석사학위 논문

ARIMA 모형과 텍스트 마이닝을 이용한 주가 등락 예측

Stock Fluctuation Prediction using the ARIMA Model and Text Mining

2015년 6월

숭실대학교 대학원

컴퓨터학과

엄 장 윤

석사학위 논문

ARIMA 모형과 텍스트 마이닝을 이용한 주가 등락 예측

Stock Fluctuation Prediction using the ARIMA Model and Text Mining

2015년 6월

숭실대학교 대학원

컴퓨터학과

엄 장 윤

석사학위 논문

ARIMA 모형과 텍스트 마이닝을 이용한 주가 등락 예측

지도교수 이 수 원

이 논문을 석사학위 논문으로 제출함

2015년 6월

숭실대학교 대학원

컴퓨터학과

엄 장 윤

엄 장 윤의 석사학위 논문을 인준함

심 사 위 원 장 이 상 준 인

심 사 위 원 도 정 인 인

심 사 위 원 이 수 원 인

2015년 6월

숭실대학교 대학원

목 차

표 목 차

그 림 목 차

국문초록

ARIMA 모형과 텍스트 마이닝을 이용한 주가 등락 예측

엄장윤

컴퓨터학과

숭실대학교 대학원

주식 시장의 분석과 주가 예측은 경제 분야뿐만 아니라 수학, 통계, 인공지능 분야에 이르기까지 다양한 분야에서 많이 연구되고 있다. 대표적인 연구 방법으로는 과거 지수를 시계열로 분석하는 통계적 예측 방법과 뉴스 데이터 등을 활용한 텍스트 마이닝 기반 예측 방법 등이 있다. 시계열 분석 기반 예측 방법은 과거와 현재 간의 상관관계를 파악하여 추세를 확인할 수 있다는 장점이 있으며, 텍스트 마이닝 기반 예측 방법은 소셜 데이터를 이용하기 때문에 주식시장의 예기치 못한 이벤트에 대한 즉시적 반영과 분석이 가능하다는 장점이 있다.

본 논문에서는 뉴스 데이터 이용한 텍스트 마이닝 기반 예측 모델과 KOSPI 데이터를 이용한 시계열 분석 기반 예측 모델을 결합하여 주가 등락을 예측하는 모델을 제안한다. 제안 방법은 주가 도메인의 감성 사전을 이용하여 계산된 감성수치 결과와 KOSPI 데이터를 ARIMA 모형에 적용한 결과를 결합하여 당일 종가 대비 익일의 종가 등락을 예측한다. 제안 방법과 기존 방법의 성능을 비교한 결과, 제안 방법의 주가 등

락 예측 정확도가 기존 방법에 비해 약 6.5% 상향된 결과를 보였다.

ABSTRACT

Stock Fluctuation Prediction using the ARIMA Model and Text Mining

UM, JANG-YUN

Department of Computer Science and Engineering

Graduate School of Soongsil University

Many studies of stock market analysis and stock price prediction have been conducted in various fields such as mathematics, statistics and artificial intelligence as well as economics. Representative stock price prediction methods include statistical prediction methods based on time-series analysis and prediction methods based on text mining using news data. The prediction method based on time-series has a merit that it can identify the relationship between the past and the present and check the trend, while the prediction method based on text mining uses social data, so it has a merit that it can instantly reflect and analyze an unexpected event in the stock market.

This study proposes a model to predict fluctuations in a day's closing price by combining a prediction model based on text mining using news data and a prediction model based on time-series analysis using KOSPI data. The proposed method can predict the following

day's stock fluctuation by combining the sentiment result calculated by using the sentiment dictionary in the stock domain and the ARIMA model for KOSPI data.

As a result of a comparison of the performance between the proposed method and the existing method, the accuracy of the prediction of fluctuations in the stock price by the proposed method increased by 6.5% as compared to that of the existing method.

제 1 장 서 론

1.1 연구 배경 및 개요

주가지수는 정부의 정책을 결정하고 기업의 투자와 고용을 결정하는 등 경제와 사회 전반에 중대한 영향을 주기 때문에 이를 예측하는 일은 중요하다[1]. 경제 전문가들은 주가예측과 관련된 다양한 이론들을 연구하고 있다. 그 중 가장 대표적인 이론으로는 효율적 시장 이론(Efficient Market Hypothesis)과 비효율적 시장 이론(Inefficient Market Hypothesis)이 있다. 효율적 시장 이론에 따르면, 주식시장에 새로운 정보가 주어졌을 때, 그 정보가 주식가격에 즉각 반영되기 때문에 투자자는 평균 이상의 수익을 얻는 것이 불가능하다고 하였다[2]. 이와 반대로 비효율적 시장 이론에 따르면, 다수의 투자자들은 불확실한 상황에서 극단적으로 최근의 정보에 높은 비중을 두고 의사 결정을 내리기 때문에 예측이 가능하다고 하였다[3]. 최근 2008년, 2011년 금융위기 및 산타랠리 등 시장의 비효율성을 보여주는 사례가 현실에서 나타나고 있고 지속적으로 엄청난 수익률을 올리고 있는 주식 전문가들이 있다는 측면에서 비효율적인 시장이론이 힘을 얻고 있다[4].

그러나, 이론과 다르게 현실 주식시장에는 정보의 비대칭성이 존재한다. 예를 들어, 주식 전문가들은 고급 정보 또는 개인투자자들이 쉽게 사용하기 어려운 캔들 분석(Candle Analysis) 및 보조지표 분석(Indicator Analysis) 등의 기술적 분석(Technical Analysis)을 통해 주식매매 전략을 취하지만 개인투자자들은 뉴스나 증권방송과 같이 기본적인 정보만을 이용하여 주식매매 전략을 취하고 있다[5]. 이와 같은 정보의 비대칭성은 어떤 정보가 시장에 긍정적인지 부정적인지 명확하게 파악하기 힘들며, 해당 정보가 주어지더라도 분석하는 사람의 주관에 따라 의견이 달라진

다. 이는 일반 개인 투자자들이 주식 시장에서 높은 수익률을 기대하기 어렵게 만드는 이유 중에 하나이다. 하지만 접근하기 쉬운 정보가 주가를 예측할 수 있는 수단으로 사용된다면, 일반 투자자들도 높은 수익률을 기대할 수 있을 것이다. 이러한 이유 때문에 뉴스나 과거 주가 지수처럼 접근하기 쉬운 정보를 이용하여 주가의 등락을 예측하려는 연구가 진행되고 있다. 대표적인 연구 방법으로는 과거 주가 데이터 기반에 시계열 예측 방법[6, 7]과 뉴스 데이터를 활용한 텍스트 마이닝 기반 예측 방법[8, 9, 10] 등이 있다.

시계열 예측 방법은 과거의 주식시장의 데이터에 근거하여 주가를 예측하는 방법으로 과거 데이터를 이용하기 때문에 현재와 과거간의 선형 관계를 찾아내 추세를 확인할 수 있다는 장점이 있다. 텍스트 마이닝 기반 예측 방법은 소셜 데이터(뉴스, 트위터, 게시판)에서 의미있는 단어에 대해 감성수치를 부여하여 주가를 예측하는 방법으로 소셜 데이터를 이용하기 때문에 실적 발표나 금리 인상과 같이 주식시장의 예기치 못한 이벤트에 즉시적 반영과 분석이 가능하다는 장점이 있다.

이러한 배경에서 본 연구에서는 뉴스 기반 텍스트 마이닝에 의한 예측 모형과 KOSPI 데이터를 이용한 ARIMA 모형을 결합하여 주가 등락을 예측하는 모델을 제안한다. 뉴스를 이용한 텍스트 마이닝 방법은 뉴스에서 명사 및 서술어를 추출하여 주가 도메인의 감성사전을 자동으로 구축하고 하루 동안의 발생한 뉴스의 감성수치를 계산하여 예측 모델을 생성한다. 또한, ARIMA 모형을 통해 과거 데이터의 일 수에 따른 예측력을 검증하고 시계열 예측 모델을 생성한다. 이를 통해 제안 방법은 당일 종가 대비 익일 종가 등락을 당일 장 마감 전에 예측한다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 소개하며, 3장에서는 제안하는 감성사전 추출 방법 및 주가 등락 예측 모델 방법을 설명하고, 4장에서 제안 방법론에 대한 실험 결과를 분석하며, 마지막 5장에서는 결론 및 향후 연구를 기술한다.

제 2 장 관련 연구

2.1 시계열 분석 기반 주가 예측

시계열 분석을 이용한 주가 예측 연구는 다음과 같다. ARIMA 모형을 통해 RMSE을 계산하여 실제값과 예측값의 차이를 계산하고 이 오차를 최소로 하는 예측 값을 SVM을 통해 분류하는 방법을 이용한 모델을 제안한 방법[6], 가중 퍼지소속함수 기반 신경망(NEWFM)과 비중복면적 분산측정법을 통해 최근 32일간의 5일 이동 평균값을 웨이블릿 변환을 통해 가장 중요도가 낮은 특징을 자동으로 제거하여 5일 간의 주가 단기 추세의 상승과 하락에 대한 예측 연구[7] 등이 있다. 시계열 분석 기반의 주가 예측은 과거 데이터를 이용하기 때문에 현재와 과거간의 선형관계를 찾아내 추세를 확인할 수 있다는 장점이 있다.

2.1.1 ARIMA 모형

ARIMA 모형은 Box and Jenkins가 고안해낸 방법으로 미래 예측을 수행하는데 많이 사용된다. ARIMA 모형은 AR(Auto Regressive) 부분과 MA(Moving Average) 부분으로 구성되어 있으며, 변수 값의 차이를 따로 모형화할 수 있는 Integrated 부분이 있다[12].

ARIMA 모형에서 AR model은 Autoregressive model의 줄임말로 전 시점의 Y가 현 시점의 Y에 영향을 주는 자기자신에 대한 함수를 뜻한다. AR model을 생성하기 위해서는 잔차가 백색잡음(White Noise)이며, 시계열 데이터가 안정적(Stationary)인지를 검토해야 하는 조건이 있다. 백색잡음은 잔차(u_t)의 평균이 0이고 분산이 σ^2인 동일분포로부터 독립적으로(iid) 얻어진 시계열 데이터를 의미하고 안정적은 각 평균과 분선이 시점에 관계없이 상수이고 t시점과 t-n시점의 공분산(Co-variance)이

t에 관계없이 일정한 조건을 만족하는 것을 의미한다. [식 2-1]은 시간 t가 n일 때의 AR(n)을 나타낸 것이다.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_n Y_{t-n} + u_t \, (u_t : iid \ white noise)$$

[식 2-1] AR(n) 모델

ARIMA 모형에서 MA model은 Moving Average model의 줄임말로 전 시점의 Y가 현 시점의 Y의 에러와 가중치를 이용한 함수를 뜻한다. MA model은 모형의 특성상 AR process처럼 안정적 조건을 확인할 필요가 없지만, 비슷한 조건인 역변환 조건(Invertibility Condition)을 만족해야 한다. [식 2-2]은 MA 모형을 n차인 MA(n)의 모델이다.

$$Y_t = a_1 u_{t-1} + a_2 u_{t-2} + \ldots + a_n u_{t-n} + u_t \, (u_t : iid \ white noise)$$

[식 2-2] MA(n) 모델

ARIMA 모형에서 차분(Integrate)은 안정적 데이터를 만들기 위해 사용되는데, 시간의 흐름에 따라 계열의 평균이 일정하지 않으면 차분을 취하여 정상적으로 만들어야 한다. 만약 한 번 차분을 해서 안정적이게 되는 데이터라면 차분은 1이라고 표현할 수 있다. [식 2-3]은 ARIMA(p,0,q)일 때를 표현한 것이다.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_p Y_{t-p} + u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \ldots + \beta_q u_{t-q}$$

[식 2-3] ARIMA(p,0,q) 모델

2.2 텍스트 마이닝 기반 주가 예측

소셜 데이터를 이용한 텍스트 마이닝 기반 주가 예측 연구는 다음과 같다. 3개월간의 뉴스에서 출현한 단어들에 대해 범용 감성사전과 비교하여 단어별 긍정/부정 극성을 태깅하고 인덱싱 된 분류정보와 스코어링 룰을 이용하여 뉴스의 긍정/부정을 판단하는 기법으로 당일 대비 익일 종가의 등락을 예측하는 지능형 투자의사결정 모델[8], 가장 활발한 종목 게시판에 올라온 글에 대해 감성 분석의 결과와 투자의견을 이용해 실제 결과를 비교하여 글쓴이별로 신뢰도(Trust Value)를 측정해 기업별 당일 대비 익일 종가의 등락을 예측하는 방법[9], 뉴스가 배포된 20분 뒤 주가의 등락이 상승인 뉴스에 대해 명사를 수집하여, Support Vector Regression(SVR) 모델을 통해 학습을 하여 산업 분야별 주가 등락을 예측한 방법[10] 등이 있다.

기존 텍스트 마이닝 기반의 주가 예측 연구들은 소셜 데이터를 이용하기 때문에 실적 발표나 금리 인상과 같이 주식시장의 예기치 못한 이벤트에 즉시적 반영과 분석이 가능하다는 장점이 있다.

2.3 시계열 분석 및 텍스트 마이닝 기반 주가 예측

시계열 분석 및 텍스트 마이닝 기반 주가 예측 연구로는 연구자가 주가 등락에 영향이 있어 보이는 단어를 PR(Probability Ratio) Table로 만든 뒤 SVR(Support Vector Regression)을 통해 가중치를 얻고 그 결과와 Moving Average(MA) 모델을 통해 분야별 지수를 예측한 방법[10], 전일 종가 대비 당일 종가의 등락율이 ±2% 이상인 뉴스를 긍정이나 부정 뉴스로 분류하여 해당 뉴스가 포함한 명사를 Naïve Bayesian 분류기를 이용해 긍/부정으로 분류하고 RSI(Relative Strength Index)를 계산하여 과매수/매도 구간일 때의 가중치를 부여하는 방법[11] 등이 있다.

2.4 관련 연구 요약

시계열 분석을 이용한 주가 등락 예측은 과거 데이터(이동평균선, 보조지표)를 어떻게 활용할 것인지에 대한 연구가 대표적이며[6], 텍스트 마이닝을 이용한 주가 등락 예측은 SNS나 게시판, 뉴스 등에 출현하는 단어에 대해 감성수치를 부여하는 방식으로 접근하는 연구[8, 9, 10]가 있다. 또한, 텍스트 마이닝과 시계열 분석을 혼합한 방법은 뉴스에서 등장하는 단어의 빈도와 과거 데이터의 보조지표를 활용한 연구가 있다[11,12]. 본 연구는 당일 장중 뉴스와 ARIMA(2,0,2) 모형을 이용하여 당일 종가 대비 익일 종가 등락을 예측하는 모델을 제안한다. [표 2-1]은 관련 연구 및 본 연구에 대한 요약이다.

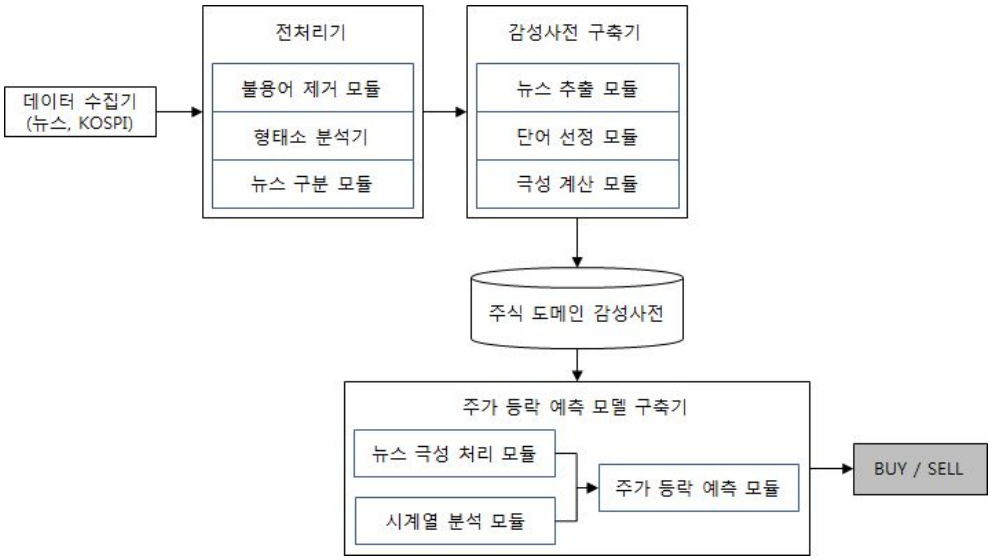
[표 2-1] 관련 연구 요약표

논문명	예측 대상	사용 데이터 범위	알고리즘	데이터 수집 기간	주식시장	정확도(%) or RMSE
[6]	당일 종가	ARIMA(0,1,0)	ARIMA, SVM	2002.10 ~ 2002.03	개별종목 NASDAQ	RMSE 0.4
[8]	당일 시초가 대비 당일 종가 등락	당일 09:00 ~ 당일 15:00 뉴스	임계치	2011.07 ~ 2011.09	KOSPI	54%
	전일 종가 대비 당일 시초가 등락	전일 15:00 ~ 당일 09:00 뉴스				
[9]	당일 종가 등락	전일 09:00 ~ 당일 09:00 종목 게시판	Naïve Bayes, Decision Tree	N/A	개별종목 NASDAQ	54%
[10]	20분 뒤 주가 등락	예측 20분 전 뉴스 데이터	SVR	2005.10 ~ 2005.12	개별종목 S&P 500	65%
[11]	당일 종가	전일 15:00 ~ 당일 09:00 뉴스	SVR	2008.03 ~ 2008.07	섹터종목 상해 A	RMSE 0.38
[12]	당일 시초가 대비 당일 종가 등락	당일 09:00 ~ 당일 15:00 뉴스	RSI, Naïve Bayes	2005.01 ~ 2008.12	개별종목 KOSPI	55%
	전일 종가 대비 당일 시초가 등락	전일 15:00 ~ 당일 09:00				
제안 방법	전일 종가 대비 당일 종가 등락	전일 09:00 ~ 전일 15:00 ARIMA(2,0,2)	Logistic Regression	2010.01 ~ 2014.12	KOSPI	58%

제 3 장 제안 방법

3.1 가설 및 시스템 구조도

본 연구는 '과거 데이터는 주가 예측에 영향을 미친다.' 라는 가설을 바탕으로 과거 뉴스와 시계열 데이터를 이용한 주가 등락 예측 시스템을 제안한다. 제안 주가 등락 예측 시스템은 데이터 수집기, 전처리기, 감성사전 구축기, 주가 등락 예측 모델 구축기로 구성되어 있다. 제안 방법에 대한 전체 시스템의 구조도는 [그림 3-1]과 같다.



[그림 3-1] 주가 등락 예측 시스템 구조도

데이터 수집기와 전처리기는 뉴스와 KOSPI 데이터를 수집하고 수집된 뉴스는 불용어 제거 및 형태소 분석을 통해 단어를 추출하는 역할을 수행하며, 감성사전 구축기는 뉴스 추출 모듈과 단어 선정 모듈을 통해 감성단어들을 선정하고 극성 계산 모듈을 통해 감성단어의 감성수치를

계산하여 주가 예측을 위해 필요한 주식 도메인의 감성사전을 구축하는 작업을 수행한다. 마지막으로, 주가 등락 예측 모델 구축기는 구축된 감성사전을 이용한 예측 모델 및 KOSPI 데이터를 이용한 ARIMA 예측 모델을 결합하여 당일 대비 익일 종가의 등락을 예측한다.

3.2 전처리기

전처리기는 수집된 데이터에서 감성단어를 추출하기 위한 기본적인 작업을 수행하며, 불용어 제거 모듈, 형태소 분석기, 뉴스 구분 모듈로 구성되어 있다.

불용어 제거 모듈은 수집된 데이터가 올바르게 분석되도록 정제하는 모듈이다. 경제 뉴스는 불필요한 광고문구와 숫자, 종목코드와 같은 단어가 많다. 이로 인해 형태소 분석기가 올바르게 동작하지 못하기 때문에 데이터를 정제해주는 작업이 필요하다. [표 3-1]은 불용어 처리 항목 및 불용어 예시를 나타낸 것이다.

[표 3-1] 불용어 처리 예시

불용어 처리 항목	불용어 예시
하나의 음절을 가진 단어 제거	저, 그, 외 등
숫자 혹은 연도와 같은 의미 없는 단어들 삭제	2013년
종목명, 종목코드 삭제	삼성전자, 현대차, 하이닉스
특수문자와 광고문구 삭제	▶ [오늘의 핫 화보]

형태소 분석기는 특정 문장이 주어졌을 때, 단어의 형태론적 구조를 기계적으로 분석해주는 모듈로 본 연구에서는 형태소 분석기를 통해 명사 및 서술어를 추출한다.

뉴스 구분 모듈은 감성사전을 구축하기 위해 필요한 데이터를 설정하여 개장일에 배포된 뉴스가 아닌 경우(주말, 공휴일)를 처리하고 장중 뉴스를 추출하는 모듈이다. 장중 뉴스는 개장 시간(09:00~15:00) 사이에 배포된 뉴스이다.

3.3 감성사전 구축기

감성사전 구축기는 주가 예측을 위해 필요한 주식 도메인의 감성사전을 구축하는 작업을 수행하며, 뉴스 추출 모듈, 단어 선정 모듈, 극성 계산 모듈로 구성되어 있다.

뉴스 추출 모듈은 등락율 임계치 이상의 장중 뉴스만을 추출하는 역할을 수행한다. 단어 선정 모듈은 형태소 분석기를 통해 추출된 단어들 중 특정 품사의 단어를 추출하고 추출된 단어의 빈도를 계산하여 빈도 임계치 이상의 감성단어를 추출하는 역할을 수행한다. 극성 계산 모듈은 전처리 작업이 완료된 장중 뉴스 데이터로부터 감성단어에 대한 극성을 계산하여 감성사전을 구축한다. [식 3-1]은 감성단어에 대한 극성을 구하는 식이다.

$TF_{w_i,t}$ = 날짜 t에 발생한 장중 뉴스에서 단어 w_i의 출현 빈도 수(중복 허용)

$$Ratio_t = \frac{KOSPI(t) - KOSPI(t-1)}{KOSPI(t-1)} * 100$$

$$Score_{word}(w_i) = \frac{\sum_t (TF_{w_i,t} * Ratio_t)}{\sum_t TF_{w_i,t}}$$

[식 3-1] 감성단어 극성 추정 식

[식 3-1]에서 $TF_{w_i,t}$는 날짜 t에 발생한 장중 뉴스에서 단어 w_i의 중복을 허용한 출현 빈도 수이다. $Ratio_t$는 특정 날짜 t의 전일 KOSPI 종가 대비 당일 종가의 등락율을 의미하는 변수이다. $Score_{word}(w_i)$는 단어 w_i가 출현한 날짜 t에서 등락율 가중치 $Ratio_t$에 w_i의 출현 빈도 수 $TF_{w_i,t}$를 가중평균하여 단어 w_i의 감성수치를 계산하는 수식이다. $Score_{word}(w_i)$는 $-\infty \sim \infty$의 범위를 가지며 ∞에 가까울수록 강한 상승을 의미한다. 감

성단어의 감성수치를 추정하는 예시는 [표 3-2]와 같다.

[표 3-2] 극성 값 추정 예시(공매도)

인덱스	뉴스 시간	전일 종가 대비 당일 종가 등락률	본문	$TF_{w_i,t}$	$Ratio_t$
1	2013-12-02 14:28	-1.05	연말 배당수익을 노린 자금이 유입되는 데다 그동안의 **공매도**를 갚아나가는 대차상환도 수급에 힘이 될 것이라는 분석이다	1	-1.05
2	2013-11-27 15:45	0.84	차거래 잔고가 줄어들 것으로 예상되면서 그간 **공매도**가 집중됐던 종목들에 대한 쇼트커버링재매입 효과가 기대돼서다	-	-
3	2013-07-16 09:28	1.13	국내 주식시장에서 대차거래가 계속 증가하고 있다.	-	-
4	2013-07-03 14:04	0.79	투자자들도 대차잔고 고공행진에 바짝 긴장하고 있다 **공매도**로 이어질 수 있기 때문이다.하지만 **공매도**가 주식시장에 미치는 영향은 제한적이다.	2	0.79
5	2011-11-09 09:10	2.77	**공매도** 금지 조치 이전에 비해 12% 줄었다	2	2.77
6	2011-11-09 10:39	2.77	**공매도**가 금지는 주식시장에 미치는 영향이 제한적이었기 때문이다.		

$$\sum_t TF_{\text{공매도},t} = 5$$

$$Score_{word}(\text{공매도}) = \frac{-1.05 + (2*0.79) + (2*2.27)}{5} = 1.014$$

[표 3-2]에서 “공매도”라는 단어는 6개의 뉴스 중에서 5개의 뉴스(인덱스 1, 2, 4, 5, 6번)에서 발생하였다. 그러나, 인덱스 2번의 뉴스는 장전에 발생한 뉴스이므로 제외된다. 하나의 뉴스에 단어가 여러번 발생하는 경우인 인덱스 4번의 뉴스는 공매도라는 단어가 2번 출현하였기 때문에 $TF_{w_i,t}$는 2가 되고 장중에 발생한 뉴스들에서 단어가 여러번 발생하는 경우인 인덱스 5번과 6번은 공매도라는 단어가 2번 발생하였기 때문에 $TF_{w_i,t}$는 2가 된다. 이와 같은 방식으로 $Score_{word}$(공매도)를 계산하면, $TF_{w_i,t}*Ratio_t$의 평균 값인 1.014가 된다. [표 3-3]은 감성사전 구축기의 극성값을 계산하는 수식을 통해 생성된 감성사전의 일부를 나타낸 것이다.

[표 3-3] 생성된 감성사전 예시

감성단어	단어 속성	$Score_{word}(w_i)$	$\sum_t TF_{w_i,t}$	$TF_{w_i,t}*Ratio_t$
공매도	명사	0.542	285	154.58
주가하락	명사	0.414	120	49.68
금리인하	명사	0.372	183	68.05
전략적	명사	0.361	165	59.55
괴리율	명사	0.295	123	36.31
불균형	명사	0.242	171	41.38
엔저	명사	0.236	221	52.07
훈풍	명사	0.226	176	39.72
낙폭과대주	명사	0.221	115	25.36
제한되다	서술어	0.212	208	44.13
구성되다	서술어	0.203	246	49.85
상향조정	명사	0.173	121	20.91
경기선행지수	명사	0.17	554	93.91
수주하다	서술어	0.165	155	25.51
갖추다	서술어	0.16	405	64.65

3.4 주가 등락 예측 모델 구축기

주가 등락 예측 모델 구축기는 주가 등락을 예측하기 위한 모델을 생성하는 역할을 수행하며, 뉴스 극성 처리 모듈과 시계열 분석 모듈, 주가 등락 예측 모듈로 구성되어 있다.

뉴스 극성 처리 모듈은 특정 날짜의 장중 뉴스에 대한 감성수치를 계산한 다음 특정 날짜에 대한 감성수치를 계산한다. [식 3-2]는 특정 뉴스에 등장하는 감성단어들의 극성 평균으로 특정 뉴스에 대한 감성수치를 계산하는 식이다. [식 3-2]에서 구축된 감성사전의 감성단어들의 집합을 SW(Sentiment Word)라고 하고 특정날짜 t의 k번째 뉴스에서 추출한 단어들의 집합을 $News_{t,k}$라고 할 때, 특정 뉴스 k의 감성수치인 $Score_{News}(News_{t,k})$는 SW와 $News_{t,k}$에 동시에 출현하는 단어들의 감성수치의 평균으로 계산된다. [식 3-3]은 특정 날짜에 발생한 뉴스들로부터 해당 날짜에 대한 감성수치를 계산하는 과정이다.

$$SW(Sentiment\,Word) = \{w_1, \cdots, w_n\}$$
$$News_{t,k} = \{w_{t,k,1} \cdots, w_{t,k,n}\}$$
$$Score_{News}(News_{t,k}) = \frac{\sum_{w_i \in SW \cap News_{t,k}} Score_{word}(w_i)}{|\,SW \cap News_{t,k}\,|}$$

[식 3-2] 특정 뉴스 k에 대한 감성수치

$$n_t = \text{특정 날짜 } t\text{에 발생한 장중 뉴스의 개수}$$
$$Score_{Day}(t) = \frac{\sum_k Score_{News}(News_{t,k})}{n_t}$$

[식 3-3] 특정 날짜 t에 대한 감성수치

하루동안의 감성수치인 $Score_{Day}(t)$는 $Score_{News}(News_{t,k})$의 평균을 이용하여 계산된다. [표 3-5]는 감성사전의 예시를 나타낸 것이며, [표 3-4]은 감성사전의 SW를 이용하여 하루동안의 감성수치인 $Score_{Day}(t)$를 계산하는 방법의 예시이다.

[표 3-4] $Score_{Day}(t)$ 계산 방법

뉴스 시간	본문	$Score_{News}(News_{t,k})$
2013-12-02 14:28	외국인은 올 국내증시 순매도 케이만아일랜드1조4331억 주식 보유금액은 되레 늘어 **공매도**로 **주가하락** 부추겨 스페인 등 유럽서도 활약 국채 공매도로 불안감 증폭 최근 증시 변동성을 키운 주범으로 헤지펀드가주목받고 있다.	$Score_{word}(\text{공매도}) = 0.542$ $Score_{word}(\text{주가하락}) = 0.414$ $\frac{0.542+0.414}{2} * 100 = 47.8$
2013-12-02 11:45	국제금융시장의 불확실성 확대로 외국인 증권자금 유입이 **제한되면서**원달러 환율이 다소 높은 수준을 보일 것이라며 다만 미 연준 **금리인하**가 금융시장의 큰 충격 없이 시작된다면 불확실성이 완화되며 경상수지 흑자 지속 등 견실한 기초경제여건이 부각돼 원화 절상압력이 점차 커질 것이라고 전망했다	$Score_{word}(\text{제한되다}) = 0.212$ $Score_{word}(\text{금리인하}) = 0.372$ $\frac{0.212+0.372}{2} * 100 = 29.2$
2013-12-02 09:28	판매채널 다양화 관점에서 ICT 기업과 사전적 제휴에 적극적으로 나서고 **전략적**으로 기존 결제 생태계에 참여하는 등 ICT 기업과의 프로세싱 협력을 구축하리란 예상이다	$Score_{word}(\text{전략적}) = 0.361$ $\frac{0.361}{1} * 100 = 36.1$
$Score_{Day}(2013-12-02) = \frac{47.8+29.2+36.1}{3} = 37.7$		

2013-12-02에 장중에 배포된 뉴스는 총 3개이고 감성사전 [표 3-5]에 존재하는 감성단어들의 감성수치를 이용하여 특정 뉴스의 감성수치를 계

산하면, 1번 뉴스는 공매도가 1번, 주가하락이 1번 나왔기 때문에 감성수치의 값인 $Score_{News}(News_{t,k})$는 47.8이라는 값을 가진다. 이와 같은 방식으로 2번 뉴스, 3번 뉴스의 감성수치를 계산하면, 각각 29.2, 36.1이 산출된다. $Score_{Day}(2013-12-02)$는 하루동안 발생한 장중 뉴스의 극성값의 평균이기 때문에 위에서 구한 47.8, 29.2, 36.1의 평균 값인 37.7이라이 계산된다.

[표 3-5] SD(Sentiment Dictionary) 예시

감성단어	단어 속성	감성수치
공매도	명사	0.542
주가하락	명사	0.414
금리인하	명사	0.372
전략적	명사	0.361
낙폭과대주	명사	0.221
제한되다	서술어	0.212

시계열 분석 모듈은 KOSPI 데이터를 사용하여 모형의 예측력을 검증하고 AR, MA, Integeration을 결정하고 결정된 ARIMA 모형을 이용하여 예측 확률 값을 계산한다. [식 3-4]는 AR이 p이고 MA가 q이며 Integeration이 0인 ARIMA 모형을 나타낸 것이다.

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} + \ldots + a_p Y_{t-p} + u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \ldots + \beta_q u_{t-q}$$

[식 3-4] ARIMA(p,0,q) 모형

주가 등락 예측 모듈은 [식 3-2]와 [식 3-3]을 통해 계산된 특정일에 대한 감성수치와 [식 3-4]에서 결정된 ARIMA 모형의 확률 값인

$ARIMA_{preds}(t)$를 이용하여 주가 등락 예측 모델(Logistic Regression)을 생성하고 당일 대비 익일 종가의 상승, 하락을 예측한다. [식 3-5]는 본 연구에서 제안하는 주가 등락 예측 모델로 x_1은 뉴스를 통해 추출된 감성사전을 이용하여 하루동안의 감성수치를 계산한 결과 값이며, x_2은 ARIMA 모형의 예측력을 검증을 통해 결정된 ARIMA 모델의 예측 확률 값이다.

$$x_1 = Score_{Day}(t)$$
$$x_2 = ARIMA_{preds}(t)$$
$$In(\frac{p}{1-p}) = a + b_1 x_1 + b_2 x_2$$

[식 3-5] 주가 등락 예측 모델

제 4 장 실험 및 결과

4.1 실험 데이터

본 연구에서는 2010년 1월부터 2014년 12월까지 '네이버>증권>뉴스>주요뉴스' 탭에 있는 경제 뉴스(총 76,300건)와 '한국증권거래소(KRX)>국내지수>일자별 지수' 탭에 있는 일별 KOSPI 데이터(총 1,239건)를 수집하였다. 뉴스 및 KOSPI 데이터 각각에 대한 수집 결과는 [그림 4-1] 및 [그림 4-2]와 같다.

(코증시 오후) 상하이종합 강세..일본 휴장	2013-12-31 17:19 이데일리	IPO 랠리에 상승 마감 김유성 기자 아시아 주요국 증시..랠리 상승 마감 김유성
(뉴욕전망대) 한산한 시장	2013-12-31 17:19 이데일리	염지현 기자 올해 마지막 장이 열리는 31일현지시간에는 염지현 기자 올해 마지막
정유株 옥자구간 진입...SK이노베이션 빛볼까	2013-12-31 16:30 매일경제	경기 회복 조짐으로 석유화학제품 수요가 많아지면서 휴부경기 회복 조짐 석유화학
증시 폐장일 오후 4시 '올빼미 공시' 속출	2013-12-31 15:50 파이낸셜뉴스	2시간30분간 102개 공시 하루 공시 절반가량 차지 부정적 시간 분간 개 공시 하루
펀드매니저들이 바라본 2014년 증시	2013-12-31 15:50 파이낸셜뉴스	올 증시 작년보다 낫다 화학조선 유망 주의를 2014년엔 2증시 작년 화학조선 유망
국민연금, 45개 종목 지분율 10% 이상 확대	2013-12-31 15:19 머니투데이	최경민 기자2013년 9월 공시의무인 10를 완화삼성물산 §최경민 기자 년 월 공시
IPO 늦출수록 힘들어진다..'실적 부진에 등급 하향?	2013-12-31 14:50 이데일리	SK루브르컨즈 등급전망 안정적으로 하향 LG실트론대성·루브르컨즈 등급전망 안
올해 코스피 '쥐락펴락', NAVER·삼성전자 주도	2013-12-31 13:36 이데일리	코스피 기여도 NAVER삼성전자 오희나 기자 올해 코스ㅍ코스피 기여도 삼성전자
현대차, 감오년 青馬처럼 달릴 수 있을까	2013-12-31 12:01 이데일리	김도년 기자 유럽 경기가 서서히 회복되는 가운데 삼성전김도 년 기자 유럽 경기
영화관객 2억명 시대..영화株 주가는 웨이래	2013-12-31 11:35 이데일리	지난 18일 누적관객수 2억41만명세계 5번째 2억명 돌파 지난 일 누적관객수 억
그룹 효자에서 천덕꾸러기 전락한 '건설'	2013-12-31 10:45 머니위크	건설부동산 경기침체에 건설사를 계열사로 둔 모기업들괴건설부동산 경기침체 건
(코증시 오전)상하이종합 약세..일본 휴장	2013-12-31 11:21 이데일리	12월 제조업PMI 발표 앞두고 관망세 성문재 기자 올해 월 제조업 발표 관망세
[원자재 시황]원유·금 등 일제히 하락	2013-12-31 09:15 이데일리	유가 재고 증가랠리 부담에 1 멸어져 경기 개선 판단에 유가 재고 증가랠리 부담
'낙관과 비관 사이' 코스피는 어디로	2013-12-31 08:05 머니위크	일찍이 없었던 밑이 잃어나는 미증유의 시대다 불확실성 ·밑이 미증유 시대 불확실

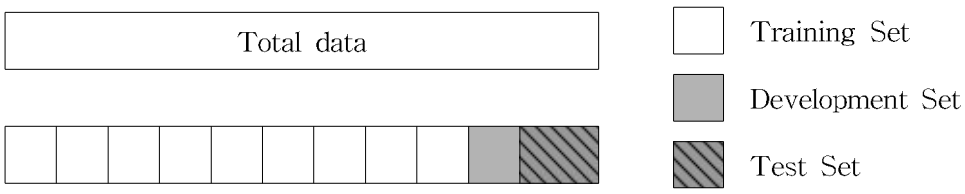
[그림 4-1] 뉴스 수집 결과

인덱스	일자	현재지수	대비	등락률	바이너리	시가지수	고가지수	저가지수
1	2010-01-04	1696.14	13.37	0.79	1	1681.71	1696.14	1681.71
2	2010-01-05	1690.62	-5.52	-0.33	0	1701.62	1702.39	1686.45
3	2010-01-06	1705.32	14.7	0.87	1	1697.88	1706.89	1696.1
4	2010-01-07	1683.45	-21.87	-1.28	0	1702.92	1707.9	1683.45
5	2010-01-08	1695.26	11.81	0.7	1	1694.06	1695.26	1668.84
6	2010-01-11	1694.12	-1.14	-0.07	0	1700.79	1705.73	1694.12
7	2010-01-12	1698.64	4.52	0.27	1	1695.83	1701.16	1683.29
8	2010-01-13	1671.41	-27.23	-1.6	0	1683.51	1687.58	1671.11
9	2010-01-14	1685.77	14.36	0.86	1	1680.68	1692.78	1677.46
10	2010-01-15	1701.8	16.03	0.95	1	1694.65	1704.43	1686.12
11	2010-01-18	1711.78	9.98	0.59	1	1696.14	1716.62	1688.89
12	2010-01-19	1710.22	-1.56	-0.09	0	1719.41	1723.22	1706.73
13	2010-01-20	1714.38	4.16	0.24	1	1723.01	1723.01	1708.58
14	2010-01-21	1722.01	7.63	0.45	1	1700.53	1722.01	1695.18
15	2010-01-22	1684.35	-37.66	-2.19	0	1696.21	1706.09	1665.6
16	2010-01-25	1670.2	-14.15	-0.84	0	1662.77	1681.83	1660.58
17	2010-01-26	1637.34	-32.86	-1.97	0	1670.47	1671.66	1626.98

[그림 4-2] KOSPI 데이터 수집 결과

4.2 평가 방법

제안 방법의 평가를 위해 Training Set, Development Set, Test Set을 구분하였다. Training Set은 주가 등락 예측 모델을 학습하는데 사용되며, Development Set은 파라미터별 실험을 진행하여 예측 검증하는 데이터이고, Test Set은 Development Set에서 가장 좋았던 파라미터를 이용하여 예측하는 데이터이다. [그림 4-3]은 학습 데이터 및 평가 데이터의 구성 방법을 나타낸 것이다.



[그림 4-3] 학습 데이터 및 평가 데이터 구성 방법

또한, [표 4-1]은 학습 데이터 및 평가 데이터를 나타낸 것이다.

[표 4-1] 학습 데이터 및 평가 데이터

항목	기간	수집된 뉴스 건 수	전처리를 통한 장중 뉴스 건 수	KOSPI 일 수
Traning Set	2010.01 ~ 2013.07	51,638건	23,504건	892일
Development Set	2013.08 ~ 2013. 12	3,937건	1,323건	102일
Test set	2014.01 ~ 2014.12	20,725건	4,030건	245일

4.3 평가 척도

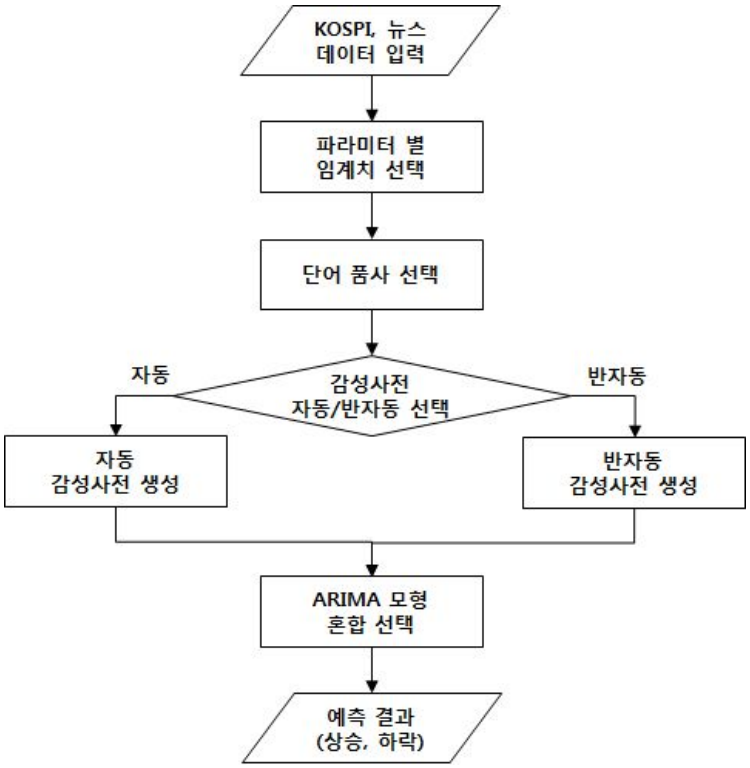
주가 등락 예측 모델에 대한 평가 척도로는 정확도(Accuarcy)를 사용하였으며, [식 4-1]과 같이 계산된다. [식 4-1]에서 정확도는 주가 등락 예측 모델에서 상승과 하락으로 예측한 결과 중에서 올바르게 예측한 비율로 정의된다. TP(True Positive)는 실제 상승인 것을 예측 모델(Logistic Regression)이 상승으로 분류한 것을 의미하며, FP(False Positive)는 실제 하락인 것을 예측 모델이 상승으로 분류한 것을 의미한다. 또한, FN(False Negative)는 실제 하락인 것을 예측 모델이 상승으로 분류한 것을 의미하며, TN(True Negative)는 실제 하락인 것을 예측 모델이 하락으로 분류 한 것을 의미한다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

[식 4-1] 정확도 계산식

4.4 실험 방법

본 연구에서는 다양한 방법을 이용하여 주가 등락의 정확도 비교 실험을 실시하였다. 감성사전을 생성하는 데 이용하는 파라미터별, 품사별로 예측 정확도를 비교하고, 수동/자동/반자동으로 생성된 감성사전을 이용한 예측 정확도를 비교하며, 감성사전 예측 모델과 ARIMA 예측 모델을 혼합했을 때의 정확도를 비교하여 평가하였다. [그림 4-4]는 실험 수행 방법에 대한 흐름도이다.



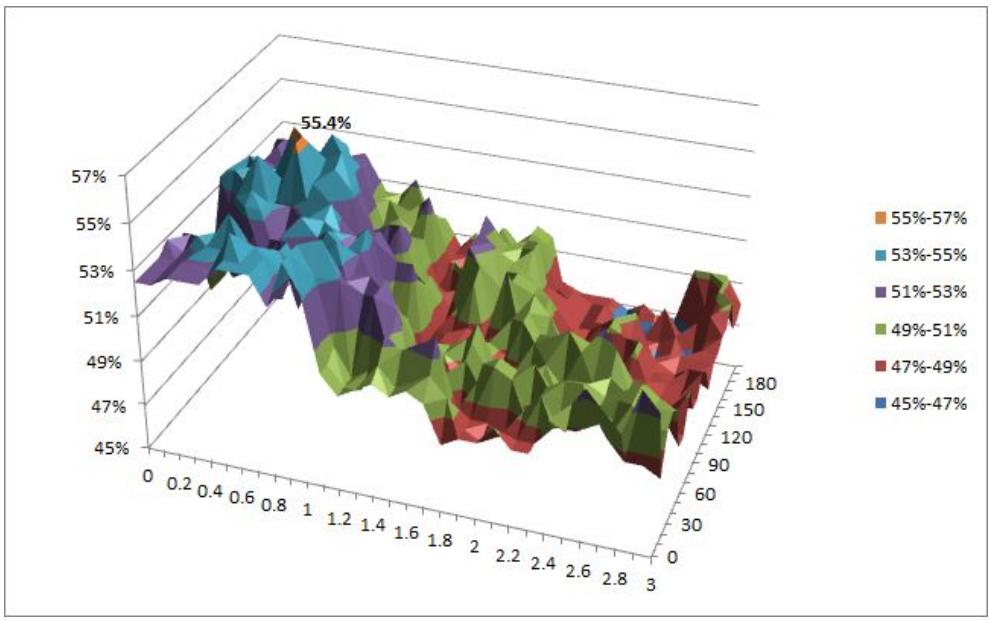
[그림 4-4] 실험 수행 방법 흐름도

4.5 제안 방법 실험

4.5.1 파라미터 임계치별 감성사전에 따른 예측 정확도

감성사전은 등락율 임계치와 빈도 임계치에 따라 단어와 감성수치 및 예측 정확도가 달라진다. 본 절에서는 등락율 임계치와 빈도 임계치에 따른 주가 등락 예측 정확도를 비교하였다.

[그림 4-5]는 등락율 임계치와 빈도 임계치별 예측 모델의 정확도를 비교한 것이다. 등락율 임계치는 0.2~0.5% 사이의 낮은 등락율 임계치를 가지면서 빈도 임계치가 100~120번 사이의 높은 빈도 임계치를 가질 때의 구축된 감성사전을 이용한 예측 모델의 예측 정확도가 가장 높았고, 등락율 임계치가 증가할수록 예측 정확도가 낮아지는 경향을 보였다.

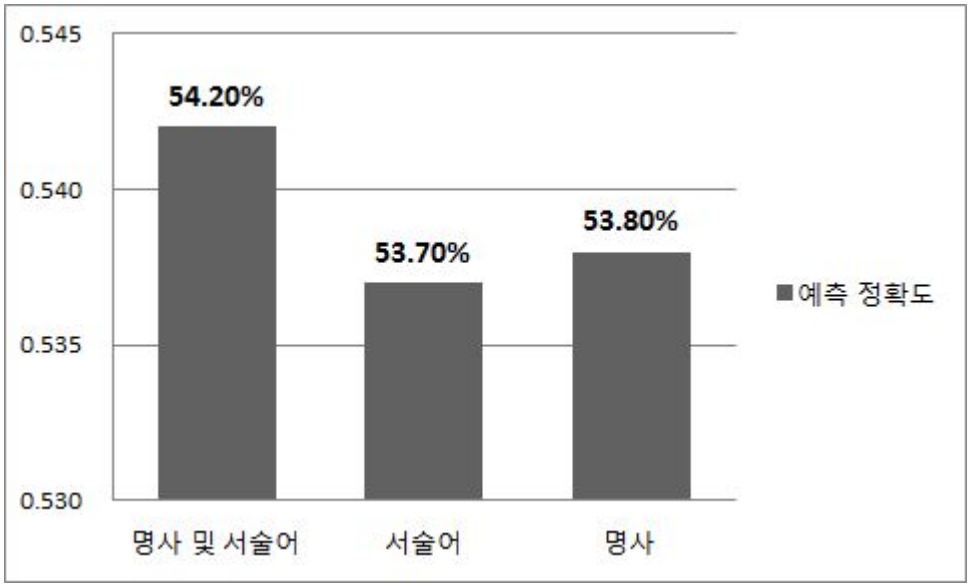


[그림 4-5] 등락율 임계치 및 빈도 임계치 변화에 따른 정확도 변화

4.5.2 단어 품사별 감성사전에 따른 예측 정확도

감성사전은 단어들로 구성되기 때문에, 단어의 품사별로 예측 정확도가 달라질 수 있다. 이에 대해 본 절에서는 4.5.1의 실험을 통해 예측 정확도가 가장 높았던 등락율 임계치(0.5%)와 빈도 임계치(110번)으로 고정하고 단어의 품사(명사, 서술어, 명사 및 서술어)에 따라 구축된 감성사전을 이용한 예측 모델의 예측 정확도를 비교하였다.

[그림 4-6]은 이에 대한 예측 정확도 결과를 나타낸 것으로 명사 및 서술어를 모두 이용한 감성사전에 따른 예측 모델이 54.2%로 가장 높은 정확도를 보였고 서술어를 이용한 감성사전에 따른 예측 모델이 53.7%로 가장 낮은 정확도를 보였다.



[그림 4-6] 단어 품사별 감성사전에 따른 예측 정확도

4.5.3 수동/자동/반자동 감성사전 구축에 따른 예측 정확도

본 절에서는 4.5.2절의 결과에 의해 생성된 품사별 감성사전을 수동/자동/반자동으로 구축한 경우의 예측 정확도를 비교한다.

수동 감성사전은 4.5.2절의 결과에 의해 자동으로 생성된 품사별 감성사전에서 연구자가 임의로 의미가 있어 보이는 단어들에 대해 -1(부정적인 단어), 1(긍정적인 단어)로 감성수치를 부여하고 의미가 없어 보이는 단어들은 감성단어에서 제외한 감성사전을 이용하는 방법이다. [표 4-2]는 수동 감성사전의 예시를 나타낸 것이며 “승인하다”와 “소비심리”는 연구자에 의해 제거된 단어이다.

[표 4-2] 수동 감성사전 예시

감성단어	단어 속성	감성수치
폭락하다	서술어	-1
승인하다	서술어	-
호조세	명사	1
수주하다	서술어	1
소비심리	명사	-
…	…	…

반자동 감성사전은 4.5.2절의 결과에 의해 자동으로 생성된 품사별 감성사전에서 연구자가 임의로 의미가 있어 보이는 단어들은 자동으로 생성된 감성수치를 그대로 유지하고 의미가 없어 보이는 단어들은 감성단어에서 제외한 감성사전을 이용하는 방법이다. [표 4-3]은 반자동 감성사전의 예시를 나타낸 것이며, “이집트”와 “해소하다”는 연구자에 의해 제외된 단어이다.

[표 4-3] 반자동 감성사전 예시

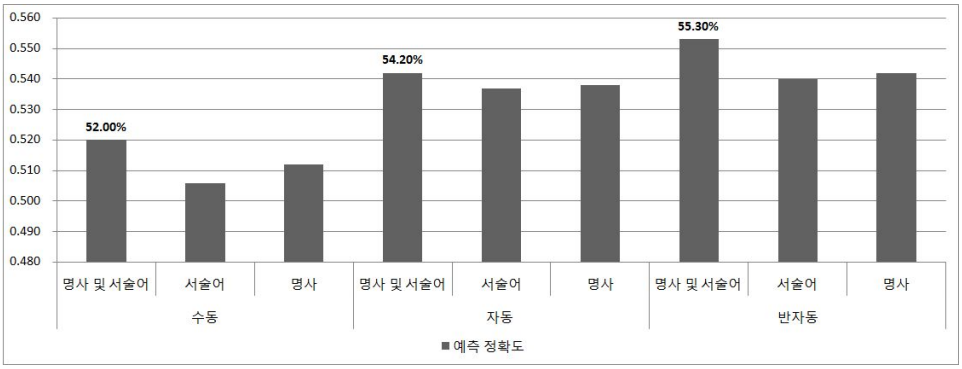
감성단어	단어 속성	감성수치
이집트	명사	0.207
해소하다	서술어	0.145
불안심리	명사	-0.252
이탈하다	서술어	0.223
어닝쇼크	명사	-0.031
…	…	..

위와 같은 방법으로 수동/반자동/자동 감성사전을 구축하였을 때, 수동과 반자동 감성사전의 감성단어 수는 235개로 동일하고, 자동 감성사전의 감성단어 수는 861개로 구성되어 있다. [표 4-4]는 수동/반자동/자동 감성사전의 감성단어 수를 나타낸 것이다.

[표 4-4] 수동/반자동/자동 감성사전의 감성단어 수

	감성 단어 수
수동 감성사전	235개
반자동 감성사전	235개
자동 감성사전	861개

[그림 4-7]은 수동/반자동/자동 감성사전을 이용한 예측 정확도 결과를 표로 정리한 것이다. 그 결과 수동/자동으로 감성사전을 구축한 것에 비해 반자동으로 감성사전을 구축했을 때 품사별 감성사전 예측 모델의 예측 정확도가 증가하였다. 그 중에서 명사 및 서술어를 이용하여 감성사전을 만든 경우가 54.2%에서 55.3%로 가장 높은 예측 정확도를 보였다.



[그림 4-7] 수동/자동/반자동 감성사전에 따른 예측 정확도

또한, 예측 정확도가 가장 높았던 명사 및 서술어를 이용한 반자동 감성사전을 살펴보면, [표 4-5]와 [표 4-6]과 같다. [표 4-5]는 명사 및 서술어를 이용한 반자동 감성사전에서 감성수치가 상위 20%인 감성단어를 나열한 것이고 [표 4-5]는 하위 20%인 감성단어를 나열한 것이다.

[표 4-5] 반자동 감성사전의 감성수치 상위 20%

감성단어	단어 속성	감성수치	빈도 수	가중치
공매도	명사	0.542	285	154.58
주가하락	명사	0.414	120	49.68
금리인하	명사	0.372	183	68.05
전략적	명사	0.361	165	59.55
괴리율	명사	0.295	123	36.31
불균형	명사	0.242	171	41.38
엔저	명사	0.236	221	52.07
훈풍	명사	0.226	176	39.72
낙폭과대주	명사	0.221	115	25.36
제한되다	서술어	0.212	208	44.13
구성되다	서술어	0.203	246	49.85
상승폭	명사	0.202	1,285	259.52
저성장	명사	0.175	175	30.55
상향조정	명사	0.173	121	20.91
경기선행지수	명사	0.17	554	93.91
수주하다	서술어	0.165	155	25.51
갖추다	서술어	0.16	405	64.65
반등세	명사	0.156	258	40.32
순매도세	명사	0.151	135	20.33
상승탄력	명사	0.146	376	54.94
소비심리	명사	0.145	275	39.96
가격제한폭	명사	0.141	197	27.83
선방하다	서술어	0.139	293	40.67
소비자신뢰지수	명사	0.135	209	28.16
매수우위	명사	0.12	409	48.9
후퇴하다	서술어	0.118	165	19.53
인정하다	서술어	0.117	218	25.58
초저금리	명사	0.109	137	14.91

[표 4-6] 반자동 감성사전의 감성수치 하위 20%

감성단어	단어 속성	감성수치	빈도 수	가중치
경기침체	명사	-0.072	508	-36.6
피로감	명사	-0.077	130	-10.07
조언하다	서술어	-0.082	1,509	-123.49
승인하다	서술어	-0.085	181	-15.42
급락세	명사	-0.088	588	-51.45
호조세	명사	-0.089	263	-23.34
어닝쇼크	명사	-0.115	346	-39.66
경계감	명사	-0.115	129	-14.89
어닝시즌	명사	-0.116	562	-64.98
안되다	서술어	-0.125	152	-19
불황	명사	-0.13	304	-39.62
증가폭	명사	-0.138	164	-22.63
디폴트채무불이행	명사	-0.141	110	-15.5
엿보다	서술어	-0.149	197	-29.38
강등	명사	-0.158	947	-149.69
주춤하다	서술어	-0.193	166	-32.03
급락	명사	-0.2	1,281	-256.31
물가상승률	명사	-0.202	247	-49.97
흑자전환	명사	-0.236	228	-53.74
원화가치	명사	-0.236	119	-28.13
강세장	명사	-0.241	161	-38.88
불안심리	명사	-0.252	135	-34
디폴트	명사	-0.285	815	-232.12
저가매수	명사	-0.355	355	-126.06
증액	명사	-0.394	285	-112.38
국가신용등급	명사	-0.419	269	-112.75
패닉	명사	-0.516	168	-86.72
더블딥이중침체	명사	-0.746	114	-85.02

그러나, [표 4-5]와 [표 4-6]의 단어들을 살펴보면 “공매도”, “주가하락” 처럼 의미적으로 감성수치가 음수가 적절하다고 생각되는 단어들이 상위 단어로 뽑힌 경우가 있었다. 그러나, 실제 뉴스 데이터를 살펴보면, [표 4-7]과 같이 상위 또는 하위로 뽑힌 이유에 대한 설명이 가능하다.

[표 4-7] 감성사전의 단어별 상/하위 설명

감성단어	감성수치	설명1	설명2
공매도	0.542	국내 공매도 금지 조치	대차거래 감소로 인해 공매도는 제한적
주가하락	0.414	주가하락은 매수의 기회	주가하락은 과대하다
괴리율	0.295	저평가 여부를 알려주는 지표	
제한되다	0.212	추가적인 차익거래 매도로 제한될 가능성이 높다	지수 상승은 제한되나 전반적인 불확실성 해소
저성장	0.175	장기불황과 저성장 탈피	저성장에 대한 인정을 해야 의미 있는 반등
순매도세	0.151	순매도세를 보였지만 막판에 매도 규모를 줄였다	순매도세를 보였다 그러나
어닝시즌	-0.116	다음 분기 어닝시즌을 기대해보자	어닝시즌에 접어든다는 사실
증가폭	-0.138	증가폭 둔화	
물가상승률	-0.202	물가상승률 둔화	물가상승률이 낮아질것으로 예상
흑자전환	-0.236	흑자전환이 가능할 것이라고 전망했다	내년 흑자전환 기대
원화가치	-0.236	주가가 급락하고 원화가치가 뚝 떨어진다	원화가치가 폭락
강세장	-0.241	강세장까지는 내년 중반이후를 기다려야 한다	강세장으로 향하기 위해선 주도주 출현이 필요하다
저가매수	-0.355	섣불리 저가 매수에 들어갔다가는 오랜 기간 고통	현 지수대라면 저가매수 나설만 지지선 자체는 의미가 없지만
증액	-0.394	미국 부채한도 증액	IMF, 구제금융 증액 요청 소식

4.5.4 ARIMA 모형 예측력 검증 및 선택

ARIMA 모형의 예측력을 검증하기 위해 AR model의 Stationarity Condition의 판별 여부를 확인하였다. [그림 4-8]은 KOSPI 데이터에 대해 Dickey-Fuller의 단위근 검정을 실시한 것이다. 그 결과 p-value 값이 0.1보다 작으므로 불안정하다는 귀무가설을 기각하기 때문에, KOSPI 지수는 Integeration을 진행하지 않아도 된다는 것이 검증되었다.

또한, ARIMA 모형을 선택하기 위해 R에 사용하는 함수 중 auto.arima를 사용하여 자동으로 최적의 ARIMA 모형을 생성하였다. 그 결과 [그림 4-9]처럼 ARIMA 모형에서 AR model이 2의 차수를 갖고 MA model이 2의 차수를 갖는 모형이 생성되었다.

```
> adf.test(train$value)

        Augmented Dickey-Fuller Test

data:  train$value
Dickey-Fuller = -3.2877, Lag order = 8, p-value = 0.07281
alternative hypothesis: stationary
```

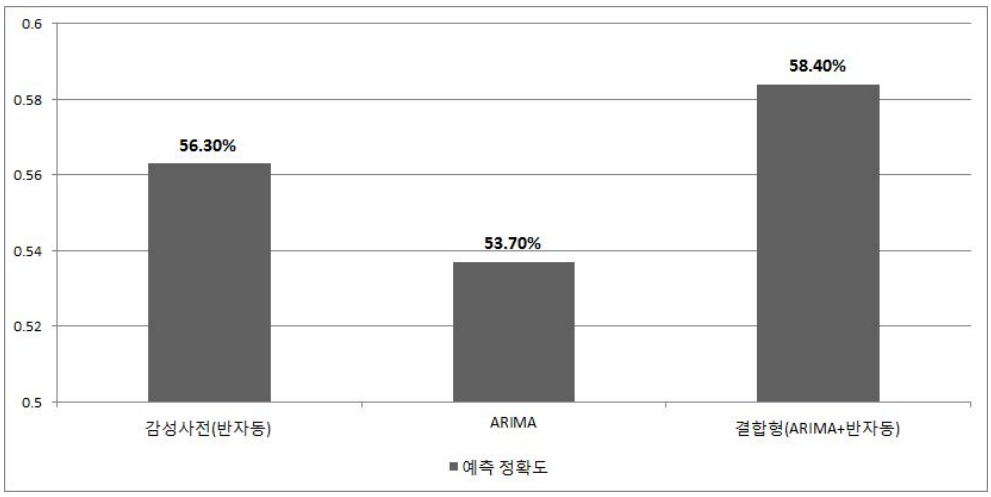
[그림 4-8] ADF 검증 결과

```
Series: train.stockRet
ARIMA(2,0,2) with zero mean

Coefficients:
          ar1      ar2     ma1     ma2
      -0.3271  -0.9179  0.3361  0.9884
s.e.   0.0399   0.0358  0.0197  0.0230

sigma^2 estimated as 0.7054:  log likelihood=-615.63
AIC=1241.25   AICc=1241.38   BIC=1262.27
```

[그림 4-9] 최적의 ARIMA 모형 선택 결과

4.5.5 예측 모델별 예측 정확도

4.5.3절의 실험을 통해 명사 및 서술어의 품사를 갖는 반자동 감성사전의 예측 정확도가 가장 높았기 때문에 본 연구에서는 이를 이용한 예측 모델과 ARIMA 모형 검증을 통해 생성된 ARIMA(2,0,2) 예측 모델, 그리고 두 모델을 결합한 결합형 예측 모델에 대한 예측 정확도 비교하였다. 그 결과 결합형 예측 모델이 58.4%로 가장 높은 예측 정확도를 보였다. [그림 4-10]은 예측 모델별 예측 정확도를 보여준다.



[그림 4-10] 예측 모델별 예측 정확도 비교

[표 4-8]은 결합형 예측 모델에 대해 실제 클래스 대비 예측 클래스의 상승/하락의 예측 건 수를 표로 정리한 것으로 실제 하락을 하락으로 예측한 경우 보다 실제 상승을 상승으로 예측한 경우가 더 높았으며, 실제 하락을 상승으로 예측한 경우가 실제 상승을 하락으로 예측한 것보다 더 높았다.

[표 4-8] 결합형 예측 모델의 예측 결과표

	예측 하락	예측 상승
실제 하락	58	67
실제 상승	34	86

4.5.6 강한 상승, 강한 하락의 예측 정확도

추가적으로 본 제안 방법을 이용하여 코스피 지수의 큰 폭의 상승이나 큰 폭의 하락인 경우를 예측하는 연구를 진행하였다. 강한 상승은 전날 종가 대비 당일 종가의 등락이 0.5%이상 상승한 경우를 의미하고 강한 하락은 전날 종가 대비 당일 종가의 등락이 -0.5%이하인 경우로 설정하였다. 그 결과 실제 강한 상승/하락을 강한 상승/하락으로 예측한 경우가 각각 3번, 2번이었으며, 실제 강한 상승/하락을 상승/하락으로 예측한 경우까지 포함하면 약 36.8%의 예측 정확도가 나타났다. [표 4-9]는 강한 상승, 강한 하락에 따른 예측 정확도를 나타낸 것이다.

[표 4-9] 강한 상승, 하락에 따른 예측 정확도

		예측 클래스			
		강한 상승	상승	하락	강한 하락
실제 클래스	강한 상승	3	18	29	0
	상승	1	36	35	3
	하락	1	24	49	5
	강한 하락	1	8	30	2

4.6 비교 연구와의 비교실험 및 평가

비교 연구[6]는 전날 15:00 ~ 당일 09:00에 배포된 뉴스로 부터 감성사전을 구축하여 전일 종가 대비 당일 시초가의 등락을 예측하였고, 당일 09:00 ~ 당일 15:00에 배포된 뉴스로부터 감성사전을 구축하여 당일 시초가 대비 당일 종가 등락을 예측하였다. 본 연구에서 제안한 결합형 예측 모델을 [6]의 데이터와 예측 범위로 변경하여 비교 실험을 진행하였다. 그 결과 비교 연구에 비해 본 연구로 주가 등락을 예측한 경우가 약 7% 더 높았다. [표 4-10]은 본 연구와 비교 연구의 실험 결과를 정리한 것이다.

[표 4-10] 본 연구와 비교 연구의 실험 결과

	본 연구(결합형 예측 모델)	비교 연구 [6]
예측 범위	전일 종가 대비 당일 시초가 등락 당일 시초가 대비 당일 종가 등락	전일 종가 대비 당일 시초가 등락 당일 시초가 대비 당일 종가 등락
예측에 사용되는 뉴스	전날 15:00 ~ 당일 09:00 당일 09:00 ~ 당일 15:00	전날 15:00 ~ 당일 09:00 당일 09:00 ~ 당일 15:00
감성 단어 추출 방법	창원대 형태소 분석기, 명사 및 서술어 추출, 등락율 임계치(0.5), 빈도 임계치(120)	창원대 형태소 분석기, 명사 추출, 등락율 임계치(0.3), 빈도 임계치(3)
예측 모델	Logistic Regression (Text mining, ARIMA)	Naïve Bayes, RSI
Trainning Set	2005년 ~ 2007년	2005년 ~ 2007년
Test Set	2008년	2008년
정확도	61.5%	55%

4.7 연구 방법별 시뮬레이션 결과

본 연구에서 제안한 결합형 예측 모델로 실제 주식시장에 투자를 하였을 때, 자산이 어떻게 변화되는지를 시뮬레이션하여 평가하였다. 이를 위해 투자 기간은 1년(2014.01 ~ 2014.12)으로 설정하였으며, 매매에 대한 수수료는 고려하지 않고 평가하였다. [표 4-11]은 예측 결과에 따른 매매 전략을 나타낸 것이다. 예를 들어, 전일의 예측 결과가 상승으로 나타난 경우, 현금을 보유하고 있다면 주식을 전량 매수하고 주식을 보유하고 있다면 보유 주식을 유지하는 매매 전략을 취한다. 전일의 예측 결과가 하락으로 나타난 경우, 현금을 보유하고 있다면 현금을 유지하고 주식을 보유하고 있다면 전량 매도하는 매매 전략을 취한다.

[표 4-11] 예측 결과에 따른 매매 전략

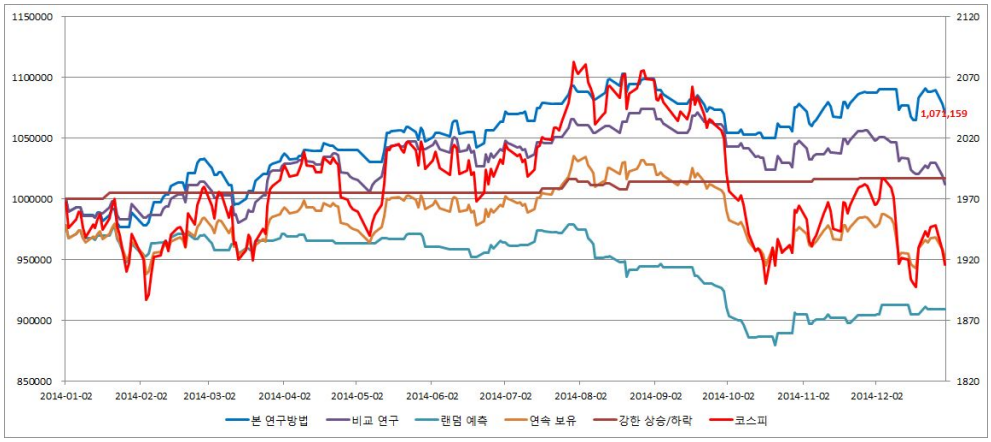
예측 결과	현금 보유 시	주식 보유 시	실제 결과	자산 변화
상승	주식 전량 매수	주식 유지	상승	**(투자 금액 * 등락율) 수익**
			하락	**(투자 금액 * 등락율) 손실**
하락	현금 유지	주식 전량 매도	상승	변화 없음
			하락	변화 없음

[표 4-12]과 [그림 4-11]은 연구 방법별 자산변화의 결과를 나타낸 것이다. 비교 연구는 [6]의 논문 실험 결과를 적용하였고 랜덤 예측은 다음 날의 주가 등락을 상승 또는 하락으로 랜덤 예측을 한 결과를 이용하였다. 또한, 연속 보유는 주식을 팔지 않고 계속 보유한 경우이며, 강한 상

승/하락은 강하게 상승하거나 하락하는 경우만을 매매하도록 하였다. 본 연구방법은 원금 100만원을 투자하였을 때, 1년 후 자산 변화는 71,159원으로 약 7%의 자산 변화율을 보였고 비교 연구는 12,367원으로 약 1.24%의 자산 변화율을 보였다. 이에 비해 코스피는 -4.39%로 나타났다. 이는, 코스피가 -4.39%인 것에 비해 본 연구방법은 약 7%의 자산 변화가 있었으므로 실제로는 KOSPI 대비 약 10% 이상의 자산변화라고 평가할 수 있다.

[표 4-12] 연구 방법별 자산 변화 결과

	자산 변화	자산 변화율
본 연구방법	1,071,159	7%
비교 연구	1,012,367	1.24%
랜덤 예측	909,054	-2.65%
연속 보유	952,278	-4.77%
강한 상승/하락	1,017,363	1.74%
코스피	-	-4.39%



[그림 4-11] 연구 방법별 자산 변화 추이

제 5 장 결론 및 향후 계획

본 연구에서는 당일 종가 대비 익일 종가 등락을 예측하기 위해 뉴스 기반 텍스트 마이닝에 의한 예측 모형과 KOSPI 데이터를 이용한 ARIMA 모형을 결합한 모델을 제안하였다.

제안 방법의 특징은 경제 뉴스로부터 추출된 단어를 이용하여 긍정/부정으로 수치화 할 수 있는 주식 도메인의 감성사전을 제시하였다는 점과 뉴스 기반의 텍스트 마이닝에 ARIMA 모형을 결합한 결합형 모델을 제안했다는 점이 있다. 제안 방법에 대한 실험 결과 뉴스 기반의 텍스트 마이닝 방법만을 이용한 것보다 ARIMA 모형을 결합한 예측 모델이 약 7%의 높은 예측 정확도를 보였다. 품사 선택에 있어서는 명사 또는 서술어를 이용하여 감성사전을 구축하는 방법 보다는 명사 및 서술어를 함께 이용하여 감성사전을 구축하는 방법이 가장 우수한 성능을 보였다.

본 연구는 일반 투자자들이 접근하기에 쉬운 뉴스와 과거 KOSPI 데이터를 이용하여 주가 등락을 예측하기 때문에 실용적이다. 또한, 주가와 밀접한 환율, 원자재와 같은 경제 지수를 예측하는데도 활용이 가능하다. 본 연구에서 제안한 주가 도메인의 감성사전을 자동으로 구축하는 방법은 상품 리뷰 기반의 감성사전을 이용한 상품 판매량 예측, 영화 리뷰 기반의 감성사전을 이용한 영화 흥행 예측과 같은 타 도메인에 적용 가능할 것으로 기대된다.

향후 연구를 통해 감성 단어의 전후 수식어를 활용하여 감성사전의 정확도를 높이고 다양한 속성들을 추가적으로 활용하여 수치 예측의 정확도를 높일 수 있는 연구가 필요하다.

참고문헌

[1] 주식과 상품, 그리고 시장, “주가 예측”, http://infopedia.usembassy.or.kr/KOR/_f2_030405.html.

[2] Eugene F. Fama(1993), “Common risk factors in the returns on stocks and bonds”, Journal of Financial Economics, vol. 33, pp. 3-56.

[3] Robert J. Shiller(1980), “Do stock prices move too much to be justified by subsequent changes in dividends?”, NBER Working Paper, No. 456.

[4] BG Malkiel(2003), “The efficient market hypothesis and its critics”, The Journal of Economic Perspectives Vol. 17, No. 1, pp. 59-82.

[5] 남달우, 박진우, 김민경, 조현, 김성희(2012), “인터넷 주식게시판을 통한 집단지성과 주식시장과의 상관관계 연구” , 인터넷전자상거래연구 제12권 제2호, pp. 149-164.

[6] Ping-Feng Pai(2005), “A hybrid ARIMA and support vector machines model in stock price forecasting”, Omega 33, pp. 497-505.

[7] 신동근, 정경용 (2011). “웨이블릿 변환과 퍼지 신경망을 이용한 단기 KOSPI 예측”, 한국콘텐츠학회논문지, 11, pp. 1-7.

[8] 유은지(2012), “주가지수 상승 예측을 위한 주제지향 감성사전 구축 방안”, 한국지능정보시스템학회 2012년 추계학술대회, pp. 42-49.

[9] V Sehgal(2007), “SOPS: Stock Prediction Using Web Sentiment”, ICDMW '07 Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, pp. 21-26.

[10] Robert P. Schumaker(2009), "A quantitative stock prediction syste m based on financial news", Information Processing and Managem ent 45, pp. 571-583.

[11] Xiangyu Tang(2009), "Stock Price Forecasting by Combining Ne ws Mining and Time Series Analysis", WI-IAT '09 Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Int elligence and Intelligent Agent Technology vol.01, pp. 279-282.

[12] 안성원(2010), "뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예 측", 한국정보과학회 2010 한국컴퓨터종합학술대회 논문집 제37권 제1호 364-369 p. 6.

[13] Ahmed, M.S. & Cook, A.R.(1979), "Analysis of freeway traffic ti meseries data by using Box-Jenkins techniques.", Transportation Research Record. , no. 722, pp. 1-9.

[14] 이종원(1998), 『계량경제학』, 박영사.