뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측

Stock Prediction Using News Text Mining and Time Series Analysis

저자 (Authors)
안성원, 조성배
Sungwon Ahn, Sung-bae Cho

출처 (Source)
한국정보과학회 학술발표논문집 37(1C), 2010.6, 364-369(6 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01474739

APA Style
안성원, 조성배 (2010). 뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측. 한국정보과학회 학술발표논문집, 37(1C), 364-369

뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측

안성원 O, 조성배

연세대학교 컴퓨터과학과 soyoja@paran.com, sbcho@yonsei.ac.kr

Stock Prediction Using News Text Mining and Time Series Analysis

Sungwon Ahn O, Sung-bae Cho

Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 뉴스 텍스트 마이닝을 수행하여 2005년 1월부터 2008년 12월까지 4년 간의 뉴스 데이터에 대해 주가에 호재인지 악재인지 여부에 대해 학습을 하고, 이를 근거로 신규 발행된 뉴스가 주가 상승 또는 하락에 영향을 미치는지를 예측하는 알고리즘을 제안한다. 뉴스 텍스트 마이닝을 위해 변형된 Bag of Words 모델과 Naïve Bayesian 분류기법을 사용하였으며, 특히 주가 예측에 있어서 뉴스 마이닝에만 의존하던 기존의 관련 연구와는 달리 예측의 정확성을 높이기 위해 주가의 시계열 데이터 분석기법인 RSI를 추가로 적용하였다. 2009년 11월부터 2010년 2월까지 4개월간 42,355건의 뉴스 데이터에 대해 실험한 결과, 기존 연구 대비 의미 있는 결과인 55.01%의 예측성공률을 얻었다.

1. 서론

주식 시장의 분석과 예측은 경제 분야뿐만 아니라 수학과 통계, 전산 분야에 이르기까지 오랜 기간 매우 중요한 연구과제로 인식되었다. 근래 들어 금융공학의 발달과 함께 과학적인 방법을 통해 주가 예측과 활용에 대한 연구는 크게 활성화되었다. 알고리즘 트레이딩(algorithm trading)의 발달, 퀀트(Quants; 금융시장 분석가)들의 등장과 이들이 개발한 시장 예측 모델의 성공은 과학적인 기법을 이용하여 주식시장을 분석하고 미래의 주가를 예측하려는 많은 연구들을 이끌어냈다.

효율적 시장 가설(efficient market hypothesis, EMH)은 효율적 시장에서 거래되는 금융자산의 가격, 특히 주식의 가격은 이미 이용 가능한 모든 정보를 반영하고 있다는 이론이다[1]. EMH에 의하면 주가는 자산의 가치를 충실히 반영하고 있으며, 자산의 가치에 영향을 줄만한 정보가 발생할 때 주가는 움직인다. EMH는 정보의 공개가 주가에 미치는 정도에 따라 약형, 준강형, 강형의 세가지 모형으로 나눠진다. 주가 예측에 대한 또 다른 시각은 랜덤워크 이론(random walk theory, RWT)이다[2]. 이는 주가의 변화는 과거의 변화나 어떤 패턴에 제약을 받지 않고 독립적으로 움직인다는 이론이다. RWT에서 개개의 가격변동은 시계열(time series) 상에서 서로 상관관계가 없이 독립적이며, 과거의 가격변동이 어떠했는가를 분석하더라도 그것이 미래의 가격변동 예측에 도움을 줄 수 없다. RWT는 모든 정보가 대중에게 공개되어 있다는 가정 하에서 준강형 EMH와 유사한 이론적 배경을 갖는다.

그러나 미래를 예측하기 위해 과거의 자료를 이용하는 것은 객관적인 통계적 개념에 근거하고 있는 것이며, EMH의 주장대로라면 경제적 분석을 포함하여 과거 자료를 근거로 미래를 예측하는 방법이 과연 잘못된 것인가에 대한 의문이 생긴다. 특히 EMH는 특별한 재료가 없음에도 발생하는 주식 시장의 이상 폭등이나 폭락과 같은 돌발적인 상황에 대한 설명을 할 수 없다는 약점을 갖고 있기도 하다. RWT 역시 주가가 일정한 추세를 갖고 움직인다고 믿는 학자들로부터 반론이 제기되었다. 인지행위적 재무론(behavioral finance)을 연구하는 학자 중 한 명인 웨버(Martin Weber)는 10 년간의 주식 시장 관찰을 통해 주식 시장의 가격이 주목할만한 수준의 일정한 흐름을 보이는 것을 확인했다. 다양한 관찰을 통해 웨버를 비롯한 RWT를 반박하는 학자들은 비랜덤 워크 가설(non-random walk hypothesis)에 대한 연구를 발표하였다[3].

이러한 배경 이론들로부터 미래의 주가 예측이 일정 수준 가능하다는 견해가 점점 힘을 얻고 있다. 기존의 주가 예측에 대한 연구는 Schumaker[11]의 분류에 따르면 수학적 예측, 통계학적 예측, 인공지능적 예측의 세가지 유형으로 나눠 볼 수 있다.

2. 관련 연구

2.1 수학적 예측 방법

경제학자들과 많은 금융투자 회사들이 즐겨 사용하는 정량적인 투자 전략(quantitative investment strategy)은 수학모델에 의거하여 수치화된 미래가치를 예측하여 포트폴리오 구축, 매매와 같은 투자 여부를 결정하는 기법이다.

피셔 블랙(Fischer Black)과 마이런 숄스(Myron Scholes)는 노벨 경제학상을 수상한 이론인 옵션 거래의 가격설정에 관한 "블랙-숄스 모델(Black Scholes Model)"을 발표하였다[18]. 이 연구로부터 투자자들은 정확하게 계산된 스톡옵션을 활용하여 주가 등락의 피해를 상쇄할 수 있게 되었다. 이후 많은 관련 연구가 진행되고 수학적 예측모델을 이용한 다양한 시도가 이루어졌다. 대표적인 수학적 예측모델로는 거래가격 범위를 제한한 상태에서 매매주문에 대해 가격이 어떻게 움직이는지를 연구하는 여과 방법(percolation method), 시계열 데이터의 움직임을 분석하여 데이터 간의 연관성과 미래

움직임 예측에 활용하는 웨이블렛 변환 방법(Wavelet transform) 등이 있다.

2.2 통계학적 예측 방법

과거의 데이터로부터 미래를 예측하는 것은 통계학적 예측의 기본적인 접근법이다. 과거의 주식시장의 데이터에 근거하여 미래를 예측하는 통계학적인 접근 방법 역시 주가 예측에 있어서 널리 사용되고 있다. 주요 통계학적 예측 방법으로는 일정 기간내의 주가를 산술 평균한 값을 나누어 평균 주가로 표현하는 이동평균(moving average) 분석, 그리고 난수를 대량으로 발생시켜 구하고자 하는 결과의 확률적 분포를 통계적으로 구하는 몬테카를로(Monte Carlo) 시뮬레이션과 같은 확률 모델이 있다.

2.3 인공지능적 예측 방법

인공지능은 주가예측에 있어서 최적화와 기계학습과 관련된 분야의 발달에 큰 공헌을 하였다. 인공지능적 주가 예측으로는 분류와 회귀 분석에 널리 사용되는 기계 학습 방법인 SVM(Support Vector Machine), 신경망(Artificial Neural Network)을 사용한 방법, 유전 알고리즘(Genetic Algorithm)을 사용한 방법 등이 있다. 인공지능적 접근방법은 대부분 예측모델에서 적용 가능한 최적화된 파라미터를 찾는 것이다. 특히 신경망이나 유전 알고리즘을 이용한 방법은 예측모델의 최적의 패턴이나 가중치 변수를 찾기 위해 널리 사용되었다. 최근에는 신경망과 유전 알고리즘 기법을 혼합하여 활용하는 방법도 연구 중이다[15].

2.4 뉴스 데이터를 이용한 예측 방법

근래에 들어서는 뉴스 데이터로부터 주가를 예측하고자 하는 연구가 활발히 이루어지고 있다[13]. 뉴스와 주가 간의 강력한 상관관계는 비교적 최근의 연구 결과[14]에서도 증명된 바 있다. 금융 뉴스를 이용한 주가 예측 기법은 대개 다음과 같은 접근법을 취한다. 뉴스 데이터를 수집한 이후 이 뉴스 데이터에 대해서 텍스트 마이닝 처리를 하여 의미있는 문서 내의 특징(feature)들을 추출한 후, 이를 이용하여 해당 뉴스가 주가에 호재인지, 악재인지를 분류한다. 그리고 분류된 결과를 이용하여 시뮬레이션 투자 및 가격 변동추이 예측을 시도한다. 뉴스 본문의 자연어 처리 방법으로는 bag of words, TF-IDF 방식이 주로 사용되었으며[5][6][7][8], 유사한 의미의 단어들을 묶어서 의미를 확장하는 명사구(noun phrase) 처리 기법과 유사한 logical AND 기법이 사용된 예도 있다[4]. 또한 자연어 문서 처리에 있어서 가장 간단한 방법인 PR(Probability Ratio) 역시 최근 연구에서 사용된 사례가 있다[10]. 그 외에 기존에 널리 사용되는 bag of words 방법보다 고유명사(proper nouns) 처리기법을 이용하여 보다 의미있는 결과를 얻었다는 보고도 있다[9]. 추출된 특징으로부터 해당 뉴스를 분류하는 방법으로는 Naïve Bayesian 분류기와 SVM(Support Vector Machine)이 많이 사용되고 있다. 표1과 2는 뉴스를 이용하여 주가 예측을 시도한 기존 관련연구들에 대한 비교이다. 이 표는 Mittermayer의 연구[13]를 토대로 작성되었다.

표 1. 뉴스를 이용한 기존의 주가예측 연구(1)

	주가 예측범위	뉴스 추출단위	분류 알고리즘	본문 카테고리	텍스트 처리방법
[4]	24 시간	Tuple (Words)	Naïve Bayes	3	Logical AND
[5]	1 시간	Terms	Naïve Bayes	5	TF-IDF
[6]	1 시간	Single Words	Naïve Bayes	3	Bag of Words
[7]	1 시간	Single Words	SVM	5	TF-IDF
[8]	15 분	Tuple	SVM	4	Bag of Words
[9]	20 분	Terms (명사구)	SVM (SVR)	3	Proper Nouns
[10]	당일종가	Words	PR Rank	2	PR Rank, Blacklist

표 2. 뉴스를 이용한 기존의 주가예측 연구(2)

	처리되는 주가간격	데이터 수집기간	학습 데이터 기간	실험 데이터 기간	대상 주식시장
[4]	거래일 종가	1997	100 일	79 일	HangSeng
[5]	10분	1999-2000	3 개월	1.5 개월	US 127 종목
[6]	10분	2001-2002	5.5 개월	2 개월	DJIA
[7]	일간주가	2002-2003	6 개월	1 개월	홍콩 614 종목
[8]	15초	2002	9 개월 Cross check	9 개월 Cross check	S&P 500
[9]	1분	2005	N/A	5 주	S&P 500
[10]	거래일간	2008	5개월	90일	상해A 중 부동산 에너지 정보통신

3. 기존 연구의 문제점

1) 정확한 마이닝 결과를 판단하기 위해 충분한 양의 학습과 실험 데이터를 사용해야 한다. 그러나 기존 연구들은 대부분 학습 데이터의 기간 및 시뮬레이션 투자에 있어서 수주 ~ 6개월 이내의 단기간의 데이터를 사용하는데 그쳤다[4][5][6][7][8][9][10]. 이는 상대적으로 오랜 기간 연구된 GA나 ANN을 이용한 주가예측 연구와 비교해 볼 때 뉴스를 이용한 주가예측 연구들의 큰 단점이라 할 수 있다.

2) 주가에 영향을 주는 변수는 무수히 많지만 기존 연구들은 뉴스가 주가에 영향을 미치는 것에 초점을 맞추어 예측에 있어서도 뉴스 데이터에만 의존하였다[4][5][6][7][8][9]. 최근에 뉴스 분석과 함께 시계열 데이터를 활용한 연구도 발표된 바 있었다[10]. 그러나 주가에 영향을 미칠만한 종합적인 변수들에 대한 복합적인 분석은 뉴스를 이용한 주가 예측 분야에 있어서 지속적으로 연구해야 할 과제임이 분명하다.

3) 주가 예측의 경우 실험 데이터가 되는 주식 데이터의 특징 중 하나는 종목, 업종 별로 예측의 편차가 심하다는 점이다. 따라서 객관적인 실험 결과를 얻기 위해서는 전 업종을 아우르는 실험이 진행되어야 한다. 그러나 일부 연구의 경우 시장 대표종목 몇 개만 선별[15]하거나, 특정 업종의 종목들만을 이용[10]하여 실험을 진행한 경우가 많다. 이는 특정 도메인에서만 검증된 주가 예측 알고리즘이라는 문제가 있다.

본 논문에서는 이러한 기존 연구들의 문제점을 개선하고자, 4년 간의 다양한 종목군의 뉴스 데이터를 확보하였다. 그리고 뉴스 데이터 분류 방법과 함께 주가 예측에서 많이 활용되고 있는 시계열 데이터 분석 기법을 함께 적용하여 보다 정확한 주가 예측 알고리즘을 제안하고자 한다. 끝으로, 기존 연구의 주류를 이루고 있는 단기(20분 ~ 1시간)의 예측이 아닌 일간 주가예측을 시도하고자 한다.

4. 제안한 주가예측 알고리즘

4.1 뉴스 텍스트의 특성추출 기법과 문서 분류

뉴스와 같이 비정형화된 문서에서 정제된 마이닝 결과를 얻기 위해, 본 논문에서는 다음과 같은 특성추출(feature extraction) 기법을 사용하고자 한다. 우선 뉴스 본문 인식을 위해서 자연어 처리 알고리즘 중 가장 인식률이 좋은 방법 중 하나로 알려진 bag of words 모델을 사용한다. 또한 보다 정형화된 텍스트 마이닝 처리를 위해서 문서 내에서 불필요하며 의미를 정규화하기 어려운 부분은 노이즈(noise)로 규정, 사전에 제거하는 작업을 거친다. 이를 위해 다음과 같은 세 가지 전처리 작업을 거친다.

1) 단음절 단어(“그,” “저,” “외,” “이,” “등,” …)를 제거한다.
2) 문서 내에서 등장하는 의미를 정규화 하기 어려운 숫자(종목코드, 매출액, 순이익 금액 등)를 제거한다.
3) 전체 문서 내에서 3회 이상 등장하는 단어에 대해서만 bag of words에 담는다.

특히 2번의 처리는 대부분의 뉴스가 “순이익 2억이 증가,” “매출이 20% 감소”와 같은 표현이 사용되는 것에 착안하여, 실제 숫자 부분을 제거하더라도 “순이익 증가,” “매출이 감소”와 같이 문장 고유의 의미는 여전히 인지할 수 있다는 생각에서 출발한다. 이 과정을 거쳐 생성된 단어 가방을 문서를 나타내는 특징이라 부른다. 이 특징들은 뉴스의 원문 묘사(textual representation)가 된다.

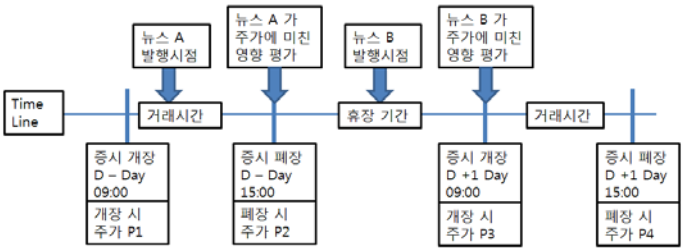

그림 1. 뉴스 발행시간에 따른 예측 기준 주가

이제 뉴스가 발행된 시간을 기준으로, 그림 1과 같이 거래시간대(09:00부터 15:00까지)와 휴장 기간(15:00부터 다음날 09:00까지) 중 어느 시간대에 뉴스가 발행되었는지를 판단한다. 예를 들어 거래 시간대에 발행된 뉴스 A는 시초가인 P1의 가격을 기준으로 하여 당일의 종가인 P2 가격이 유의미한 수준으로 변동이 있었는지 여부를 판단한다. 휴장 기간에 발행된 뉴스 B는 종가인 P2의 가격을 기준으로 D+1 일의 개장 시점의 가격인 P3가 P2에 비해 유의미한 수준으로 변동이 있었는지를 확인한다.

뉴스가 발행된 이후 주가가 유의미한 수준으로 변동된 경우 해당 뉴스는 가격에 영향을 미칠 가능성이 있는 후보 군으로 판단하여 특성추출을 시도한다. 유의미한 수준의 가격 변동은 ±2%로 정의하였다. 뉴스가 발행된 이후의 가격이 이전 가격에 비해 ±2%로 변동이 생긴 경우 이 뉴스가 주가에 영향을 줄 수 있는 뉴스로 판단하여 긍정(Positive)이나 부정(Negative) 뉴스로 분류하여 해당 뉴스가 포함한 특징들을 Naïve Bayesian 분류기를 이용하여 긍정 / 부정으로 분류 작업을 수행한다. 만약 긍정, 부정 어디에도 해당하지 않는 경우는 미결정 상태로 판단하며 측정에서 제외한다.

4.2 시계열 데이터 분석 기법

RSI(Relative Strength Index)는 1978년 와일더(Welles Wilder)에 의해 개발된 보조지표로 현재의 주가 추이가 어느 정도의 강도를 갖고 있는지를 백분율로 표시한 것이다. RSI는 본 논문과 같이 단기간의 주가 예측에 있어 널리 활용되는 기술적 분석의 방법 중 하나이다. RSI 숫자가 높을수록 과매수, 즉 주가가 과열되어 있다는 의미이고 RSI 숫자가 낮을수록 주가가 과매도 국면에 접어든다는 의미이다. 식 (1)과 (2)는 과매수/과매도 상태를 판단하는 일반적인 기준을 설명하고 있다.

$$\text{과 매수 상태: } RSI \geq 70 \quad (1)$$

$$\text{과 매도 상태: } RSI \leq 30 \quad (2)$$

RSI 계산방법은 주어진 매매 기간 중 종가를 기준으로 전일 대비 상승한 경우의 가격차를 U라 하며, 하락한 경우의 가

격차를 D라 한다. 가격의 상승 변동폭을 U, 하락 변동폭을 D라 할 때 식 (3)과 같이 표현할 수 있다.

$$\begin{aligned} U &= Close_{now} - Close_{previous} \\ D &= 0 \\ D &= Close_{previous} - Close_{now} \\ U &= 0 \end{aligned} \quad (3)$$

이제 주어진 N일(본 논문에서 N = 9)에서의 U의 이동평균과 D의 이동평균을 구한다. 이것이 식 (4)가 되고 그 결과로 식 (5)를 구할 수 있다.

$$RS = \frac{EMA\,[n]of\ U}{EMA\,[n]of\ D} \quad (4)$$

$$RSI = 100 - 100 \times \frac{1}{1+RS} \quad (5)$$

RSI 값은 0부터 100 사이의 백분율 값으로, 이 값이 나타내는 의미는 현 주가의 매매 강도 수준을 나타낸다. 본 논문에서는 매 거래일간의 일일 주가를 예측하고 있기 때문에 과매도 / 과매수 상태가 일정기간 유지되는 추세를 이용하여 주가 추이를 예측하고자 한다. 휴리스틱 반복 실험을 통해 N일에 RSI 수치가 높은 경우(과매수 상태)에서는 N+1일에도 상승할 확률이 높으며, N일에 RSI 수치가 낮은 경우(과매도 상태)에는 N+1일에도 하락할 확률이 높다는 사실을 확인하였다. 이를 통해 다음과 같이 RSI 예측을 적용하는 경우 가장 예측 성공률을 높일 수 있음을 확인하였다.

U : 상승 가중치 D : 하락 가중치

```
if(RSI>=0.70&&RSI<0.95) D *= 1.48;
else if (RSI >= 0.95)          D*= 1.25;
else if (RSI<=0.30&&RSI>0.05)  U *= 1.48;
else if (RSI <= 0.05)          U *= 1.25
```
(6)

4.3 텍스트 마이닝과 시계열 분석 방법의 결합

주가는 다양한 변수가 영향을 주므로 주가에 영향을 줄 수 있는 주요 변수들을 활용하여 주가 예측을 한다는 것은 합리적인 접근이라 할 수 있다. 이를 위해 주가를 만드는 주요 변수들을 선별하는 작업이 필요하다. 이는 기존 연구에서 많이 다루어지던 부분이다. 특히 GA와 ANN 기법을 통해 주가에 영향을 주는 주요 요인들의 가중치를 찾는 모델은 기존 연구에서도 발표되어, 다음과 같은 공식이 제안된 바 있다[16].

$$Model\ T = x \times V + y \times Q + z \times E + w \times P \quad (7)$$

- x, y, z, w: 가중치 변수
- V (Valuation): 주식 가치, PER, PBR
- Q (Quality): 자산가치, 매출, 순이익, 영업이익
- E (Analyst): 애널리스트 평가, 공시, 시장환경
- P (Price momentum): 가격 모멘텀, 거래량, 차트

이 모델에서, V 와 Q 는 분기별 내지 연도별로 변동되는 변수이므로 단기간의 주가예측모델에는 큰 영향을 미치지 못한다. 그러므로 본 논문에서는 이 모델을 변형하여, E와 P만을 사용한 단기 주가모델을 새롭게 제안하고자 한다. 이 모델에서 E는 뉴스 마이닝 데이터의 결과값이 될 것이며, P는 RSI 방법이 되어 최종적인 제안 분류 알고리즘은 아래와 같다.

```
Set T = Test Document
Using Corpus A
    Set A = log(pc) //pc=prior probability of corpus A
    For each word in Test-Document that is in Bag
        A = A + log(pi)// pi = prior probability of word i
    End For
End Using
Repeat for Corpus B
If RSI >= 70 || RSI <= 30 // RSI 값 적용 가능 기간이면
// Naïve Bayesian 분류 결과에 식 6 의 RSI 가중치 추가
A *= fRSI(A) and B *= fRSI(B)
// 결과를 비교하여 주가 상승(A)/하락(B) 여부를 판정한다.
// Naïve Bayes 분류 값이 작을수록 Corpus 와 유사하다
If A < B choose A else choose B
```

5. 실험 환경

5.1 실험 데이터

실험을 위해 2010년 3월 현재 거래소 시장에 상장된 전 종목에 대한 뉴스 데이터를 수집하였다. 뉴스 데이터는 포털 사이트인 네이버의 증권정보 사이트(http://stock.naver.com)에서 제공하는 종목별 관련 뉴스 데이터를 웹 페이지에서 다운받은 후, 필요한 뉴스 본문만 파싱하여 추출하는 형식으로 수집하였다.

수집된 전체 뉴스 데이터 중에서 앞서 언급한 바와 같이 전일 종가 대비 시초가가 2% 이상 차이 나는 경우 및 시초가 대비 종가가 2% 이상 차이 나는 경우에 해당하는 구간에 발행된 뉴스들만 다시 선별하여 학습에 활용하였다. 한글의 텍스트 마이닝을 위한 형태소 분석을 위해 국민대학교 자연어처리 연구실에서 제공하는 KLT 2010 (Korea Language Technology 2010)을 사용하였다[16].

5.2 실험 수행방법

학습을 위해 2005년 1월부터 2008년 12월까지 4년간의 뉴스 데이터 중, 해당 구간 내에서 발행 후 2% 이상 주가에 영향을 미친 뉴스 164,812건을 추출하였다. 이 뉴스들은 상승/하락 2 가지로 구분될 수 있다. 이 뉴스데이터는 bag of words 모델로 특징 추출을 한 다음 상승/하락인 경우에 대한 기준 카테고리로 Naïve Bayes 분류기의 학습에 적용시킨다. 학습이 완료되면 학습한 결과를 바탕으로 다량의 뉴스를 입력받아 이 뉴스가 발행된 시간을 기준으로 현재 거래일의 종가(뉴스가 거래시간 중에 발행된 경우) 혹은 다음 거래일의 시초가(뉴스가 거래시간 외에 발행된 경우)의 상승/하락 여부를 예측한다. 테스트를 위해 2009년 11월 1일부터 2010년 2월 28일까지 4개월간 발행된 뉴스를 수집하였다.

2010년 2월말 기준으로, KOSPI에 상장된 KOSPI 종목은 총 714개이다. 그러나 이 종목들 중에서 아래에 해당하는 종목들은 충분한 실험 데이터의 확보가 미비하거나 훈련 데이터의 숫자가 부족한 경우가 있어 올바른 실험의 결과를 도출하기 위해 제외하였다.

1) 2009년 11월 1일부터 2010년 2월 28일 사이에 관련 뉴스가 한 건도 발행되지 않은 종목 17개 제외
2) 2009년 1월 이후로 상장된 종목 26개 제외((1)과 2개 종목이 겹침)

상기 종목들을 제외한 총 673개 종목의 뉴스 42,355건에 대해서 실험을 수행하였다. 이 673개 종목들은 훈련용 뉴스 데이터의 개수와 종목별 실험용 뉴스 데이터의 개수의 많고 적음에 차이가 난다. 그러나 본 논문에서는 보다 일반적인 실험 결과를 얻고자 이 상태에서 추가적인 실험 데이터의 첨삭은 가하지 아니하였다. 실험항목은 종목에 대한 해당 뉴스가 발행된 시간을 기준으로, 당일 종가(거래시간 외에 발행된 뉴스라면 다음 개장시의 시초가)가 이전 가격보다 올랐는지 내렸는지 여부를 예측하는 주가의 방향성 정확도(directional accuracy) 예측도에 대해 실험하였다.

6. 실험 결과

6.1. Naïve Bayes 분류기만 사용하여 예측한 경우

Naive Bayes 분류기만 이용하여 주가 예측을 시도한 경우의 예측 성공 결과는 표 3과 같다. 미결정이란 Naïve Bayes 분류기의 스코어가 완전히 동일하여 주가 상승과 주가 하락 어느 경우에도 완전히 동일한 확률(50%)을 갖기 때문에 예측을 하지 못하는 경우를 말한다. 표 4를 보면 종목별로 예측 성공률의 편차가 심한 것을 볼 수 있다. 좀더 정확한 결과를 얻기 위하여, 대상 기간 중에 발행된 뉴스 건수가 50건 미만인 종목들의 예측 성공률의 표준편차와 뉴스 건수가 50건 이상인 종목들의 예측성공률의 표준편차를 비교해 보았다. 그 결과, 입력 데이터(기간 중 발행된 뉴스)의 개수가 적을수록 예측 성공률의 표준편차가 크게 달라지는 것을 알 수 있다. 즉, 안정적인 결과를 확인하기 위해서는 평가해야 하는 뉴스 데이터의 개수가 충분해야 한다.

표 1. 실험 6.1의 예측성공률

실험 데이터	예측성공	예측실패	미결정	예측성공률
42355	21693	19909	753	52.14 %

표 2. 실험 6.1의 최대/최소 예측성공률 종목

종목	예측성공률	전체뉴스건수	미결정건수
고려개발(004200)	81.25%	50	2
OCI㈜(010060)	31.81%	94	6

표 3. 실험 6.1의 건수에 따른 표준편차

	종목 개수	표준편차
뉴스 50건 미만 종목	502	27.62%
뉴스 50건 이상 종목	171	7.08%

결론적으로, 전체 42,355건에 대해 주가의 방향성 예측을 한 결과, 텍스트 마이닝 후에 bag of words 모델에 Naïve Bayes 분류기만 적용한 경우의 예측 정확도는 52.14%였다. 기존 연구 중 20분 후의 주가 예측을 하는 경우에 bag of words만을 사용하여 52.4%의 예측 성공률을 보인 사례가 보고된 바 있다[13]. 이를 감안할 때 본 논문에서의 Naïve Bayes 분류기와 bag of words 모델을 사용한 방법은 적절한 성능을 보여주었다고 할 수 있다.

6.2. RSI 만 사용하여 예측한 경우

RSI만 사용하여 예측한 경우 표 6과 같이 60.12%의 상대적으로 높은 예측성공률을 보이고 있지만, 미결정 케이스가 전체 실험 데이터 대비 62.19%가 발생하여 전체 예측대상 중 40% 미만의 경우에 대해서만 적용할 수 있음을 알 수 있다. 미결정 상태가 다수 발생하며, 일반적으로 사용하기에는 역시 무리가 있는 보조 지표라는 점을 확인할 수 있다. 표 7은 RSI 시계열 데이터 분석만 적용한 경우의 최대/최소 예측성공률이 발생한 종목에 대한 비교이다. 마찬가지로 전체 뉴스 건수가 50건 이상인 종목에 한하여 검출하였다. RSI 기법을 사용한 경우 실험 6.1에 비해 그 표준편차가 매우 심하여 안정적인 예측 방법이 아니라는 사실을 알 수 있다.

표 4. 실험 6.2의 예측성공률

실험데이터	예측성공	예측실패	미결정	예측성공률
42,355	9,627	6,384	26,344	60.12 %

표 5. 실험 6.2의 최대/최소 예측성공률 종목

종목	예측성공률	전체뉴스건수	미결정건수
㈜코오롱(002020)	94.28%	68	33
㈜고제(002540)	21.73%	60	13

표 6. 실험 6.2의 뉴스 건수에 따른 표준편차

	종목 개수	표준편차
뉴스 50건 미만 종목	502	39.91%
뉴스 50건 이상 종목	171	13.91%

6.3 Naïve Bayes 분류기 사용 후 RSI 를 적용한 경우

Naïve Bayes 분류기로 상승/하락 뉴스를 분류 후 RSI 분석 기법을 적용한 경우의 예측 성공률은 6.1의 실험에 비해 3% 정도 높은 55.01%를 기록하였으며 어느 종목에서든 가장 안정적인 예측성공률의 표준편차를 보여준다.

표 7. 실험 6.3의 예측성공률

실험데이터	예측성공	예측실패	미결정	예측성공률
42,355	23,064	18,863	428	55.01 %

표 8. 실험 6.3의 최대/최소 예측성공률 종목

종목	예측성공률	전체뉴스건수	미결정 건수
㈜코오롱(002020)	80.88%	68	0
㈜고제(002540)	30.51%	59	0

표 9. 실험 6.3의 뉴스 건수에 따른 표준편차

	종목 개수	표준편차
뉴스 50건 미만 종목	502	27.01%
뉴스 50건 이상 종목	171	7.40%

본 논문의 예측성과를 기존 연구와 비교한 것은 표 12와 같다. [4] 보다 좋은 예측성공률을 기록했지만, [9]와 [10] 보다는 낮은 결과를 얻었다. 하지만 Schumaker[9] 의 연구는 20분 후의 주가 예측을 시도한 것으로 본 논문에서 예측하는 일간 단위의 예측과는 차이가 있다. 한편, 본 연구와 동일하게 일간 단위 예측을 시도한 Tang[10]의 연구는 상해 A 증시의 3개 업종(부동산, 정보통신, 에너지)군에 대해서만 실험을 했다는 점을 고려해야 한다. 그림 2에서 볼 수 있듯이 종목별 성공률의 편차가 심한 주가 예측의 특성상 특정 업종만을 실험 데이터를 사용한 것은 정확한 결과를 왜곡할 소지가 있다는 점에서 본 실험과는 그 의미가 다르다 할 수 있다.

표 12. 기존 연구와 본 연구의 예측성공률 비교

	[4]	[9]	[10]	본 연구
예측성공률	46.8%	58.2%	63%	55.01%
예측 단위	일간	20 분	일간	일간
실험 데이터	전 업종	전 업종	업종 선별	전 업종

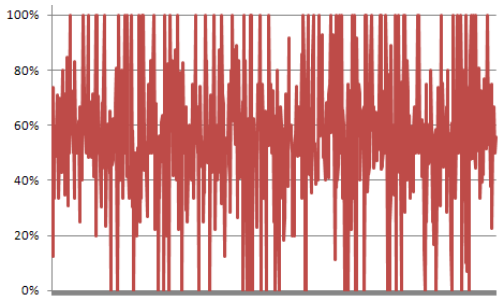

그림 2. 본 연구의 KOSPI 전 종목 예측성공률 분포

7. 결론

뉴스를 이용한 주가예측은 상당히 어려운 예측 기법이며, 뉴스 문서의 텍스트 마이닝 처리와 특성 추출을 통한 문서 분류가 효율적으로 이루어진다 해도 많은 변수를 내포하고 있다.

또한, 문서 내용(종목별, 뉴스의 종류별)에 따라 예측편차가 심한 것도 확인할 수 있다. 이러한 이유 중 하나는 종목마다 특정 분야에 대한 내용이 수록되기 때문에 어느 하나의 종목에서 상승/하락의 강한 신호가 되는 특정 키워드가 다른 종목에서는 거의 영향을 미치지 못하는 것을 들 수 있다. 그러므로 기존의 일부 연구들처럼 특정 종목을 선별하여 실험을 할 경우 왜곡된 결과가 나올 수 있기에 향후 주가 예측과 관련된 연구에서는 S&P 500, KOSPI 200과 같이 일반적으로 널리 사용되는 투자 종목 지표를 이용하여 폭넓게 여러 종목들에 대해 실험을 하는 것이 적절할 것이다.

한편, 뉴스를 이용한 주가예측에 시계열 데이터 분석 기법을 추가한 것은 매우 좋은 효과를 보여주었다. 뉴스만을 이용하여 예측을 하는 경우에 비해 시계열 데이터 분석 기법을 함께 활용하여 평균 3% 만큼의 예측성능 향상을 보였다. 기존의 뉴스를 이용한 주가 예측 방법은 뉴스 자체의 정보만을 이용하여 주가를 어느 정도 수준으로 예측할 수 있는지에 대해 주로 연구되었다. 더 높은 예측률을 거두기 위해서는 본 논문에서와 같이 주가에 영향을 주는 다른 변수들에 대해 추가적인 연구가 필요하며, 이는 향후의 주가 예측 연구에 있어 반영해야 할 사항이라고 생각된다.

참 고 문 헌

[1] E. F. Fama, The Behavior of Stock Market Prices, The Journal of Business, The University of Chicago Press, 1965.

[2] B. G. Malkiel, A Random Walk Down Wall Street, W. W. Norton & Company, 1973.

[3] H. Fromlet, "Behavioral finance-theory and practical application," Business Economics, 2001

[4] V. Cho et al., “Text processing for classification,” Journal of Computational Intelligence in Finance, vol. 7, no. 2, 1999.

[5] V. Lavrenko, et al. "Language models for financial news recommendation," Proc. of the Ninth Int. Conf. on Information and knowledge manage, 2000.

[6] G. Gidofalvi, Using News Articles to Predict Stock Price Movements, University of San Diego, Computer Science Dept. 2003.

[7] G. Fung, J. Yu, and W. Lam. "Stock prediction: Integrating text mining approach using real-time news," Computational Intelligence for Financial Engineering, 2003.

[8] M.A. Mittermayer, and G. Knolmayer, "NewsCATS: A news categorization and trading system," Proc. of the Sixth Int. Conf. on Data Mining, 2006.

[9] R. Schumaker, and H. Chen. "Textual analysis of stock market prediction using breaking financial news: The AZFinText system," ACM Transactions on Information Systems, vol. 27, no. 2, 2009.

[10] X. Tang, C. Yang, and J. Zhou. "Stock price forecasting by combining news mining and time series analysis," Proc. of the 2009 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology, 2009.

[11] R. Schumaker, and H. Chen. "A discrete stock price prediction engine based on financial news," IEEE Computer Society, vol. 43, no. 2, pp. 51-56, 2010.

[12] Y. Becker, and U.M. O'Reilly. "Genetic programming for quantitative stock selection," Proc. of the First ACM/SIGEVO Summit on Genetic and Evolutionary Computation, 2009.

[13] M.A. Mittermayer, and G. Knolmayer, "Text mining systems for predicting market response to NEWS," Proc. of IADIS European Conf. on Data Mining, 2007.

[14] T. Fu, et al. “Discovering the correlation between stock time series and financial news,” Proc. of Web Intelligence and Intelligent Agent Technology, 2008.

[15] Y. Kwon, and B. Moon, "A hybrid neuro-genetic approach for stock forecasting," IEEE Trans. on Neural Networks, vol. 18, no. 3, pp. 851-864, 2007.

[16] http://nlp.kookmin.ac.kr/HAM/kor/download.html