ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2017 June: 23(2): 123~138 http://dx.doi.org/10.13088/jiis.2017.23.2.123

카테고리 중립 단어 활용을 통한 주가 예측 방안: 텍스트 마이닝 활용*

이민식
연세대학교 정보산업공학과
(salvia0413@gmail.com)

이홍주
가톨릭대학교 경영학부
(hongjoo@catholic.ac.kr)

주식 시장은 거래자들의 기업과 시황에 대한 기대가 반영되어 움직이기에, 다양한 원천의 텍스트 데이터 분석을 통해 주가 움직임을 예측하려는 연구들이 진행되어 왔다. 주가의 움직임을 예측하는 것이기에 단순히 주가의 등락 뿐만이 아니라, 뉴스 기사나 소셜 미디어의 반응에 따라 거래를 하고 이에 따른 수익률을 분석하는 연구들이 진행되어 왔다. 주가의 움직임을 예측하는 연구들도 다른 분야의 텍스트 마이닝 접근 방안과 동일하게 단어-문서 매트릭스를 구성하여 분류 알고리즘에 적용하여 왔다.

문서에 많은 단어들이 포함되어 있기 때문에 모든 단어를 가지고 단어-문서 매트릭스를 만드는 것보다는 단어가 문서를 범주로 분류할 때 기여도가 높은 단어들을 선정하여야 한다. 단어의 빈도를 고려하여 너무 적은 등장 빈도나 중요도를 보이는 단어는 제거하게 된다. 단어가 문서를 정확하게 분류하는 데 기여하는 정도를 측정하여 기여도에 따라 사용할 단어를 선정하기도 한다.

단어-문서 매트릭스를 구성하는 기본적인 방안인 분석의 대상이 되는 모든 문서를 수집하여 분류에 영향력을 미치는 단어를 선정하여 사용하는 것이었다. 본 연구에서는 개별 종목에 대한 문서를 분석하여 종목별 등락에 모두 포함되는 단어를 중립 단어로 선정한다. 선정된 중립 단어 주변에 등장하는 단어들을 추출하여 단어-문서 매트릭스 생성에 활용한다. 중립 단어 자체는 주가 움직임과 연관관계가 적고, 중립 단어의 주변 단어가 주가 상승에 더 영향을 미칠 것이라는 생각에서 출발한다. 생성된 단어-문서 매트릭스를 가지고 주가의 등락 여부를 분류하는 알고리즘에 적용하게 된다.

본 연구에서는 종목 별로 중립 단어를 1차 선정하고, 선정된 단어 중에서 다른 종목에도 많이 포함되는 단어는 추가적으로 제외하는 방안을 활용하였다. 온라인 뉴스 포털을 통해 시가 총액 상위 10개 종목에 대한 4개월간의 뉴스 기사를 수집하였다. 3개월간의 뉴스 기사를 학습 데이터로 분류 모형을 수립하였으며, 남은 1개월간의 뉴스 기사를 모형에 적용하여 다음 날의 주가 움직임을 예측하였다. 본 연구에서 제안하는 중립 단어 활용 알고리즘이 희소성에 기반한 단어 선정 방안에 비해 우수한 분류 성과를 보였다.

주제어 : 주가 예측, 중립 단어, 텍스트 마이닝, 온라인 뉴스

논문접수일 : 2017년 4월 9일 논문수정일 : 2017년 5월 27일 게재확정일 : 2017년 5월 29일
원고유형 : 일반논문 교신저자 : 이홍주

1. 개요

소셜 미디어의 포스트, 온라인 리뷰, 뉴스 기사 등의 텍스트를 분석하여 특정 이벤트, 제품, 서비스 등에 대해 느끼는 점을 파악하는 텍스트 마이닝이 여러 분야에 적용되어 왔다(Cao et al.,

* 본 연구는 2017년도 가톨릭대학교 교비연구비의 지원을 받아 수행되었습니다.

2011.; Jeong et al., 2015; Oh and Sheng, 2011). 사람들이 생각하고 느끼는 바에 따라 결과가 정해지거나 결과가 변화하는 분야에서는 해당 분야에 대한 사용자들의 텍스트를 분석하여 결과 변화와 텍스트 간의 관계를 연구하여 왔다. 영화제 수상 예측이나(Bothos et al., 2010), 선거 결과 예측(Tumasjan et al., 2010), 주식 시장 예측(Schumaker, 2009)이 대표적인 적용 분야이다.

텍스트 마이닝의 가장 기본적인 접근 방안은 문서에서 단어를 추출하여 단어-문서 매트릭스(Term-Document Matrix)를 만드는 것이다(Perkins, 2014). 단어-문서 매트릭스를 입력으로 하여 특정 단어들이 포함된 문서들을 어떤 범주로 분류할지 정하는 모형을 학습하고 모형의 성과를 측정하는 방식으로 연구가 이루어진다. 문서에 많은 단어들이 포함되어 있기 때문에 모든 단어를 가지고 단어-문서 매트릭스를 만드는 것보다는 단어가 문서를 범주로 분류할 때 기여도가 높은 단어들을 선정하여야 한다. 단어의 빈도, TF-IDF 등을 고려하여 너무 적은 등장 빈도나 중요도를 보이는 단어는 제거하게 된다(Lee and Lee, 2016; Perkins, 2014). 단어가 문서를 정확하게 분류하는 데 기여하는 정도를 측정하여 기여도에 따라 사용할 단어를 선정하기도 한다(Choeh et al., 2015; Zhang and Tran, 2011).

주식 시장도 거래자들의 기업과 시황에 대한 기대가 반영되어 움직이기에, 다양한 원천의 텍스트 데이터 분석을 통해 주가 움직임을 예측하려는 연구들이 진행되어 왔다(Jeong et al., 2015; Kim et al., 2012; Oh and Sheng, 2011; Schumaker, 2009). 주가의 움직임을 예측하는 것이기에 단순히 주가의 등락뿐만이 아니라, 뉴스 기사나 소셜 미디어의 반응에 따라 거래를 하고 이에 따른 수익률을 분석하는 연구들이 진행되어 왔다(Ding et al. 2014). 주가의 움직임을 예측하는 연구들도 다른 분야의 텍스트 마이닝 접근 방안과 동일하게 단어-문서 매트릭스를 구성하여 분류 알고리즘에 적용하여 왔다.

단어-문서 매트릭스를 구성하는 기본적인 방안인 분석의 대상이 되는 모든 문서를 수집하여 분류에 영향력을 미치는 단어를 선정하여 사용하는 것이었다. 본 연구에서는 개별 종목에 대한 문서를 분석하여 종목별 등락에 모두 포함되는 단어를 중립 단어로 선정한다. 선정된 중립 단어 주변에 등장하는 단어들을 추출하여 단어-문서 매트릭스 생성에 활용한다. 중립 단어 자체는 주가 움직임과 연관관계가 적고, 중립 단어의 주변 단어가 주가 상승에 더 영향을 미칠 것이라는 생각에서 출발한다. 생성된 단어-문서 매트릭스를 가지고 주가의 등락 여부를 분류하는 알고리즘에 적용하게 된다.

본 연구에서는 종목 별로 중립 단어를 1차 선정하고, 선정된 단어 중에서 다른 종목에도 많이 포함되는 단어는 추가적으로 제외하는 방안을 활용하였다. 온라인 뉴스 포털을 통해 시가 총액 상위 10개 종목에 대한 4개월간의 뉴스 기사를 수집하였다. 3개월간의 뉴스 기사를 학습 데이터로 분류 모형을 수립하였으며, 남은 1개월간의 뉴스 기사를 모형에 적용하여 다음 날의 주가 움직임을 예측하였다. 본 연구에서 제안하는 중립 단어 활용 알고리즘이 희소성에 기반한 단어 선정 방안에 비해 우수한 분류 성과를 보였다.

2. 관련 연구

주가와 관련하여 다양한 데이터를 기반으로 주식과 관련된 정보 분석과 주가 예측 연구가 진

행되었다. 시계열 데이터에 기반한 분석 방법으로 Jeantheau(2004)는 ARCH 모형, Amilon(2003)는 GARCH 모형을 통해 주가 예측을 진행했다. Park and Shin(2011)은 기존의 시계열 분석에 사용하는 정보와 함께 타기업이나 각종 경제지표들을 바탕으로 기계학습을 통한 주가 예측을 진행했다. 기계학습의 발전과 함께 인공신경망 기반의 주가 예측 연구도 진행되었다. Kim and Lee(2008)은 기업의 재무 데이터를 분석하여 인공신경망 모형을 구축하였다. Lee(2008)은 회귀모형, 인공신경망, SVM 모형을 결합하여 결합모형을 통한 주가 예측을 진행했다.

기업 관련 뉴스들을 분석하여 포함된 텍스트와 기업들의 주식 가치 변화 간의 상관관계를 파악하는 연구들도 진행되어 왔다(Jeong et al., 2015; Kim et al., 2012, Schumaker, 2009). 기업에 대한 뉴스 기사를 수집한 이후에 기사에 포함된 단어들의 감성 극성(sentiment polarity)를 측정하여 해당 기업 주가의 등락에 미치는 영향을 학습하고, 이를 기반으로 향후 주가를 예측하는 것이 기본적인 접근 방안이다. 소셜 미디어에서 주식시장이나 기업에 관련된 메시지를 수집하여 기업 주가를 예측하는 연구들도 기본 접근 방안은 비슷하였다(Oh and Sheng, 2011).

뉴스 텍스트에서 Bag of Words 방식의 단어 추출을 진행하고 상승, 하락, 변화 없음 등으로 분류하는 방안을 활용하였으며, Naïve Bayesian 분류기(Ahn and Cho, 2010; Seo et al., 2002), SVM(Fung et al., 2002; Mittermayer, 2004), Genetic Algorithm(Thomas and Sycara, 2002)를 활용하였다. Schumaker (2009)는 뉴스 기사에서 추출한 단어와 주식 가격 데이터를 SVM을 통해 수립한 모형을 통해 뉴스 출시 20분 후의 주가 움직임을 예측하였다. 뉴스 기사에 포함된 단어, 명사, Named Entities을 활용하여 다양한 실험을 수행하였으며, 트레이딩 시뮬레이션 및 주가 움직임 방향성 정확도는 명사만 활용한 경우의 성과가 제일 좋았다.

뉴스 기사들의 감성 극성 파악을 위해 범용적인 사전을 활용하여 뉴스 내용의 긍정, 부정 정도를 측정하였으며, 감성 극성과 주가 움직임 간의 연관관계를 예측에 활용하였다(Kim et al., 2012; Jeong et al. 2015). 일상에서 쓰이는 범용 사전을 기업 뉴스의 감성 극성 파악에 활용하는 것은 기업의 주가 상승에 긍정적인 내용과 부정적인 내용을 파악하는 데 어려움이 있다(Jeong et al., 2015). 범용적인 단어 사전을 활용하는 문제를 해결하기 위하여 연구들은 수집한 기업 뉴스 이후에 주가가 상승하였는지 하락하였는 지를 활용하여 주가에 영향을 미치는 단어사전을 직접 만들어 활용하였다(Yu et al., 2013). Kim et al. (2012)는 뉴스의 긍정/부정 오피니언 마이닝을 통해서 감성분석을 진행하고 이를 기반으로 하는 지능형 투자의사결정 모형을 형성했다. Jeong et al. (2015)는 단어사전을 직접 만들 때 여러 기업들의 뉴스에 기반하여 기업들에 공통적으로 활용할 수 있는 사전을 만들지만, 개별 기업별로 단어들이 미치는 영향이 다르기에 개별 기업별로 단어사전을 만들어 주가를 예측하는 연구를 수행하였다.

온라인 리뷰를 통해 제품이나 서비스에 대한 평가를 수행하거나 유용한 리뷰를 파악하는 연구들에서도 제품 특성 별로 추출할 단어를 선정하거나(Cao et al., 2011), 단어-문서 매트릭스 생성에 활용하는 단어를 선정하는 방안을 제안하였다(Lee and Lee, 2016). Lee and Lee (2016)은 제품별로 유용한 리뷰와 유용하지 않은 리뷰에 비슷한 비율로 등장하는 단어를 중립 단어로 정

의하고, 이를 다양하게 제거하여 분류 성과를 향상시켰다.

Ding et al. (2014)와 Ding et al. (2015)는 뉴스 기사에 포함된 단어를 전부 사용하는 것보다 이벤트 속성에 해당하는 단어들만 사용하는 것이 더욱 예측 정확도가 높은 것을 보였으며, 예측에는 SVM, 신경망과 딥 러닝 기법을 활용하였다. SVM이나 신경망을 사용한 경우보다 딥 러닝 기법을 적용한 모형의 성과가 더 좋았다.

3. 중립 단어 활용 알고리즘

본 연구에서 제안하는 방안은 개별 종목의 뉴스 기사를 분석하여 종목별 등락에 모두 포함되는 단어를 중립 단어로 선정하는 것이다. 그리고 선정된 중립 단어 주변에 등장하는 단어들을 추출하여 단어-문서 매트릭스 생성에 활용한다. 예를 들어, 현대차의 뉴스에 SUV라는 단어가 주가가 상승하는 경우와 하락 경우에 비슷한 비율로 등장한다면 중립 단어가 된다. 뉴스 기사에서 SUV 주변에 등장하는 단어들이 단어-문서 매트릭스 생성에 활용된다. 즉, SUV라는 단어 자체는 주가상승에 영향력이 적고, 주변에 주가 상승에 긍정적이거나 부정적인 내용이 포함되어 있을 것이라는 생각에서 출발한다. 생성된 단어-문서 매트릭스를 가지고 주가의 등락 여부를 분류하는 알고리즘에 적용하게 된다.

단어-문서 매트릭스를 구성하는 기본적인 방안은 분석의 대상이 되는 모든 뉴스 기사를 수집하여 분류에 영향력을 미치는 단어를 선정하여 사용하는 것이었다. 본 연구에서는 종목 별로 중립 단어를 1차 선정하고, 선정된 단어 중에서 다른 종목에도 많이 포함되는 단어는 추가적으로 제외하는 방안을 활용하였다.

3.1 구조적 단어 추출 방안

구조적 단어를 추출하는 방법은 중립 단어를 선정하고 중립 단어가 포함된 문장에서 나머지 단어들을 추출하는 방법이다. 본 연구에서는 10가지 주식 종목과 관련한 뉴스 데이터를 교차 사용하는 방법으로 용언, 체언, 외국어만 사용하여 불용어를 제거하고 중립 단어를 설정하였다.

3.1.1 불용어 설정

단어의 출현 빈도에 근간하여 불용어를 설정하거나 단어의 정보 기여도를 산출하여 정보 기여가 낮은 단어들은 제거하는 방식이 사용되어 왔다(Perkins, 2014). 본 연구에서는 온라인 뉴스의 특성을 고려하여 뉴스의 출처별로 불용어를 설정했다. 뉴스의 출처별로 불용어를 설정하는 이유는 <Table 1>과 같이 온라인 뉴스의 출처에 따라서 뉴스의 내용과 관계없이 등장하는 특정한 단어들이 있기 때문이다.

아래 알고리즘 1과 같이 뉴스 출처별 불용어가 설정되었다. 먼저 주식 종목(S_k)과 뉴스의 출처(P_n)를 구분하여 종목과 출처에 따라서 뉴스에 등장하는 단어($W_{(S_k,\ P_n,\ w)}$)를 추출한다. 추출된 단어가 해당 종목과 출처의 조합에서 등장하는 확률이 0.5 이상인 단어를 종목과 출처에 따른 불용어($Stopwords_{(S_k,\ P_n,\ w)}$)로 설정한다. 또 주식 종목과 관계없이 출처별로 모든 종목에 빈번히 등장하는 단어를 뉴스 출처별 불용어($Stopwords_{(P_n,\ r)}$)로 정하였으며, 이는 주식 종목 수의 90%이상에 등장한 단어들로 구성된다. 출처별 불용어의 예시는 <Table 1>과 같다.

Algorithm 1. Setting up stopwords removed according to news sources

Input. 주식종목 S_k

뉴스의 출처 P_n

종목과 출처에 따른 뉴스에 등장하는 단어 $W_{(S_k,\ P_n,\ w)}$

Output. 뉴스 출처별 불용어 $Stopwords_{(P_n,\ r)}$

Method.

If $\Pr(W_{(S_k,\ P_n,\ w)}) > 0.5$ **Then** // 종목과 출처에 따라서 등장하는 단어의 확률

$Stopwords_{(S_k,\ P_n,\ w)} = W_{(S_k,\ P_n,\ w)}$

End If

For j = 1 **to** w // 단어

For i = 1 **to** n // 뉴스 출처

count = 0

For q = 1 **to** k // 종목

If $Stopwords_{(S_q,\ P_i,\ j)}$ not NULL **Then**

count ++

End If

Next q

If count >= (k*0.9) **Then**

$Stopwords_{(P_i,\ i)} = Stopwords_{(S_q,\ P_i,\ j)}$

Next i

Next j

〈Table 1〉 Stopwords removed according to news sources

news sources	stopwords
YTN	PLUS YTN 금지 무단 재배포 저작권자 전재
Yonhap news	co kr yna 금지 무단 서울 연합뉴스 재배포 저작권자 전재 핫클릭
JTBC	All Co Copyright DramaHouse JTBC JcontentHub Ltd Reserved Rights SNS by 공식 금지 기다리 기자 뉴스 무단 생생하 앵커 여러분 유튜브 재배포 전재 제보 카카오스토리 트위터 페이스북
Finanacial news	com fnnews 금지 무단 재배포 저작권자 전재 파이낸셜뉴스

3.1.2 중립 단어 선정

중립단어는 다음과 같은 알고리즘을 통해 선정되었다. 각 종목별 뉴스에 등장하는 단어를 추출하여 해당 단어의 등장 확률($\Pr(W_{(S_k,\ news_{(d,\ n)},\ w)})$)을 구한다. 이후 단어의 희소성을 구하여 등장하는 빈도가 α 보다 희소한 단어들을 제거하고 나머지 단어들의 가격 등락에 따른 등장 확률을 구한다. 가격 상승에 따른 단어 출현 확률($\mathrm{P}(Pluswords_{(S_q,\ news_{(p,\ i)},\ j)} \mid Price(S_q,\ Day_p))$)과 가격 하락에 따른 단어 출현 확률($\mathrm{P}(Minuswords_{(S_q,\ news_{(p,\ i)},\ j)} \mid Price(S_q,\ Day_p))$)을 비교하여 확률 값이 유사한 경우, 가격과 상관없이 등장하는 단어로 판단하여 1차 중립 단어로 설정한다.

1차 중립 단어의 경우 많은 단어들이 일반적인 뉴스에서 볼 수 있는 단어들이며 주식 종목과 관련 없는 단어들이다. 예를 들어 <Table 2>는 α가 0.2일때의 삼성전자와 현대차의 1차 중립 단어이다. 삼성전자는 '갤럭시 S7', '모바일', '스마트폰', 현대차는 '하이브리드', 'SUV', '자동차' 등 종목과 관련된 단어들이 등장한다. 하지만 '밝히', '올해', '이상', '한국', '일부'등 일반적인 뉴스에서 볼 수 있는 단어들도 중립 단어로 포함된다.

따라서 주식 종목과 관련 없는 중립 단어를 제거하기 위한 절차를 추가하였다. 10개의 주식 종목에서 선정된 1차 중립 단어들을 누적하여 빈도가 β이상의 값을 갖는 공통 단어들을 제거한 후 최종 중립 단어($NeutralW_{(S_q,\ j)}$)를 추출한다. 종목과 상관없이 빈번히 등장한다는 것은 구체적인 종목에 관련된 단어가 아닌 일반적인 단어로 본 것이다. <Table 3>은 β가 3인 경우의 최종 중립 단어이다. 삼성전자의 경우 141개의 1차 중립 단어가 60개로, 현대차의 경우 147개의 중립 단어가 46개로 축소되었다.

Algorithm 2. Defining the neutral terms

Input. 주식종목 S_k

날짜별 뉴스 $news_{(d,\ n)}$

종목과 날짜에 따라서 뉴스에 등장하는 단어 $W_{(S_k,\ news_{(d,\ n)},\ w)}$

Output. 종목별 중립 단어 $NeutralW_{(S_k,\ w)}$

Method.

```
If Pr(W_(Sk, news(d, n), w)) < α Then        // 종목별 뉴스 단어 등장 확률
   W = W ∩ W_(Sk, news(d, n), w)^c
End If
Pluswords = 0
Minuswords = 0
For j = 1 to w                               // 단어
   Neutralcount = 0
   For q = 1 to k                            // 종목
      For p = 1 to d                         // 날짜
         For i = 1 to n                      // 뉴스
```

If $Price(S_q,\ Day_p) > 0$ & $W_{(S_q,\ news_{(p,\ i)},\ j)}$ not NULL **Then** // 주가 상승 날

$Pluswords_{(S_q,\ news_{(p,\ i)},\ j)} ++$

Else If $Price(S_q,\ Day_p) < 0$ & $W_{(S_q,\ news_{(p,\ i)},\ j)}$ not NULL **Then** // 주가 하락 날

$Minuswords_{(S_q,\ news_{(p,\ i)},\ j)} ++$

Next i

Next p

If $P(Pluswords_{(S_q,\ news_{(p,\ i)},\ j)} \mid Price(S_q,\ Day_p)) \approx P(Minuswords_{(S_q,\ news_{(p,\ i)},\ j)} \mid Price(S_q,\ Day_p))$ **Then**

$NeutralW_{(S_q,\ j)} = NeutralW_{(S_q,\ j)} \cup W_{(S_q,\ news_{(p,\ i)},\ j)}$ // 1차 중립 단어로 선정

$NeutralWcount ++$

Next q

If $Neutralcount \geq \beta$ **Then**

$NeutralW_{(S_q,\ j)} = NeutralW_{(S_q,\ j)} \cap W_{(S_q,\ news_{(p,\ i)},\ j)}{}^{c}$

Next i

Next j

〈Table 2〉 The first neutral terms according to stocks (α = 0.2)

Stock	The first neutral terms
Samsung Electronics	삼성전자 LG 국내 기업 따르 때문 밝히 보이 삼성 세계 스마트폰 시장 올해 이상 전망 제품 지나 지난해 2014년 IT SK 가격 가능 가능성 가운데 가지 강화 개발 개선 갤럭시S7 결과 경우 경쟁 계획 공개 관계자 관련 관심 규모 글로벌 기능 기대 기록 기술 기존 기준 나오 높이 다르 다양한 대비 대표 만들 모바일 못하 미국 반면 발표 분석 분야 빠르 사업 사용 사진 상승 상황 새롭 서비스 서울 선보이 설명 성장 수준 시작 실적 열리 애플 어렵 업계 업체 열리 예상 예정 오르 이날 이루 이르 이번 이후 적용 전략 전체 정도 제공 주목 주요 중국 중심 지원 진행 최대 출시 투자 판매 평가 필요 확대 회사 1위 3월 갖추 갤럭시 경기 경쟁력 기반 나서 내놓 내리 대부분 대상 떨어지 모델 모습 방식 사람 사장 산업 시간 올리 운영 으로 일부 전하 지난달 최고 추가 카메라 프리미엄 한국 행사 확보
Hyundai Motor	현대차 계획 관계자 국내 기록 기아차 때문 미국 밝히 보이 시장 올해 이상 이후 자동차 지나 지난해 판매 현대자동차 1월 2014년 3월 LG SK SUV 가격 가능성 가운데 강화 개발 개선 개인 거래 결과 경개 경우경쟁 관련 규모 글로벌 기관 기대 기술 기존 기준 나오 나타나 내리 다르 다양한 대비 대표 떨어지 만들 모델 못하 문제 반면 발표 분석 브랜드 사업 상승 상황 새롭 서울 설명 성장 세계 수준 시가총액 시작 이번 적용 전망 전체 정도 정부 제공 주요 중국 중심 증가 지난달 지원 진행 차량 최대 추가 출시 코스피 투자 하락 한국전력 현대 확대 회사 가능 강조 경쟁력 공개 관심 나서 나타내 내놓 달하 동안 마감 목표 미래 분야 비중 비하 삼성물산 상위 선보이 약세 업종 연구원 우려 이유 일본 일부 전날 증시 처음 최초 평가 프로그램 필요 하이브리드 한국 해외 환율 회장

〈Table 3〉 The final neutral terms (β = 3)

Stock	The final neutral terms
Samsung Electronics	삼성 스마트폰 제품 IT 기능 가능 가지 갤럭시S7 경쟁 공개 기능 기술 높이 모바일 분야 빠르 사용 사진 새롭 게 서비스 선보이 알리 애플 업체 이루 이르 적용 전략 정도 주목 출시 판매 평가 필요 회사 1위 갖추 갤럭시 경쟁력 기반 나서 내놓 모델 모습 방식 사람 사장 산업 시간 올리 으로 일부 전하 지난달 최고 카메라 프리미엄 한국 행사 확보 삼성전자
Hyundai Motor	자동차 판매 현대자동차 1월 SUV 경쟁 기술 모델 문제 브랜드 새롭 신차 업체 적용 정도 지난달 차량 출시 현대 회사 가능 강조 경쟁력 공개 나서 내놓 달하 동안 목표 미래 분야 비중 비하 선보이 우려 일본 일부 처음 최초 평가 필요 하이브리드 한국 해외 회장 현대차

3.1.3 문장기준 단어 추출

중립 단어가 포함된 문장에서 중립 단어의 주변 단어를 추출한다. 예를 들어 현대차의 중립 단어가 SUV일 때 “2016년 하반기 소비자들의 SUV 시장 점유율이 증가 하고 있다.”라는 문장에서 주변 단어({‘2016년’, ‘하반기’, ‘소비자’, ‘시장’, ‘점유율’, ‘증가’})를 추출한다. 만약 위 문장과 약간 다른 “2016년 하반기 소비자들의 자동차 시장 점유율이 증가 하고 있다.”라는 문장이 있다면, 여기에는 중립 단어가 포함되어있지 않기 때문에 어떤 단어도 추출하지 않는다.

위와 같은 단계를 걸쳐 추출된 단어를 가지고 단어-문서 매트릭스를 구성하였으며, 분류 알고리즘 모형 생성의 입력 자료로 활용하였다.

4. 자료 수집 및 실험

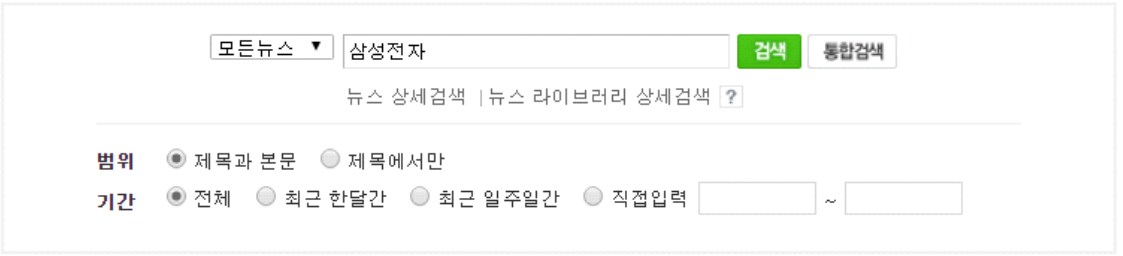
본 연구는 2016년 6월 시점에서 KOSPI 시가총액 기준 상위 10개 종목을 선정하였다. 해당 기업의 뉴스를 2016년 2월 1일부터 2016년 5월 31일까지 수집하였다. 뉴스 수집은 네이버 뉴스를 통해 이루어 졌으며, R에서 N2H4 패키지를 활용하였다(https://github.com/forkonlp/N2H4). KOSPI 시가총액 상위 10개 종목은 삼성전자(005930), 한국전력(015760), 현대차(005380), 현대모비스(012330), 아모레퍼시픽(090430), 삼성물산(028260), NAVER(035420), SK하이닉스(000660), 삼성생명(032830), POSCO(005490) 이다. 삼성전자우선주(005935)의 경우 삼성전자와 매우 유사한 주식이기 때문에 이를 제외하고 다음 순위 주식 종목을 포함하여 10개의 종목을 선정하였다.

4.1.1 주가 정보

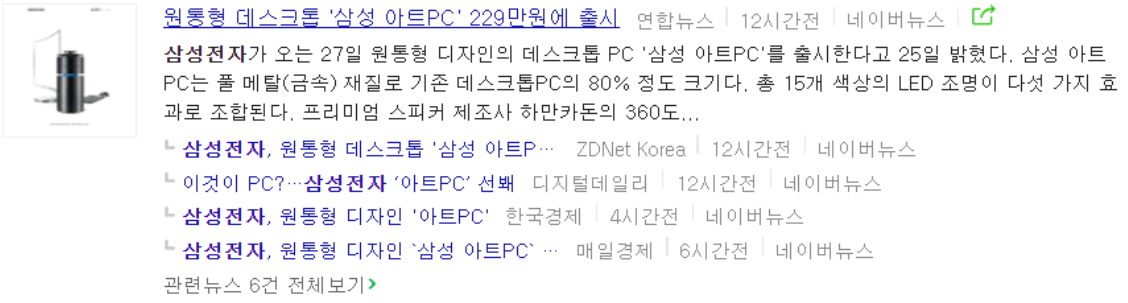
4개월(2016/02/01 ~ 2016/05/31) 동안 주식 시장이 개장한 날은 총 80일이며 학습 집합으로 초기 60일, 테스트 집합으로 나머지 20일을 활용하였다. 데이터 수집 시기의 주가 상승, 하락의 경우가 보합인 경우 보다 무척 많았기에 보합인 날은 제외하고, 주가 상승과 하락이라는 두 개의 카테고리로 분류하였다(<Table 4> 참조).

〈Table 4〉 Number of data except for steadiness

Stock	Number of days		Number of news	
	Train data	Test data	Train data	Test data
Samsung Electronics	58	20	24,715	6,580
Korea Electric Power Corporation	58	20	4,978	1,863
Hyundai Motor	54	18	8,729	2,518
Hyundai Mobis	53	18	2,038	668
AmorePacific	58	19	3,341	1,152
Samsung C&T	53	17	4,571	1,062
NAVER	56	20	13,320	5,054
SK hynix	53	18	2,700	782
Samsung Life Insurance	51	14	2,053	567
POSCO	60	16	7,779	1,863



〈Figure 1〉 News article search screen



〈Figure 2〉 Duplicate news article

4.1.2 온라인 뉴스

주가 예측을 위한 온라인 뉴스는 국내 웹 포털 사이트인 네이버를 통해 수집하였다 <Figure 1>.

예를 들어 삼성전자 주식과 관련된 뉴스 기사를 수집하기 위해서 종목의 이름에 해당하는 삼성전자를 키워드로 하여 나타나는 모든 뉴스를 수집하였다. 단, 현대차, NAVER, POSCO의 경우 각각 현대자동차, 네이버, 포스코가 동일한 종목을 의미한다는 점을 고려하여 키워드를 확대하여 뉴스 기사를 수집하였다.

해당 포털 사이트에서 수집한 뉴스 기사에는 출처가 다르지만 동일한 내용의 기사가 중복 되어 있었다. 예를 들어 <Figure 2>의 '삼성전자가 아트PC를 출시했다.'는 기사는 연합뉴스, ZDNet Korea, 디지털데일리 등에서 같은 뉴스가 반복적으로 등장하였다. 중복되는 기사를 제거하기 뉴스의 제목을 분석하여 이미 수집한 뉴스 기사와 유사한 경우 제거하였다. 유사한 뉴스를 파악하는 방법은 뉴스의 제목에 포함된 단어들을 추출하여 코사인 유사도를 측정한 후 0.5 이상인 뉴스들을 중복 뉴스로 판단하여 제거하였다 (Huang, 2008).

4.1.3 실험 결과

본 연구에서 제안하는 구조적 단어 추출 방법과 기존의 희소성 기반의 단어 추출 방법을 활용하여 생성된 단어-문서 매트릭스를 가지고 분류 알고리즘에 적용하였다. SVM(Meyer et al, 2012),

〈Table 5〉 Prediction accuracy

	The word selection method based on sparsity			The word selection method based on structure		
Stock	SVM	Boosting	RandomForest	SVM	Boosting	RandomForest
Samsung Electronics	0.60	0.45	0.60	0.60	0.65	0.60
Korea Electric Power Corporation	0.60	0.50	0.60	0.60	0.63	0.60
Hyundai Motor	0.60	0.66	0.70	0.60	0.66	0.60
Hyundai Mobis	0.40	0.50	0.60	0.70	0.77	0.70
AmorePacific	0.50	0.57	0.60	0.60	0.63	0.60
Samsung C&T	0.50	0.64	0.70	0.50	0.70	0.50
NAVER	0.70	0.60	0.50	0.70	0.70	0.70
SK hynix	0.40	0.50	0.50	0.60	0.66	0.60
Samsung Life Insurance	0.60	0.78	0.40	0.60	0.78	0.60
POSCO	0.40	0.56	0.40	0.60	0.56	0.40
Average	0.53	0.58	0.56	0.61	0.67	0.59

Boosting(Tuszynski, 2012), RandomForest(Liaw and Wiener, 2002) 알고리즘을 활용하여 Table 4의 학습 데이터를 가지고 학습 모형을 생성하였으며, 테스트 데이터에 적용하여 정확도(accuracy)를 측정하였다. 학습 모형 생성은 반복적인 수행과정을 거쳐 선정된 최적의 변수를 활용하였다.

<Table 5>는 단어 추출 방안과 적용 알고리즘에 따른 예측 정확도이다. 구조적 단어 추출을 통한 방안의 평균 정확도가 희소성 기반 단어 추출 방안의 평균 정확도보다 높았으며, Boosting 분류 알고리즘을 적용한 경우의 정확도 평균이 제일 높았다.

5. 결론

본 연구는 시가총액 상위 10개의 종목의 뉴스 기사를 수집, 분석하여 주가 등락을 예측하였다. 주가 등락 예측 방법은 단어-문서 매트릭스 기반의 분류 모형을 활용하였으며, 단어-문서 매트릭스에서 단어를 제거하는 방안으로 기존의 희소성 기반의 단어 추출 방법과 구조적 단어 추출 방법의 성과를 비교했다. 구조적 단어 추출 방법은 주가 예측을 위해서 해당 종목 뉴스만을 사용하는 것이 아니라 다른 종목 뉴스들도 활용하여 추출할 단어를 결정하는 것이 차이점이다. 즉 상승, 하락에 모두 등장하는 단어를 제거하는 것뿐만 아니라 여러 종목의 뉴스에 공통적으로 등장

하는 단어들도 제거하였다. 예측 정확도를 비교하였을 때 구조적 단어 추출 방법을 활용한 것이 더 높은 정확도를 보였다.

기존 주가 예측 연구들은 하나의 주가를 예측하기 위해서 해당 종목의 뉴스를 사용하였다. 하지만 구조적 단어 추출 방법은 기존 방법과 다르게 여러 종목의 뉴스를 종합적으로 교차 사용하여 분석한 것에 의의가 있다. 또한 기존 연구들은 감성 분석을 기반으로 주가를 예측하는데 이 경우 감성이 긍정 혹은 부정으로 존재하는 단어를 사용했다. 반면 본 연구에서는 감성 분석에서 사용되지 않던 감성이 없는 중립 단어를 중심으로 예측한 방법이라는 점에서 의의가 있다.

주가 예측은 해당 주식을 둘러싼 환경에 대해서 객관적인 이해를 기반으로 분석할 필요가 있다. 하지만 주식을 분석하는 것은 사람이기 때문에 주관성을 배제하는 것은 상당히 어렵다. 본 연구는 단어를 추출하는 방법에서부터 주가를 예측하는 모형 모두 이런 분석가의 주관적 분석을 배제하며 분석이 가능하다.

본 연구의 한계점은 상승, 하락을 분류하는 문제로 주가 예측을 설정하여 보합인 날의 데이터는 사용하지 않은 점과 시가 총액 상위 10개 종목에 대해서만 실험이 행해졌다는 것이다. 실험에 활용한 10개 주식이 주식시장 전체를 대표하지는 않는다. 또한 주가의 등락과 수익률은 다를 수 있기 때문에 투자 성과를 나타내기 어렵다. 따라서 더 많은 표본을 활용한 연구와 트레이딩 시뮬레이션을 통한 수익률 예측 연구가 필요하다.

또한 종목별 정확도 차이와 분류 알고리즘별 정확도차이를 가져오는 요인에 대한 추가적인 분석이 필요하다. 이를 통해 데이터의 특성에 적합한 분류 알고리즘 선정이 가능해 질 수 있다.

참고문헌(References)

Ahn, S. W and S. B. Cho, “Stock Prediction Using News Text Mining and Time Series Analysis”, Proceedings of Korea Computer Congress, Vol.37, No.1(2010), 364~369

Amilon, H., “GARCH estimation and discrete stock prices: an application to low-priced Australian stocks”, *Economics Letters*, Vol.81, No.2(2003), 215~222.

Bothos, E., D. Apostolou, G. Mentzas, “Using Social Media to Predict Future Events with Agent-Based Markets”, *IEEE Intelligent Systems*, Vol.25, No.6(2010), 50~58.

Cao, Q., W. Duan, and Q. Gan, “Exploring determinants of voting for the “helpfulness” of online user reviews: A text mining approach”, *Decision Support Systems*, Vol.50, No.2(2011), 511~521.

Choeh, J. Y., H. J. Lee, and S. J. Park, “A Personalized Approach for Recommending Useful Product Reviews Based on Information Gain”, *KSII Transactions on Internet and Information Systems*, Vol.9, No.5(2015), 1702-1716.

Ding, X., Y. Zhang, T. Liu, and J. Duan, “Using Structured Events to Predict Stock Price Movement: An Empirical Investigation”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, 1415~1425.

Ding, X., Y. Zhang, T. Liu, and J. Duan, “Deep Learning for Event-Driven Stock Prediction”, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, 2015, 2327~2333.

Fung, G. P. C., J. X. Yu, X. Yu and W. Lam, "News Sensitive Stock Trend Prediction", Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Taipei, Taiwan, 2002.

Huang, A. "Similarity measures for text document clustering." Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, 2008.

Jeantheau, T., "A link between complete models with stochastic volatility and ARCH models," *Finance and Stochastics*, Vol. 8, No. 1(2004), 111~131.

Jeong, J. S., D. S. Kim, and J. W. Kim, "Influence analysis of Internet buzz to corporate performance: Individual stock prediction using sentiment analysis of online news," *Journal of Intelligence and Information Systems*, Vol. 21, No. 4(2015), 37~51.

Kim, K. Y., and K. R. Lee, "A Study on the Prediction of Stock Price Using Artificial Intelligence System", *Korean Journal of Business Administration*, Vol.21, No.6 (2008), 2421~2449

Kim, Y. S., N. G. Nim, and S. R. Jeong, "Stock-Index Invest Model Using News Big Data Opinion Mining," *Journal of Intelligence and Information Systems*, Vol. 18, No. 2(2012), 143~156.

Lee, H. Y., "A Combination Model of Multiple Artificial Intelligence Techniques Based on Genetic Algorithms for the Prediction of Korean Stock Price Index(KOSPI)", *Entrue Journal of Information Technology*, Vol.7, No.2(2008), 33~43.

Lee, M. and H. J. Lee, "Increasing Accuracy of Classifying Useful Reviews by Removing Neutral Terms", *Journal of Intelligence and Information Systems*, Vol. 22, No. 3(2016), 129~142.

Liaw, A. and M. Wiener, "Classification and regression by randomForest", R News, 2(3), 18~22, 2002.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2012. URL http://CRAN.R-project.org/package=e1071. R package version 1.6-1.

Mittermayer, M. A., "Forecasting Intraday Stock Price Trends with Text Mining Technique", Proceedings of the 37th Hawaii International Conference on Social Systems, Hawaii, 2004.

Oh, C. and O. R. L. Sheng, "Investigating Predictive Power of Stock Micro Blog Sentiment in Forecasting Future Stock Price Directional Movement", Proceedings of ICIS 2011, Shanghai, China.

Park, K. H and H. J. Shin, "Stock Price Prediction Based on Time Series Network", *Korean Management Science Review*, Vol.28, No.1(2011), 53~60

Perkins, J., *Python 3 Text Processing with NLTK 3 Cookbook*, Packt Publishing, 2014.

Schumaker, R. P. 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), Article No. 12.

Seo, Y. W., J. Giampapa and K. Sycara, "Text Classification for Intelligent Portfolio Management", Carnegie Mellon University, Robotics Institute, 2002.

Thomas, J. D. and K. Sycara, "Integrating Genetic Algorithms and Text Learning for Financial Prediction", Proceedings of Genetic and Evolutionary Computation Conference (GECCO), Las Vegas, NV, 2002.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, I. M. Welpe, "Election Forecasts With Twitter", *Social Science Computer Review*, Vol. 29, Issue 4, 2011, 402~418.

Tuszynski, J., caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc., 2012. URL http://CRAN.R-project.org/package=caTools. R package version 1.13.

Yu, E. J., Y. S. Kim, N. G. Kim, and S. R. Jeong, "Prediction the Direction of the Stock Index by Using a Domain-Specific Sentiment Dictionary," *Journal of Intelligence and Information Systems*, Vol. 19, No. 1(2013), 95~110.

Zhang, R. and T. Tran, "An information gain-based approach for recommending useful product reviews", *Knowledge Information Systems*, Vol. 26, No. 3(2011), 419~434.

Abstract

Stock Price Prediction by Utilizing Category Neutral Terms: Text Mining Approach

Minsik Lee* · Hong Joo Lee**

Since the stock market is driven by the expectation of traders, studies have been conducted to predict stock price movements through analysis of various sources of text data. In order to predict stock price movements, research has been conducted not only on the relationship between text data and fluctuations in stock prices, but also on the trading stocks based on news articles and social media responses. Studies that predict the movements of stock prices have also applied classification algorithms with constructing term-document matrix in the same way as other text mining approaches.

Because the document contains a lot of words, it is better to select words that contribute more for building a term-document matrix. Based on the frequency of words, words that show too little frequency or importance are removed. It also selects words according to their contribution by measuring the degree to which a word contributes to correctly classifying a document.

The basic idea of constructing a term-document matrix was to collect all the documents to be analyzed and to select and use the words that have an influence on the classification. In this study, we analyze the documents for each individual item and select the words that are irrelevant for all categories as neutral words. We extract the words around the selected neutral word and use it to generate the term-document matrix. The neutral word itself starts with the idea that the stock movement is less related to the existence of the neutral words, and that the surrounding words of the neutral word are more likely to affect the stock price movements. And apply it to the algorithm that classifies the stock price fluctuations with the generated term-document matrix.

In this study, we firstly removed stop words and selected neutral words for each stock. And we used a method to exclude words that are included in news articles for other stocks among the selected words. Through the online news portal, we collected four months of news articles on the top 10 market cap stocks.

* Department of Information and Industrial Engineering, Yonsei University
** Corresponding Author: Hong Joo Lee
Department of Business Administration, Catholic University of Korea
43 Jibong-ro, Bucheon, Gyeonggi 14662, Korea
Tel: +82-2-2164-4009, Fax: +82-2-2164-4280, E-mail: hongjoo@catholic.ac.kr

We split the news articles into 3 month news data as training data and apply the remaining one month news articles to the model to predict the stock price movements of the next day. We used SVM, Boosting and Random Forest for building models and predicting the movements of stock prices. The stock market opened for four months (2016/02/01 ~ 2016/05/31) for a total of 80 days, using the initial 60 days as a training set and the remaining 20 days as a test set. The proposed word - based algorithm in this study showed better classification performance than the word selection method based on sparsity.

This study predicted stock price volatility by collecting and analyzing news articles of the top 10 stocks in market cap. We used the term - document matrix based classification model to estimate the stock price fluctuations and compared the performance of the existing sparse - based word extraction method and the suggested method of removing words from the term - document matrix. The suggested method differs from the word extraction method in that it uses not only the news articles for the corresponding stock but also other news items to determine the words to extract. In other words, it removed not only the words that appeared in all the increase and decrease but also the words that appeared common in the news for other stocks. When the prediction accuracy was compared, the suggested method showed higher accuracy.

The limitation of this study is that the stock price prediction was set up to classify the rise and fall, and the experiment was conducted only for the top ten stocks. The 10 stocks used in the experiment do not represent the entire stock market. In addition, it is difficult to show the investment performance because stock price fluctuation and profit rate may be different. Therefore, it is necessary to study the research using more stocks and the yield prediction through trading simulation.

Key Words : Stock Price, Neutral Terms, Text Mining, Online News

Received : April 9, 2017 Revised : May 27, 2017 Accepted : May 29, 2017
Publication Type : Regular Paper Corresponding Author : Hong Joo Lee

저 자 소 개

이 민 식
현재 연세대학교 정보산업공학과 석사과정에 재학 중이다. 가톨릭대학교 경영학부에서 학사학위를 취득하였다. 주요 관심분야는 데이터 분석, 인간-컴퓨터 상호작용, 최적화 등이다.

이 홍 주
현재 가톨릭대학교 경영학전공 교수로 재직 중이다. KAIST 산업경영학과를 졸업하고 KAIST 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 관심분야는 데이터 분석, 지능형 정보시스템, 온라인 사용자들의 상호작용 등이다.