# 인터넷 뉴스를 활용한 개별 주식 수익률 예측 모델 연구
## 오피니언 마이닝 기법의 활용

저자
(Authors)

문하늘, 김종우

출처
(Source)

한국지능정보시스템학회 학술대회논문집 , 2014.5, 387-393(7 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06240754

APA Style

문하늘, 김종우 (2014). 인터넷 뉴스를 활용한 개별 주식 수익률 예측 모델 연구. 한국지능정보시스템학회 학술대회논문집, 387-393

# 인터넷 뉴스를 활용한 개별 주식 수익률 예측 모델 연구
## : 오피니언 마이닝 기법의 활용

**문하늘[a], 김종우[b]**

*[a] 한양대학교 경영대학 경영학부*
*[b] 한양대학교 경영대학 경영학부 (교신저자)*
*[a,b] 서울특별시 성동구 행당동 17번지*
*Tel: 02-2220-4050, Fax: 02-2220-1169, E-mail:[a] neulmoon@gmail.com,[b] kjw@hanyang.ac.kr*

**Abstract**

본 논문에서는 개별 주식 수익률 예측을 위한 감성 사전을 구축하고 이를 이용하여 학습 및 예측하는 주식 수익률 예측 모델을 제안한다. 제안하는 모델은 2011년부터 2013년 9월까지 웹 상에 게재된 KOSPI 시가총액 상위 10개 종목의 뉴스 데이터를 이용하여 설계되었다. 데이터 중 2011년부터 2012년까지 게재된 데이터는 감성사전 구축에 활용되었으며 2013년 이후의 데이터는 학습 및 테스트에 활용되었다. 제안하는 모델은 4가지의 주요한 단계를 거쳐 일별 방향성 예측 점수를 제시한다. 첫 단계에서는 전체 기간 동안 발생한 뉴스를 자동으로 수집하고 분석 가능한 형태로 전환시키며 두 번째 단계에서는 이 중 일부를 이용하여 종목별 감성사전을 구축한다. 이 때 각 단어에는 고유한 감성 점수가 주어지게 된다. 세 번째 단계에서는 감성사전의 점수를 이용하여 일별 오피니언 점수를 구하며 마지막 단계에서는 훈련 집합을 이용하여 테스트 집합의 오피니언 점수를 표준화한다. 표준화 된 오피니언 점수는 그 자체가 향후 주가의 방향성 예측 값이며, 이 때 중립의견을 고려하기 위해 다양한 임계 값을 적용하게 된다. 본 연구에서는 0.0, ±0.5, ±1.0 의 세 가지 임계 값을 적용하여 결과를 비교하였으며, 그 결과 임계 값 ±0.5 에서 정확도 55.21%, 표준편차 8.07%로 가장 좋은 성과를 보였다.

*Key word: 오피니언 마이닝, 감성 사전, 주가 방향성 예측, 뉴스*

## 1. 서론

효율적 시장 가설에 따르면 금융 자산의 가격에는 모든 정보가 충분히 반영되어 있어, 평균 이상의 수익을 얻는 것이 불가능하다. 주식시장이 완전히 효율적이라면 모든 이용 가능한 정보가 즉각적으로 주가에 반영되어 주가는 항상 내재가치와 동일할 것이기 때문이다[1]. 그러나 주식시장이 폭등과 폭락 및 장기간에 걸친 버블형성과 같이 설명이 불가능한 등락을 거듭한다는 측면에서 이상적인 효율적 시장이라 보기 어렵다는 비판이 제기되었고 이에 따라 비효율적 시장가설이 등장하였다. 비효율적 시장가설은 정보에 대한 주가의 지연 반응 및 과잉 반응을 설명하고자 하는 가설로, 단기적으로 새로운 뉴스에 과잉 반응하고자 하는 모멘텀 효과, 투자자의 심리로 주가를 설명하는 행동재무론 등을 통해 주식 시장의 비효율성을 설명한다[2][3].

비효율적 시장가설에 따르면 대다수는 불확실한 상황에서 극단적으로 최근의 정보에 높은 비중을 두고 의사결정을 내린다. 즉 다수의 투자자가 최근에 접한 정보에 따라 주가가 등락하게 되는 것이다[4][5]. 이 같은 비효율적 시장가설의 등장으로 인해 학계에서는 주가를 예측하고자 하는 노력이 다방면에서 진행되었다[6].

본 연구에서는 이러한 행동재무적 관점에 근간을 두고 일간 주식 수익률을 예측하고자 하였다. 새로운 정보가 발생하였을 때 투자자들이 실제 정보보다 과민하게 반응한다는 점에 착안하여 증권 관련 뉴스를 이용하여 일간 주식 수익률을 예측하고자 하는 것이다. 그러나 하루에도 수없이 많은 관련 뉴스가 발행되고 있는 현대 사회에서 투자자가 실시간으로 증권 관련 뉴스를 모으고 분석하는 것은 불가능하며, 가능하다 하더라도 많은 시간을 필요로 한다. 따라서 재무적 측면에서 텍스트의 감성을 인식하는 오피니언 마이닝 기법은 활용도가 높을 것으로 기대된다. 오피니언 마이닝 기법을 활용하면 웹 상에서 다량으로 발생하는 주식 관련 뉴스를 자동 수집하여 감성을 인식하고 이를

활용해 수익을 얻는 것이 가능하기 때문이다. 때문에 오피니언 마이닝을 활용하여 뉴스와 주가의 상관관계를 밝히고자 하는 노력은 국내외를 망라하여 오래 전부터 계속되어 왔으며, 최근에는 다양한 데이터 마이닝 및 오피니언 마이닝 툴이 제공됨에 따라 관련연구가 더욱 활발히 진행되고 있다. 하지만 상대적으로 복잡한 언어체계 때문에 한국어 기반의 뉴스와 주가의 상관관계는 여전히 괄목할만한 상관관계를 보여주지 못하고 있다. 이는 언어 특성상 어휘의 감성을 알기 어렵고, 긍정어휘의 수나 부정어휘의 수가 전체 문서의 의견과 다를 가능성이 높기 때문이다.

따라서 본 연구에서는 다수의 투자자에게 쉽게 노출되는 웹 뉴스를 실시간으로 스크래핑하고 분석·활용하여 주가의 향후 움직임을 예측하는 주가 방향성 예측 모델을 제안함에 있어 이러한 한계들을 극복하고자 한다. 첫째로 보다 세밀한 어휘의 감성을 태깅하기 위해 기존의 명사 중심 감성 태깅에서 벗어나 감성 분석 대상 품사를 보통명사, 고유명사, 형용사, 동사를 포함하도록 확장하고자 하며, 둘째로 종목별로 상이한 영향을 주는 어휘들을 고려하여 종목별 감성사전을 구축하고자 한다. 또한 마지막으로 하루에 양산되는 전체 뉴스의 오피니언 지수 총량을 점수화하여 점수 그 자체가 주가의 방향성을 알려줄 수 있도록 하고자 한다. 이는 투자자가 편리하게 다량으로 양산되고 있는 정보들의 종합적 의견을 파악할 수 있도록 함으로써 폭넓은 투자의 기회를 제공하고자 함이다. 분석 대상 종목은 충분한 뉴스 데이터 확보를 위해 KOSPI 시가총액 상위 10개 종목으로 제한하고자 하며, 웹 크롤러(Web Crawler)를 이용하여 2011년부터 2013년9월까지 웹 상에 게시된 뉴스 수집하여 분석하고자 한다.

## 2. 관련 연구

웹 뉴스 및 온라인 데이터를 활용하여 주가의 향후 방향성을 파악하고자 하는 연구는 오래 전부터 진행되어 왔으며, 근래에는 더욱 활발히 진행되는 양상을 보이고 있다. 이러한 연구의 주요 논점은 자연어 처리 및 감성 극성 태그 기법과 이를 활용하여 뉴스와 주가의 상관관계를 파악하는 모델링 기법이다.

### 2.1. 국내 연구 사례

국내에서는 뉴스 및 온라인 데이터의 감성을 분석하여 주가지수를 예측하기 위한 다양한 연구[7][8][9][10]와 함께 한국어 기반 문서의 정확한 감성 인식을 위한 자연어 처리 및 감성 극성 태그 기법에 대한 연구[10][11][12]도 활발히 진행되어 왔다. 구체적으로는 비교적 초기에는 단순히 문서로부터 추출된 어휘의 감성을 긍정/부정으로 이분하여 예측에 이용하였으나[7][8][9], 최근에는 어휘의 감성 극성이 그 정도를 포함하는 감성사전을 구축하여 이용하는 방향으로 연구가 이루어 지고 있다[10]. 일반적으로는 개별 뉴스 별로 점수를 산출하고 이를 취합하여 일별주가를 예측하는 방식을 띄고 있다[7][8][9][10]. 자연어 처리 및 감성 극성 태그 방식으로는 특정 조건에서 출현하는 어휘에 감성 극성을 부여하는 방식[10]과 기본적으로 사용되는 어휘들을 몇 가지의 감정 상태로 구분하여 감성을 태그하는 방식[11], 그리고 문서 전체의 패턴을 추출하여 어휘의 감성 극성 정도를 태그하는 방식[7][12] 등을 이용하고 있다.

### 2.2. 해외 연구 사례

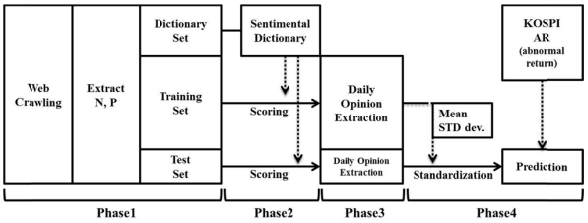
해외에서는 오피니언 마이닝을 이용하여 웹 뉴스나 트위터 등 온라인 데이터를 분석·분류하고 주가를 예측하고자 하는 연구가 비교적 오랜 기간 동안 진행되어 왔다. 뉴스와 주가의 상관관계에 대한 연구는 1990년대부터 본격적으로 이루어졌으며[13][14], 2000년대에 들어서면서 이러한 연구는 더욱 활발히 이루어지기 시작하였다[15][16][17]. 그러나 최근에는 웹 뉴스보다는 트위터 게시물의 감성을 인식하고 이를 이용하여 주가를 예측하고자 하는 연구가 활발히 진행되고 있으며[18][19], 일부 기관에서는 이를 이용한 실제 투자 전략을 실행하고 있다[20]. 분석 대상으로는 S&P500[17] 혹은 DJIA(Dow Jones Industrial Average)[15][20]와 같은 인덱스 지표를 사용하고 있다. 감성 극성 태그방식으로는 GPOMS(Google Profile of Mood States)[20]등의 정형화 된 감성 태그 사전을 이용하고 있다.

본 연구는 개별 주식의 초과수익률 방향성을 예측하고자 한다는 점과 예측 단위를 일간 오피니언으로 하고자 한다는 점, 그리고 초과수익률을 이용하여 감성 극성의 정도를 점수화한 감성사전을 사용하고자 한다는 점에서 인덱스 예측과 개별 뉴스 중심으로 이루어진 선행연구와는 차이가 있다.

## 3. 주식 수익률 예측 방안

제안하는 주식 수익률 예측 모형은 예측점수를 제시하기까지 네 단계로 구성되어 있다(<그림 1> 참조). 첫 번째 단계는 뉴스를 수집하고, 이를 분석하기 위해 전처리 하는 단계로 한국어의 어휘 특성을 반영하는 단계이다. 두 번째 단계는 종목별 감성사전을 구축하는 단계로, 각 종목별·업종별로 상이하게 사용되는 어휘의 감성을 파악하고

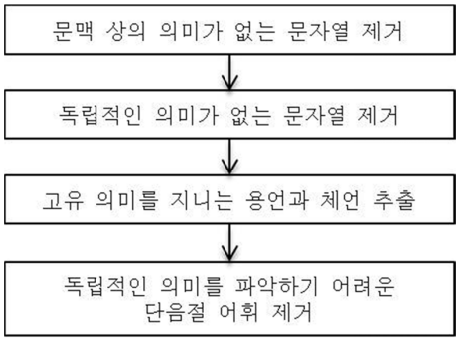
반영하는 단계이다. 세 번째 단계는 일별 오피니언을 점수화하는 단계로 하루 동안 양산된 전체 뉴스의 오피니언을 종합하고 향후 주가의 방향성을 예측하는 단계이다. 마지막으로 네 번째 단계는 일별 오피니언 점수를 표준화하여 직접적으로 예측에 사용하는 표준화 값을 이용한 예측 단계이다.




<그림 1> 주식 수익률 예측 모형

### 3.1. 뉴스 수집 및 전처리

본 단계에서는 보다 정제된 결과를 얻기 위해 비정형 데이터인 웹 뉴스를 분석 가능한 데이터로 변환하기 전에 아래의 전처리 과정을 시행한다(<그림 2> 참조). 각 전처리 기준은 한국어의 어휘특성을 고려하였다. 구체적인 전처리 과정은 다음과 같다.




<그림 2> 뉴스의 전처리 과정

1) 먼저 문맥상의 의미를 지니지 않는 URL, 메일주소, 문장부호 및 특수기호 등 제거한다.

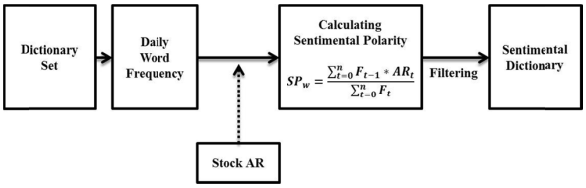
2) 문맥상의 의미는 있으나 독립적인 의미파악이 어려운 숫자(주가, 등락률, 종목코드 등)를 제거한다.

3) 한국어의 언어체계상 독립적인 의미를 지니는 체언과 용언 중 그 자체로 고유한 의미를 가진다고 판단되는 보통명사, 고유명사, 동사, 형용사만을 추출한다.

4) 단순히 어휘만으로 문맥상의 의미를 가늠하기 어려운 “이”, “그”, “하-( ‘하다’ )” 등 단음절 체언 및 용언 제거한다.

### 3.2. 구축

본 연구에서 사용된 모형에서는 추출된 어휘들의 감성 극성을 점수화하기 위해 종목별로 독립적인 감성사전을 구축하였다. 선행 연구에서 증명된 바 있듯 주제지향 감성사전을 구축함으로써 결과의 정확도를 높이고자 함이다. 감성 사전의 구축은 전처리 과정을 거친 2011년부터 2012년까지의 데이터를 대상으로 이루어 졌으며, 선행연구와 달리 출현 빈도에 가중치를 두고 해당 어휘의 출현 이후 발생한 초과수익률에 근간하여 감성 정도를 점수화하였다. 구체적인 구축과정은 <그림 3>과 같다.




<그림 3> 감성사전 구축 알고리즘

각 단어의 감성 극성 점수를 $SP(W_i)$라 할 때 감성 극성 점수의 계산 방법은 다음의 식(1)과 같이 표현할 수 있다.

$$SP(W_i) = \frac{\sum_{t=0}^{n} F_{t-1} * AR_t}{\sum_{t=0}^{n} F_t} \quad (1)$$

$SP(W_i)$: Word i의 감성 점수
$F_t$: t일의 Word i 출현 빈도
$AR_t$: t일의 초과이익

감성 점수 $SP(W_i)$가 0보다 클 경우 Word i는 주가에 긍정적인 영향을 주는 것으로 볼 수 있으며 반대의 경우에는 Word i가 주가에 부정적인 영향을 미치는 것으로 볼 수 있다. 또한 $SP(W_i)$의 절대값이 높을수록 Word i는 주가에 많은 영향을 미치는 어휘라 볼 수 있으며, $SP(W_i)$가 0인 경우에는 Word i는 주가와 무관한 것으로 볼 수 있다. 그러나 이러한 논리적인 구분과는 달리 실제 데이터에서 0이 나오기는 매우 어려운 일이다. 따라서 0에 가까운 감성점수를 보이는 단어의 경우에도 주가와 무관한 것으로 파악하는 것이 더 합리적이라고 볼 수 있을 것이다. 본 연구에서는 반복적인 실험을 통해 중립어휘로 간주하기 위한 임계 값이 $\pm 0.1$이상이 되면 유의미한 차이를 보이지 않는다는 사실을 확인하였기에 제안 모델에서는 $\pm 0.1$ 을 기준으로 하여 0.1이상은 긍정 어휘, -0.1이하는 부정 어휘, 나머지는 중립 어휘로 간주하고자 한다.

이 같은 분류를 하기에 앞서 해당 어휘의 총

출현 빈도에 따라 데이터를 필터링하였다. 이는 전체 기간(730일)동안 유의하게 출현한 어휘만을 포함하기 위함이다. 종목별로 어휘의 빈도 범위가 상이하지만 모든 종목의 어휘빈도가 좌경분포(skewed right)된 형태를 보이기 때문에 전체적인 빈도분포를 상대적으로 고르게 만드는 것이다. 본 모델에서는 총 출현 빈도가 상위 3%이내인 경우 이상수치(outlier)로 간주하였으며 전체평균에서 3% 이상 낮은 경우 유의한 횟수만큼 출현하지 않은 단어로 간주하였는데, 이는 반복실험을 통해 결과값에 부정적인 영향을 주지 않는 수준에서 결정하였다. 총 출현 빈도가 1인 데이터는 무의미하므로 우선적으로 제외하며 그 외 데이터 중 누적확률 0.47이상 0.97이하인 데이터만이 최종적으로 감성사전에 사용된다.

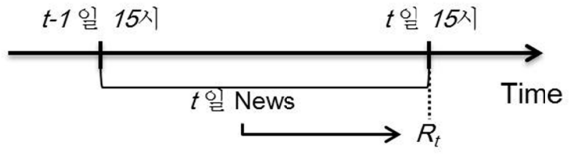
### 3.3. 일별 오피니언 점수화

구축한 감성사전을 바탕으로 일별로 수집 및 전처리 된 데이터를 단어의 감성점수와 매칭하여 일별 오피니언을 점수화하였다. 각 날짜 별로 추출된 전체 어휘 중 감성사전과 매칭되는 어휘에 해당 어휘의 감성 점수를 부여하고 이들 점수의 평균값을 구하는 방식이다. 일별 오피니언 점수(DO: Daily Opinion)의 산출 식은 아래의 식(2)와 같다.

$$DO_{it} = \frac{\sum_{i=1}^{n} F_i * SP(w_i)}{\sum_{i=1}^{n} F_i} \quad (2)$$

$DO_{it}$: 개별 주식 X의 t일 오피니언 지수
$F_t$: t일 발행된 개별 주식 X의 뉴스에서 Wi가 출현한 빈도
$SP(w_i)$: Wi의 감성 점수

일별 오피니언 점수를 계산하기 위한 뉴스 데이터의 날짜구분은 정규장 종료 시각인 15시를 기준으로 하였다. (t-1)일 15시 이후부터 t일 15시 이전에 발행된 뉴스를 t일의 종가 변동 예측에 사용한 것이다. 비개장일의 경우 익개장일의 주가 변동 예측에 사용하였다. t일 전 이틀이 비개장일이라면 (t-3)일 15시부터 t일 15시까지의 데이터가 $R_t$를 예측하는데 사용되는 것이다. (<그림 4> 참조)




*<그림 4> 뉴스 게재 시간에 따른 예측 기준 일자*

### 3.4 표준화 값을 이용한 예측

상기 일별 오피니언 점수는 종목별로 그 범위와 분포가 상이하여 표준화가 필요하다. 따라서 본 연구에서는 훈련 데이터 집합과 테스트 데이터 집합을 구분하고 훈련 데이터 집합으로부터 계산된 통계량을 테스트 집합의 DO (Daily Opinion)을 표준화하는 데 사용하였다. 즉, 훈련 집합의 DO평균과 표준편차를 각각 $\overline{DO_{it}}$, $std(DO_{it})$ 라 할 때 테스트 집합의 표준화된 DO값, $Z(DO_{it})$ 은 식(3)과 같이 표현할 수 있다.

$$Z(DO_{it}) = \frac{DO_{it} - \overline{DO_{it}}}{std(DO_{it})} \quad (3)$$

표준화된 DO값은 그 범위와 분산이 일정하므로 감성점수와 유사하게 중립 오피니언으로 간주하기 위한 임계 값을 설정함으로써 보다 향상된 결과를 얻을 수 있으며 그 결과값 자체가 방향성을 의미하게 된다. 본 연구에서는 다양한 임계 값에 대해 반복 실험하여 최적의 임계 값과 결과를 얻고자 하였으며, 최종적으로는 0.0, 0.5, 1.0 세 가지의 임계 값을 모델의 실험 값으로 선택하였다.

## 4. 실험 수행

### 4. 이

충분한 뉴스 데이터를 가지고 있는 종목으로 한정하기 위하여 한국 거래소 시장에 상장된 종목 중 시가총액 상위 10개 종목을 대상으로 실험을 진행하였다. 각 종목별로 2011년 1월 1일부터 2013년 9월 30일까지 총 130,576개의 웹 뉴스 데이터를 수집하였으며, 포탈 사이트 네이버의 증권 섹션(http://finance.naver.com)에 게시된 뉴스만을 대상으로 하였다. 실험데이터는 과적합(overfit) 현상을 방지하기 위해 기간별로 용도를 달리하여 1회만 사용하였으며, 감성 사전 구축, 훈련용, 테스트용으로 사용한 데이터의 기간은 <표 1>과 같다.

*<표 1> 기간별 데이터 집합 분류*

| 용도 | 데이터 기간 |
|---|---|
| Dictionary | 2011-01-01 ~ 2012-12-31 |
| Training | 2013-01-01 ~ 2013-06-30 |
| Test | 2013-07-01 ~ 2013-09-30 |

주식 수익률 데이터로는 해당 기간의 일간 초과수익률(Abnormal Return)을 사용하였다. 초과수익률은 개별 증권의 수익률이 당일의

시장전체의 수익률 변동과 유의적으로 다르게 나타난 그 차이를 측정한 것으로 식(4)와 같이 계산된다[21]. 시장 수익률로는 KOSPI 인덱스 수익률을 사용하였다.

$$AR_{it} = R_{it} - (\hat{\alpha}_i + \hat{\beta}_i R_{mt}) \quad (4)$$

$AR_{it}$ : t 시점에서의 표본기업 i의 초과수익률
$R_{it}$ : t 시점에서의 표본기업 i의 주가수익률
$(\hat{\alpha}_i + \hat{\beta}_i R_{mt})$ : t 시점에서의 표본기업 i의 기대수익률

오피니언 마이닝을 위한 한국어 형태소 분석은 KLDP (Korean Linux Documentation Project)에서 제공하는 자연어 처리 패키지 KoNLP (Korean NLP)를 사용하여 시행하였다. 품사태깅을 위한 사전으로는 KAIST 태그 셋을 기초로 KOSPI 및 KOSDAQ시장에 상장된 전 종목과 대상 종목의 주요 제품 등은 고유명사로 태그 셋에 추가하여 사용하였다[22].

### 4.2. 실험 절차

실험은 앞서 제시된 모형에 따라 진행되었다. 2011년 1월 1일부터 2013년 09월 30일까지 총 1003일동안 네이버 포탈 증권 섹션에 게재된 거래소 시가총액 상위 10개 종목의 웹 뉴스를 수집하였으며 전처리 과정을 거쳐 일별 출현 어휘 및 빈도를 데이터화하였다. 이 중 2011년부터 2012년까지 2년간의 데이터는 감성사전 구축에 사용되었으며, 감성사전은 각 종목별로 개별적으로 구축하였다. 감성 사전 구축에 활용된 종가기준 초과이익률은 식(4)에 따라 계산되었으며, 감성사전은 빈도 및 점수를 기준으로 필터링 되었다. 총 빈도가 1인 데이터는 우선적으로 제외되었으며 그 이외의 데이터는 빈도가 누적확률 47%이상 97%이하인 경우에만 감성사전에 포함되었다. 또한 이 중 감성 점수가 -0.1보다 크고 0.1보다 작은 어휘는 중립 어휘로 간주하여 제외하였다. 감성사전 구축 이후에는 감성 사전의 감성 극성 점수를 이용하여 식(2)에 따라 2013년 1월 1일부터 2013년 9월 30일까지의 일별 오피니언 점수를 계산하였다. 산출된 일별 오피니언 점수 데이터 중 2013년 1월 1일부터 2013년 6월 30일까지의 181일 중 영업일 123일간의 데이터는 후기의 테스트 데이터(2013.06.01~2013.09.30)를 표준화하기 위한 통계량을 계산하는데 사용되었다. 이는 일별 오피니언 점수 자체는 분포가 고르지 않고 범위가 상이하여 표준화가 필요하기 때문이다. 최종적으로 계산된 일별 오피니언은 0을 기준으로 0보다 클 경우 상승, 0을 포함하여 0이하일 경우 하락으로 판단하였다. 종속변수로는 해당 종목의 초과수익률을 사용하였으며 당일 초과수익률이 0을 초과할 경우 상승, 0 이하일 경우 하락으로 판단하였다.

*<표 2> 예측 성공 및 실패 구분 조건*

| 구분 | 조건 |
|---|---|
| 예측성공 | $stdDO_{it} > 0,\ AR_{it} > 0$<br>$stdDO_{it} \leq 0,\ AR_{it} \leq 0$ |
| 예측실패 | $stdDO_{it} > 0,\ AR_{it} \leq 0$<br>$stdDO_{it} \leq 0,\ AR_{it} > 0$ |

## 5. 실험 결과

본 실험에서는 일별 오피니언이 임계 값 내의 수치로 나타나 날 경우 중립 오피니언으로 간주하여 투자하지 않는다고 가정하고, $0.0, \pm0.5, \pm1.0$ 의 세 가지 값을 임계 값으로 설정하여 그 결과를 비교해보았다. 일별 오피니언 점수 값이 중립의 범위 내에 있지 않은 총 데이터의 수를 Total, 예측에 성공한 데이터의 수를 True, 중립이 아닌 데이터 중 예측에 성공한 비율(True/Total)을 정확도라 할 때 각 종목의 상승/하락 예측 정확도는 <표 3>와 같다.

*<표 3> 종목별 상승/하락 예측 정확도*

| 종목 | 임계 값 | Total | True | 정확도 |
|---|---|---|---|---|
| 삼성전자 | 0.0 | 62 | 28 | 45.16 |
| | ±0.5 | 31 | 14 | 45.16 |
| | ±1 | 45 | 24 | 53.33 |
| 현대차 | 0.0 | 62 | 27 | 43.54% |
| | ±0.5 | 28 | 11 | 39.29% |
| | ±1 | 10 | 2 | 20.00% |
| 현대 모비스 | 0.0 | 62 | 34 | 54.84% |
| | ±0.5 | 44 | 27 | 61.36% |
| | ±1 | 22 | 14 | 63.64% |
| POSCO | 0.0 | 62 | 32 | 51.61% |
| | ±0.5 | 43 | 22 | 51.16% |
| | ±1 | 27 | 12 | 44.44% |
| 기아차 | 0.0 | 62 | 32 | 51.61% |
| | ±0.5 | 41 | 22 | 53.66% |
| | ±1 | 24 | 13 | 54.17% |
| SK 하이닉스 | 0.0 | 62 | 36 | 58.06% |
| | ±0.5 | 35 | 18 | 51.43% |
| | ±1 | 21 | 12 | 57.14% |

| | | | | |
|---|---|---|---|---|
| 현대 중공업 | 0.0 | 59 | 37 | 62.71% |
| | ±0.5 | 35 | 22 | 62.86% |
| | ±1 | 21 | 13 | 61.90% |
| NAVER | 0.0 | 62 | 30 | 48.39% |
| | ±0.5 | 32 | 16 | 50.00% |
| | ±1 | 14 | 8 | 57.14% |
| 신한지주 | 0.0 | 62 | 34 | 54.84% |
| | ±0.5 | 43 | 24 | 55.81% |
| | ±1 | 24 | 14 | 58.33% |
| 한국전력 | 0.0 | 59 | 38 | 64.41% |
| | ±0.5 | 38 | 26 | 68.42%% |
| | ±1 | 22 | 16 | 72.73% |

각 종목별로는 예측 정확도의 범위나 최적 임계값에 편차가 존재하였다. 종목별 평균 예측정확도는 한국전력이 68.52%로 가장 높았으며 현대 중공업이 62.50%로 뒤따랐다. 또한 한국전력과 현대 중공업을 포함하여 과반수의 종목이 55%를 상회하는 높은 평균 예측정확도를 보였다. 그러나 삼성전자와 현대차의 경우 평균 예측정확도가 47.89%, 32.28%에 그쳐 50%를 하회하였기에 이는 논의의 여지가 있다고 판단하였다.

임계 값에 따른 결과 값을 비교해 보면 10개 종목 중 6개의 종목에서 임계 값과 예측 정확도가 정비례하여 임계 값의 유의성을 확인 할 수 있었다. 하지만 임계 값이 높아질수록 미결정 건수(중립 의견으로 간주하는 건수)가 급격하게 증가하였으며 이에 따라 임계 값을 ±1.0이상으로 높일 경우 일부 종목은 예측 정확도가 오히려 떨어지는 결과를 보였다. 임계 값 변화에 따른 데이터 수의 변화는 <표 4>와 같았으며, 임계 값으로 ±0.5를 적용하였을 경우에는 전체 데이터 중 63.54%, ±1.0 을 적용하였을 경우에는 전체 데이터 중 35.18%가 중립이 아닌 예측 값을 나타냈다. 일별 오피니언 점수 값이 중립의 범위 내에 있지 않은 총 데이터의 수를 Total, 예측에 성공한 데이터의 수를 True, 중립이 아닌 데이터 중 예측에 성공한 비율(True/Total)을 정확도라 할 때 전체 데이터의 상승/하락 예측 정확도는 <표 4>와 같다.

전체 데이터의 상/하락 예측 정확도와 표준편차는 임계 값 별로 차이가 있었으나 ±0.5 일 때 정확도 55.21%, 표준편차 8.07%로 가장 좋은 결과 값을 보였다. 비록 이는 선행 연구방식들과 비교하여 볼 때 괄목할 만한 정확도라 하기는 어렵다. 하지만 일부 종목의 낮은 예측 정확도에도 불구하고 표준편차가 매우 낮고 전체 예측 정확도가 높은 수준이기에 향후 방향성 예측 모델의 활용가능성이 높을 것으로 판단된다.

*<표 4> 전체 데이터의 상승/하락 예측 정확도*

| 임계 값 | Total | True | 정확도 | 평균편차 |
|---|---|---|---|---|
| 0.0 | 614 | 328 | 53.42% | 6.91% |
| ±0.5 | 384 | 212 | 55.21% | 8.07% |
| ±1 | 216 | 119 | 55.09% | 14.19% |

## 6. 논의 사항

본 연구의 결과는 뉴스 데이터를 이용하여 시장 INDEX가 아닌 개별 주가의 방향성을 알아보고자 하였다는 점에서 의의를 가진다. 또한 제시한 모델의 고유한 특성은 다음과 같이 정의할 수 있다. 첫 째, 종목 별로 독립적인 감성사전을 구축하여 사용하였다. 업종이나 재무적 상황이 다를 경우 같은 어휘가 사용되었다 하더라도 주가의 등락에 있어 그 영향은 상이하다. 따라서 본 모델에서는 종목별로 고유의 감성사전을 구축함으로써 어휘가 해당 종목에 미치는 영향을 파악하고 적용하였다.

둘 째, 감성 사전을 구축하거나 뉴스를 분석함에 있어 어휘의 범위에 명사뿐만 아니라 동사와 형용사까지도 포함하였다. 한국어에서 실질적인 감성을 나타내는 어휘는 주로 동사나 형용사이다. 하지만 한국어의 언어 체계상 용언은 문맥에 따라 형태를 달리하기 때문에 명확히 구분하기 어려워 주로 그 형태를 명확히 구분할 수 있는 명사만을 오피니언 마이닝의 대상으로 해왔다. 그러나 본 모델은 파싱의 결과로 나온 어휘에 점수를 부여하는 방식을 사용하기에 동사나 형용사의 형태변화에 관계없이 어간에 점수를 부여하여 카운팅하였다.

셋 째, 어휘가 등락에 미치는 영향의 정도가 감성 극성 점수에 포함되었다. 본 모델에서는 어휘의 감성 극성 점수를 계산하는 과정에서 해당 어휘의 출현 이후 AR값에 어휘의 출현빈도를 가중치로 하는 가중평균 방식을 사용하였다. 주식 수익률이 높은 날에 많이 출현할수록 어휘의 감성 극성 점수가 높아지는 것이다. 따라서 별도의 학습 과정이 없이 이전의 데이터에서 추출한 통계량으로 일별 오피니언 점수를 표준화하기만 하면 그 자체로 방향성을 예측할 수 있다.

## 7. 결론

본 논문에서는 웹에서 발생하는 뉴스 데이터를 이용하여 익일의 개별주식의 초과수익률의 방향성을

예측하는 모델을 제안하였다. 네 단계로 구성되어 있는 본 모델은 언어적 특성을 고려하여 전처리 하고 이 중 일부 데이터를 이용하여 종목별로 감성사전을 구축하며, 해당 감성사전을 이용한 일별 주가 예측 점수를 계산하고 이를 토대로 하여 향후 해당 주식의 초과 수익률 방향성을 예측한다. 모델의 성과를 평가하는 데는 2013년7월부터 2013년 9월까지의 데이터가 사용되었으며 그 결과 평균 예측정확도는 중립 오피니언을 구분하기 위한 임계 값 수준에 따라 53.42%, 55.21%, 55.09%로 유의한 수준에서 나타났고, 표준 편차를 고려하였을 때 통상적으로 가장 유의한 임계 값은 $\pm 0.5$ 라고 판단되었다. 아울러 본 연구의 예측 정확도 결과는 타 연구와 유사한 수준의 수치를 보였으나 종목별 예측 정확도의 편차가 낮기에 향후 활용가능성이 높을 것으로 기대된다.

본 연구의 한계점은 시가총액 상위 10개 종목에 한정하여 실험을 진행하였기에 표본종목이 전체를 대표한다고 할 수 없으며, 모델 자체의 성과가 유의하였다고 단정하기에는 무리가 있다. 또한, 개별 주식 수익률에 다소 영향을 미치는 재무적 사항에 대해 고려하지 않고 있어 향후 개선의 여지가 있다. 제시한 모델의 세부적인 각 단계에서도 한계점이 다소 존재하였다. 우선 전처리 과정에 있어 자연어 처리가 완전하지 못했다는 점이다. 어간에 점수를 부여하는 방식이기는 했지만 일부 같은 어간을 다르게 인식하는 문제가 있었다. 감성사전의 구축과 활용에 있어서도 실험 데이터에 비해 장기간 동안 수집된 데이터를 사용하여야 하는 것은 분명하지만 그 기간이 적합했는가 하는 부분에서 한계가 존재하였다. 하루가 다르게 변화하는 시장에서 지난 2년간의 단어 극성이 현재에도 적용될 수 있는가라는 점에서 의문이 남아있으며 이는 추후 별도의 검증과 개선이 필요할 것으로 생각된다. 셋째로는 어휘를 점수화하는 과정에서 문맥이나 어휘간의 상관관계를 반영하지 못했다는 점이다. 실제 뉴스기사에 출현하는 다수의 어휘는 서로 상관성을 가지고 있기에 이러한 한계를 극복하는 연구를 수행할 필요가 있다.

## 참고문헌


[1] Eugene F.Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work”, *journal of finance*, Vol 25, Issue2(1970), pp.383-417

[2] Burton G. Malkiel, “The Efficient Market Hypothesis and Its Critics”, *journal of economic perspectives*, Vol17, No.1-winter(2003), pp.59-82

[3] Robert j. Shiller, “From efficient markets theory to Behavioral Finance”, *journal of economic perspectives*, Vol17, No.1-winter(2003), pp.83-104

[4] Jegadeesh, N., and S. Titman, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.", *Journal of Finance*, Vol.48, No.1-March(1993), pp.65-91

[5] 위한종, 곽세영, “투자자 심리와 주가 과민반응”,*대한경영학회*, Vol51(2005), pp.1623-1642

[6] R. Schumaker, and H. Chen, “A discrete stock price prediction engine based on financial news”, *IEEE Computer Society*, Vol43, No.2, pp.51-56 (2010)

[7] 안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측”, *한국컴퓨터종합학술대회* , Vol37, No.1(2010), pp.364-369

[8] 김유신, 김남규, 정승렬, “뉴스와 주가: 빅데이터 감성분석을 통한 지능형 의사결정모형”, *지능정보연구*, 제18권 제2호(2012), pp.143~156

[9] 김유신, 주가 지수 예측을 위한 뉴스 빅데이터 오피니언마이닝 모형, 국민대학교, 2013

[10] 유은지, 김유신, 김남규, 정승렬, “주가 지수 방향성 예측을 위한 주제지향 감성사전 구축 방안”, *지능정보연구*, Vol19, No.1(2013), pp.95-110

[11] 김명규, 김정호, 차명훈, 채수환, “텍스트 문서 기반의 감성 인식 시스템”, *감성과학*, Vol12, No.4(2009), pp433-442

[12] 김진옥, 한글 텍스트의 오피니언 분류 자동화 기법, 이화여자대학교, 2010

[13] Mitchell M. L, Mulherin J. H, “The Impact of Public Information on the Stock Market”, *The journal of Finance,* Vol.49, N0.3(1994),923-950

[14] V. Cho , B. Wüthrich , J. Zhang, Text Processing for Classification, The Hongkong University, 1998

[15] G.Gidofalvi, Using News Articles to Predict Stock Price Movements, University of San Diego, 2003

[16] G. Fung, J. Yu, W.Lam, “Stock Prediction: Integrating Text mining approach using real-time news”, *IEEE Computational intelligence for financial engineering*, (2003), pp.395-402

[17] Robert P. Schumaker, Hsinchun Chen, “Textual analysis of stock market prediction using breaking financial news: The AZFin text system”, *ACM Transactions on Information Systems,* Vol.27, No.2(2009), article no.12

[18] Liangfei Qiu, Huaxia Rui, Andrew Whinston, “A Twitter-Based Prediction Market: Social Network Approach”, *ICIS* (2011)

[19] Johan Bollena, Huina Maoa, Xiaojun Zeng, “Twitter mood predicts the stock market”, *Journal of Computational Science*, Vol2, No1(2011), 1-8

[20] 삼성경제연구소, “빅데이터: 산업 지각변동의 진원”, 제851호(2012)

[21] 한봉희, “비기대이익과 초과수익률의 측정”, *한국증권학회,* Vol29, No1(2001), pp.183-214

[22] KAIST 시멘틱 웹 첨단연구센터, “한나눔 한국어 형태소 분석기”, 2010