ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 December: 21(4): 37~51 http://dx.doi.org/10.13088/jiis.2015.21.4.037

온라인 언급이 기업 성과에 미치는 영향 분석 : 뉴스 감성분석을 통한 기업별 주가 예측*

정지선
한양대학교 일반대학원 경영학과
(skyhee84@hanyang.ac.kr)

김동성
한양대학교 일반대학원 경영학과
(paulus82@hanyang.ac.kr)

김종우
한양대학교 경영대학 경영학부
(kjw@hanyang.ac.kr)

인터넷 기술의 발전과 인터넷 상 데이터의 급속한 증가로 인해 데이터의 활용 목적에 적합한 분석방안 연구들이 활발히 진행되고 있다. 최근에는 텍스트 마이닝 기법의 활용에 대한 연구들이 이루어지고 있으며, 특히 문서 내 텍스트를 기반으로 문장이나 어휘의 긍정, 부정과 같은 극성 분포에 따라 의견을 스코어링(scoring)하는 감성분석과 관련된 연구들도 다수 이루어지고 있다. 이러한 연구의 연장선상에서, 본 연구는 인터넷 상의 특정 기업에 대한 뉴스 데이터를 수집하여 이들의 감성분석을 실시함으로써 주가의 등락에 대한 예측을 시도하였다. 개별 기업의 뉴스 정보는 해당 기업의 주가에 영향을 미치는 요인으로, 적절한 데이터 분석을 통해 주가 변동 예측에 유용하게 활용될 수 있을 것으로 기대된다. 따라서 본 연구에서는 개별 기업의 온라인 뉴스 데이터에 대한 감성분석을 바탕으로 개별 기업의 주가 변화 예측을 꾀하였다. 이를 위해, KOSPI200의 상위 종목들을 분석 대상으로 선정하여 국내 대표적 검색 포털 서비스인 네이버에서 약 2년간 발생된 개별 기업의 뉴스 데이터를 수집·분석하였다. 기업별 경영 활동 영역에 따라 기업 온라인 뉴스에 나타나는 어휘의 상이함을 고려하여 각 개별 기업의 어휘사전을 구축하여 분석에 활용함으로써 감성분석의 성능 향상을 도모하였다. 분석결과, 기업별 일간 주가 등락여부에 대한 예측 정확도는 상이했으며 평균적으로 약 56%의 예측률을 보였다. 산업 구분에 따른 주가 예측 정확도를 통하여 '에너지/화학', '생활소비재', '경기소비재'의 산업군이 상대적으로 높은 주가 예측 정확도를 보임을 확인하였으며, '정보기술'과 '조선/운송' 산업군은 주가 예측 정확도가 낮은 것으로 확인되었다. 본 논문은 온라인 뉴스 정보를 활용한 기업의 어휘사전 구축을 통해 개별 기업의 주가 등락 예측에 대한 분석을 수행하였으며, 향후 감성사전 구축 시 불필요한 어휘가 추가되는 문제점을 보완한 연구 수행을 통하여 주가 예측 정확도를 높이는 방안을 모색할 수 있을 것이다.

주제어 : 주가 예측, 감성분석, 예측 분석

논문접수일 : 2015년 11월 5일 논문수정일 : 2015년 12월 6일 게재확정일 : 2015년 12월 6일
교신저자 : 김종우

1. 서론

정보기술의 발전과 함께 대용량 데이터의 급속한 증가로 인하여 목적에 맞는 데이터의 선택과 효과적인 분석 방안에 대한 연구들이 활발히 이루어지고 있다. 이러한 대용량 데이터의 활용과 분석은 의사결정 과정에서 현재 상황에 대한 이해와 통찰력을 제공할 수 있을 뿐만 아니라, 급변하는 미래에 대하여 효과적 대응을 위한 정보 제공이 가능하다. 이에 따라, 미래 예측 및 동향 파악을 위한 방안으로써 데이터 마이닝과 같은 통계적 분석 기법들을 기반으로, 고객 구매

* 이 논문은 2013년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (NRF-2013S1A5A2A01019228)

정보를 활용한 제품 수요 예측에서부터 자사의 서비스를 이용하는 고객들의 이탈 예측, 생산 정보를 활용한 불량 제품 예측까지 다양한 분야에서 예측 분석(predictive analytics)과 관련된 연구들이 꾸준히 이루어지고 있다(Waller and Fawcett, 2013).

최근 예측 분석에 대한 연구들은 분석 가능한 다양한 데이터들의 증가로 인하여 마케팅이나 고객 관리와 같은 특정 분야로만 국한되지 않으며, 온라인상에서 발생된 데이터를 활용한 질병 예측, 선거 예측, 주식 시장의 변화 예측 등 다양한 분야로 확대되어 연구되고 있다(Kim et al, 2012; LaValle et al., 2013; Lee et al., 2013). 이러한 연구 동향의 일환으로 본 논문에서는 온라인상에서 발생하는 개별 기업에 대한 주식 관련 뉴스들의 수집 및 분석을 통하여 개별 기업들의 향후 주식 가치 변화에 대한 효과적인 예측 방안의 모색을 꾀하였다.

주식시장에서 기업의 미래 가치 평가 및 예측에 대한 연구들은 예측 가능성에 대한 논의부터 예측 방안에 대한 연구까지 꾸준히 이루어져 왔으며, 최근에는 온라인상에서 발생된 정보를 활용한 주가 예측 방안에 대한 연구들이 다수 수행되고 있다(Bollen et al., 2011; de Fortuny et al., 2014; Schumaker et al., 2012). 온라인 커뮤니케이션 채널의 발전은 기업 경영 활동에 대한 다양한 정보들이 빠르게 확산되고, 다수에게 접근의 용이성을 가져다주게 되었다. 더불어 뉴스를 비롯한 다양한 정보들은 주식 시장에서 특정 기업에 대한 미래 투자 의사 결정에 영향을 미치는 중요한 요소 중 하나로써, 적절한 데이터 분석 방안을 통해 주가 변동을 예측하는데 활용이 가능하다(Schumaker et al., 2009).

뉴스 분석을 통한 주가 등락 예측과 관련된 기존의 많은 연구에서는 어휘별 극성이 규정되어 있는 범용사전을 사용하였으며, 이러한 범용사전 내의 어휘의 극성을 활용하여 해당 뉴스의 극성을 산출하는 것은 분석의 목적에 따라 어휘의 의미가 다르게 적용될 수 있음을 간과한다. 따라서 본 논문에서는 이러한 점을 보완하여 개별 기업의 주가 등락 예측을 위한 기업별 감성사전을 구축하고자 하였으며, 사람의 주관적인 판단의 개입을 통제하고 자동화하기 위해 일별 주가 등락 정보를 기반으로 해당일 발생한 뉴스 내의 어휘 극성을 판별하였다. 또한 기존의 연구들이 다수의 기업들을 대상으로 종합주가지수와 같은 시장 인덱스의 예측에 초점을 맞춘 반면에, 개별 기업의 감성분석 결과를 바탕으로 실제 기업과 관련된 산업군의 시장수익 대비 해당 기업 초과수익률의 방향성을 예측하고자 하는 것이 기존 연구들과의 가장 큰 차이점이라고 할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 감성분석(sentiment analysis)과 주식 시장에서의 예측 분석에 대한 기존 선행 연구들을 검토한다. 3장에서는 온라인상에서 발생된 각각의 기업 주가 관련 뉴스들의 감성분석을 통한 개별 기업의 주가 예측 방안에 대하여 제시하며, 4장에서는 실증 분석을 바탕으로 본 연구에서 제안한 방안의 유용성에 대하여 검증한다. 마지막 5장에서는 결론과 본 연구의 한계점 및 추후 연구 방향에 대하여 제시한다.

2. 관련 연구

2.1 주가 예측 분석(Predictive Analytics) 연구

주식 시장에서 기업의 미래 가치 변화에 대한

분석 및 예측에 대한 연구들은 지속적으로 여러 학문에서 수행되어져왔다. 기업의 가치 평가 및 변화에 대한 예측은 경영 활동을 통해 산출되는 다양한 지표들로 측정 될 수 있으며, 이러한 지표들 중에서 주식 시장에서 해당 기업에 대한 주식 가격 또는 주식수익률은 기업 가치에 관한 정보가 반영된 지표로써 활용이 가능하다. 주식 시장에서 특정 기업의 주식가격 변화는 기업의 경영 활동뿐만 아니라 기업이 속한 경제 상황의 변화에도 영향을 받으며, 이로 인하여 주식 시장에서 투자자들의 의사 결정은 국내외 경제 상황 및 해당 기업과 관련 된 공시, 뉴스 정보 등이 중요한 요소로써 작용 될 수 있다(Bank et al., 2011). 이에 따라, 최근에는 온라인상에서 발생되는 기업 뉴스 및 사용자 의견 분석을 활용한 주가 등락의 예측 분석에 대한 연구들이 다수 수행되고 있다.

국내 감성분석을 통한 주가 예측과 관련 된 연구로는 소셜 미디어 상에서의 사용자 의견에 대한 감성분석 시 구글의 API (Application Program Interface)를 활용하여 한글 텍스트를 영문으로 번역 후 영문에 특화된 감성사전인 SentiWordNet을 활용한 주가 예측 연구(Kim et al., 2014), 주식 관련 뉴스 데이터에 오피니언 반의법 규칙(Opinion Antonym Rule: OAR) 알고리즘을 적용하여 감성사전 구축 후, 이를 통한 주식 상승·하락 분석 연구(Jo et al., 2015), 주식 시장에 특화 된 감성사전 구축을 통한 주가 지수의 방향성 예측 연구(Yu et al., 2013) 등이 있다. 이외에도 온라인상에서의 텍스트 데이터 분석을 활용한 주가 예측 분석과 관련된 국내외 대표적 연구들은 다음과 같다(Table 1 참조).

기존 온라인상의 텍스트 데이터를 활용한 주가예측 관련 연구들의 검토를 통하여 미래 주가

〈Table 1〉 Review of Related literature on Stock Prediction with Sentiment Analysis

Authors (year)	Research overview
Jo et al., 2015	Construct of sentiment dictionary using Opinion Antonym Rule (OAR) algorithm
Kim et al., 2012	Suggestion of Investment decisions model throughout the logistic regression based on the sentiment analysis for the online news information
Evangelopoulos et al., 2012	Suggestion of stocks yield prediction models for individual businesses based on the analysis of text mining technique
Bank et al., 2011	Analysis of relationship between the Google search volume and stock returns of the individual companies
Bollen et al., 2011	Research on ways of predicting stock market changes through sentiment analysis for tweets of twitter users

등락여부 예측에 텍스트 데이터에 대한 감성분석을 활용한 방안이 유용함을 확인 할 수 있었다. 그러나 기존 다수의 국내 연구들의 경우, 개별 기업에 대한 감성사전의 구축을 바탕으로 한 개별 주가 예측 연구나 장기간에 걸친 예측 방안의 검증에 대한 연구는 미비한 편이다. 또한 기업 주가 변화 예측 연구에서 기업의 주당 가격을 예측 대상으로 하였으나, 기업의 실제 주가 변화의 확인은 전반적 경제 상황 또는 산업군의 특성이 반영된 상대적 측정이 선행적으로 요구된다. 이에 따라 본 연구에서는 보다 정확한 기업별 주가 변화의 확인을 위하여 시장 변화율 대비 분석 대상 기업의 변화율 측정이 가능한 초과수익률을 예측 값으로 선정하였다. 이는 주식 시장 전반적 상승, 또는 하락의 모멘텀이 어느 정도 반영된 예측 값으로써 보다 실질적인 주식 가격 변화의 예측이라고 볼 수 있다. 또한 본 논문에서는 기존 연구들의 단기적인 측면에서 수행 되었던 분석 기간의 한계점을 보완하고자 기업별 약

2년 4개월간의 뉴스 데이터를 수집하여 감성사전 구축 및 예측 성능을 확인하였다. 이에 따라 본 연구가 갖는 기존 연구와의 차이점은 다음과 같다. 첫째, 기업별 주가 등락 예측이라는 연구 목적에 부합하도록 주식시장에 특화된 기업별 감성사전 구축을 통하여 기업별 온라인 뉴스 정보에 대해 보다 정확한 확인을 도모하였다. 범용적으로 사용되는 사전의 어휘는 연구 대상과 목적에 따라 발생 빈도와 의미 면에서 차이가 발생할 수 있으므로 이를 보완하기 위해 주식 시장 등락에 따른 어휘를 중심으로 감성사전을 구축하였다. 둘째, 기업별 주식의 초과수익률을 예측 대상으로 하여 기업 경제적 외부 요인들을 고려하고자 하였다. 초과수익률을 산업군의 지수를 활용하여 산출함으로써, 각 종목의 단순 등락률이 아닌 동일 산업군의 시장수익률 대비 해당 종목의 등락을 고려하였다. 마지막으로 다양한 산업군에 속한 여러 기업들을 연구 대상으로 선정하고 장기간의 뉴스 데이터를 활용하여 제안하는 연구 방안에 대한 실증적인 유용성 검증을 꾀하였다.

2.2 감성분석(Sentiment Analysis) 연구

감성분석(Sentiment Analysis) 기법은 사람이 사용하는 자연어에 대하여 기계 학습 기법을 바탕으로 문장의 긍정・부정과 같은 문맥 정보 등을 추출하거나 분류하고자 하는 기법을 의미한다. 감성분석 방안으로는 감성사전을 활용하여 문서 내 단어의 빈도를 기반으로 한 분석 방안과 문맥 전체의 정보를 해석하여 분석하고자 하는 방안들이 존재한다. 감성사전을 기반으로 한 문장의 긍정・부정 또는 선호・비선호 같은 문맥 정보의 확인은 문장 내 출현한 용어들의 종류나 관계, 빈도에 따라 확인 가능하며, 이러한 이유로 감성 용어 사전의 구축은 감성분석 성능에 높은 영향을 미치는 중요한 요소 중 하나이다(Kim and Kim, 2014). 감성사전을 구축하는 방안으로는 대표적으로 문장 또는 단어의 극성이 사전에 정의된 감성사전을 바탕으로 새로운 어휘들의 유사성 및 거리 관계 등의 비교를 통해 감성사전을 구축하는 방안이 있으며, 실제로 수집된 문장들에 대한 형태소 분석을 바탕으로 개별 단어의 극성을 정의하여 구축하는 방안이 있다(Jo, 2012). 구축 된 감성사전을 바탕으로 실제 텍스트에 대한 감성분석 방안으로는 단순 문서 내 단어의 출현 빈도만을 계산하는 방안, 문서 내 단어들의 어의적 관계 및 구문적 관계 등의 언어규칙을 PMI (Pointwise Mutual Information), Navie Bayes, SVM (Support Vector Machine) 등의 통계적 분석 기법을 사용하여 분류하는 방안들이 존재하며, 이 밖에도 언어학적 방법론들을 적용한 연구들이 다양한 형태로 수행되고 있다(Song et al., 2010; Jo, 2012; Kim and Kim, 2014).

본 연구에서는 기존 다수 연구에서 수행된 감성분석 방안에서 나아가 이를 활용한 주가 방향성 예측을 위한 연구를 수행하고자 한다. 감성분석 방안의 활용을 위해 총 8개 산업군에 속한 상위 종목을 선정하였으며, 온라인 뉴스를 기반으로 주식시장에 특화된 기업별 감성사전을 구축하는데 중요성을 두었다.

3. 연구 방안

3.1 연구 데이터

본 연구에서는 국내 대표 포털 사이트인 네이

버의 증권 정보 서비스 중, 기업별 '종목뉴스' 게시판에서 2013년 1월부터 2015년 4월까지 수집한 데이터를 활용하였다. 2013년 1월부터 2014년 4월까지 총 123,985개의 데이터를 훈련 데이터로 기업별 감성사전 구축에 활용하였으며, 2014년 5월부터 2015년 4월까지 총 90,817개의 데이터는 본 연구에서 제안한 기업별 주가 예측 방안의 유용성 확인을 위한 검증 데이터로 사용하였다. 분석 대상 기업의 선정은 연구 방안의 실증적 검증을 위하여 'KOSPI 200 지수'에 속하는 종목을 8개 산업군별로 분류한 'KOSPI 200 섹터지수'를 기준으로 하였으며, 사전 구축을 위한 훈련 데이터와 예측 성과 확인을 위한검증 데이터의 기간에 따라서 'KOSPI 200 섹터지수'에 편입되거나 제외되는 일부 기업들이 존재하였다. 이에 따라 검증 데이터의 기간을 기준으로 다음과 같이 산업군별 시가총액 순으로 총 40개 기업을 최종 연구 대상으로 선정하였다(Table 2 참조).

〈Table 2〉 The number of research data based on the online news related to companies

Stock item	Training data set	Validation data set	Stock item	Training data set	Validation data set
Hyundai Engineering & Construction	1,284	976	Samsung Electronics	22,224	16,367
Doosan Heavy Industries & Construction	827	538	SK Hynix	5,607	4,747
Daelim Industrial	1,084	740	NAVER	4,504	3,888
Daewoo Engineering & Construction	1,293	869	LG Electronics	6,632	4,594
Doosan Corporation	676	492	LG	1,339	841
Hyundai Heavy Industries	3,573	1,978	Shinhan Financial Group	3,673	2,565
Samsung Heavy Industries	1,222	1,166	Samsung Life Insurance	2,877	2,120
Hyundai Glovis	403	737	KB financial	3,217	2,468
Daewoo Shipbuilding & Marine Engineering	1,386	1,176	Hana Financial Group	1,591	994
Hyundai Mipo Dockyard	586	610	Samsung Fire & Marine Insurance	1,202	992
POSCO	5,382	4,140	Korea Electric Power Corporation	3,964	3,661
Hyundai-Steel	1,193	961	SK Telecom	5,635	4,657
Korea Zinc Company	430	261	KT&G	780	851
Hyundai Hysco	661	479	LG Household & Health Care	1,010	972
Young Poong Corporation	74	64	KT	4,034	2,716
LG Chemical	3,411	1,788	Hyundai Motor Company	12,171	8,671
SK Innovation	1,686	1,097	Hyundai Mobis	5,143	3,197
SK	1,568	959	Kia Motors	7,984	4,477
S-Oil	683	881	Lotte Shopping	1,850	1,124
Lotte Chemical Co., Ltd	644	611	Kangwon Land	482	392

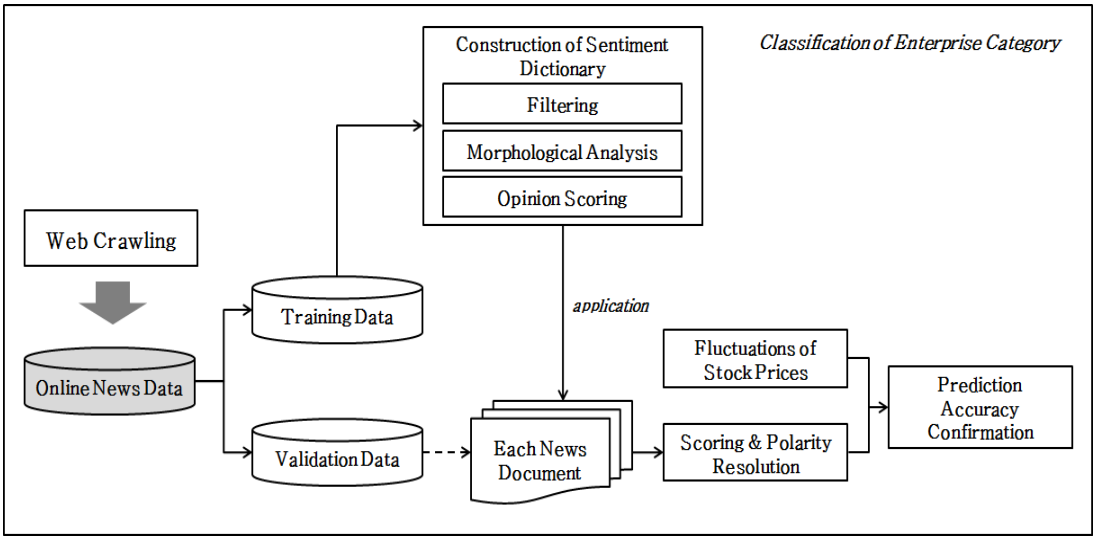
3.2 연구 절차

본 연구는 다음과 같은 연구절차를 통하여 수행되었다(Figure 1 참조). 온라인 포털 서비스 업체인 네이버의 증권 관련 뉴스 게시판에서 기업별 뉴스 정보를 웹 크롤링(web crawling)하였다. 수집된 2013년 1월부터 2015년 4월까지의 뉴스를 바탕으로 2013년 1월부터 2014년 4월까지의 뉴스를 사전 구축용 데이터로, 2014년 5월부터 2015년 4월까지의 뉴스는 주가 예측을 위한 검증 데이터로 구분 하였다. 훈련용 데이터와 검증용 데이터 각각 R에서 패키지로 제공되는 'tm'[1] 과 'KoNLP'[2]를 활용하여 텍스트 분석을 수행하였으며, 최종적으로는 본 연구에서 제안하는 방안을 통하여 개별 기업의 주가 예측 결과에 대한 정확도를 확인하였다.

3.2.1 감성사전 구축

감성사전을 구축하기 위해 본 연구에서는 수집된 온라인 뉴스에서 '명사' 만을 활용하였다. 명사 추출에 앞서, '무단전재 및 재배포금지', '기자', '편집자주', '저작권자', '증권시황' 등 빈번하게 발생하는 불필요한 어휘와 의미를 알 수 없는 단음절 체언 및 용언을 제거하였으며, 이외에 뉴스 작성자의 메일주소와 특수기호 등 불용어들을 제거하였다. 최종적으로 추출된 명사의 감성 점수화를 통해 기업별 감성사전을 구축한다. 각 어휘의 감성 점수의 계산 방법은 다음 식 (1), (2), (3)과 같다.

$$TermScore(i_p) = \frac{Num(i \in PosDocs)}{Total\,Num(i)} \quad (1)$$



〈Figure 1〉 Stock price prediction using individual sentiment dictionary of each companies

1) Ingo Feinerer [aut, cre], Kurt Hornik [aut], Artifex Software, Inc. [ctb, cph] (2014). tm: Text Mining Package. R Package Version 0.6. http://CRAN.R-project.org/package=tm

2) Jeon, H.(2013). KoNLP: Korean NLP Package. R Package Version 0.76.9. http://CRAN.R-project.org/package=KoNLP

$$TermScore(i_n) = \frac{Num(i \in NegDocs)}{Total\,Num(i)} \quad (2)$$

식 (1)과 (2)에서 $TermScore(i)$는 어휘 i의 감성 점수이며, 0에서 1사이의 값을 가진다. $TotalNum(i)$는 뉴스 전체에서 나온 i의 출현 빈도이며, $Num(i \in PosDocs)$는 긍정적 영향을 갖는 뉴스에서 발생한 i의 출현 빈도이다. 여기서 긍정적 영향을 갖는 뉴스는 각 종목의 주가가 상승한 날 발생한 뉴스를 의미한다. 즉, 주가가 상승한 날의 뉴스는 긍정적 의미가 있는 것으로 간주한다. 반대로 $Num(i \in NegDocs)$는 주가가 하락한 날의 뉴스에서 발생한 i의 출현 빈도이다.

$$TermScore(i) = TermScore(i_p) - TermScore(i_n) \quad (3)$$

이러한 과정에서 주가가 상승한 날의 뉴스와 하락한 날의 뉴스에 동시에 출현하는 중복 어휘가 발생한다. 식 (3)과 같이 해당 어휘의 긍정점수와 부정점수를 합산하여 최종 감성 점수를 부여한다. 또한, 본 연구에서는 기업의 주가 등락을 확인하기 위하여 일간 초과수익률을 사용하였다. 초과수익률은 개별 기업의 수익률을 당일의 시장 수익률 변동과의 유의적인 차이를 측정한 것으로 다음의 식 (4)와 같다(Moon and Kim, 2014).

$$AR_{jt} = R_{jt} - (\widehat{\alpha_j} + \widehat{\beta_j} R_{mt}) \quad (4)$$

AR_{jt}는 t시점에서 기업 j의 초과수익률이며, R_{jt}는 t시점에서 기업 j의 수익률이다. $\hat{\alpha}_j + \hat{\beta}_j R_{mt}$는 t시점에서 기업 j의 기대수익률이며, R_{mt}는 t시점의 시장수익률이다. 본 연구에서는 전체 KOSPI 지수가 아닌 KOSPI200의 각 섹터별 지수를 활용하여, 해당 기업이 속한 특정 산업군 시장수익률 대비 기대수익률을 도출함으로써 보다 정확하고 세분화된 기업 초과수익률을 구하고자 시도하였다.

3.2.2 감성사전을 활용한 개별 기업의 주가 예측 방안

훈련용 데이터를 바탕으로 기업별 주가 등락과 관련된 감성사전을 구축 후, 검증용 데이터 적용을 통한 개별 기업의 주가 예측 방안은 다음 식 (5)와 같다. 이를 바탕으로 기업별 일별 뉴스에 대하여 점수화하여 실제 해당일의 주가 등락과 일치하는지 확인한다.

$$ComScore(j_t) = \frac{\sum_{i=1}^{n} Num(i_t) \times TermScore(i)}{\sum_{i=1}^{n} Num(i_t)} \quad (5)$$

$ComScore(j_t)$는 t시점의 기업 j에 대한 오피니언 점수이다. 즉, 해당일에 발생한 기업별 전체 뉴스의 극성을 의미하며 이를 기업 j에 대한 오피니언 평가 기준으로 활용한다. $Num(i_t)$는 t시점에 발생한 모든 뉴스에서의 어휘 i의 출현 빈도이며, $TermScore(i)$는 어휘 i의 극성 점수이다. 일별 개별 기업에 대한 뉴스에서 출현한 전체 어휘에서 사전에 구축된 감성사전과의 비교를 통하여 개별 어휘에 점수를 부여하고 평균화한다. 일별 기업에 대한 평가 대상 뉴스의 선정 기준은 주식시장의 개장 시간부터 마감 시간을 기준으로 하였다. 이에 따라, 주가 예측일

에 사용되는 기업 뉴스는 전일 15시에서 예측일 15시 사이의 게시된 시간을 반영하였으며, 개장되지 않는 날(휴일 및 공휴일)에는 이전 개장일 15시 이후의 뉴스를 예측 분석에 사용한다.

4. 연구 결과

본 논문에서 제안한 개별 기업에 대한 감성사전 구축을 통한 주가 등락 예측 방안의 실제 예측 정확도는 다음과 같다(Table 3 참조). 본 연구에서 기업별 주식 예측일의 기준은 뉴스가 발생한 시점 이후로써, 기업별로 뉴스 발생일에 따라 예측 대상일 수에 차이가 존재한다. 또한 발생되는 뉴스의 양도 기업별로 차이가 있으며, '삼성전자'가 16,367개로 가장 많은 뉴스가 분석 대상 기간 동안 발생되었으며, 다음으로 '현대차'가 8,671개, 'SK하이닉스'가 4,747개, 'LG전자' 4,594개로 많은 양의 뉴스가 발생되었다. 이와 반대로, '영풍'이 64개로 분석 대상의 기간과 비교하여도 적은 수의 뉴스가 발생되었으며, '두산', '현대하이스코', '강원랜드', '고려아연'이 500개 이하의 뉴스 수가 발생됨을 확인 할 수 있었다. 산업군 구분에 따라서는 '에너지/화학', '생활소비재', '경기소비재'는 다른 산업군과 비교하여 뉴스의 양이 많은 것을 확인 할 수 있다. 전반적으로 기업별 뉴스의 양에 따른 주가 예측 정확도는 차이가 없는 것으로 확인된다.

4.1 기업별 주가 예측 정확도

개별 기업 수준에서의 예측 정확도는 '강원랜드', 'KT&G', 'SK이노베이션'이 각각 66.92%, 66.84%, 66.82%로 타 기업들과 비교하여 높은 예측 정확도를 보임을 확인할 수 있다. 다음으로는 '고려아연' 63.73%, '롯데케미칼' 61.85%, '현대모비스' 62.70%, 'SK하이닉스' 61.07%, '두산중공업' 60.65%, '삼성화재' 60.28%, 'LG화학' 60.17%, '현대하이스코' 60.14% 순으로 평균적으로 약 60%대의 예측 정확도를 보임을 확인할 수 있다. 그 밖의 다수의 기업들이 약 50%의 예측 정확도를 보이는 것을 확인 할 수 있었으나, '영풍'이 41.30%로 가장 낮은 예측 정확도를 보였으며, 'LG' 44.55%, '삼성생명' 47.54%, '두산' 48.80%로 약 40%대의 낮은 예측 정확도를 보이는 것을 확인 할 수 있다.

또한, 본 연구에서 제안한 예측 방안의 분석 결과 성능을 비교하기 위해 Random work 방안을 활용한 주가 등락 예측 실험을 수행하였다(Table 3 참조). Random work 방안에 따른 예측력은 최소 42.21%에서 최대 53.68%의 값을 보였으며, 본 연구에서 제시한 방안이 대부분의 기업에서 더 높은 주가 등락 예측률을 보임을 확인할 수 있다. 특히, '고려아연', '현대하이스코', 'LG화학', 'SK이노베이션', '롯데케미칼', 'SK하이닉스', 'KT&G', '현대모비스', '강원랜드' 등에서 상당한 성능 차이를 보이는 것으로 확인되었다. 반면, '영풍', '삼성전자', 'LG', '삼성생명'은 상대적으로 다소 낮은 예측력을 보임을 알 수 있다.

4.2 산업별 주가 예측 정확도

산업별 기준에 대한 정확도 결과를 확인하면, '에너지/화학' 산업에 속하는 기업들의 정확도가 60.52%로 타 산업과 비교하여 상대적으로 정확도가 가장 높은 것으로 확인되며, 다음으로는 '생활소비재'와 '경기소비재' 산업에 속하는 기

〈Table 3〉 Stock price prediction accuracy of individual companies

Industry fields	Stock item	The number of days for prediction	The number of generated news	Random (%)	Accuracy (%)	Average (%)
Cconstruction/ Machinery	Hyundai Engineering & Construction	219	976	49.18	54.34	55.68
	Doosan Heavy Industries & Construction	155	538	49.18	60.65	
	Daelim Industrial	196	740	53.68	55.61	
	Daewoo Engineering & Construction	205	869	47.13	59.02	
	Doosan Corporation	166	492	47.13	48.80	
Shipbuilding/ Transportation	Hyundai Heavy Industries	242	1978	45.90	57.44	53.85
	Samsung Heavy Industries	209	1166	42.21	51.67	
	Hyundai Glovis	172	737	50.00	51.16	
	Daewoo Shipbuilding & Marine Engineering	219	1176	50.00	56.62	
	Hyundai Mipo Dockyard	170	610	47.54	52.35	
Steel/ Material	POSCO	244	4140	46.72	52.46	55.17
	Hyundai-Steel	213	961	49.59	58.22	
	Korea Zinc Company	102	261	48.36	63.73	
	Hyundai Hysco	143	479	47.95	60.14	
	Young Poong Corporation	46	64	49.18	41.30	
Energy/ Chemical	LG Chemical	241	1788	45.49	60.17	60.52
	SK Innovation	223	1097	50.81	66.82	
	SK	226	959	52.04	56.19	
	S-Oil	198	881	50.00	57.58	
	Lotte Chemical Co., Ltd	173	611	51.63	61.85	
Information Technology	Samsung Electronics	244	16367	52.86	49.59	52.19
	SK Hynix	244	4747	47.54	61.07	
	NAVER	244	3888	47.54	52.05	
	LG Electronics	244	4594	46.72	53.69	
	LG	220	841	51.22	44.55	
Finance	Shinhan Financial Group	244	2565	45.90	58.20	54.56
	Samsung Life Insurance	244	2120	50.40	47.54	
	KB financial	240	2468	50.00	54.58	
	Hana Financial Group	226	994	46.72	52.21	
	Samsung Fire & Marine Insurance	214	992	50.81	60.28	
Consumer goods for living	Korea Electric Power Corporation	244	3661	47.13	59.43	59.48
	SK Telecom	244	4657	50.00	56.56	
	KT&G	193	851	47.95	66.84	
	LG Household & Health Care	211	972	48.77	59.24	
	KT	244	2716	48.36	55.33	
Consumer discretionary	Hyundai Motor Company	244	8671	49.18	56.15	59.16
	Hyundai Mobis	244	3197	48.77	62.70	
	Kia Motors	244	4477	52.04	54.92	
	Lotte Shopping	225	1124	48.36	55.11	
	Kangwon Land	130	392	48.36	66.92	

업들의 주가 예측 정확도가 각각 59.48%, 59.16%로 확인된다. ‘건설/기계’와 ‘철강/소재’, ‘금융’ 산업의 주가 예측 정확도는 55.68%, 55.17%, 54.56%로서 3개의 산업이 유사한 예측 정확도를 보임을 확인할 수 있었다. 마지막으로 ‘조선/운송’과 ‘정보기술’ 산업은 각각의 예측 정확도가 53.85%, 52.19%로 다른 산업과 비교하여 전반적으로 낮은 정확도를 보이는 것으로 확인된다.

5. 결론

5.1 연구 결과 정리

본 연구에서는 국내 개별 기업 뉴스에 대한 감성분석과 이를 활용한 주가 예측 연구를 수행하였다. 분석 결과, 기업별 예측 정확도는 상이했으며 평균적으로 약 56%의 예측률을 보였다. ‘강원랜드’, ‘KT&G’, ‘SK이노베이션’과 같이 최대 66%의 예측률을 보이는 기업에 반해 ‘영풍’, ‘LG’, ‘삼성생명’, ‘두산’과 같이 약 40%대의 예측 정확도를 보이는 기업들도 존재하였다. 특히, 데이터의 수가 가장 적었던 ‘영풍’은 약 41%의 예측 정확도를 보였으며, 반대로 발생된 뉴스가 가장 많았던 삼성전자 역시 약 49%의 예측 정확도를 보여 추후 기업별 뉴스의 양과 주가 예측과의 상관관계에 대하여 확인할 필요가 있다. 기업에 대한 뉴스가 많을수록 포괄하는 정보가 많아 주가 예측에 효과적일 수 있다고 예상되었으나 수집된 데이터의 확인을 통해, 실제 기업의 경영활동과 직접적인 관련이 없는 뉴스 또는 단순 언급으로 인하여 포함된 뉴스들도 함께 수집됨에 따라 주가 예측 정확도에 영향을 미친 것이라고 확인된다. 이와 반대로, 뉴스가 매우 적은 경우에는 충분한 정보가 확보되지 않아 기업 주가 변화에 대한 예측 정확도가 낮아진 것이라 생각된다.

산업 구분에 따른 주가 예측 정확도 확인을 통하여 ‘에너지/화학’, ‘생활소비재’, ‘경기소비재’의 산업군의 경우 다른 산업과 비교하여 상대적으로 높은 주가 예측 정확도를 보였으며, ‘정보기술’과 ‘조선/운송’ 산업의 경우는 예측 정확도가 낮은 것으로 확인되었다. 수집된 산업별 대표 기업의 수가 각각 5개로 일반화의 어려움은 다소 있으나, 주가 예측 정확도에 산업별 차이가 존재하는 것을 확인 할 수 있었다. 이를 바탕으로 향후 산업별 다수의 기업들을 대상으로 한 주가 예측에 대한 연구 수행을 통하여 온라인 뉴스 정보를 활용한 주가 예측에 적합한 산업 분야 도출이 가능할 것이다.

5.2 연구 한계점 및 향후 연구 방안

본 연구에서는 국내 주식 시장의 개별 기업에 대한 뉴스 정보를 수집 후, 이에 대한 감성분석을 활용한 주가 예측 방안에 대하여 연구하였다. 본 연구를 통하여 온라인 뉴스 정보를 활용한 주가 예측 성능에 뉴스의 양과 산업군별 차이가 존재함을 확인한 것이 주요 학문적 기여라고 할 수 있다. 그러나 개별 기업의 주가 예측에 대한 정확도 향상을 위해 본 연구가 갖는 연구 한계점과 이에 따른 추후 연구 방안은 다음과 같다.

첫째, 본 연구에서는 수집한 개별 기업별 증권 뉴스 기사에서 감성사전 구축 시, ‘수출확대’, ‘실적개선’, ‘강세’, ‘악재’, ‘불황’, ‘적자’ 등과 같이 명사 자체가 갖는 상징적 의미를 활용하여 용어의 품사를 명사로 한정하여 사용하였다. 그러

나 국내외를 비롯한 복잡한 기업 경영 환경 하에서 발생된 뉴스의 의미 분석은 문장 전체에 대한 의미 파악을 위한 감성분석 방안 적용이 더욱 효과적일 것이라 생각된다.

둘째, 수집 된 뉴스에 대하여 해당 기업과 직·간접적 연관이 존재하는 뉴스들의 구분이 필요하다. 실제 수집 된 증권 관련 뉴스의 일부는 해당 기업과는 직접적 관련이 없는 단순 언급만으로 게시된 뉴스이거나 경제적 의미를 갖지 않는 간접적으로 연관된 뉴스들이 일부 확인되었다. 이러한 직·간접적 관련 뉴스들의 사전 선별 방안의 모색을 통하여 보다 실제적인 주가 관련 뉴스의 수집 및 감성분석이 필요하다.

셋째, 본 연구에서는 주가 등락 예측을 위해 기업별 주가와 상당히 관련 있는 정보가 포함된 증권 뉴스를 선별적으로 제공하는 대표적인 국내 포털사이트 NAVER 증권 서비스에서 데이터를 수집 및 활용하였다. 이러한 뉴스 데이터는 기업의 경영활동이나 기업 상황 등을 판단할 수 있기 때문에, 기업 주가 변화에 영향을 미치는 외생적 요인들이 부분적으로 포함된 정보라고 할 수 있다. 또한 시장수익률 대비 상대적 측정치인 초과수익률을 주가 등락 값으로 활용하여 주식시장의 경제 상황을 반영하려 하였다. 그러나 기업 특성, 외부충격과 같은 통제 변수를 직접적으로 고려하지 못하는 한계가 존재한다. 이러한 외생적 요인의 통제 노력에도 불구하고 주가 등락의 예측력이 다소 떨어지는 기업들이 존재함에 따라 향후 관련 선행연구의 확인과 분석을 통하여 기업 주가와 관련된 변수의 추가 활용을 통한 예측성과 향상 방안 등의 연구가 필요할 것으로 생각된다. 향후 이러한 연구 한계점을 보완한 연구를 통하여 보다 정확한 개별 기업의 주가 예측 방안 제안이 가능할 것이다.

참고문헌(References)

Bank, M., M. Larch, and G. Peter, “Google Search Volume and Its Influence on Liquidity and Returns of German Stocks,” *Financial Markets and Portfolio Management*, Vol.25, No.3(2011), 239～264.

Bollen, J., H. Mao, and X. Zeng, “Twitter Mood Predicts the Stock Market,” *Journal of Computational Science*, Vol.2, No.1(2011), 1～8.

de Fortuny, E. J., T. De Smedt, D. Martens, and W. Daelemans, “Evaluating and Understanding Text-Based Stock Price Prediction Models,” *Information Processing & Management*, Vol.50, No.2(2014), 426～441.

Evangelopoulos, N., M. J. Magro, and A. Sidorova, "The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices?,” *Informing Science: the International Journal of an Emerging Transdiscipline*, Vol.15(2012), 247～268.

Jo, E. K., “The Current State of Affairs of the Sentiment Analysis and Case Study Based on Corpus,” *The Journal of Linguistic Science,* Vol.61(2012), 259～282.

Jo, H. J., J. H. Seo, and J. T. Choi, “OAR Algorithm Technology Based on Opinion Mining Utilizing Stock News Contents,” *Journal of Korea Institute of Information Technology,* Vol.13, No.3(2015), 111～119.

Kim, D. S. and J. W. Kim, “Public Opinion Sensing and Trend Analysis on Social Media: A Study on Nuclear Power on Twitter,” *International Journal of Multimedia and Ubiquitous Engineering*, Vol.9, No.11(2014),

373～384.

Kim, S. W. and N. G. Kim, "A Study on the Effect of Using Sentment Lexicon in Opinion Classification," *Journal of Intelligence and Information Systems,* Vol.20, No.1(2014), 133～148.

Kim, Y. M., S. J. Jeong, and S. J. Lee, "A Study on the Stock Market Prediction Based on Sentiment Analysis of Social Media," *Enture Journal of Information Technology,* Vol.13, No.3(2014), 59～70.

Kim, Y. S., N. G. Kim, 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.

LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big Data, Analytics and the Path from Insights to Value," *MIT Sloan Management Review*, Vol.52, No.2(2013), 21～31.

Lee, J., E. Lapira, B. Bagheri, and H. A. Kao, "Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment," *Manufacturing Letters*, Vol.1, No.1(2013), 38～41.

Moon, H. N. and J. W. Kim, "A Study on the Individual Stock Price Prediction Using the Internet News(written in Korean)," *Proceedings of 2014 Korea Intelligent Information Systems Society Spring Conference,* (2014), 387～393.

Schumaker, R. P. and H. Chen, "A Quantitative Stock Prediction System Based on Financial News," *Information Processing & Management*, Vol.45, No.5(2009), 571～583.

Schumaker, R. P., Y. Zhang, C. N. Huang, and H. Chen, "Evaluating Sentiment in Financial News Articles," *Decision Support Systems,* Vol.53, No.3(2012), 458～464.

Song, S. I., D. J. Lee, and S. G. Lee, "Identifying Sentiment Polarity of Korean Vocabulary Using PMI," *Proceeding of Korea Computer Congress,* Vol.37, No.1(2010), 260～265.

Yu, E. J., Y. S. Kim, N. G. Kim, and S. R. Jeong, "Predicting 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.

Waller, M. A. and S. E. Fawcett, "Data Science, Predictive Analytics, and Big Data: a Revolution That Will Transform Supply Chain Design and Management," *Journal of Business Logistics*, Vol.34, No.2(2013), 77～84.

Abstract

Influence analysis of Internet buzz to corporate performance : Individual stock price prediction using sentiment analysis of online news

Ji Seon Jeong* · Dong Sung Kim* · Jong Woo Kim**

Due to the development of internet technology and the rapid increase of internet data, various studies are actively conducted on how to use and analyze internet data for various purposes. In particular, in recent years, a number of studies have been performed on the applications of text mining techniques in order to overcome the limitations of the current application of structured data. Especially, there are various studies on sentimental analysis to score opinions based on the distribution of polarity such as positivity or negativity of vocabularies or sentences of the texts in documents. As a part of such studies, this study tries to predict ups and downs of stock prices of companies by performing sentimental analysis on news contexts of the particular companies in the Internet. A variety of news on companies is produced online by different economic agents, and it is diffused quickly and accessed easily in the Internet. So, based on inefficient market hypothesis, we can expect that news information of an individual company can be used to predict the fluctuations of stock prices of the company if we apply proper data analysis techniques. However, as the areas of corporate management activity are different, an analysis considering characteristics of each company is required in the analysis of text data based on machine-learning. In addition, since the news including positive or negative information on certain companies have various impacts on other companies or industry fields, an analysis for the prediction of the stock price of each company is necessary. Therefore, this study attempted to predict changes in the stock prices of the individual companies that applied a sentimental analysis of the online news data. Accordingly, this study chose top company in KOSPI 200 as the subjects of the analysis, and collected and analyzed online news data by each company produced for two years on a representative domestic search portal service, Naver. In addition, considering the differences in the meanings of vocabularies for each of the certain economic subjects, it aims to

* Dept. of Business Administration, Graduate School, Hanyang University
** Corresponding Author: Jongwoo Kim
School of Business, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea
Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail: kjw@hanyang.ac.kr

improve performance by building up a lexicon for each individual company and applying that to an analysis. As a result of the analysis, the accuracy of the prediction by each company are different, and the prediction accurate rate turned out to be 56% on average. Comparing the accuracy of the prediction of stock prices on industry sectors, ‘energy/chemical’, ‘consumer goods for living’ and ‘consumer discretionary’ showed a relatively higher accuracy of the prediction of stock prices than other industries, while it was found that the sectors such as ‘information technology’ and ‘shipbuilding/transportation’ industry had lower accuracy of prediction. The number of the representative companies in each industry collected was five each, so it is somewhat difficult to generalize, but it could be confirmed that there was a difference in the accuracy of the prediction of stock prices depending on industry sectors. In addition, at the individual company level, the companies such as ‘Kangwon Land’, ‘KT & G’ and ‘SK Innovation’ showed a relatively higher prediction accuracy as compared to other companies, while it showed that the companies such as ‘Young Poong’, ‘LG’, ‘Samsung Life Insurance’, and ‘Doosan’ had a low prediction accuracy of less than 50%. In this paper, we performed an analysis of the share price performance relative to the prediction of individual companies through the vocabulary of pre-built company to take advantage of the online news information. In this paper, we aim to improve performance of the stock prices prediction, applying online news information, through the stock price prediction of individual companies. Based on this, in the future, it will be possible to find ways to increase the stock price prediction accuracy by complementing the problem of unnecessary words that are added to the sentiment dictionary.

Key Words : Stock Prediction, Sentiment Analysis, Predictive Analytics

Received : November 5, 2015 Revised : December 6, 2015 Accepted : December 6, 2015
Corresponding Author : Jong Woo Kim

저 자 소 개

정지선

현재 한양대학교 일반대학원 경영학과 경영정보시스템 전공 박사과정에 재학 중이며, 홍익대학교 컴퓨터 정보통신 공학과에서 학사를 마쳤으며, 한양대학교 경영학과에서 석사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 추천 시스템, 사회 네트워크 분석, 오피니언 마이닝 등이다.

김동성

현재 한양대학교 일반대학원 경영학과 경영정보시스템 전공 박사과정에 재학 중이며, 협성대학교 경영정보학과에서 학사를 마쳤으며, 한양대학교 경영학과에서 석사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 빅 데이터, 오피니언 마이닝, 사회 네트워크 분석, 등이다.

김종우

현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 서울대학교 수학과에서 학사를 마쳤으며, 한국과학기술원에서 경영과학으로 석사학위를, 산업경영학으로 박사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 오피니언 마이닝, 상품 추천기술, 지능형 정보시스템, 집단지성, 사회 네트워크 분석, 클라우드 컴퓨팅 서비스 등이다.