Feldman et al., 2010), the opinion columns of the Wall Street Journal (Dougal et al., 2012) and the finance section of the New York Times (Garcia, 2013), to earning press releases (Huang et al., 2014). Specific words, phrases, paragraphs, and documents appearing in textual data convey certain sentiments, and the analysis of these sentiments is gaining growing attention as a new method for stock price prediction tasks in the financial field.

An investor may claim ownership on business by purchasing shares, hence making stocks a very important component of the free market economy. Stock investors are striving to land on the optimal investment strategies in order to generate profits as large as possible from the subsequent transactions by investigating various types of new data continuously provided. Given the breath and variety of data available, there is a growing need to extract valuable information from the oceans of data provided. More recently, efforts have been made to incorporate different modes of data in the analysis to extract more valuable information that is more meaningful and applicable in practice. Textual data, in addition to numeric indicators, has begun to be used in stock price prediction models. It has been shown that textual data can provide a very rich set of insights in terms of market research in both qualitative and quantitative manners (Mostafa, 2013; Sun et al., 2016). Incorporation of textual data has proven useful in the sense that it adds intuitive descriptions to the numeric predictions of machine learning models, hence serving as a data-driven decision-making support tool for active traders in the field as they make split-second trading

decisions.

On the other hand, as discussed in Section 1.1, there are only few studies that focuses on the analysis of the bond market despite the importance of its role in understanding the market movements (J. Wang & Wu, 2015). Bond market research on emerging markets is even rarer. In many cases, transactions decisions are often made based on personal assessments rather than data-driven analysis. However, the bond market is affected by various domestic and overseas economic indicators, monetary policy, and the supply and demand trends of foreign investors. Recently, efforts have been made to collectively summarize patterns implied by various indicators and use them in bond prices or interest rate predictions (Kung, 2015). Studies have shown that bond prices are less affected by noise generated from the reckless speculation of individuals, while being more responsive to numeric variables such as macroeconomic indicators (Cheng et al., 1996). Given so, precise mathematical modeling of bond prices predictions, based on observations from massive historic data, may provide greater prediction power as compared to the conventional stock price prediction models. More importantly, past studies focused on countries with large bond market. Thus, their results may not be applicable for markets of different characteristics or history. It calls for analysis specifically designed for the market of interest so that the empirical findings can reflect the distinct domain characteristics peculiar to the subjective market.

As traders face the vast amount of data available as they make split-second decisions in the market, they will benefit from data-driven analysis utilizing advanced machine learning techniques in the following aspects. Firstly, by ap-

plying various approaches to understand and derive insights from data, they can identify the characteristics of the financial markets of a particular country as well as the characteristics of the global financial market. In addition, based on data analysis techniques such as prediction models and association analysis, it will be possible to detect signals that can generate revenue. Finally, results from data-driven analysis may add objectivity to the intuitive decision-making process.

1.3 Outlook of this Dissertation

In this dissertation, we develop a financial prediction model using multiple modes of financial market data from both numeric and textual sources. We propose methodologies that predict the target value of the stock market and the bond market and extract intuitions that may prove meaningful to the active business agents. The diagram of this dissertation is shown in Figure 1.1.

As shown in Figure 1.1, we use sentiment analysis, aggregation of sentence sentiments, and interpretational data mining framework methods using stock and bond data to predict various target variables of the financial market. We support the data-driven decision-making processes in the business by providing the results from the machine-learning-based predictions, and visualization tools, from which meaningful interpretations of data can be drawn.

The rest of this dissertation is organized as follows. In Chapter 2, past studies on the financial markets are briefly reviewed in terms of predictability and interpretability, while introducing different methodologies for analyzing the

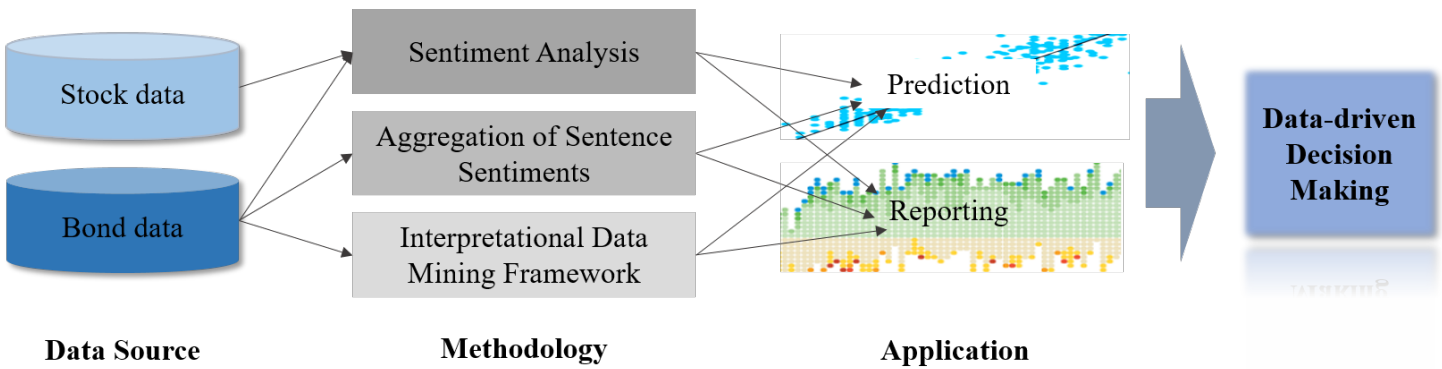


Figure 1.1: Diagram of this dissertation

data. Chapter 3 and 4 introduce various ways to exploit financial textual data in solving the market prediction tasks. In Chapter 3, in particular, we provide both quantitative and qualitative ways to predict stock price movements using corporate disclosures. We additionally provide new venues for reporting results through visualization. In Chapter 4, we try to characterize the Korean bond market by analyzing and forecasting the vote result of the base rate decision, the most important factor in the bond market, based on the Korean monetary policy documents. In Chapter 5, we define a data mining framework with characteristics necessary for financial prediction problems. Finally, we discuss contributions and future work of this dissertation in Chapter 6.

Chapter 2

Literature Review

2.1 Financial Predictability Modeling

Various hypotheses and models have been proposed and applied to explain financial markets. Particularly in the stock market, research is actively conducted to predict the value or direction (upward or downward movement) of stock prices or stock indexes. According to the efficient market hypothesis of Fama (1970), it is impossible to predict the market perfectly since publicly available information is fully reflected in stock price and the market will respond only to the new information. However, Lo and MacKinlay (1988) insisted that the market could be predicted to some extent. Apart from the traditional autoregressive integrated moving average (ARIMA) models (Ariyo et al., 2014), more recent work began to look to artificial intelligence algorithms to solve stock price prediction tasks. For instance, Sitte and Sitte (2002) used a neural network(NN) to detect weak signals in the S&P500 time series, while Ballings et al. (2015) employed random forest (RF) model to predict the direction of stock price movements. Support vector machine (SVM) was another technique utilized in various studies to predict stock price indices (Tay & Cao, 2001). In ad-

dition, in the financial market and industry, researchers have actively employed both textual and numerical data in order to predict stock price changes.Gálvez and Gravano (2017) analyzed stock messages on stock bulletin boards to predict asset returns, and there was research that predicted DJIA's stock price movement by analyzing Yahoo!Finance and Google Finance news (Beckmann, 2017).Bollen et al. (2011) studied the relationship between Dow Jones Industrial Average (DJIA) and public mood measured by quantifying twitter data using OpinionFinder. Google-Profile of Mood States(GPOMS) measured public mood in terms of six dimensions: calm, alert, sure, vital, kind, and happy. The correlation between public mood and DJIA was analyzed by Granger causality analysis, and it was observed that "Calm" mood was the most significant feature. The self-organizing fuzzy neural network method predicted a daily up and down change of DJIA closing-values and obtained 86.7% accuracy and a 6% decrease in mean average percentage error (MAPE). Lee et al. (2014) reported that the performance of stock price prediction improved when using linguistic features of financial reports rather than the existing analysis using quantitative indicators such as earning surprise, recent movement, volatility, and event category for S&P 1500 companies. The unigram features, extracted from documents, were expressed using non-negative matrix factorization (NMF) and used as input variables. For three classification problem through RF, the model with quantitative indicators was used as a baseline with 50.1% accuracy, on the other hand, when linguistic factors were added, the prediction accuracy was 55.5% and improved by more than 10% over baseline performance. Sun et al. (2016) explored the effect of text information from user-generated microblogs on the

market. On a financial communications platform called StockTwits ®[1], textual data were collected for five years. The author created term-document matrix and dense input variables through sparse matrix factorization model. Using the latent space model by Ming et al. (2014) improved performance and predicted accuracy of 51.37% compared to the baseline regression model. It was found that StockTwits® contain information useful to asset managers and investors and contributed to the use of high-volume social media data without using news and text sentiment.

On the other hand, in the bond market, there are relatively few studies predicting the target values of bond such as bond prices and the base rate. Kim and Noh (1997) used a neural network (NN) to predict the interest rates of both Korea and the United States (US) and to establish whether they have a significant impact on the countries' economic network and financial markets. A total of 36 input variables were used to predict US interest rates, with lag 0 to lag 5 applied to treasury bills with one-year maturity, the money stock, the consumer price index, the industrial production index, housing starts, and the Standard & Poor's 500 respectively. In order to predict Korea's interest rates, the author once again used 36 input variables, with lag 0 to lag 5 applied to corporate bond yields with three-year maturity, the money stock, the consumer price index, the industrial production index, permits for building construction, and the Korean stock price index respectively. The results showed that applying the integrated NN to the Korean interest rate was not significantly different to applying the random walk model; however, Kim and Noh (1997) found that

[1]http://www.StockTwits.com

applying the integrated NN to US interest rates surpassed the random walk model.

2.2 Financial Interpretability Modeling

Due to the nature of financial markets, models with both predictive and interpretational power are called for. Many researches have exploited textual data, such as news, online blogs, and financial reports in order to provide intuitive descriptions beyond the numerical prediction results. Chen and Liu (2009) and Schumaker and Chen (2010) make use of both news articles and text from social network services as well as online blogs, where the former represents market sentiment, while the latter, individual investor's sentiments. Druz et al. (2015) collected earning transcripts from 2004 to 2012 for all stocks belonging to the $S\&P$ 500 index and studied the relationship between the managerial tone and the investor's variables of interest, including future stock returns. A "negativity" score, defined as the number of negative words minus the number of positive words divided by the sum of negative and positive words plus 1, was used to standardize the "one surprise", the excessive components of managerial tone. Druz et al. (2015) reported that tone surprise was a significant factor in predicting earnings per share adjustments to the sell-side trader. Furthermore, from the adjusted long-short strategy from the regression framework - taking long stocks with positive tone surprise and short stocks with negative tone surprise – they obtained 1% return within 60 days after the earning call. Jegadeesh and Wu (2013) found a meaningful relationship between market re-

action and the tone quantified from the Form 10-K documents by counting negative and positive words. The results from multivariate regressions showed that both negative and positive words were significant features in explaining market reaction, and the effect of the tone of documents were observed quickly in the market, mostly within two weeks. The same methodology was applied to the IPO prospectuses to examine the relationship with IPO underpricing and found that tone of IPO prospectuses adversely affected IPO underpricing. Heston and Sinha (2017) predicted stock prices exploiting almost 1 million news stories. The sentiment of news stories was measured by using the Thomson Reuters sentiment engine, and the sentiment of companies was calculated by subtracting negative sentiment score from the positive. Results showed that the duration of stock return predictability depended on the temporal aggregation of news. Sentiment of news over a day has predictability only a few days, while the sentiment of news over a week lasts the predictability for up to a quarter. In addition, positive news causes stock returns to rise quickly, while negative news affects stock returns with a long-delayed reaction. Liu (2015) inspected the relationship between investor sentiments and the time-series variation in stock market liquidity. Market liquidity is defined as the absolute price change per dollar of daily trading volume for each stock each day, and calculated as $\frac{|R^i_{td}|}{\$VOL^i_{td}}$, where R^i_{td} is stock i's return on day d of month t and $\$VOL^i_{td}$ is the same day dollar trading volume of this stock. They sentiment index also indicated market professional read about 150 newsletters each week and marked the future market movement as bullish, bearish, and neutral. Two indices were verified to be significant by the Granger-cause test. In other word, as the sentiment index

increases, the stock market becomes more liquid. In addition, the time series regression revealed that investor sentiment, in particular, had an effect on stock market volatility and price effect, and the investor's strong sentiment was found to directly and indirectly affect stock prices.

In addition to these statistical analyses, data mining techniques have been used to predict interest rates. In particular, text-mining methods have provided intuitive results in the form of graph, wordcloud, and tree visualizations (Bholat 2015).

Whereas, in the bond market, ddata mining techniques have been used to predict interest rates.In particular, text-mining methods have provided intuitive results in the form of graph, wordcloud, and tree visualizations (Bholat et al., 2015). In addition, Various studies have examined the effect of monetary policy on stock prices, exchange rates, and interest rates in financial markets (Hughen & Beyer, 2015; B. Bernanke & Gertler, 2000). Research has also been conducted to find relevant insights by statistically analyzing the effects of monetary policy on interest rates (Goodfriend, 1991; Fuhrer & Moore, 1995). Kuttner (2001) analyzed the monetary policy data published by the federal funds and tried to classify the anticipated and unanticipated effect elements. As a result, the author statistically confirmed that the movement of the interest rate of the federal funds had a small effect on the anticipated element and a significant effect on the unanticipated element. Moreover, a sudden change in the target rate did not affect the actions of the Federal Reserve System, but the author confirmed that such a change provided mainly explanatory power for short-term curves. Other studies have considered the monetary policies adopted by central banks

to address the 2008 financial crisis (Gagnon et al., 2011; D'Amico & King, 2013; Li & Wei, 2013). The central bank of each country implemented quantitative easing policies, whereby the bank purchased government or corporate bonds in advance to stimulate the economy, resulting in almost zero interest rates in developed countries. Indeed, an analysis of asset price movements on the day of a Federal Open Market Committee (FOMC) announcement confirmed that monetary policy has a significant impact on long-term Treasury yields (Kiley, 2016).

B. S. Bernanke et al. (2005) used factor-augmented vector autoregression (FAVAR) to determine the impact of the monetary policy process on the economy. By applying FAVAR to diverse variables, he identified the impact of monetary policy through two steps: finding the major components of diverse variables and accelerating the calculation through Bayesian methods based on Gibbs sampling. This model found variables where monetary policy has a significant impact on the macroeconomy; for example, when a contractionary monetary policy shock occurred, responses such as a decrease in the real activity measure, a decrease in monetary aggregates, and an increase in the dollar price were confirmed. The study was able to provide a comprehensive picture of the impact of monetary policy. Moreover, diverse variables that responded to monetary policy were used in the macroeconomic model; thus, they provided empirical plausibility. In fact, prior studies have mainly used various economic indicators to analyze the impact of monetary policy on market interest rates. Alternatively, they have focused on studies in developed countries such as the US and the United Kingdom. (Burger et al., 2017) analyzed the impact of US

monetary policy on the bond markets of emerging countries; however, there is limited research on such markets. It is necessary, though, to study monetary policy in emerging countries because their financial markets are expanding (Tiemei, 2010).

2.3 Data-driven Modeling Techniques

Prediction models have been developed exploiting massive amount of historic data, in order for knowledge discovery. This section introduces representative prediction models, namely, logistic regression (LR), random forest (RF), multinomial Naïve Bayes (MNB), support vector machine (SVM), support vector regression (SVR), Naïve Bayes SVM (NBSVM), and neural network (NN).

Logistic regression (LR) is widely used when the output variable is a categorical variable. The probability of response can be estimated through logistic function (Cox, 1958; Walker & Duncan, 1967). The logarithm of the odds ratio, y_i, which is the ratio the probability of $Y = 0$ and $Y = 1$ at $X = x$, is predicted by the linear regression:

$$y_i = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon_i, \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{2.1}$$

where p_i is the probability that the response variable equals a case i, β_0 is the intercept from the linear regression equation, and $\beta_1, \ldots, \beta_n$ are the regression coefficients. The Equation 2.1 can be expressed as an equation for p_i, defined

Algorithm 1 Random forest

Given a training set $\mathcal{S} = (x_1, y_1), \ldots, (x_n, y_n)$, feature $\mathcal{F}$, and the number of trees $\mathcal{B}$

function RANDOM FOREST($\mathcal{S}$, $\mathcal{F}$)
  $\mathcal{T} \leftarrow \phi$
  for $i \leftarrow \sim 1 : \mathcal{B}$ **do**
    $\mathcal{S}^{(i)} \leftarrow$ *random sample set from* $\mathcal{S}$
    $t_i \leftarrow DecisionTreeLearn(\mathcal{S}^{(i)}, \mathcal{F})$
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{t_i\}$
  end for
end function
function DECISIONTREELEARN($\mathcal{S}, \mathcal{F}$)
  for At each node **do**
    $f \leftarrow$ *small subset of* $\mathcal{F}$
    Split on best feature in f
  end for
  return *Learned tree*
end function

as a logistic function:

$$\hat{p} = \hat{T}(Y = 1|x) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n)} \tag{2.2}$$

The resulting probability $\hat{p}$ ranges between 0 and 1. If $\hat{p}$ is greater than the predetermined criteria, x is classified as 1; otherwise, the model classifies the observation as class 0. LR is easy to implement and provides intuitive interpretation of the relationship between the response variable and predictor variables. LR, however, is often less accurate due to overfitting.

Random forest (RF), which is an ensemble learning method for classification and prediction, is a combination of a number of decision or regression trees. Algorithm 1 presents the pseudocode of this study's RF.

After training, unseen samples, x^{new}, are predicted by accepting a majority vote among the trees for classification or by averaging the predictions from all trees on x^{new} in the case of the regression tree, as follows (Breiman, 2001):

$$\hat{T} = \frac{1}{B} \sum_{b=1}^{B} \hat{T}_b(x^{test}) \tag{2.3}$$

where $\hat{T}_b$ is the prediction from tree B. RF is one of the most accurate and widely known learning algorithms. For many data sets, it not only makes a highly accurate classifier but also provides estimates of the variables that are important in the classification. However, random forests have been observed to overfit some data sets with noisy classifications or regression tasks.

Multinomial naïve Bayes (MNB) is a classification model that uses conditional probability of belonging to a class, assuming that each variable is independent (Rennie et al., 2003). Given data x_i, $i = 1, \ldots, n$ with K classes, the conditional probability of belonging to class k is as follows:

$$p(C_k|x_1, x_2, \ldots, x_n) = p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{X}|C_k)}{p(\boldsymbol{x})} \tag{2.4}$$

where, the denominator is a constant value, regardless of the class, which is calculated only from the observed data. Since each variable is independent, a probability model can be expressed using only the numerator:

$$p(C_k, x_1, x_2, \ldots, x_n) = p(C_k)p(x_1|C_k)\ldots p(x_n|C_k) = p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{2.5}$$

We can now define the class of each variable that maximizes the Equation 2.5 as follow:

$$\hat{y} = \arg \max_{k \in \{1,\dots,K\}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{2.6}$$

Because MNB assumes that all variables are independent of each other, it can lead to inaccurate results depending on the data set. However, it is possible to estimate the distribution of a class as a one-dimensional distribution, so the model has good performance and exhibits fast speed even with large datasets.

Support Vector Machine (SVM) is a supervised learning model which constructs a hyperplane in a high dimensional space for classification, regression, or other tasks. The hyperplane, $w^T x + b = 0$ where x is dataset, w is the normal vector to the hyperplane, and b is the bias, separates the vector space (Cortes & Vapnik, 1995). A graphic illustration of SVM model is shown below in Figure 2.1.

We search for a decision boundary shown in the solid line in Figure 2.1 and the ones closest to the that boundary are called the support vectors. The distance between these support vectors and the decision boundary is $1/w$, and the range $2/w$ is called the margin. Finally, learning the SVM can be formulated as an optimization:

$$\begin{aligned} \max_{w} \quad & \frac{2}{||w||} \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 \end{aligned} \tag{2.7}$$

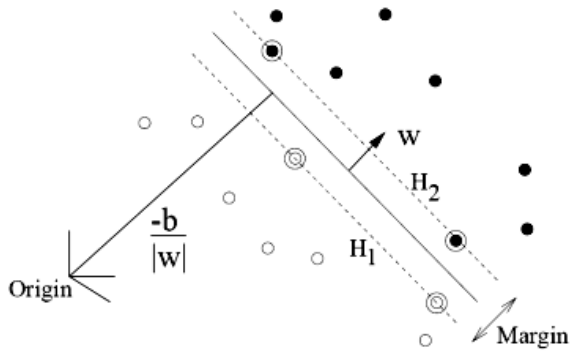


Figure 2.1: Support vector machine

We classify the data using a discriminant function, $f(x) = w^T x + b$, that has optimized parameter w, b. If a discriminant function $f(x)$ is greater than 1, classify as class 1; if a discriminant function $f(x)$ is less than -1, as class 0. SVM is the proposed model based on structural risk minimization, as it has better prediction and a wide range of application, while optimization requires a long time for learning with large datasets.

SVR finds the regression equation by using epsilon-band contained lots data. Given a training set $\{x_i, y_i\}$, $i = 1, ..., N$, standard SVR can be formulated as

follows (Nocedal & Wright, 2006):

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^*} \quad & \tfrac{1}{2} w^T w + C(\textstyle\sum_i \xi_i + \sum_i \xi_i^*) \\ \text{s.t.} \quad & y_i - (w^T \varphi(x_i) + b) \leq \epsilon + \xi_i, \\ & (w^T \varphi(x_i) + b) - y_i \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, i = 1, ..., N \end{aligned} \tag{2.8}$$

where C determines the tradeoff between the training error and flatness, ϵ is the parameter of ϵ-insensitive loss function of SVR, and ξ_i, ξ_i^* are the slack variables, which are zero when the training pattern is inside the ϵ-insensitive tube. This primal problem can commonly be solved by the dual quadratic programming problem (Wolfe, 1961). We can obtain the non-linear regression function by employing a kernel function, which maps the original input space into a high-dimensional feature space (Slavkovic et al., 2013). SVR is the proposed model based on structure risk minimization, as it has better prediction and a wide range of application, while quadratic programming optimization requires a long time for learning.

Naïve Bayes SVM (NBSVM) is very similar to SVM, except that we use transformed input variables x_i, where $x_i = r \circ f$ is the elementwise product. r, the log count ratio, is defined as follows:

$$r = \log(\frac{p/||p||_1}{q/||q||_1}) \tag{2.9}$$

where $p = \alpha + \sum_{i:y_i=1} f_i$, $q = \alpha + \sum_{i,y_i=-1} f_i$, and α is smoothing parameter.

Interpolation w' is calculated using an interpolation parameter(β) as follows:

$$w' = (1 - \beta)\bar{w} + \beta w \tag{2.10}$$

where $\bar{w} = ||w||_1/|V|$ is the average magnitude of w. Finally, a discriminant function for NBSVM is defined as: $f(x) = \bar{w}^T x + b$. If a discriminant function $f(x)$ is greater than 1, then x is assigned to class 1; if a discriminant function $f(x)$ is less than -1, to class 0. By applying these transformed parameters to SVM, we can improve the performance.

Among various kinds of Neural Network (NN), we introduce multilayer perceptron architecture for the regression problem. Regression function of the NN is based on a linear combination of nonlinear basis functions $h_i(x)$ in the form

$$f(x) = w_0 + \sum_{i=1}^{N_{hn}} w_i h_i(x) \tag{2.11}$$

where w_i are weights, and $h_i(x)$ are generally chosen to be sigmoidal functions. In order to derive optimal weights in the training phase, a back propagation algorithm is generally employed. Back propagation algorithms update weights iteratively to minimize an error function. After training, the output of test data x can be estimated using weights (Rumelhart et al., 1988). NN is applicable to a number of problems because it is not affected by variable attributes, and it can find the nonlinearity between response variables and predictor variables. However, the training time is considerable on account of model complexity and

it is difficult to interpret the model.

As compared to numerical data, textual data requires much more intensive preprocessing, since it is less structured. There exists a number of methods to represent text as numbers, including a technique as simple as bag-of-words (BoW) approach. The BoW model is the method of mapping a document to a feature vector. A simple overview is shown in Figure 2.2 (Harris, 1954).

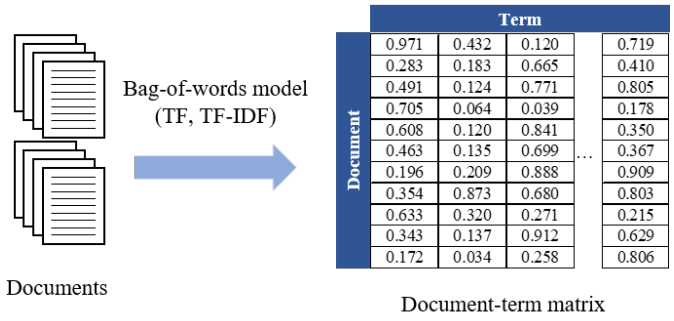


Figure 2.2: Bag-of-words model

The BoW can be represented by a document-term matrix (DTM). The DTM, the two-dimensional array that represents documents in rows and terms in columns, is a method of expressing the frequency of occurrences of words in documents as a matrix, and each cell value represents a score for the term frequency. Among the term-weight methods of representing documents, the current study described the most frequently used: term frequency (TF) and term frequency-inverse document frequency (TF-IDF). TF identifies a document by assigning weights to it, assuming that other documents with related words are similar. Further, each cell in a row represents the number of terms

in one document. In other words, if $a_{(t,d)}$ is a cell that corresponds to the term t in document d, $a_{(t,d)}$ can be defined with various weights. This study used the number of times that term t occurs in document $d(f_{(t,d)})$. The equation is expressed as follows:

$$TF(t,d) = f_{(t,d)} \tag{2.12}$$

TF-IDF is a method that includes not only TF information but also the importance of a term to a document in a corpus. The value that reflects the commonness of a term in the entire document set is expressed as follows:

$$IDF(t,D) = \log \frac{|D|}{1 + |d \in D : t \in d|} \tag{2.13}$$

D is the total number of documents and $|d \in D : t \in d|$ is the number of documents containing the term t. We add one to the denominator so that it is not zero; we then combine Equation 2.12 and 2.13 to obtain TF-IDF as follows.

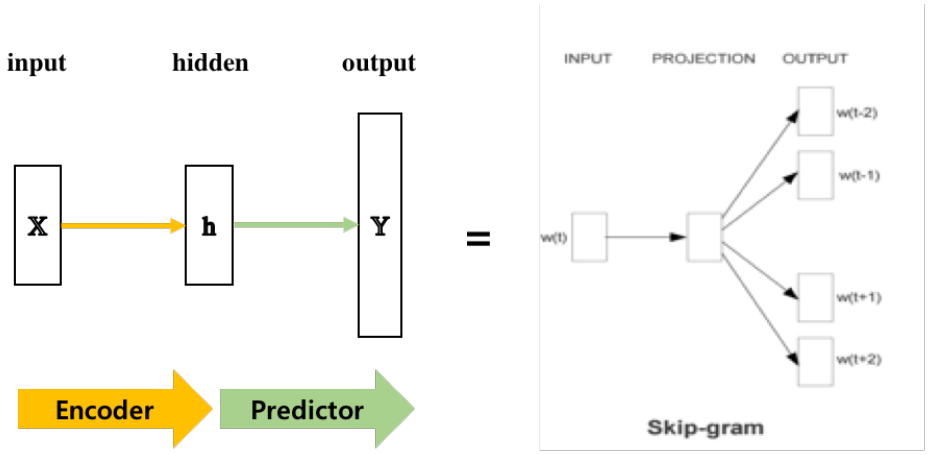
$$\begin{aligned} TFIDF(t,d,D) &= TF(t,d) \times IDF(d,D) \\ &= f_{(t,d)} \times \log \frac{|D|}{1 + |d \in D : t \in d|} \end{aligned} \tag{2.14}$$

A high weight in Equation 2.14 is obtained through a high-term frequency in a particular document and a low document frequency of the term in the entire document set. If a term appears in other documents, the value inside the

logarithm approaches 1; thus, the logarithmic value becomes 0. As a result, it is possible to filter out common words in every document.

However, the biggest shortcoming of BoW model is in that as the number of words increases, the dimension of the representations explodes astronomically. Distributed representation addresses such a limitation by projecting input data onto a continuous space of a smaller dimension (Rumelhart et al., 1985). The similarity between words and documents can easily be calculated, because each document is projected on to the same, continuous embedding space. The resulting representation then can be used in various ways, from extracting words that are "closer" to a given word to clustering similar documents. Moreover, it has been reported that performance, to some extent, is guaranteed even when working with a small dataset. In this study, we employ word2vec (Mikolov et al., 2013) to obtain distributed representations of documents. Word2vec is a simple NN model that embeds words onto a continuous embedding space, and it has become the most widely used word embedding model by reducing the computational time and enabling learning several times faster than, for instance, calculating a sparse matrix as required by the conventional BoW method. One interesting feature of word2vec is that linguistic regularities can be applied to the representation vectors. For example, one may express Rome as a combination of Paris, France, and Italy in the following way: v("Rome")= v("Paris")-v("France")+v("Italy"). There are two ways through which word2vec learns distributed representations of words: Skip-gram and Continuous Bag-of-Words (CBOW). Skip-gram predicts the context words given a target word; CBOW predicts the probability of observing the target word given its context words.

Skip-gram model is reportedly suitable for learning distances in discrete data that is difficult to define similarity. It represents textual data as a continuous vector, with which words within similar contexts are placed close to each other on the same embedding space. Figure 2.3 shows the structure of the skip-gram model.



$x \in R^{V\times 1}$ One representation word vector. Input layer size equals with no. of terms

$h \in R^{d\times 1}$ Encoded word vector

$y \in R^{V\times 1}$ Context prediction prob. Output layer size also equals with no. of terms

Figure 2.3: The structure of the skip-gram model

Skip-gram utilizes a NN, which consists of an encoder and a predictor. The encoder converts the input words, coded discretely, into a continuous vector; the predictor, then, uses the resulting vector from the encoder to predicts context words. Mathematically, given training words $w_1, w_2, \ldots, w_T$, the objective

function of skip-gram is defined as:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-r\leq j\leq r, j\neq 0}\log p(w_{t+j}|w_t) \tag{2.15}$$

where r is the size of the training context. The conditional probability $p(w_{t+j}|w_t)$ is calculated by the softmax function:

$$p(w_{t+j}|w_t) = \frac{\exp({v'_{w_{t+j}}}^T v_{w_t})}{\sum_{w=1}^{W}\exp({v'_{w_{t+j}}}^T v_{w_t})} \tag{2.16}$$

where $v'_{w_{t+j}}$ and v_{w_t} are the input and output vector representations of w, and W is the number of words in the vocabulary. Skip-gram treats each context-target pair as a new observation, and this tends to work better with large datasets.

Textual data was used based on the above feature engineering methods, thus many studies have confirmed that the prediction performance was improved by using textual data rather than numerical data only. Therefore, various algorithms have been developed and applied to better represent textual data to improve the prediction performance.

Maas et al. (2011) proposed a method to capture sentiment and semantic term-document information by combining unsupervised and supervised techniques to overcome the problem of not capturing sentiment information when representing documents. In order to capture semantic similarities, they constructed a probability model of a document that uses a continuous mixed dis-

tribution of words, $p(d) = \int p(d, \theta)d\theta = \int p(\theta)\Pi_{i=1}^{N} p(w_i|\theta)d\theta$, where N is the number of words in d and w_i is the i^{th} word in d. For capturing the word sentiment, used was the logistic function, $p(s = 1|w, R, \psi) = \sigma(\psi^T \phi_w + b_c)$, where R is a word representation matrix, ψ is regression weights, ϕ_w is w's vector representation, and b_c is a scalar bias. The objective function was set up by combining the two probabilities in order to search for optimal parameters. It was shown that the proposed method extracted semantically and sentimentally similar words using cosine similarity better than the model that use only the semantic objective function, or latent semantic analysis (LSA)(Turney & Pantel, 2010). For example, when five words similar to the word "romantic" were extracted, the proposed method extracted "romance", "love", "sweet", "beautiful", and "relationship", whereas LSA extracted "romance", "screwball", "grant", "comedies", and "comedy". The performance of the proposed algorithm was tested against linear SVM on bag of words features, latent dirichlet allocation (LDA), and LSA, using the widely used IMDB data of movie reviews. Document-level sentiment polarity classification was conducted, and the proposed method outperformed other vector space models (VSMs). It is meaningful that the performance is improved only by simple sentiment information, so the scope of application is wide.

Naïve Bayes (NB) and SVM, commonly used as the baseline for classifying text, were in some cases found to perform better than state-of-the-art models, which showed that complex models do not always perform well (S. Wang and Manning (2012)). Wang also found that the bigram features have more consistent gains than the unigram features. The study used various dataset

widely used for data analysis such as RT-s(Pang & Lee, 2005), CR(Hu & Liu, 2004), MPQA(Wiebe et al., 2005), Subj(Pang & Lee, 2004), RT-2k (Pang & Lee, 2004), and IMDB(Maas et al., 2011). The performance of the algorithms proposed in the latest researches were tested using the snippet dataset. Algorithms considered included: Recursive Autoencoder (RAE)(Socher et al., 2011), namely, RAE-pretrain (Collobert & Weston, 2008) which trains on Wikipedia, and Voting and RulesNakagawa et al. (2010), which use sentiment lexicon and hard-coded reversal rules. Experiment results showed that MNB with bigram feature and NBSVM tend to outperform other models. In addition, the performance of full-length reviews was compared with the result of BoW and LDA from Maas et al. (2011), the result of tf.Δidf(Martineau & Finin, 2009), and word presentation restricted boltzmann machine(Dahl et al., 2012). The result showed that NBSVM achieved 91.22% performance in the IMDB classification problem and showed the best performance in other full-length datasets. Furthermore, it provides robust results for snippets and full-length text.

Socher et al. (2013) proposed Recursive Neural Tensor Network(RNTN) as a detection algorithm that captures the compositional effects of text more accurately, given that many semantic vector space methods can't represent the meaning of a longer phrase. For a recursive neural network (RNN), parent vectors are calculated as all their children vectors. For example, if a is a parent vector of b and c, a can be calculated as $a = f(W\left[\begin{smallmatrix} b \\ c \end{smallmatrix}\right])$, where $f = \tanh$ is a standard element-wise nonlinearity, $W \in \mathbb{R}^{d \times 2d}$ is the main parameter to learn. In the RNN, the input vectors only interact with the nonlinearity function, while the RNTN adds more powerful interaction information between the

input vectors. In other words, RNTN was able to better express compose aggregate meaning by adding a tensor-based composition function when calculating the parent vector. Applying the above example to RNTN, the parent vector is equal to $a = f(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:d]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix})$, where $V^{[1:d]} \in \mathbb{R}^{2d \times 2d \times d}$ is the tensor that defines multiple bilinear forms. In this equation, each slice $V^{[i]}$ of $V^{[1:d]}$ can be interpreted as the reflection of a specific type of composition. This way, each word can be represented as a vector. The data used the Movie review corpus and consisted of two types of output variables. The first was labeling with five sentiments: negative, somewhat negative, neutral, positive or somewhat positive, and the second labelling with two classes: positive vs. negative. They compared the performance of the proposed method to various algorithms such as NB and SVM with bag of words features, NB with bigram features, averages neural word vectors (VecAvg), RNN, and Matrix-Vector RNN (MV-RNN) (Socher et al., 2012), which represents every word and longer phrase in a parse tree as both a vector and a matrix. As a results, RNTN was found to outperform all other algorithms. In the five-class classification problem, the accuracy rate reached 80.7% for all nodes and 45.7% at the sentence-level. For the two-class classification task, the accuracy of all nodes was 87.6% and that of the sentence-level was 85.4%. The sentiment accuracies for the binary classification for single sentence have not exceeded 80% for several years, but this study is significant in that RNTN achieved 85.4% accuracy. In particular, the sarcastic sentiment polarity expressed in the form of contrastive conjunction such as 'X but Y' or in the form of negating positive or negative sentences such as 'less negative' were identified.

So far, methodologies for achieving high performance using various sources of data have been proposed, and many studies have also been conducted incorporating textual data. However, results have lacked intuitive interpretations that can be directly applied to the active trading management side of the industry, provided in a form that can be of practical assistance to the decision-making process. In this dissertation, we will focus on reporting the results that is accompanied by prediction performance and effective interpretation.

Chapter 3

Prediction of Stock Price through Sentiment Analysis of Corporate Disclosures

3.1 Background

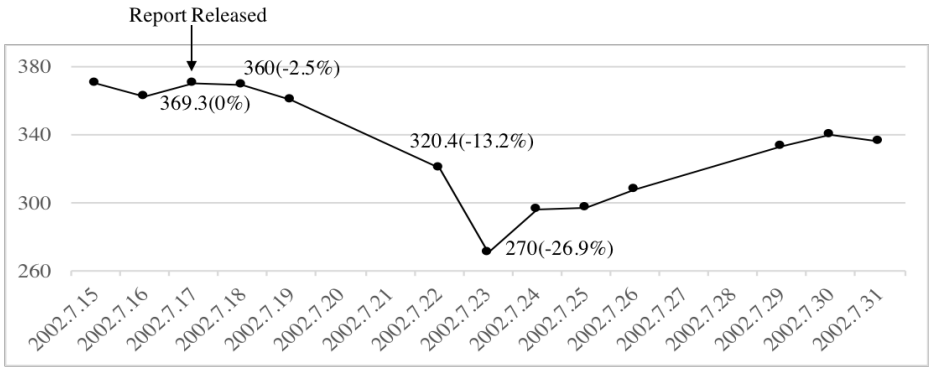
The study of the prediction of stock prices has been one of the major branches in the field of financial research. Previous studies relied on mathematical models to predict stock prices using quantitative variables such as historical time-series of stock prices(Kumar et al., 2004; Ahn et al., 2009), as well as microeconomic and macroeconomics indicators (Ghosn & Bengio, 1997). More recent studies began to incorporate textual data, such as comprehensive reports and financial columns, in the analysis of market behaviors. However, same words may convey different sentiments in different sectors, and performance can potentially degrade if the prediction model learns documents from multiple sectors simultaneously. For example, words such as war, attack, and terrorism are generally associated with negative sentiments, while they may positive implications for firms in military supply or military intelligence industry. Therefore, it is necessary to design a model that can accurately account for different industry effects

when considering textual information in solving the prediction tasks.

The objective of this paper is to predict stock price movements in both quantitative and qualitative ways by analyzing the sentiments of 8-K financial reports based on the means of distributed representation. For example, consider the excerpts from 8-K reports by Citigroup Inc. and JP Morgan Chase & Co and their respective stock price movements presented in panels (a) and (b) of Figure 3.1. Citigroup's 8-K report, presented in panel (a) was released on July 17, 2002; JP Morgan's, presented in panel (b), was released on Jan 15, 2010. Words highlighted in blue are associated with negative sentiments. The values in parentheses show the percentage decline in stock prices following the 8-K announcement date. Figure 3.1 shows that the stock price has fallen by 26.9% in four business days after the announcement for Citigroup, while it drops by 7.2% within three business days after the announcement for JP Morgan. Observations from Figure 3.1 indicates that the sentiment of financial report announcements may be a great tool to explain the subsequent stock price movements. Our goal is to automate such a process and produce meaningful results by employing distributed representation method.

In this study, we propose a prediction model that uses the 8-K financial reports as the input and stock price movements as the output. Distributed representation expresses documents as vectors, which are then used to embed documents along with class information on the same space. This enables calculation of distances between documents and sentiment classes, hence allowing identification of the sentiment class of given documents. Moreover, since documents are now in the form of vectors, one can easily visualize the sentiment class

(a) *"continues to be **depressed** by the ongoing **economic crisis in Argentina** as well as by a **slowdown** in Brazil precipitated by the upcoming elections in that country. Proprietary investment activities and Corporate/Other Citigroup's proprietary investment activities recorded a **loss of $ 190 million** in the second quarter, which included a $ 132 million **after tax loss** in the life and annuities investment portfolios due to the **impairment of WorldCom securities**, as well as **lower market values** in the company's publicly - traded proprietary investment portfolio, as major market indexes **declined sharply**."*



(b) *"The provision for **credit losses** was $4.2 billion, **an increase of $653 million** from the prior year and $241 million from the prior quarter. **Weak economic conditions** and **housing price declines** continued to drive **higher estimated losses** for the mortgage and home equity portfolios. The provision included an addition of $1.5 billion to the allowance for **loan losses**, compared with additions of $1.9 billion in the prior year and $1.4 billion in the prior quarter. Included in the $1.5 billion addition to the allowance was a $491 million increase in the fourth quarter of 2009 related to estimated **deterioration in the Washington Mutual** purchased **credit-impaired portfolio**;"*

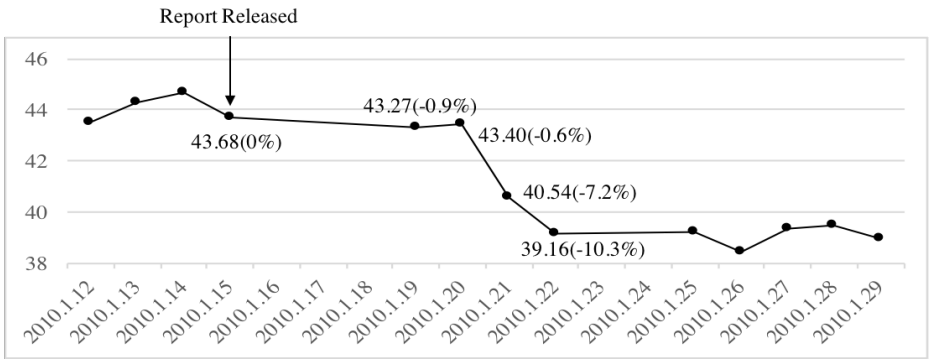


Figure 3.1: (a) 8-K announcement of Citigroup Inc. and subsequent stock price movement (b) 8-K announcement of JP Morgan Chase & Co and subsequent stock price movement

of the document after some dimension reduction process. A data flow diagram of the methodology is shown in Figure 3.2.

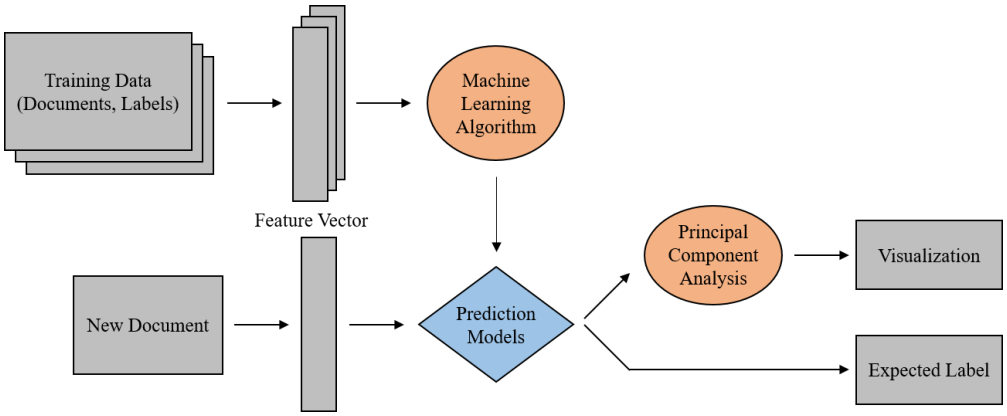


Figure 3.2: Data flow diagram

Distributed representation method incorporated in this study does not only improve predictability of the model as compared to the conventional one-hot encoding approach but also provide visualization of the prediction results, hence adding interpretability. Improvement in predictability of the model, as well as the visualizations produced as the result of the analysis, can then be used to help active traders in the field when making time-split trading decisions or data-driven detection of promising companies. On the other hand, the visualization produced by the model may directly benefit the decision-making agents of the industry by providing intuitive illustration of the relationship between the sentiment and the stock price movement. Because our visualization results place stock price movements side-by-side to the sentiment trend of documents reporting about the respective company, which help enhance the understanding of the readers and assist with their decision-making by providing intuitive

insights.

The remainder of this chapter is structured as follows: We introduce proposed method of this study in Section 3.2. Section 3.3 describes experiment settings and reports results. Finally, we conclude this chapter in Section 3.4.

3.2 Proposed Method

This study predicts stock price movements in two ways: model-based and visualization -based. The framework of our methodology is shown in Figure 3.3.

Model-based analysis calculates the sentiment of 8-K financial reports via distributed representation method (Park, 2016) to predict stock price changes, while the visualization-based analysis provide qualitative assessment of the prediction results.

3.2.1 Distributed Representation

In order to form the stock price prediction problem into a classification model with text as input, it is important to appropriately and clearly represent documents for the task in hand. Textual data such as financial reports and news is represented using the word2vec method described in Section 2.3. One of the extensions from word2vec is the paragraph vector(PV) (Le & Mikolov, 2014). The PV model adds a paragraph ID when sliding through each word of the corresponding paragraph, which allows learning word and document vectors simultaneously in the same embedding space. Graphical illustration of PV model is presented in Figure 3.4.

In the Figure 3.4, D is a PV matrix, which is expressed as a set of unique

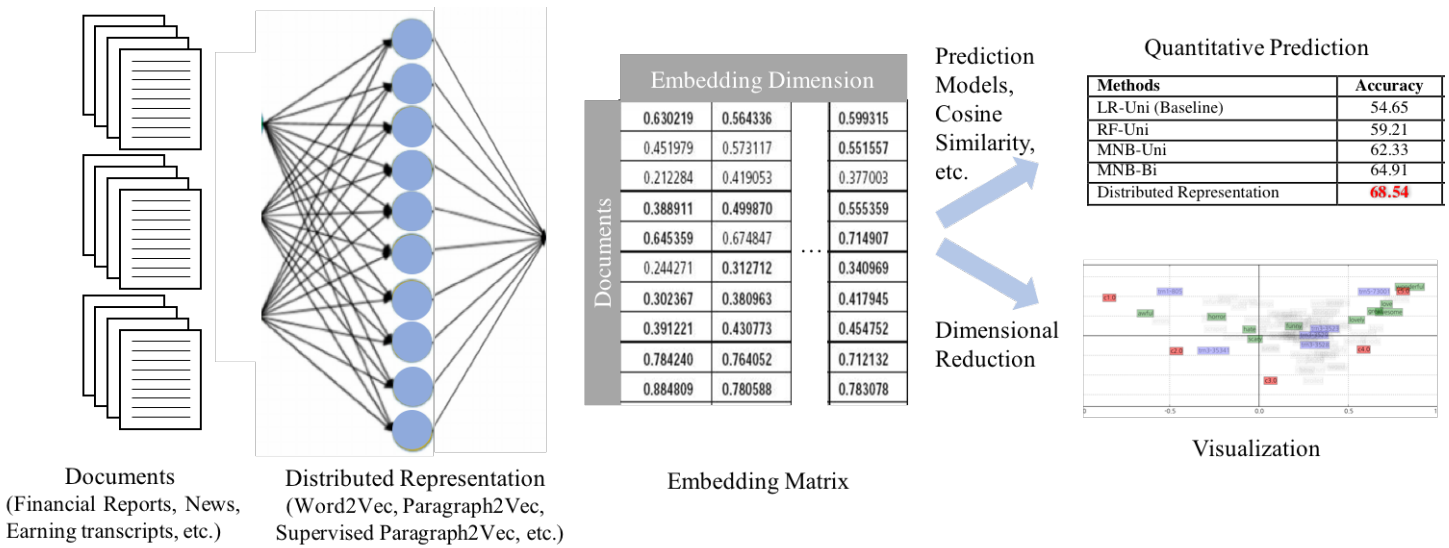


Figure 3.3: Diagram of stock price change prediction

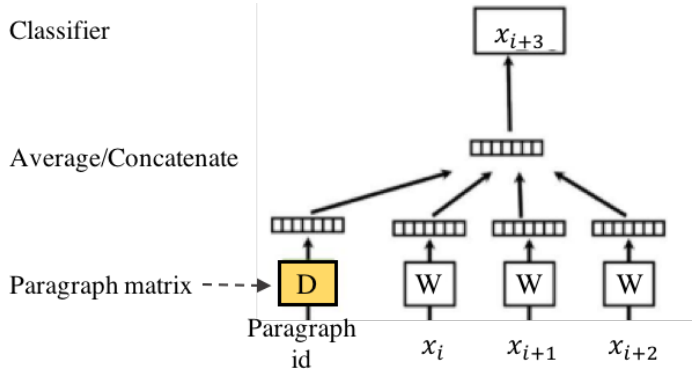


Figure 3.4: Distributed model using paragraph vectors (Le & Mikolov, 2014)

vector for each paragraph, and W is a unique vector mapping of words. PV is inserted as an input vector just as a word vector is. PV provides information about current context of words belonging to the same paragraph; therefore, it may be thought of as a memory model in some sense. PV allows consideration of word orders as the bag-of-n-gram models do but with a much smaller dimension.

On top of the PV vector, we enrich the model by adding the class vectors when learning the document. This is because PV model, without the class vector, may produce results that may not be suitable for sentiment analysis of financial documents. Since word2vec slides through the input text word-by-word, when the set of context words are similar, the target words will end up being placed very closely to each other on the embedding space even if they may convey very different sentiments. For example, suppose there are two documents evaluating the operational performance of two different firms, Firm A and Firm B, namely. Document 1 evaluates Firm A well by stating: "... showed good performance." On the other hand, Firm B receives a harsher

remark, and Document 2 has reported: "... showed bad performance." In both cases, the context words surrounding the target words are exactly the same, "showed" and "performance"; hence, word2vec will place "good" and "bad" at the same position on the embedding space when the learning is complete. This is not appropriate, especially for sentiment analysis, since, "good" and "bad" are clearly the exact opposites of each other in terms of sentiments. In this study, we address this issue by setting up a supervised learning framework for PV model by enforcing the model to learn class information simultaneously with PV as input variables.

Class information is the sentiment labeling of documents using the direction of the closing price movement the next business day. Each document is assigned to one of the two labels ("UP" and "DOWN") as its class depending on the stock price shift the next business day. "UP" means that the price went up more than the criterion, while "DOWN" means that the price went down more than the criterion. Because stock price movements are already known at the time of learning, hence we design our model as supervised learning. The main difference of Supervised PV (SPV) models versus PV models is that the class label information of documents is added on to the input vector at the time of training. The objective function of the SPV model is defined as:

$$\frac{1}{T}\sum_{t=1}^{T}\{\sum_{-r\leq j\leq r, j\neq 0}\log p(w_t|w_{t+j}) + \log p(w_t|d) + \log p(w_t|c)\} \tag{3.1}$$

where c indicates the class of the document, d is the document and all other notations are equal to those of the skip-gram model. Above framework can be considered as a form of SPV model introduced in (Park, 2016) applied in the financial setting.

3.2.2 Visualization

In this study, we visualize the result of sentiment analysis via distributed representation. By expressing a document with a distributed representation, words, sentences, documents and class information in the text document can be represented by the vectors of same dimension. It is, however, very difficult to visualize vectors with large dimensions. There exists a number of dimensionality reduction techniques for visualization purposes, such as LDA, t-Distributed Stochastic Neighbor Embedding (t-SNE), and Principal component analysis (PCA). LDA calculates a linear combination of variables to categorize them into two or more groups, and its performance is reported to be fairly good. However, because it is a supervised method, and its applications are limited (Fisher, 1936). t-SNE is one of the newer methods, which reduces dimension by maintaining the relative distance between observations based on the non-linear relationship in between. Nonetheless, it is a non-linear method, hence it takes a long time to compute (Maaten & Hinton, 2008). PCA, which is similar to LDA but unsupervised, is a multivariate technique that analyzes data in which observations are described by several inter-correlated dependent variables. PCA extracts important information from the data in the form of orthogonal variables called principal components (Abdi & Williams, 2010). PCA has a mathematical rela-

tionship to other popular machine learning methods such as k-means clustering and factor analysis, while being simple. For such reasons, this study chooses PCA to reduce dimension and visualize vectors resulting from the supervised learning PV stage. We linearly map data to a lower-dimensional space using PCA and find two eigenvectors with the greatest variances. These eigenvectors are then used to visualize class information, words, and documents into two-dimensional space. Figure 3.5 presents an example of our approach for visualization.

In the above figure, we represent class information, namely 'c1.0', 'c2.0', ... ,'c5.0', in a red box, words in green, and documents in blue. 'c1.0' is a negative class and 'c5.0' is a positive class. 'trn5-73001' document is located close to class 5 and appears as a positive document, and words such as 'love', 'wonderful', 'great', and 'awesome' are classified as positive words. In this way, we will visualize the sentiment of documents by embedding the word, document and class into the same space.

3.2.3 Model-based Prediction

The framework of model-based prediction is outlined in Figure 3.6.

An input document is parsed to extract unigram and bigram features, and the document term matrix is created using the term frequency (TF) and term frequency-inverse document frequency (TFIDF). We use six prediction models: logistic regression (LR), random forest (RF), multinomial Naïve Bayes (MNB), support vector machine (SVM), Naïve Bayes SVM (NBSVM), and supervised

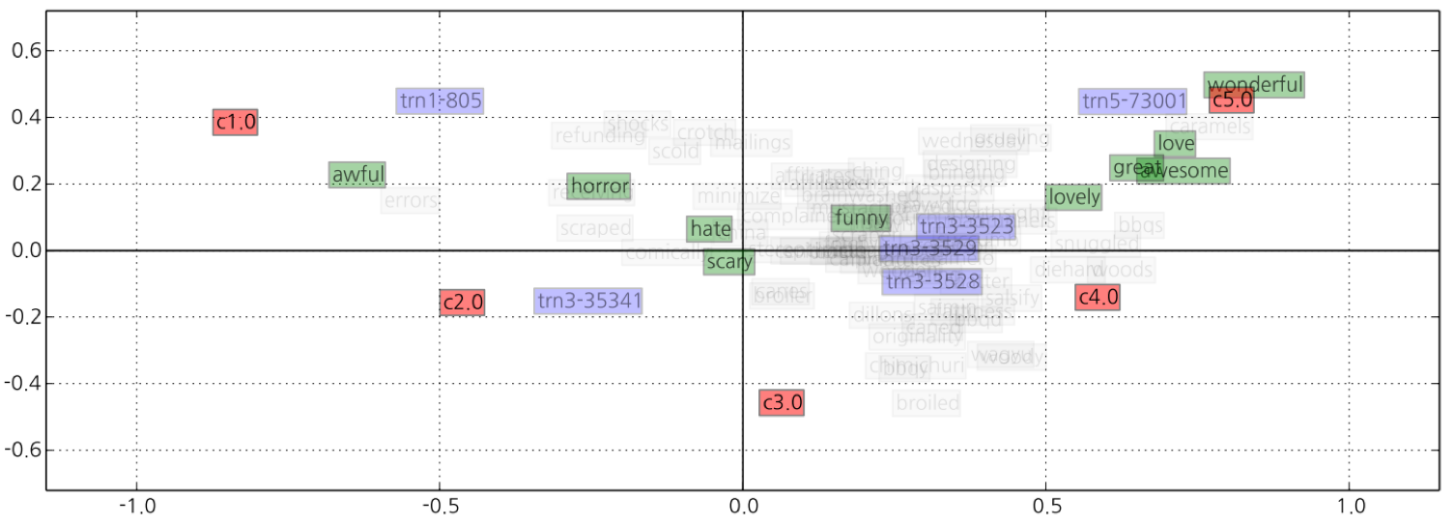


Figure 3.5: Example of visualization(Park, 2016)

Figure 3.6: The framework of model-based prediction

PV (SPV) as introduced in Subsection 3.2.1. LR and RF serve as baseline models, and they use unigram features only as input variables. A detailed description of each algorithm is given in Section 2.3.

3.3 Experimental Results

3.3.1 Data Descriptions

We use 8-K financial reports[1] as the primary data source. 8-K financial reports are reports of unscheduled material events or corporate changes at a company that could be of importance to the shareholders or the SEC (Lee et al., 2014). Data contains company ID, time of report, and relevant business events such as bankruptcies, layoffs, the election of a director, a change in credit, etc. and main contents. For the sentiment analysis, we collected 8-Ks from 2002 to 2012 for the four companies in financial sector, as listed in Table 3.1.

Preprocessing included removing stopwords and changing all numbers in various meaning to the word 'num'. We gather company's daily stock prices from Yahoo! Finance[2] and use them as the target variable. "UP" and "DOWN" classes are used as the output variable, calculated by taking the difference in

[1]https://www.sec.gov/edgar.shtml
[2]http://finance.yahoo.com/

Table 3.1: Company lists

Ticker Symbol	Company Name	Number of Doc
C	Citigroup Inc	513
WFC	Wells Fargo & Co	427
GS	Goldman Sachs Group Inc	257
JPM	JP Morgan Chase & Co	835
Number of Total Document		2,032

the company's stock price before and after the report is released. It is assumed that the news announced in the middle of the day affects the stock price of the next day. We used the closing price of the date of the news announcement as the stock price before the report is released and the open price of the next day as the stock price after the report is released (Lee et al., 2014). We normalize this difference by subtracting the difference of S&P500 index for the same period to remove the effect of market conditions (bull or bear) from the influence of news on stock movements. The equation is:

$$\Delta = \frac{SP_{T+1} - SP_T}{SP_T} - \frac{S\&P500_{T+1} - S\&P500_T}{S\&P500_T} \tag{3.2}$$

where Δ is the normalized difference, T is the announcement date of financial report and SP indicates the individual stock price. We used the closing price at time T and the opening price at time T+1. For instance, if company's stock price rises 2% and S&P500 index goes up 1% after event, the normalized difference equals 1%. We set the criteria at 1% and assigned to "UP" class if the difference is greater than the criteria, otherwise assigned to "DOWN" class.

3.3.2 Experimental Settings

We set the parameters the same as used by (S. Wang & Manning, 2012) for direct comparison of results to his work. The full list of parameters for each algorithm are reported in Table 3.2.

Table 3.2: Parameters

Algorithm	Parameters	Values
Support Vector Machine	tradeoff	0.1
Naïve Bayes SVM	tradeoff	1
	α	1
	β	0.25

Tradeoff is the tradeoff between the training errors and the model complexity; α is the smoothing parameter, and β is the interpolation parameter. We use ten-fold cross-validation for performance evaluation.

3.3.3 Quantitative Prediction

Based on the assumption that the day after the announcement of a financial report will have the greatest impact on the company's stock price, we applied all algorithms mentioned in Subsection 3.2.2. The results are shown in Table 3.3.

We used two types of input variables extracted from unigram and bigram models and set the output variable as the stock price change after 1 day of the announcement. 'Uni' and 'Bi' are abbreviations of the input variable used in the prediction models with the unigram or bigram features, respectively. Using

Table 3.3: Prediction accuracy of each algorithm

Method	Accuracy	Method	Accuracy
Unigram(LR, Baseline)	54.65	SVM-Uni	63.41
Unigram(RF)	59.21	SVM-Bi	63.80
MNB-Uni	62.33	NBSVM-Uni	62.87
MNB-Bi	64.91	NBSVM-Bi	65.67
SPV model	68.54		

unigram features as input variables, results show that RF performs better than LR. SVM outperforms MNB, and NBSVM has been found to improve SVM results. When using bigram features, models tend to perform better in general as compared to using unigram features. In case of the SPV model, the improvement over the LR baseline amounts to 25.4%.

In order to compare the performance of prediction models, we conducted an independent 2-sample t-test assuming unknown standard deviation. The test was run 28 times for 28 pairs of combinations among 8 different models. A null hypothesis, $H_0 : \mu_A = \mu_B$, was rejected if the difference in means of the pair are statistically significant. We calculated the probability of rejecting the null hypothesis, and the results showed 96.4% probability of rejecting the null hypothesis on significance level 0.05. That is, except for one, all combinations rejected the null hypothesis.

3.3.4 Qualitative Prediction

Unlike other areas, documents in financial markets are not independent of each other and affect stock prices for a certain period of time. Based on this property, we visualize the sentiment of 8-K report over time and confirmed that the

relationship between the sentiments and the stock price movements makes sense. The sentiment of the published documents and the stock prices of Wells Fargo Company from 2008 to 2009 is shown in Figure 3.7. During this time period, stock prices of Wells Fargo were volatile due to financial crisis. Documents from the same time period were split every two months due to the limit number of documents.

Panel (a) of Figure 3.7 exhibits a graph representing the sentiment of documents every two months, whereas (b) plots the 10-day moving average of the stock prices. As the sentiment of documents changes, stock price moves accordingly to the sentiment trend. For example, in panel (a), the sentiment of the document in March-April 2008 grew relatively more negative, which is reflected in the stock price is affected from the end of April, 2008 where it fell moderately steeply, as plotted in panel (b). In addition, the sentiment of the documents from October 2008 to February 2009 is negative, and it can be seen that stock price has been steadily decreasing since December 2008. Above illustration shows that the sentiment of the documents is closely related to the stock price and that, since the direction of the stock price appears after the document announcement, the sentiment of documents can be used as the leading indicator. Furthermore, plots like Figure 3.7 can serve as qualitative evaluation of the model, providing the practitioners with more interpretable insights.

We plot the sentiment trend of documents over time for the four selected companies as listed in Table 3.1. Furthermore, we compare the sentiment trend with stock price movements side-by-side for each selected company from 2002 to 2012. The resulting visualization is shown in Figure 3.8.

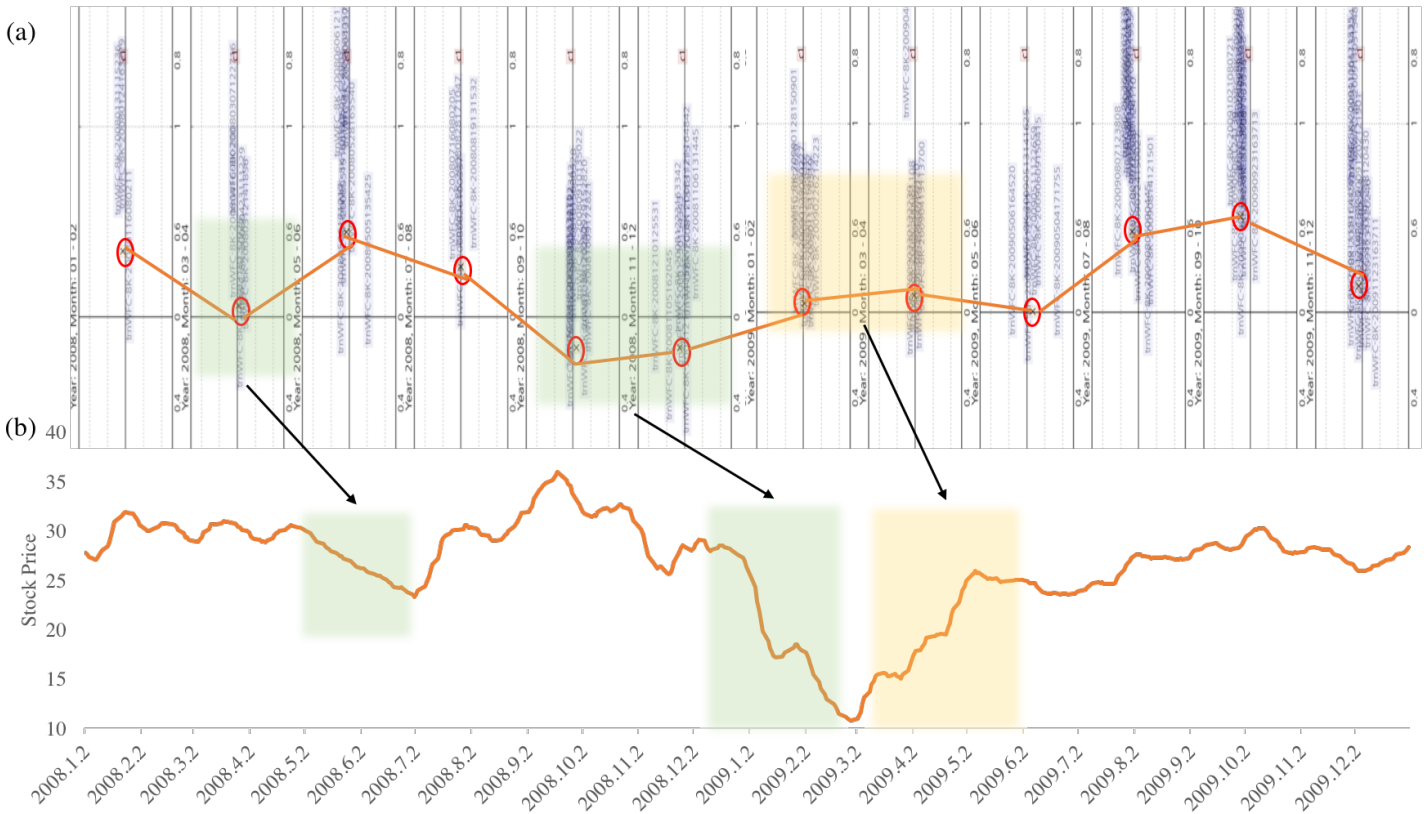


Figure 3.7: Sentiment of 8-K reports and stock price for Wells Fargo Company

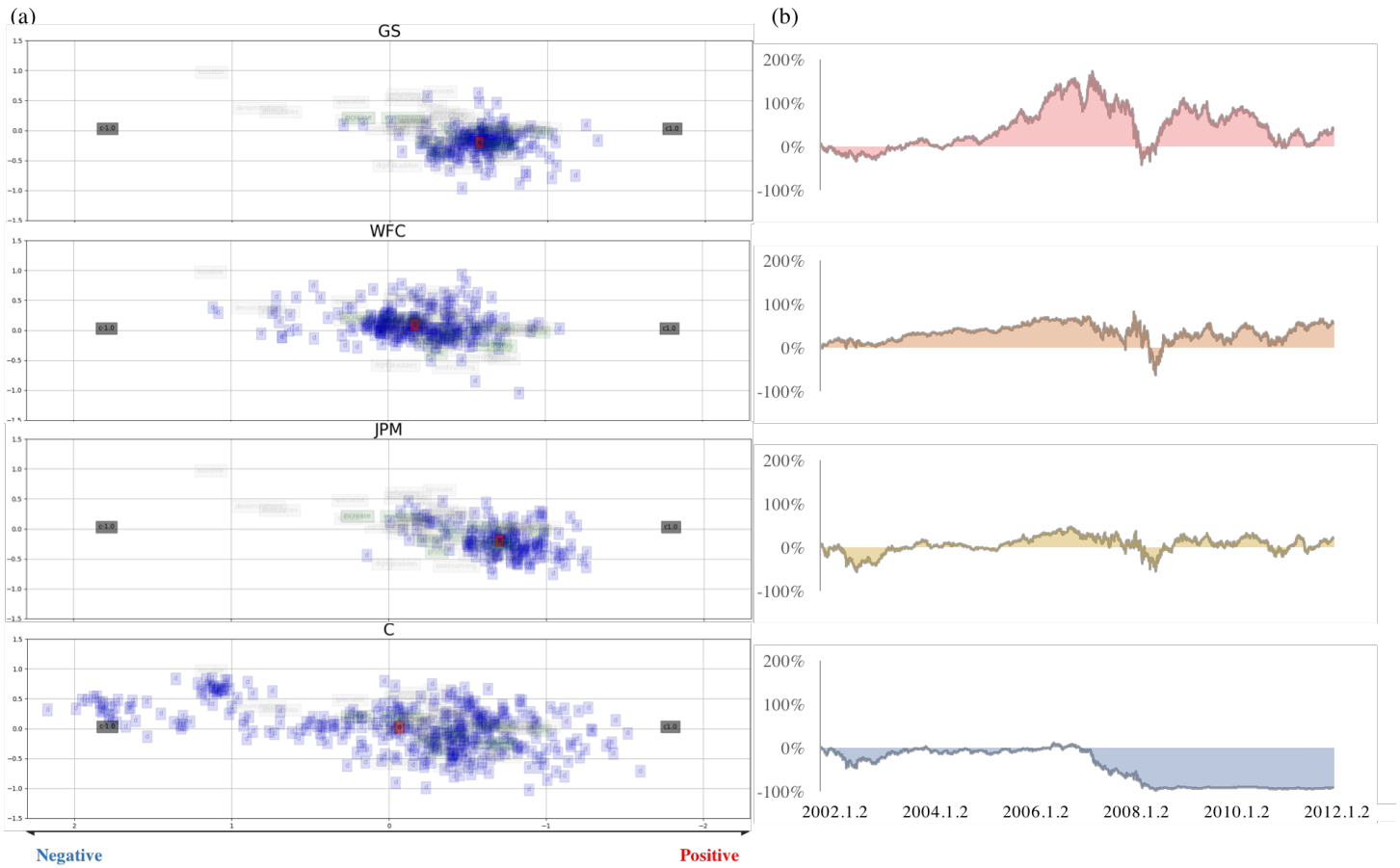


Figure 3.8: The sentiment of entire documents(left) and stock price trend(right) for each company

Panel (a) of Figure 3.8 exhibits a graph representing the sentiment of documents belonging to each company as considered in analysis. The more to the right the documents are distributed, the more positive the document is. The center of the documents in space is represented by a red dot. Whereas plots in the panel (b) represent the volatility in stock prices as compared to the first business day in 2002. For example, if the y-value is 30%, it means that the stock price has risen 30% compared to the first business day in 2002. As shown in the lots on panel (a) of the figure, the sentiment trend is different for each company. There is a handful of documents with neutral or positive sentiment for Goldman Sachs, Wells Fargo and Company, and JP Morgan, while lots of documents of Citigroup Inc. exhibit relatively negative sentiment. In the meantime, Goldman Sachs shows the steepest positive volatility of stock price as presented in panel (b), while and Citigroup displays negative volatility. Observations from Figure 3.8 indicates that Citigroup had a number of issues that had greatly reduced stock prices since 2002, and these issues appeared to have been associated with negative sentiment conveyed by the financial reports. We represent the negative index of sentiment and volatility of stock price as shown in Figure 3.9.

In the above graph, x-axis is the negative sentiment index and y-axis is the negative stock price index. Both indices are calculated as follows:

$$Negative\ Sentiment\ Index = \frac{\sum_{i \in \mathbf{N}} |d_i|}{\sum_{j \in \mathbf{T}} |d_j|} \tag{3.3}$$

where $\mathbf{N}$ is the set of negative sentiment documents, $\mathbf{T}$ is the set of total

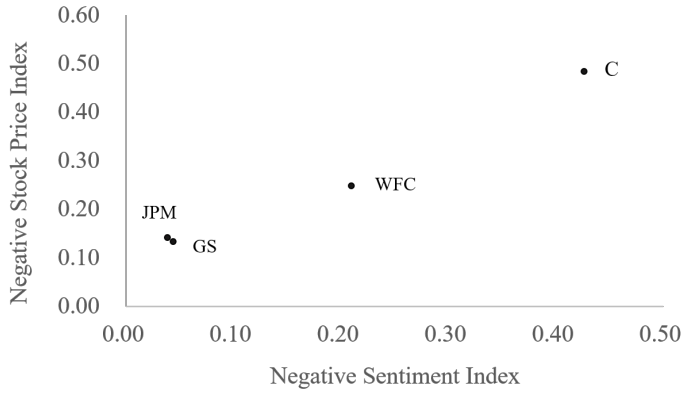


Figure 3.9: Correlation between the negative sentiment index and the negative stock price index

documents, and d is the distance from 0.

$$Negative\ Stock\ Price\ Index = \sum_{i \in \mathbf{V}_N} |v_i| \tag{3.4}$$

where $\mathbf{V}_N$ is the set of negative volatility of stock price and v is the volatility in stock prices as compared to the first business day in 2002. The correlation between the two indices is 0.9918, which means that the negative sentiment index and the negative stock price index are positively correlated. Therefore, the polarity of the report sentiment is consistent with the movements of the stock price trend.

3.4 Summary

This chapter predicts the direction of stock price changes using 8-K financial reports for the four selected companies in the financial sector. We propose

two methods to solve the prediction task: the model-based method and the visualization-based method. For the model-based stock price prediction, unigram features were extracted from financial reports and applied to the LR and RF models, and the results were used as baselines. The same unigram and bigram features were applied to MNB, SVM, and NBSVM models for comparison. Finally, we predicted stock price change prediction using distributed representations. Experiment results show that distributed representation produced most accurate predictions, and the improvement in prediction accuracy over the baseline model amounts to an impressive 24.5%. On the other hand, we visualize the sentiment of the document by projected the class information and the document on to the space of the same dimension. The benefits of this visualization helps one easily understand the sentiment changes in financial documents of a selected company, while providing rich illustration of the relationship between sentiment trends and stock price trend movements. Visualization results confirmed that when the sentiment of documents was positive, the stock price movement showed an upward trend. On the other hand, when the sentiment of the document was negative, the stock price fell. The mean values of sentiment were calculated to create a sentiment index for the entire document. Visualization results showed that different sentiment trends of documents for selected companies were well reflected in the stock price movements of the corresponding companies. For GS firms, the sentiment of the entire documents was positive, of which the trend was consistent with the stock price moving mostly in the positive direction. On the other hand, in the case of company C, there were many documents with negative sentiment, and the stock

price had continuously decreased since 2002.

Our proposed model does not only improve accuracy, but it also provides interpretability by producing visualizations that can show the sentiment trends of associated documents. It allows a trader to visualize the sentiment of documents of the company of interest, to view the trend of the sentiment at a glance, which help the active traders and decision-making agents in the field to make more data-informed decisions. In the financial market, traders are often required to make split-second decisions, and our proposed model can potentially provide them with opportunities to monitor new investment companies objectively and detect companies at crisis.

In this study, we analyze only a few companies in the finance sector, but in the future we expect to gain more insights by analyzing a wider range of companies. In addition, by analyzing public documents and private document such as SNS and online blogs through distributed representation, it may be possible to extract sentiment words that are specific to the financial sector using our method to build the sentiment dictionary specifically designed for the financial domain. Finally, since visualizing the sentiment of documents is a powerful tool, research is needed on more effective visualization tools.

Chapter 4

Predicting the Korean Monetary Policy Committee's Vote Results with Monetary Policy Decision Text

4.1 Background

The Monetary Policy Committee(MPC) of the Bank of Korea(BOK) applies a look-at-everything approach, and determines interest rates by taking into consideration inflation trends, domestic and overseas economic and financial market conditions and so on. It is called the Base Rate. The MPC of BOK announced the Base Rate on the second Thursday of each month until 2016 unless there were special circumstances, but from 2017 it is reduced to eight times a year to determine the Base Rate, as with the FOMC, and the date of the meeting is announced in advance. In this way, the central bank of each country sets the Base Rate as a benchmark for various rates. In other words, the Base Rate has a prompt impact on the call rate, and this leads to changes in short and long term market rates and deposit and loan rates, thus ultimately influencing economic activities. For example, if the Base Rate is raised, short term market rate such as call rate rise immediately, and then deposits and loan

rates rise mostly as well as the long term market rates rise. The rise of various rates will not only reduce household consumption due to increased interests in deposits and loans, but also lead to a reduction in investment because of increased financial costs for companies. The Base Rate is usually set at a value of 3%, 2.75%, or 2.5% depending on the increase or decrease of 25 basis points (bp). The MPC consists of seven people: six MPC members and the governor of BOK. The MPC decides on the base rate through a majority vote. In the event of a tie, the governor of BOK has the casting vote. The base rate decisions are categorized as follows: a rate hike (RISE), a frozen rate (FREEZE), and a rate cut (FALL). Each base rate decision is reached by a unanimous vote, five votes, four votes, or three votes in accordance with the majority rule. The distribution of votes for each base rate decision from May 1999 to July 2017 is shown in Figure 4.1.

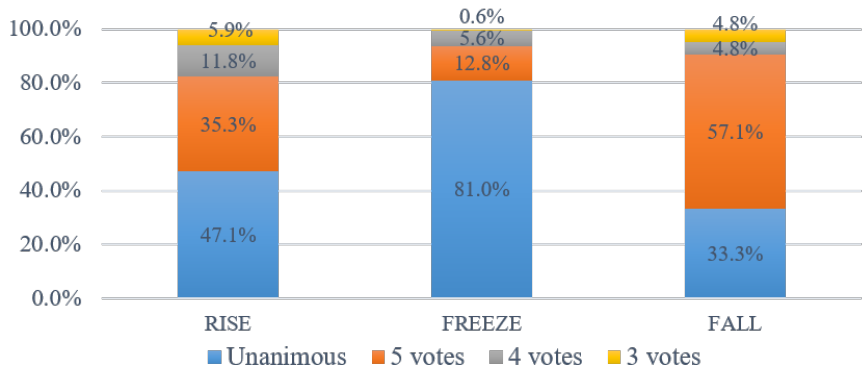


Figure 4.1: Distribution of the vote results, May 1999 to July 2017

Figure 4.1 shows that the distribution of the majority results for each in-

terest rate decision differs. When the base rate decision is RISE or FALL, the proportion of one or more members opposed to the decision is relatively large. Moreover, future base rates change direction in accordance with the results of prior votes. For example, even though the decision about the base rate is FREEZE, future base rate trends differ between unanimous decisions and one or more FALL or RISE votes (see Figure 4.2).

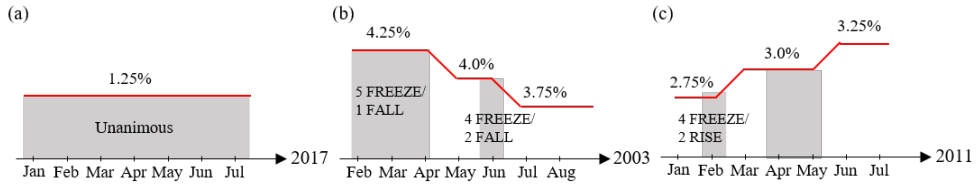
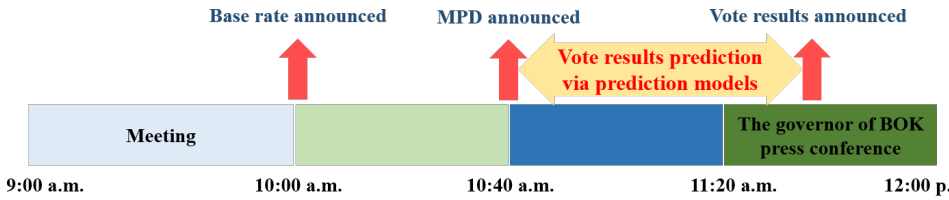


Figure 4.2: The vote results and future base rate trends

Figure 4.2 (a) shows that from January to July 2017, the base rate does not fluctuate after FREEZE receives a unanimous vote. However, Figure 4.2 (b) shows that although the base rate decision is FREEZE from February to April 2003, one member of the MPC continues to insist on a base rate cut. Subsequently, a 25 bp rate cut occurs in May 2003. In June, the base rate decision is FREEZE; however, two members of the MPC insist on a rate cut. Following this, in the next month, an additional interest rate cut occurs. In Figure 4.2 (c), two members insist on a rate hike in February 2011; however, the base rate decision is FREEZE, although thereafter, in March 2011, a 25 bp rate hike occurs. Then, a four:two vote in favor of FREEZE from April to May 2011 leads to a 25 bp rate hike the following month. Most market participants

are confident that a strong correlation exists between the MPC's vote results and future base rate trends. This confidence leads to real transactions that directly affect the market, such as bond price fluctuations.

The MPC's vote result is announced in the middle of a press conference held by the governor of BOK after the MPC meeting. The schedule of an MPC meeting for monetary policy decision-making is shown in Figure 4.3.



Note: MPD refers to monetary policy decision.

Figure 4.3: Schedule of the MPC's meeting for monetary policy decision-making

The meeting usually starts at 9 a.m. The base rate is announced at approximately 10 a.m. Then, the monetary policy decision (MPD), the statement for which is written during the meeting, is disclosed at 10:40 a.m. on BOK's website.[1] From 11:20 a.m., the governor of BOK holds a press conference on monetary policy. During the conference, the governor explains the detailed background to the committee's base rate decision, provides any new comments on monetary policy, and announces the vote result. The important point to make here is that market volatility increases, as shown in Figure 4.4, depending on the vote result and the governor's stated economic outlook.

Figure 4.4 shows an example of the effect of vote unanimity on 10-year bond

[1] http://www.bok.or.kr/broadcast.action?menuNaviId=89

Figure 4.4: Example of the increase in market volatility of 10-year bond futures' prices (April 19, 2016) following the announcement of a base rate decision

futures' prices. At the April 19, 2016 meeting, one member of the MPC insisted on a base rate cut although the decision was FREEZE. Right at the moment when this was announced through the press conference following the meeting, 10-year bond futures' prices rose sharply due to the possibility of interest rate cut.

This study's main purpose is to analyze the MPD that is announced before the press conference in order to determine whether the vote result is predictable before the press conference. Because the MPD text is the only information that is available to analyze the opinion of each MPC member before the press conference, this study uses the MPD's wording to estimate the vote distribution, as shown in Figure 4.5.

Figure 4.5 (a) and (b) are part of the MPD for a base rate FREEZE. In

(a) 4월중 소비자물가 상승률은 석유류 제외 공업제품가격의 상승폭이 확대되었으나 서비스요금의 오름세가 둔화됨에 따라 전월과 같은 1.0%를 나타내었다. 농산물 및 석유류 제외 근원인플레이션율은 전월의 1.7%에서 1.8%로 소폭 상승하였다. 앞으로 소비자물가 상승률은 저유가의 영향 등으로 낮은 수준을 이어갈 것으로 보인다. 주택매매가격은 전월 수준을 유지하였으며 전세가격은 낮은 오름세를 나타내었다.

(b) 3월중 소비자물가 상승률은 석유류 가격 하락폭 확대 등으로 전월의 1.3%에서 1.0%로 낮아졌다. 농산물 및 석유류 제외 근원인플레이션율도 전월의 1.8%에서 1.7%로 소폭 하락하였다. 소비자물가 상승률은 저유가의 영향 등으로 당분간 물가안정목표 2%를 상당폭 하회할 것으로 전망된다. 주택매매가격은 전월 수준을 유지하였으며 전세가격의 오름세는 둔화되었다.

(In translation)

(a) "Despite expansions in the extents of increase in prices of industrial products other than petroleum, consumer price inflation registered 1.0% in April, the same as in March, in line mainly with a slowdown in the rate of service fee increase. Core inflation excluding agricultural and petroleum product prices rose slightly to 1.8%, from 1.7% in March. Looking ahead the Board forecasts that consumer price inflation will continue at a low level, under the influence of the low oil prices for example. In the housing market, sales prices maintained their level of the previous month while leasehold deposit prices showed low rates of increase."

(b) "Consumer price inflation fell from 1.3% the month before to 1.0% in March, owing chiefly to increases in the extents of decline in petroleum product prices. Core inflation excluding agricultural and petroleum product prices also decreased slightly to 1.7%, from 1.8% in February. Looking ahead the Board forecasts that consumer price inflation will fall considerably short of the 2% inflation target for the time being, due mainly to the low oil prices. In the housing market, sales prices maintained their level of the previous month while the uptrend in leasehold deposit prices slowed."

Figure 4.5: (a) MPD with a vote of 6:0 in favor of FREEZE in May 2016; (b) MPD with a vote of 5:1 in favor of FREEZE in April 2016

Figure 4.5 (a), opinions about the Korean economy are relatively neutral based on the comments that consumer price inflation is the same as the prior month and core inflation, excluding agricultural and petroleum product prices, rose to 1.8% from 1.7% in the prior month. There is no mention of a base rate fall or rise; hence, the base rate in May 2016 was unanimously frozen. However, in Figure 4.5 (b), there is a sentence about the decline in consumer price inflation and core inflation, excluding agricultural and petroleum product prices, and an overall opinion that consumer price inflation is likely to fall considerably below the 2% inflation target for the time being; hence, there are generally negative views about the Korean economy. These views are reflected by one member of the committee who proposes a rate cut based on the negative economic outlook. In practice, traders consider the MPD text of the last three to five months and infer unanimous information by detecting changes in adjectives and nuances. Then, they strive to generate profits through transactions with financial instruments such as various bond futures.

In this chapter, we propose a prediction model that uses the MPD text as the input and the vote result as the output. If the MPC's vote result can be predicted before its announcement using the MPD text, traders will be able not only to generate additional profits through one-day volatility but also create competitive, long-term portfolio management based on future base rate direction. We first use a BoW model, based on a unigram and bigram model, to represent the documents; then, we use LR and a RF tool based on a decision tree to predict the sentence sentiment. The value obtained by aggregating this sentence sentiment can be seen as document sentiment. Accordingly, the

document sentiment is used to predict the vote result.

The rest of this study is organized as follows. Section 4.2 describes the data and proposed method, and Section 4.3 presents the experimental results. Finally, we discuss our conclusion and future related work in Section 4.4.

4.2 Proposed Method

4.2.1 Sentence Representation

This study uses the BoW model to map MPD text and the minutes of MPC meetings to sentence-level feature vectors. A sentence-term matrix is shown in Figure 4.6.

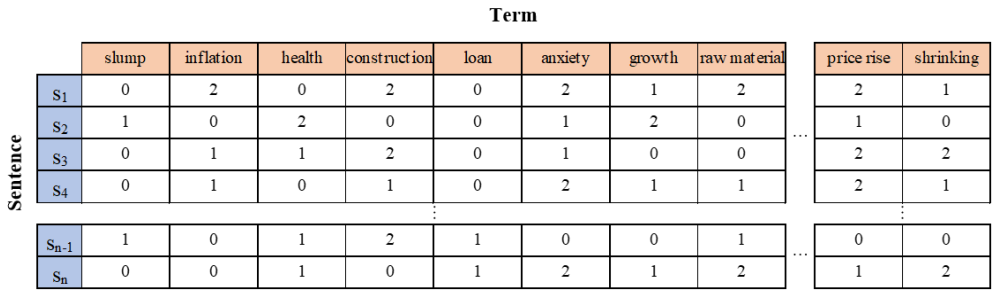


Figure 4.6: Sentence-term matrix with weights using TF

The total number of sentences that compose the matrix in Figure 4.6 is 18,606. The number of terms is 656 when only the unigram feature is used; when the bigram feature is used as well, the number of terms is 1,178. We used the unigram model and the bigram model because the bigram model may have a sparsity problem when expressing a sentence-term matrix because of the

number of documents.

4.2.2 Prediction Models of Sentence Sentiment

The sentence-term matrix constructed in 4.2.1 is used as the input for this study's prediction models. The base rate decisions of the MPC meetings, namely RISE, FREEZE, and FALL, are used as the output in order to predict sentence sentiment using LR and RF (see Figure 4.7).

Figure 4.7 is a framework for predicting sentence sentiment using each sentence of the MPD text in January 2013 as the input. Monetary policy documents $\mathcal{D} = \{D_1, D_2, \ldots, D_{N_d}\}$, where N_d is the number of documents. The base rate decision $c_i \in \{1, \ldots, C\}$ for $\forall i = 1, \ldots, N_d$, where c_i is a class of the i^{th} document. Thus, the dataset $\mathcal{D} = \{s_{i,j}, c_i\}$, for $\forall i = 1, \ldots, N_d$, $j = 1, \ldots, N_{S_i}$, where $s_{i,j}$ is the j^{th} sentences of the i^{th} document and N_{s_i} is the number of sentences of the i^{th} document, is composed by labeling the base rate decision for every sentence in each document collectively. The following sentences are excerpts from the minutes of the MPC meeting at which the base rate decision was FREEZE.

> *- "It is also necessary to refine the monetary policy measures such as flexible liquidity control in case the liquidity in the market is not sufficient for a limited time because of the surge in overseas interest rates and large-scale capital outflows."*
>
> *- "Looking at the domestic economy, the domestic economy is expected to recover weakly as exports continue to decline. However, as the economic sentiment has gradually improved, domestic demand*

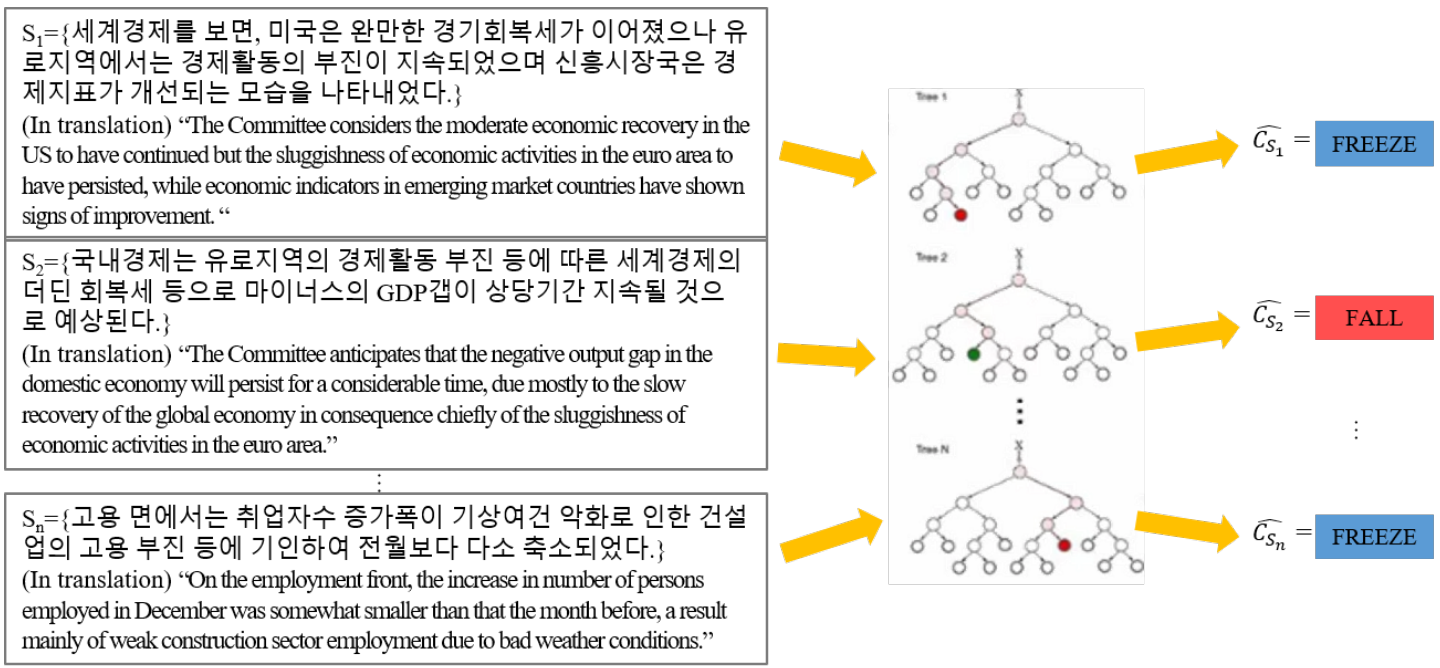


Figure 4.7: Framework of the prediction models for sentence sentiment

is showing signals of improvement."

In these sentences, some passages indicate the need for the base rate to rise or fall, even though the document overall advocates a base rate freeze. Thus, we solved the two-class classification problem except in the case of FREEZE, which has a sentence where it is difficult to assess the interest rate decision. Suppose that a data set $\mathcal{D} = \mathcal{D}^R \cup \mathcal{D}^F = \{s_{i,j}, c_i\}$ for $\forall i = 1, \ldots, N_d$, $j = 1, \ldots, N_{s_i}$ is given, where $\mathcal{D}^R$ is the data set of documents with RISE decisions and $\mathcal{D}^F$ is the data set of documents with FALL decisions. Further, the minutes of the associated MPC meetings contain personal comments; however, if the interest rate decisions are not unanimous, there may be content that proposes different interest rate directions in one document. Thus, we used only unanimously determined documents, $\mathcal{D}^{R,U}$ and $\mathcal{D}^{F,U}$, where $\mathcal{D}^{R,U}$ and $\mathcal{D}^{F,U}$ are documents where the RISE and FALL decisions respectively were unanimously determined. The pseudocode of the prediction models for sentence sentiment (PMSS) is shown in Algorithm 2.

We used the minutes of the MPC meetings as training data to improve performance. Some of the MPD text and the minutes of the MPC meetings were employed as a training set and the rest of the MPD text as a test set.

4.2.3 Aggregation of Sentence Sentiment

We can predict sentence sentiment with the sentence feature vector of the MPD text released on a particular day as the input for the prediction model. We can then calculate the MPC's vote result using sentence sentiment. A diagram of sentence sentiment aggregation is shown in Figure 4.8.

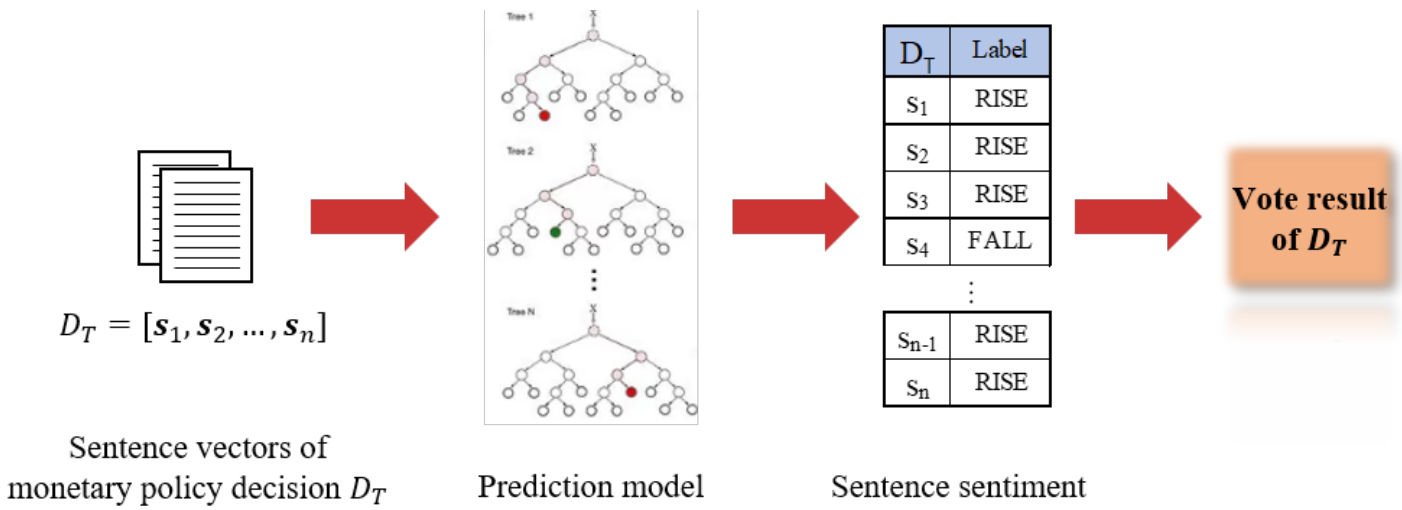


Figure 4.8: Aggregation of sentence sentiment

Algorithm 2 Prediction model for sentence sentiment (PMSS)

Input: dataset $\mathcal{D} = \mathcal{D}^R \cup \mathcal{D}^F = \{s_{i,j}, c_i\}$, for $\forall i = 1, \ldots, N_d$, $j = 1, \ldots, N_{S_i}$, $c_i \in \{1, \ldots, C\}$, $\mathcal{D}^R = \{s^R_{i,j}, c_i\}$, $D^F = \{s^F_{i,j}, c_i\}$, candidate classification algorithm $\mathcal{A}_1, \ldots, \mathcal{A}_L$

Output: set of sentence classifier $\mathbb{C}$

1: **procedure** PMSS
2: $\mathbb{C} \leftarrow \phi$
3: $\mathcal{D}^{R,U} \leftarrow$ *sampling documents of "RISE" with unanimity from* $\mathcal{D}^F$
4: $\mathcal{D}^{R,U} \leftarrow$ *sampling documents of "FALL" with unanimity from* $\mathcal{D}^R$
5: $\mathcal{D}^U \leftarrow \mathcal{D}^{R,U} \cup \mathcal{D}^{F,U}$
6: $c_{s_{i,j},\mathcal{A}_k} \leftarrow$ *candidate classifier trained from* $\mathcal{D}^U$ *using* $\mathcal{A}_k$, *for* $\forall k = 1, \ldots, L$
7: $\mathcal{A}_{best} \leftarrow \arg\min_{\mathcal{A}_k} \sum_{(s_{i,j},c_i) \in \mathcal{D}^U} \mathbf{1}_{c_{s_{i,j},\mathcal{A}_k} \neq c_i}$
8: $\mathbb{C} \leftarrow \mathbb{C} \cup \{c_{s_{i,j},\mathcal{A}_{best}}\}$
9: **end procedure**

MPD text consists of 5–17 sentences. The feature vector of each sentence is used as the input data of the prediction model and is classified into two classes (RISE or FALL). The document sentiment of each MPD is calculated as follows:

$$document\ sentiment = \begin{cases} \dfrac{n(RISE)}{n(RISE) + n(FALL)}, & \text{if } c_{D_T} = RISE, \\ \dfrac{n(FALL)}{n(RISE) + n(FALL)}, & \text{if } c_{D_T} = FALL \end{cases} \tag{4.1}$$

where $n(RISE)$ is the number of sentences with a RISE decision, $n(FALL)$ is the number of sentences with a FALL decision, and c_{D_T} is the class of the MPD text(D_T). The foregoing document sentiment has values from 0 to 1. When all sentences have the same label, the document sentiment is 1. Thus, it can be said that the base rate decision is unanimous because all sentences in the MPD text propose the same direction for the base rate. Conversely, if the document sentiment is approximately 0.5, there may be members who propose a different

direction for the base rate because some of the sentences include suggestions about the rate's direction that differ from the current base rate decision.

4.3 Experimental Results

4.3.1 Data Descriptions

We collected 217 MPD documents, together with minutes of MPC meetings, from May 1999 to July 2017 from BOK's website. Preprocessing included the removal of specific morphemes and words that directly indicated interest rate direction, such as upward, downward, maintain, present level, and base rate. An MPD text is composed of five to seven paragraphs. It includes references to the global economic growth of the US, China, and emerging countries and to the Korean economy. These references include exports, employment-to-population, the number of persons employed, consumer price inflation, the trend of housing sales and leasehold deposit prices, market interest rates, stock prices, and foreign exchanges. The minutes of an MPC meeting are released on BOK's website two weeks after the meeting. The minutes are usually composed of 10 to 40 pages with three major components, as shown in Figure 4.9

The "Summary of Discussions" component contains the individual opinions of each member about monetary policy together with the result of the majority vote. In the past, opinions from government ministers, such as the Deputy Minister of the Ministry of Strategy and Finance, have been considered. Such opinions represent the government's view on the current economic situation, inflation trends, policy direction, et cetera. However, recently, the individual

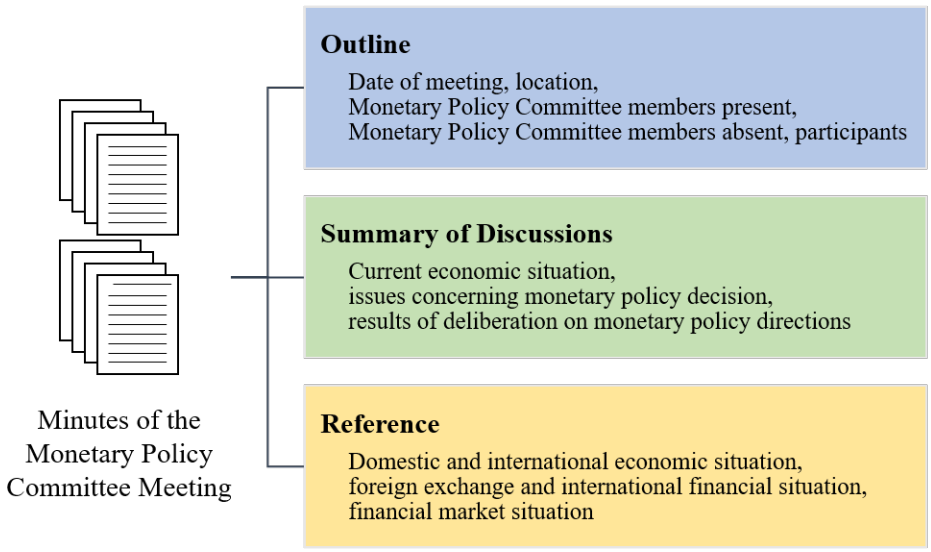


Figure 4.9: Composition of the minutes of an MPC meeting

opinions of each MPC member about issues such as the domestic and international economic situations are mainly recorded.

4.3.2 Sentence Sentiment Prediction of a Monetary Policy Decision

Before predicting an MPC's vote result, we can predict its sentence sentiment using the MPD text and the minutes of the MPC meeting. Thus, the proposed method was applied to monetary policy documents. The results are shown in Table 4.1. The general methods used to judge the performance of a classification problem are accuracy, precision, recall, and F1 measure (Yang, 1999). The purpose of this study is to reduce the misclassification rate and clearly separate the two classes; thus, we use the accuracy measure (see Table 4.1). The values are the average of test prediction accuracy over 10 repetitions. The values in

parentheses are standard deviations. The baseline model denotes the accuracy when classifying all the test data into one class.

Table 4.1: The accuracy of two-class classification

		Accuracy
Baseline		0.519
Unigram	LR-TF	0.788 (0.000)
	LR-TF-IDF	0.788 (0.000)
	RF-TF	**0.846 (0.000)**
	RF-TF-IDF	**0.846 (0.000)**
Unigram + Bigram	LR-TF	0.692 (0.000)
	LR-TF-IDF	0.750 (0.000)
	RF-TF	0.788 (0.000)
	RF-TF-IDF	**0.846 (0.000)**

These prediction models are quite robust without being influenced by initial conditions because the standard deviations of all the experiments are small. No difference exists between the unigram and bigram features; moreover, the RF performance outperforms the LR performance. The RF improvement compared with the baseline amount is 63%. It can be confirmed that prediction performance is greatly improved by using the minutes of the MPC meetings as a training set. Some of the words selected as the critical variables are listed in Table 4.2.

As shown in Table 4.2, most words illustrate not only the economic situation but also contain positive or negative meanings. However, when the input variables extracted from the unigram features are used, it is hard to judge whether the words have positive or negative meanings because there is no object. With regard to bigram features, selected critical variables clearly indicate the con-

Table 4.2: Significant words

Unigram	강화(reinforcement), 건전성(health), 격차(gap), 경색(stringency), 과잉(surplus), 구조조정(restructuring), 꾸준하다(consistent), 낮아지다(get lower), 둔화(slowdown), 반도체(semiconductor), 부각(relief), 부동산(real estate), 부진(slump), 불안(anxiety), 상승(upturn), 상승세(be on the upturn), 성장(growth), 심화(deepen), 악화(deterioration), 억제(control), 원자재(raw materials), 위축(shrinking), 인플레이션(inflation), 취업(employment), 침체(recession), 호조(being in good condition), 회복(recovery)
Unigram + Bigram	가격 상승(price rise), 경기 상승(business upturn), 경기 침체(business recession), 경기 회복(business recovery), 경색 현상(crunch), 기업 구조조정(industrial restructuring), 금융 안정(financial stability), 금융 완화(easy money), 금융시장 불안(unstable financial market), 기대 심리(psychology of expectation), 대출 증가(increase loan), 물가 상승(inflation), 물가 오르다(prices go up), 상승 기조(ascending mood), 상승세 지속(continue to climb), 신용 경색(credit crunch), 신용 위험(credit risk), 유가 하락(lower oil prices), 위축 되다(be daunted), 환율 하락(falling exchange rate)

text in which economic conditions have changed. In addition, when the input matrix is created using TF-IDF weights, the words in red in Table 4.2 appear only in a small number of documents; however, it can be seen that the words that represent important economic situations are well chosen.

4.3.3 Vote Result Prediction

The document sentiment for each MPD was calculated in Subsection 4.2.3 using the respective MPD text as input for the prediction model. $n(RISE)$ and $n(FALL)$ are the average numbers of sentences that are repeated 10 times.

In order to predict the MPC's unanimity, document sentiment was assigned to two classes, as shown in Figure 4.10. Table 4.3 presents the confusion matrix of unanimity classification.

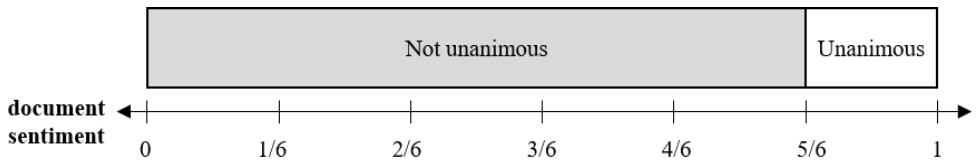


Figure 4.10: Assignment of document sentiment with two classes

Confusion matrix can represent the decision made by the classifier in a structure. The confusion matrix consists of four categories: True positives (TP) are examples correctly predicted as unanimous. False positives (FP) refer to not unanimous examples incorrectly labeled as unanimous. True negatives (TN) correspond to not unanimous correctly labeled as not unanimous. Finally, false negatives (FN) refer to unanimous examples incorrectly labeled as not unan-

Table 4.3: Confusion matrix of unanimity classification

Models	Confusion matrix			
LR			actual	
			unanimous	not unanimous
	predicted	unanimous	0.270	0.054
		not unanimous	0.135	0.541
RF			actual	
			unanimous	not unanimous
	predicted	unanimous	0.270	0.135
		not unanimous	0.135	0.459

imous. Accuracy is calculated from $(TP + TN)/(TP + FP + TN + FN)$ of confusion matrix. We suppose that the baseline model is the accuracy obtained when classifying all the test data into one class. The accuracy of the baseline model is 0.595. The LR and RF improvements compared with the baseline amount are 36.8% and 22.8% respectively. These results imply a somewhat accurate prediction of unanimity when an MPD is announced.

Next, the vote result of an MPC meeting is unanimous or has an opposing minority opinion of one, two, or three members; consequently, this study assigned document sentiment to four classes in order to compare the actual vote results of MPC meetings, as shown in Figure 4.11.

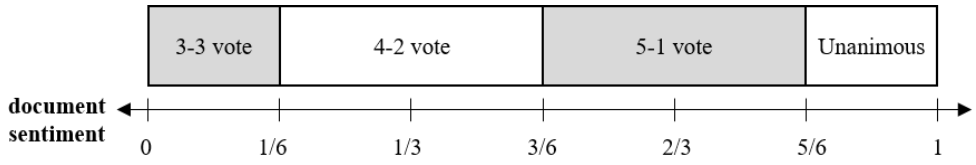


Figure 4.11: Assignment of document sentiment in terms of four classes

The correlation between the actual vote results and the class assigned to the document sentiments was calculated as shown in Table 4.4 According to the

Table 4.4: Correlation results

Correlation		LR	RF
Unigram	TF	0.52	0.30
	TF-IDF	0.49	0.31
Unigram	TF	0.28	0.20
+ Bigram	TF-IDF	0.43	0.29

correlation values in many studies, there are weak correlations, moderate correlations, high correlations, and very high correlations (Taylor, 1990; Lee Rodgers & Nicewander, 1988). Based on the various methods of interpreting correlations in prior research, when the combination of unigram features and the TF matrix is applied to the LR, the correlation, 0.52, can be interpreted as moderate. In addition, it can be confirmed that the correlation value is not large but positively correlated overall.

4.4 Summary

We have proposed a methodology for predicting the MPC's vote results using monetary policy documents. The sentence feature vector of MPD text was applied to a prediction model in order to predict sentence sentiment. The sentence sentiment was then aggregated to predict the vote result through document sentiment. First, we predicted the sentence sentiment using the MPD text and the minutes of the corresponding MPC meeting. In consequence, we found that the RF performance compared with the baseline improved by 63%. In addition,

we confirmed that the extracted words are important variables when using the bigram model and can clearly judge the economic situation. We also found that when using the TF-IDF method, words that appear only in a small number of documents were extracted as important variables. Finally, document sentiment was aggregated with sentence sentiment. We used document sentiment to classify unanimity; as a result, a classification performance of 81.1% was obtained with the LR. Further, we applied the four-class problem because there are four types of vote results. We confirmed through a correlation analysis that a positive correlation exists between document sentiment and an actual vote result.

By extracting meaningful information such as document sentiment from monetary policy documents through prediction modeling, it is possible to provide various insights to practitioners who interpret the documents empirically. In particular, when an MPD is released, the vote result can be quickly established based on the prediction model. This insight facilitates the provision of a data-driven decision-making process for practitioners who use their business experience to assess minority opinions. Such an approach enables practitioners to build a competitive position through fast and accurate bond trading. Since the base rate is the benchmark for various government bonds, traders can forecast the movement of bond rates and gain additional profits if they can predict the base rate's long-term direction. Thus, it may be possible not only to generate additional profits through competitive transactions, but also to provide portfolio management tools based on future base rate information.

In the future, base rate predictions could be improved by using information

such as interviews with the MPC and news about monetary policy in addition to monthly MPD text and minutes of the MPC meetings. Further, such information could be used to construct a domain-specific sentiment dictionary by extracting important words related to monetary policy. Each document's sentiment score could also be obtained using the dictionary, following which the score could be compared with the base rate trend and visualized to see whether it is a leading, coincident, or lagging indicator. Lastly, it is expected that various studies will be undertaken to interpret the Korean financial market. Such studies could include assessments of changes in the wording of monetary policy documents after the appointment of a new governor of BOK.

Chapter 5

Modeling the 3-10 Year Spreads with Economic Indicators

5.1 Background

Bonds issued in different countries vary in terms of characteristics and regulations involved, and a variety of market participants use these bonds as a means of asset management. Bonds are issued not only by the government but also by the national banks, high-endowed companies, and other institutions. The issue interest rates are calculated according to the credit rating of the issuing entity. After the bond issuance, bond yields are affected by a set of economic indicators such as the exchange rate, oil price, the composite price index of stocks, and employment index. Among many different bonds, those issued by the government mainly have maturities of 3, 5, 10, 20, and 30 years, and an efficient bond portfolio should be constructed since risks are calculated differently at maturity. When trading bonds, a market-associated risk is recorded as "delta". Generally, each company allocates risk limits for each maturity and total risk limits to the traders. Risk is, ultimately, directly related to the amount of the invested asset, which has a direct impact on profitability. For these reasons, it

is necessary to manage risks effectively in order so as to maximize operating profits in the financial market. One of the examples of the risk management strategies include spread trading. Spread trading uses the interest rate difference between the 3-year and the 10-year government bonds that affect risk limits by maturity yet do not affect the total risk limits. In fact, the interest rate difference between 3-year and 10-year government bonds, called the 3-10 year spreads, has functioned as an important market indicator, and traders use a strategy that generates additional profits by predicting the 3-10 year spreads. Figure 5.1 plots an example of the 3-10 year spreads trend.

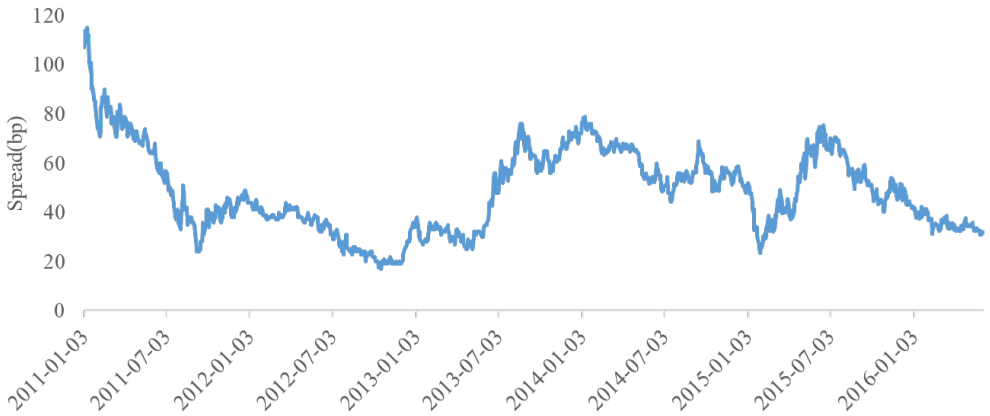


Figure 5.1: The 3-10 year spreads

The graph shows that the spreads are volatile, oscillating between 20 and 120 basis points (bp). Long-term and short-term spreads usually are of positive values. However, the spreads of negative values reflect the investor's pessimistic outlook for the long-term economic growth and inflation latently expected. Apparently, in May 2017, the spreads of 5-year and 10-year Chinese bond were of

negative values. Since the cap and the floor of the spreads are not fixed, and since the volatility of the spreads is huge, it is difficult to predict the spreads.

This chapter propose a data mining framework that uses various economic data and quantitative indicators as the input and the 3-10 year spreads of Korean government bonds as the output. In spite of that if traders could exploit different quantitative indicators that can help effectively predict the spreads of interest rates, then it can be used as an important indicator in the bond market. Firstly, wrapper methods such as forward selection, backward elimination, genetic algorithm should be used among dimensional reduction algorithms for interpretation and three regression algorithms, specifically random forest (RF), neural network (NN), and support vector regression(SVR) are used. Finally, we evaluate the performance with two metrics such as percentage of absolute range error (PARE) or mean absolute error (MAE).

In this chapter, we discuss the proposed framework, which reflects the characteristics of the spread prediction problem in Section 5.2, and Section 5.3 describes experimental settings and reports results. Finally, we conclude this chapter in Section 5.4.

5.2 Proposed Method

As for spread prediction, data mining framework for financial problem should focus on three major aspects: firstly, model selection and dimensionality reduction should be carried out with caution, because it is important to be able to determine which variables are the main drivers of the analysis result; hence,

wrapper methods such as forward selection, backward elimination, genetic algorithm should be used among dimensional reduction algorithms. Second, it is more appropriate to use metrics such as PARE or the commonly used MAE as the numeric estimate of financial market performance. In certain cases, it may be more important to approximate the range of a target measure of interest rather than computing the exact value for it. For example, as one chooses to sell a stock at a given time, one bases the decision on a range around the target price rather than the exact amount, because transaction cost may incur. Thus, we propose a framework that utilizes both MAE, a measure indicative of how well the target value is predicted, and PARE, an evaluation metric considering the tolerances of the target value. Finally, real-time transactions in the financial market are based on intuitive and rapidly produced analysis results. Hence, the delivery of the analysis results should include various modes of result descriptions, such as tables, charts, and diagrams, with detailed interpretations, such as a list of variables that contributed significantly in deriving such results. There may exists a hierarchical relationship among variables, about which the investors and active traders must be informed before making selling/buying decisions. For example, in the financial market, the convention is that macroeconomic indicators, microeconomic indicators, exchange rates, and stock prices are defined as the parent node set, and unemployment rate, consumer price index, LIBOR, USDKRW, KOSPI, the child node set. Hence, hierarchical presentation of meaningful variables using various forms of visualization may serve as an effective assistant tool for the financial agents in the market with their decision-making process.

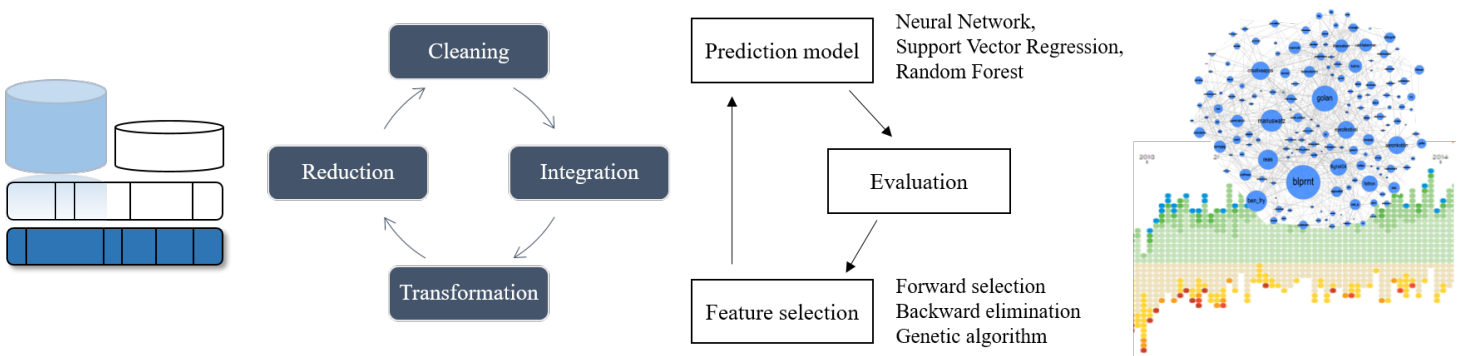


Figure 5.2: Data mining framework for spread prediction

5.2.1 Preprocessing

In general, the process of collecting and preprocessing data in data analysis is very important and time-consuming. In case of numerical data, relatively simple preprocessing is required, such as null value removal and categorical variable processing. However, for textual data, there are somewhat complicated preprocessing such as parsing, stopword elimination, and tagging. In this section, we will look into the details of preprocessing when the collected real-word data is numerical data. Most of the collected real-world data contains null values, if a certain variable has a large number of null values, the variable itself can be removed to facilitate analysis, otherwise records with null values are deleted or replaced with average values. Next, we regarded the variables having the same values of all records as no information, so we removed these variables. The categorical variables were converted to dummy variables and all variables were normalized by scaling between 0 and 1 for all variables.

5.2.2 Prediction Models

The prediction models for domain-specific framework were built using three regression algorithms, specifically RF, NN, and SVR. The descriptions of prediction models are in Section 2.3.

5.2.3 Feature Selection

For ease of interpretation, the subset selection algorithms were used when performing dimensionality reduction. Among variable selection algorithms, forward selection, backward elimination, and GA were used. Forward selection starts

with no variables in the model, tests the addition of each variable by using a chosen model comparison criterion, adds the variable that improves the model the most, and repeats this process until none improves the model. Backward elimination starts with all candidate variables, tests the deletion of each variable by using a chosen model comparison criterion, deletes the variable that improves the model the most by being deleted, and repeats this process until no further improvement is possible. GA is a meta-heuristic optimization algorithm, which has been applied to various combinatorial optimization problems, such as layout optimization and reliability/redundancy allocation (Izui et al., 2013; Kanagaraj et al., 2013). GA is based on the evolutionary process of natural selection and genetics. GA simulates the survival of the fittest over consecutive generation in each reproduction. Each individual in the population is encoded into a string representing a candidate solution to the problem, and good performers having the better opportunities are generated by genetic operations on the selected individuals from the current population in order to reproduce the next generation (Shukla & Tiwari, 2012). Such process is repeated until it meets a stopping criterion, and then the best performing solution is found. Figure 5.3 shows the main genetic operators of GA.

In the GA wrapper feature selection procedure, each feature is considered a gene and the selected features set is considered a chromosome (Yu & Cho, 2006). Candidate feature sets are generated through the evolutionary process, and when converged, the new feature sets are selected in the population. They are evaluated by prediction models and the model with the best validation performance is selected. This process repeats until stopping criteria is reached. The

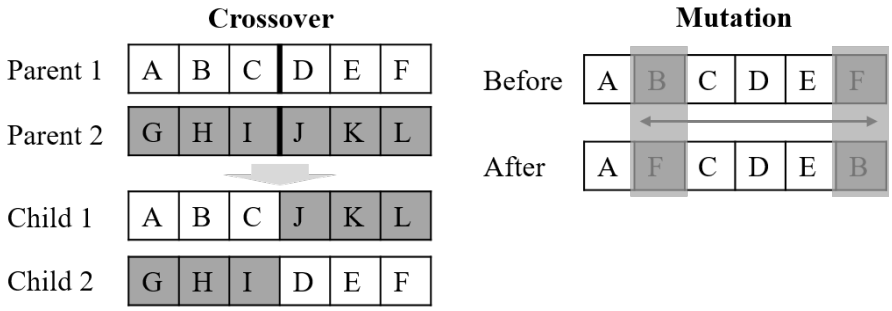


Figure 5.3: The main genetic operator of GA

procedure of GA wrapper is outlined in Figure 5.4.

We experimented a total of 9 different ways by applying dimension reduction algorithms to three prediction models, and the list is shown in Table 5.1.

5.2.4 Evaluation

We use two accuracy metrics for financial markets to evaluate the prediction performance: MAE and PARE which is an appropriate indicator for the target value with error tolerance. Firstly, the best performance model was selected by using the MAE criterion, which indicates how close the predicted values are to actual values; the MAE is calculated as follows,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|. \tag{5.1}$$

where y_i, f_i, and e_i are the actual value, predicted value, and absolute error. In addition, since the process tolerances exist due to the nature of the treating

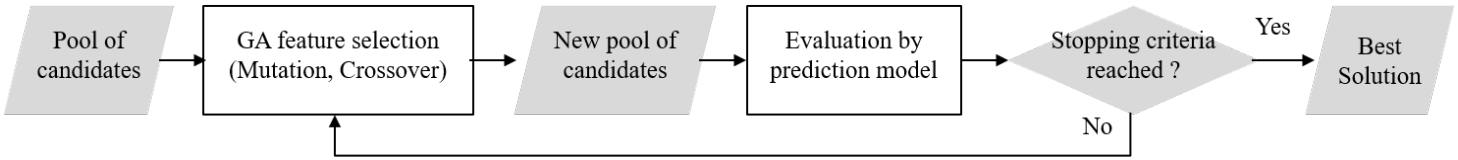


Figure 5.4: The procedure of GA wrapper approach

Table 5.1: Prediction models used in the experiments

Prediction Models	**Description**
RF-F	Random forest, selected variables using forward selection
RF-B	Random forest, selected variables using backward elimination
RF-GA	Random forest, selected variables using genetic algorithm
NN-F	Neural network, selected variables using forward selection
NN-B	Neural network, selected variables using backward elimination
NN-GA	Neural network, selected variables using genetic algorithm
SVR-F	Support vector regression, selected variables using forward selection
SVR-B	Support vector regression, selected variables using backward elimination
SVR-GA	Support vector regression, selected variables using genetic algorithm

process, the PARE criterion is calculated for ratio within the error tolerance. The equation of the PARE is as follows,

$$PARE = \frac{1}{n}\sum_{i=1}^{n} I(|f_i - y_i| \leq \theta) \tag{5.2}$$

where y_i, f_i, θ, and $I(x)$ are the actual value, predicted value, error tolerance, and indicator function, having values 1 and 0 for true x and false x, respectively. PARE has a value 1 when all the predicted values are within error tolerance. In this study, we used error tolerances determined by domain expert, and ten-fold cross validation was used to calculate MAE and PARE. In addition, we compare the results of the prediction models and find the best model among the prediction models using an independent 2-sample t-test with unknown standard deviation.

5.2.5 Reporting

There exists a myriad of ways to report analysis results based on the objective of the analysis and the modeling method employed. Visualization allows intuitive problem identification and assists decision making process in a practical sense. For example, based on the actual data, the value of each variable can be expressed as a graph over time. Figure 5.5 shows an example of a specific variable collected in the real-world. As you can see from the graph above, the values at some point lie outside the range commonly shared between other values. The field worker can easily detect this jumping data point from the visualization provided and suspect that there exists a problem involved with the

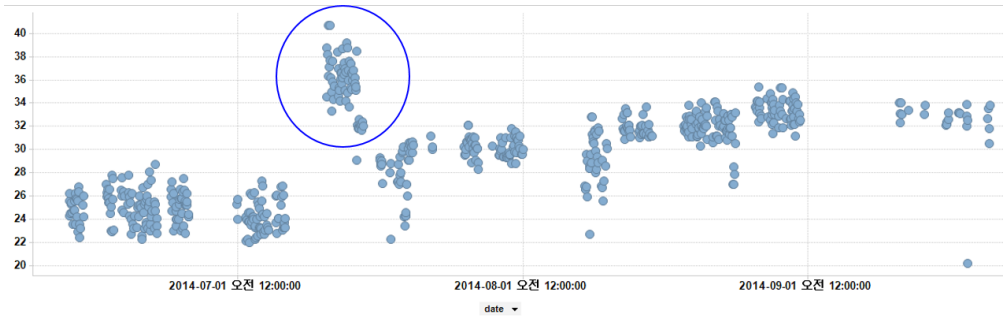


Figure 5.5: The values of a specific variable over time

variable of the data recorded. Another effective reporting method is to provide a hierarchical listing of the set of input variables observed to be meaningful in the analysis. In this way, it is possible to provide various information to the business through the reporting form which analyzes the category among the input variables having the characteristic of the hierarchical variables as well as the visualization.

5.3 Experimental Results

5.3.1 Data Descriptions

We used real-world data collected from January 1, 2011 to December 31, 2015, and the variables are selected based on the opinions of the domain expert. Firstly, we used the INFOMAX terminal to collect 20 variables such as exchange rates, WTI, interest rates, and composite stock price index. In addition, we collected seven economic indicators, which are announced monthly, on the website of the National indicator system[1]. A total of 27 input variables were

[1]http://www.index.go.kr/main.do

used as listed in Table 5.2, and output variables were used from the difference in interest rates for 3-year and 10-year government bonds, which are called 3-10 year spread.

Table 5.2: Lists of input variables

Frequency	Data Categorization			
Daily	Policy interest rate	Exchange Rate/Oil Price	Variable Interest Rate	The Price Index of Stocks
	Korea, USA, Japan, China, Australia, UK	USDKRW, USDJPY, EURUSD, WTI	LIBOR, EURIBOR	KOSPI, KOSDAQ, DOW, S&P, FTSE, HSCEI, Shanghai, NIKKEI
Monthly	Domestic		Overseas Country	
	Unemployment rate, consumer price index, index of mining and manufacturing industrial product, total production of small manufacturing firm, trade balance		Unemployment rate of USA, Nonfarm payroll	

5.3.2 Experimental Settings

Because there is no data for each country on each national holiday, in the case of public holidays, the value was replaced by the value of the previous business day. Monthly data was also converted to daily data based on the date of the announcement. The tolerance θ of PARE was used as 3bp. All data values were preprocessed from -1 to 1 through min-max normalization. First of all, we conducted experiments to find optimal parameters for four prediction models mentioned in Subsection 5.2.3. For each algorithm, the optimal parameters were

found only for the parameters shown in Table 5.3, and for the other parameters, the default values provided by MATLAB were used. The minimum leaf

Table 5.3: Search ranges of parameters of each algorithm

Algorithm	Parameters	Search Space
Random Forest	minimum leaf	{1, 2, ..., 5}
Neural Network	hidden node	{2, 4, ..., 20}
Support Vector Regression	tradeoff(C)	$\{10^{-1}, 10^{0}, ..., 10^{2}\}$
	epsilon(ϵ)	$\{2^{-5}, 2^{-7}, ..., 2^{-3}\}$
	sigma(σ)	$\{2^{-2}, 2^{-1}, ..., 2^{2}\}$

of random forest is the minimum number of observations per tree leaf for each decision tree, and the hidden node of neural network is the number of hidden nodes in the hidden layer. For SVR, tradeoff is the tradeoff between the training errors and the model complexity, epsilon is the parameter of epsilon-insensitive loss function, and sigma is the sigma in kernel function. Before applying the dimension reduction algorithm, the experiments were repeated ten times for each algorithm, so we decided the best parameters by using the smallest MAE value. Table 5.4 summarize the values of parameter estimation.

Table 5.4: The values of parameters of each algorithm

Algorithm	Parameters	Values
Random Forest	minimum leaf	1
Neural Network	hidden node	18
Support Vector Regression	tradeoff(C)	100
	epsilon(ϵ)	2^{-5}
	sigma(σ)	1

5.3.3 Spread Prediction

Base on those optimal parameters shown in Table 5.4, we performed experiments with 9 different models. The results of MAE and PARE are shown in Table 5.5.

Table 5.5: MAE and PARE results of each algorithm for spread prediction

Method	MAE	PARE
RF-F	1.706 (0.040)	0.832 (0.008)
RF-B	1.671 (0.013)	0.833 (0.004)
RF-GA	1.636 (0.009)	0.840 (0.003)
NN-F	2.222 (0.047)	0.738 (0.009)
NN-B	2.126 (0.094)	0.760 (0.016)
NN-GA	2.017 (0.011)	0.780 (0.011)
SVR-F	2.002 (0.011)	0.778 (0.005)
SVR-B	1.982 (0.011)	0.780 (0.004)
SVR-GA	1.985 (0.000)	0.778 (0.005)

The values shown in Table 5.5 are calculated by averaging across ten repetitions, and the values shown in parentheses are standard deviations of MAEs and PAREs. The LR model is a poor result compared with other prediction models, but the performance of RF is best. For both accuracy metrics, the MAE and PARE of spread are at minimum and maximum value, respectively, when they chose the RF prediction model and GA variable selection. Moreover, those prediction models are stable without being influenced by initial conditions, as standard deviations of all experiments are very small. Next, we conducted an independent 2-sample t-test when the standard deviation is unknown, as described in Subsection 5.2.4, to determine whether there was a performance

difference between the above prediction models. We calculated the proportion of rejecting the null hypothesis among all combinations of prediction models. The results of test are as below on Table 5.6.

Table 5.6: Proportions of significant test for spread data

Output Variables	Metric	Proportion of Significant Test (%)
Spread	MAE	97.22%
	PARE	80.56%

As in the results listed above, MAE metric showed more that 97.22% of proportion of rejecting the null hypothesis on significance level 0.05, and PARE showed p-value that are less than 0.05 on around 80.56% of the tests. The p-value was small enough in most single test conducted, and one could lead to a conclusion that the best model is not only outperform the other models, but ten-repetition is enough to verify whether the performance of prediction models differed significantly.

So far we understood the performances of prediction models with MAE and PARE criteria. Although both results are meaningful, we selected relevant input variables and the best prediction model by using PARE criterion for target variable since process characteristics required error tolerances. The best model is RF-GA for spread variable, and we chose the relevant input variables that were selected more than 50% for the ten repetitions and input variables divided as interest rate variables, exchange rate & oil price variables, the price index of stock variables, domestic economic indicator variables, and overseas country economic indicator variables shown in Table 5.7.

Table 5.7: The result of variable selection

	Interest Rate	Exchange Rate & Oil Price	The Price Index of stocks	Economic Indicator	Total
Spread (RF-GA)	5/8	3/4	3/8	6/7	17/27

In the Table 5.7, we can see that most economic indicators have been selected as important variables, which is consistent with paying attention to the announcement of indicators that are considered to affect bond rates in the field, and only a part of the price index of stock variables have been selected as important variables. Unexpectedly, most exchange rates & oil prices are also important variables. In the field, traders are paying attention to announcement of economic indicators and monitoring interest rates, but tend to consider that exchange rate fluctuations have little impact on spreads.

5.4 Summary

In this chapter, we propose a data mining framework suitable for spread prediction in financial markets with the following three characteristics: interpretation, prediction evaluation metrics, and visualization. First of all, a given prediction model should provide both predictive power and interpretability. We propose that such an objective may be achieved by employing the wrapper approaches when constructing the prediction model, through which the extent of contribution of each variable in the analysis can be observed. Secondly, the spread prediction model needs to be evaluated using appropriate metrics. For example, it may be more important to predict the range of the spread, rather than

the exact value, so as to reflect tolerance. In this case, PARE will be a more appropriate measure of evaluation. Finally, provision of the analysis results in various reporting modes is called for, especially for the sake of offering assistance to the active traders in the market on making split-second decisions with detailed explanations. Taking these characteristics into account, we have applied the proposed framework to the task of predicting the 3-10 year spreads. When traders manage their assets, they can use a strategy that does not affect the total amount of risk, such as the 3-10 year spread transactions. Various quantitative financial data and economic indicators such as policy interest rate, exchange rate, and the price index of stocks were collected, and evaluated through the prediction models. We used three different prediction models, RF, NN, and SVR, respectively, for building an interpretable data mining framework. In addition, we used a dimensionality reduction algorithm to remove less significant variables in order to shorten training time. Among many existing dimensionality reduction algorithms, we choose the subset selection algorithms – forward selection, backward elimination, and GA – for convenience of providing interpretation over the results. Our experiments showed that the RF-GA produced the best record of the MAE and PARE, Particularly, the PARE metric recorded 84% when using the RF-GA model. Finally, statistical tests were performed to test whether the performance of each prediction model showed significant differences, and more than 95% of the prediction model pairs for the MAE and PARE metrics rejected the null hypothesis. Finally, it can be seen that most of the economic indicators were shown to be important variables among other categorical variables, and traders have identified that the new predictor variables

such as the exchange rate, which had been treated as having little influence, also have a significant effect on the spread.

Although only 27 input variables were used in this study, we expected to improve the performance of prediction model using various economic indicators in the future. Thereby, we were able to pin down new spread indicators, as well as a set of variables empirically proven to be meaningful in the analysis. The results may lead to an effective monitoring activity by focusing on these variables intensively, instead of the conventional overall monitoring. We pose the results from our pre-defined data mining framework as an assistant tool to help traders make data-driven decisions with the proper reporting methods.

Chapter 6

Conclusion

6.1 Contributions

Many researches have exploited data of various sources and forms in order to predict financial market behaviors, such as stock or bond prices. Previous studies have shown great performance in predicting the target value by exploiting prediction modeling. However, practitioners may find the results limiting when applied to the actual trading. In other words, because it is directly linked to profits, they want to be able to quantitatively predict the performance of prediction models, while being offered the underlying evidence, or intuitive explanations, based on which their transaction decisions is to be made. Thus, this dissertation proposed methodologies suitable for distinct characteristics across different financial data and for the purposes of prediction modeling. We presented the result of analysis via visualization indifferent formats.

For the stock market, we proposed a methodology that uses distributed representations of text to improve the power of predictability through sentiment analysis of corporate disclosures. In addition, we successfully provide the underlying evidence for our prediction results through visualization of docu-

ments considered in our experiment. Methodologically, we employed the SPV model, a method using distributed representation embedding documents and class information in the same embedding space for the task of predicting stock price movements. We sampled four companies in the financial sector and used respective corporate disclosures in order to predict the associated stock price movements by forming the task into a two-class classification problem. Prediction performance was compared among a rich selection of prediction models, such as LR, RF, MNB, SVM and NBSVM. As a result, was the SPV-model not only the most accurate model in terms of prediction power, but it also recorded a 25.4%-point improvement in the prediction accuracy as compared to the LR model. Furthermore, via visualization, we confirmed that stock prices are affected by the polarity of the sentiment derived from the published documents. For example, if the sentiment of the associated documents is positive, it could be qualitatively verified that the stock price shows an upward trend. In contrast, if the sentiment of the associated documents is negative, our analysis has shown that the stock price shows a downward trend. The proposed method provided interpretability over the results with sentiment visualizations, on top of the improved prediction accuracy, so that the active traders without sufficient knowledge in data mining can still take full advantage of our study results.

A methodology was proposed for predicting the vote results of the base rate decision by using monetary policy documents from the Korean bond market. Monetary policy documents were represented through BoW and were used as input of LR and RF models to predict the sentence sentiment. As a result, we achieved better performance compared to the baseline model. In addition, we

predicted the vote result using document sentiments, computed by aggregating sentence sentiments. We achieved 81% classification accuracy for two-class classification problem that predicts unanimity. For four-class classification problem predicting the type of vote results, we conducted correlation analysis between the document sentiment of the MPD and the actual vote result, and the result showed moderate correlation of 0.52 Finally, we found that the bi-grams extracted from monetary policy documents were helpful in characterizing the economic situation, while words extracted via TF-IDF method played an important role in describing the Korean economy at the corresponding time, such as semiconductor, employment, and household loans.

We defined a data mining framework for predicting the spread, the difference between two bond rates with different maturities. The major issues involved in analyzing the spread are: interpretability, proper prediction metrics, and various reporting methods. The proposed framework used wrapper approaches which selects meaningful variables that affect the performance of prediction models so that the effect of financial variables can be directly interpreted. PARE, in combination of MAE, was used to quantify model performance, taking into account the tolerances of the target value. Lastly, we provided visualization and hierarchical information of significant variables so as to deliver analysis results in an intuitive and steady-fast manner. As a result of applying this framework to the spread prediction, the PARE of the RF-GA model recorded 84%, which may serve as the benchmark prediction accuracy for business decision making in the field. Furthermore, tests have shown that the difference in performance across different models employed were statistically significant. Finally, it can

be seen that domestic and foreign economic indicators are important, among the various category variables available, and the exchange rate, which had been treated as having little influence on the spread, in explaining the fluctuations of the spread.

In this dissertation, we applied a variety of prediction modeling to financial data and suggested methodologies that serves the domain-specific characteristics of data as well as the purpose of predictions. Agents in the financial market faces great competition in terms of generating profits, especially given the rapidly evolving nature of the market. Therefore, the proposed methodologies can be delivered to the real-site as a data-driven decision making support tool that can give various types of insight to the practitioner who handle the data empirically. Using financial data, the practitioners can monitor critical financial information such as finding new investment companies, identifying crisis companies and finding interest rate change signals. By extension, the practitioners can make split-second decisions based on quantitative and qualitative prediction. This dissertation contributes to the body of applied research by combining existing methodologies effectively in order to investigate markets previously under-studied. Methodologies we propose reflects domain-specific characteristics peculiar to the market of interest adequately, and the results revealed interesting insights.

6.2 Future Work

Data we used was limited, since we were constrained to choose companies in a selected field for the purpose of providing interpretability over our analysis results. More specifically, in this dissertation, we analyzed only a handful of companies in the finance sector when predicting stock price movement using corporate disclosures. In the future, we plan to analyze individual sentiment using SNS and blogs as well as corporate disclosure documents to gain greater insights. In addition, we aim to extract sentiment scores for a wide range of words and compile them together into a sentiment dictionary, designed specifically for finance and accounting research. The trajectory of the sentiment scores may be visualized by scoring the sentiment for each document using the mentioned sentiment dictionary and plotting them again time. By doing so, we expect to verify whether the sentiment score is a leading indicator, coincident indicator, or lagging indicator describing market movements. Finally, it is possible to develop a powerful, data-driven assistant tool for making transaction decisions by creating and simulating strategies based on prediction models.

Bibliography

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Ahn, J. J., Lee, S. J., Oh, K. J., & Kim, T. Y. (2009). Intelligent forecasting for financial time series subject to structural changes. *Intelligent Data Analysis*, *13*(1), 151–163.

Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Stock price prediction using the arima model. In *In UKSim-AMSS* 16^{th} *IEEE international conference on computer modelling and simulation (UKSim)* (pp. 106–112).

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046–7056.

Beckmann, M. (2017). *STOCK PRICE CHANGE PREDICTION USING NEWS TEXT MINING* (Unpublished doctoral dissertation). Universidade Federal do Rio de Janeiro.

Bernanke, B., & Gertler, M. (2000). *Monetary policy and asset price volatility* (Tech. Rep.). National bureau of economic research.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422.

Bholat, D. M., Hansen, S., Santos, P. M., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *Centre for Central Banking Studies Handbook*, *33*, 1–19.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Burger, J. D., Warnock, F. E., & Warnock, V. C. (2017). *The effects of us monetary policy on emerging market economies' sovereign and corporate bond markets* (Tech. Rep.). National Bureau of Economic Research.

Chen, C.-M., & Liu, C.-Y. (2009). Personalized e-news monitoring agent system for tracking user-interested chinese news events. *Applied Intelligence*, *30*(2), 121–141.

Cheng, W., Wagner, W., & Lin, C.-H. (1996). Forecasting the 30-year us treasury bond with a system of neural networks. *Journal of Computational Intelligence in Finance*, *4*.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the* 25^{th} *international conference on machine learning* (pp. 160–167).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215–242.

Dahl, G. E., Adams, R. P., & Larochelle, H. (2012). Training restricted boltzmann machines on word observations. In *In international conference on machine learning.*

D'Amico, S., & King, T. B. (2013). Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply. *Journal of Financial Economics*, *108*(2), 425–448.

Dougal, C., Engelberg, J., Garcia, D., & Parsons, C. A. (2012). Journalists and the stock market. *The Review of Financial Studies*, *25*(3), 639–679.

Druz, M., Wagner, A. F., & Zeckhauser, R. J. (2015). *Tips and tells from managers: How analysts and the market read between the lines of conference calls* (Tech. Rep.). National Bureau of Economic Research.

Fama, E. F. (1970). Multiperiod consumption-investment decisions. *The American Economic Review*, 163–174.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, *7*(2), 179–188.

Fuhrer, J. C., & Moore, G. R. (1995). Monetary policy trade-offs and the correlation between nominal interest rates and real output. *The American Economic Review*, 219–239.

Gagnon, J., Raskin, M., Remache, J., Sack, B., et al. (2011). The financial market effects of the federal reserve's large-scale asset purchases. *International Journal of Central Banking*, *7*(1), 3–43.

Gálvez, R. H., & Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*, *19*, 43–56.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267–1300.

Ghosn, J., & Bengio, Y. (1997). Multi-task learning for stock selection. In *Advances in neural information processing systems* (pp. 946–952).

Goodfriend, M. (1991). Interest rates and the conduct of monetary policy. In *Carnegie-rochester conference series on public policy* (Vol. 34, pp. 7–30).

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Heston, S. L., & Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, *73*(3), 1–17.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177).

Huang, A. H., Zang, A. Y., & Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review*, *89*(6), 2151–2180.

Hughen, J. C., & Beyer, S. (2015). Stock returns and the us dollar: the importance of monetary policy. *Managerial Finance*, *41*(10), 1046–1058.

Izui, K., Murakumo, Y., Suemitsu, I., Nishiwaki, S., Noda, A., & Nagatani, T. (2013). Multiobjective layout optimization of robotic cellular manufacturing systems. *Computers & Industrial Engineering*, *64*(2), 537–544.

Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*(3), 712–729.

Kanagaraj, G., Ponnambalam, S., & Jawahar, N. (2013). A hybrid cuckoo search and genetic algorithm for reliability–redundancy allocation problems. *Computers & Industrial Engineering*, *66*(4), 1115–1124.

Kiley, M. T. (2016). Monetary policy statements, treasury yields, and private yields: before and after the zero lower bound. *Finance Research Letters*, *18*, 285–290.

Kim, S. H., & Noh, H. J. (1997). Predictability of interest rates using data mining tools: a comparative analysis of korea and the us. *Expert Systems*

with Applications, *13*(2), 85–95.

Kumar, A., Agrawal, D., & Joshi, S. D. (2004). Multiscale rough set data analysis with application to stock performance modeling. *Intelligent Data Analysis*, *8*(2), 197–209.

Kung, H. (2015). Macroeconomic linkages between monetary policy and the term structure of interest rates. *Journal of Financial Economics*, *115*(1), 42–57.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics*, *47*(3), 523–544.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on machine learning (ICML-14)* (pp. 1188–1196).

Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014). On the importance of text analysis for stock price prediction. In *LREC* (pp. 1170–1175).

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*(1), 59–66.

Li, C., & Wei, M. (2013). Term structure modelling with supply factors and the federal reserve's large scale asset purchase programs. *International Journal of Central Banking*, *9*(1), 3–39.

Liu, S. (2015). Investor sentiment and stock market liquidity. *Journal of Behavioral Finance*, *16*(1), 51–67.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, *1*(1), 41–66.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the* 49^th *annual meeting of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 142–150).

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Martineau, J., & Finin, T. (2009). Delta TFIDF: An improved feature space for sentiment analysis. *In Proceedings of ICWSM*, *9*, 106.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Ming, F., Wong, F., Liu, Z., & Chiang, M. (2014). Stock market prediction from wsj: text mining via sparse matrix factorization. In *Data mining (ICDM), 2014 IEEE international conference on* (pp. 430–439).

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying wikipedia usage patterns before stock market moves.

Scientific Reports, *3*, 1801.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), 4241–4251.

Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 786–794).

Nocedal, J., & Wright, S. J. (2006). *Sequential quadratic programming*. Springer.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the* 42^{nd} *annual meeting on association for computational linguistics* (p. 271).

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the* 43^{rd} *annual meeting on association for computational linguistics* (pp. 115–124).

Park, E. L. (2016). *Supervised feature representation for document classification* (Unpublished doctoral dissertation). Seoul national university.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific Reports*, *3*, 1684.

Qian, B., & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, *26*(1), 25–33.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the* 20^{th} *international conference on machine learning (icml-03)* (pp. 616–623).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (Tech. Rep.). California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3), 1.

Schumaker, R. P., & Chen, H. (2010). A discrete stock price prediction engine based on financial news. *Computer*, *43*(1).

Shukla, S. K., & Tiwari, M. K. (2012). Ga guided cluster based fuzzy decision tree for reactive ion etching modeling: a data mining approach. *IEEE Transactions on Semiconductor Manufacturing*, *25*(1), 45–56.

Sitte, R., & Sitte, J. (2002). Neural networks approach to the random walk dilemma of financial time series. *Applied Intelligence*, *16*(3), 163–171.

Slavkovic, R., Jugovic, Z., Dragicevic, S., Jovicic, A., & Slavkovic, V. (2013). An application of learning machine methods in prediction of wear rate of wear resistant casting parts. *Computers & Industrial Engineering*, *64*(3), 850–857.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211).

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 151–161).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., ... others (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (Vol. 1631, p. 1642).

Sun, A., Lachanski, M., & Fabozzi, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis*, *48*, 272–281.

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

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39.

Tiemei, L. (2010). An asymmetrical analysis of inflation, inflation expectations

and monetary policy in china [j]. *Journal of Financial Research*, *12*, 005.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167–179.

Wang, J., & Wu, C. (2015). Liquidity, credit quality, and the relation between volatility and trading activity: Evidence from the corporate bond market. *Journal of Banking & Finance*, *50*, 183–203.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the* 50^{th} *annual meeting of the association for computational linguistics: Short papers* (Vol. 2, pp. 90–94).

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2), 165–210.

Wolfe, P. (1961). A duality theorem for non-linear programming. *Quarterly of Applied Mathematics*, *19*(3), 239–244.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, *1*(1), 69–90.

Yu, E., & Cho, S. (2006). Ensemble based on ga wrapper feature selection. *Computers & Industrial Engineering*, *51*(1), 111–116.

국문초록

금융 시장에서 다양한 예측 모델링을 바탕으로 시장 행동을 분석하는 방법론들이 제안되고 있다. 예측 모델링은 직접적으로 크게 거래 수익으로 이어질 수 있어 더욱 다양하게 연구가 이루어지고 있다. 금융 상품을 시장에서 직접 거래하는 실무자가 예측 모델링을 거래에 활용하기 위해서는 예측력과 해석력 두가지 측면이 모두 중요하다. 본 논문에서는 이 두가지 측면을 함께 제공하기 위해 다양한 금융 데이터의 특성과 목적에 맞는 방법론들을 제안한다. 주식 시장의 개별 주식 가격을 예측하기 위해서 기업 공시 문서의 감성을 분석하여 주식 가격 방향을 정량적 및 정성적으로 예측하는 방법론을 제안하고자 한다. 이 방법은 분산 표상 기반으로 문서와 문서의 클래스를 같은 공간으로 전사하여 정량적으로 주식 가격을 예측하고, 기존 연구에서 제안된 다양한 모델들과 예측 성능을 비교한다. 또한, 문서 감성의 시각화를 통해 예측력의 정성적인 근거를 함께 제공한다. 또한, 채권 시장에서 가장 중요한 요소인 다수결 방식의 기준 금리 결정을 예측하기 위한 방법론을 제안하고, 한국의 채권 시장에 적용한다. 다수결 결과가 발표되기 전에 공개되는 통화정책방향문의 문장 별 감성을 예측한다. 문장 감성을 종합하여 통화정책방향문의 문서 감성을 계산하고, 이 값을 이용하여 기준 금리 결정의 다수결 결과를 예측한다. 마지막으로, 만기가 다른 두 채권의 금리 차이인 스프레드를 예측하는 분석 프레임 워크를 정의한다. 이 프레임 워크는 해석력, 적합한 예측 성능 기준, 그리고 다양한 리포팅 방법 세가지 측면을 고려했다. 중요 변수의 직접적인 해석을 위해 래퍼 방법(wrapper approach)를 사용하고, 스프레드의 공차(tolerance)를 고려할 수 있는 예측 성능 기준을 사용하고, 시각화 및 변수의 계층 정보 등을 제공하는 리포팅 방법을 제공한다. 본 논문에서는 다양한 금융 시장의 문제를 정의하고

분석 방법론을 제안하여 정량적 예측력을 비교 분석 하였고, 정성적인 방법으로 그 근거를 제시하였다. 이러한 분석 방법론을 통해 신속하고 정확한 데이터 기반 의사 결정 지원 도구의 기반을 구축하였다.

주요어: 데이터마이닝(data mining), 기계학습(machine learning), 단어 임베딩(word embedding), 분산 표상(distributed representation), Bag-of-Words, 감성 분석(sentiment analysis), 주가 예측(stock price prediction), 금융통화위원회의 다수결 결과 예측 (vote result prediction of monetary policy committee), 채권 스프레드 예측(bond spreads prediction), 공시 문서(corporate disclosures), 통화 정책 문서(monetary policy documents), 경제 지표(economic indicators)
학번: 2015-30235