DBpia

온라인상의 뉴스 감성분석을 활용한 개별 주가 예측에 관한 연구

저자 (Authors)	정지선, 김동성, 김종우
출처 (Source)	한국지능정보시스템학회 학술대회논문집 , 2015.5, 45-58(14 pages)
발행처 (Publisher)	한국지능정보시스템학회 Korea Intelligent Information Systems Society
URL	http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE06366393
APA Style	정지선, 김동성, 김종우 (2015). 온라인상의 뉴스 감성분석을 활용한 개별 주가 예측에 관한 연구. 한국지능정보시스템학회 학술대회논문집, 45-58
이용정보 (Accessed)	가천대학교 203.249.***.201 2019/09/23 10:44 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s)and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s)for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

온라인상의 뉴스 감성분석을 활용한 개별 주가 예측에 관한 연구

정지선 (한양대학교 일반대학원 경영학과, 주저자 skyhee84@hanyang.ac.kr)
김동성 (한양대학교 일반대학원 경영학과, paulus82@hanyang.ac.kr)
김종우 (한양대학교 경영대학 경영학부 교수, 교신저자 kjw@hanyang.ac.kr)

● 목차 ●

··· Abstract ···

정보기술의 발전에 따른 비정형 데이터의 급속한 증가로 인하여 데이터 활용 방안 및 분석 기법에 대한 다양한 연구들이 활발히 진행되고 있다. 특히, 최근에는 기존의 정형 데이터 활용의 한계점 보완을 위한 텍스트 마이닝 기법에 대한 활용 방안에 대한 연구들이 다수 이루어지고 있으며, 문서 내의 텍스트를 기반으로 문장이나 어휘의 긍정, 부정과 같은 극성 분포에 따라 의견을 스코어링(scoring)하는 감성분석 연구 또한 다수 이루어지고 있다. 본 연구는 소셜 미디어 상의 특정 주식 종목과 관련된 뉴스 데이터를 수집하여 이들의 감성 분석을 실시함으로써 주가의 등락에 대한 예측성과를 확인하고자 한다. 경제 주체의 다양한 정보들이 온라인상에서 쉽게 확산이 되고 있으며 이러한 정보는 주식 시장에 영향을 미치는 요인으로, 적절한 데이터 분석을 통해 주가 변동을 예측하는데 유용할 것이다. 이에 따라, 본 연구에서는 KOSPI200의 상위 종목들을 분석 대상으로 선정하여, 국내 대표적 검색 포털 서비스인 네이버에서 약 2년간 발생된 개별 기업별 뉴스 데이터를 수집·분석하였다. 또한 특정 경제 주체별로 나타나는 어휘의 의미의 상이함을 고려하여 각 개별 기업의 어휘사전을 구축하여 분석에 활용함으로써 감성분석의 성능을 향상시켰다. 이를 바탕으로 온라인 뉴스 데이터를 활용한 개별 기업의 주가 변화 예측성능 향상을 꾀하였다.

Key Words : Stock Prediction, Sentiment Analysis, Predictive Analytics

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

Ⅰ. 서 론

정보기술의 발전에 따른 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, 미래에 대한 대응을 위한 방안으로써 데이터 마이닝과 같은 통계적 분석 기법들을 바탕으로 고객 구매 정보를 활용한 제품 수요 예측에서부터 자사의 서비스를 이용하는 고객들의 이탈 예측, 생산 정보를 활용한 불량 제품 예측까지 다양한 분야에서 예측 분석(predictive analytics)에 대한 관심과 활용 방안에 대한 연구들이 꾸준히 발전하고 있다(송민정, 2013).

최근 예측 분석에 대한 연구들은 분석 가능한 다양한 데이터들의 증가로 인하여 기업 경영 활동과 같은 특정 분야로만 국한되지 않으며, 온라인상에서 발생된 데이터를 활용한 질병 예측, 선거 예측, 주식 시장의 변화 예측 등 다양한 분야로 확대되어 연구되고 있다(김유신 외, 2012; LaValle et al., 2013; Lee et al., 2013). 이러한 연구 동향의 일환으로 본 논문에서는 온라인상에서 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).

기업의 경영 활동의 범위가 다각화됨에 따라 기계학습 기반의 텍스트 데이터 분석 시 개별 기업의 특성을 고려한 분석이 요구되며, 기업 또는 산업 간 미치는 영향이 높은 관계에 위치한 경우 특정 기업에 대한 긍정적 정보를 나타내는 뉴스는 다른 기업 또는 산업 분야에는 부정적 영향에 대한 정보를 의미하는 뉴스가 될 수 있을 가능성도 고려해야 한다. 이에 따라 본 논문에서는 특정 경제 주체에 따라 쓰이는 어휘의

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

상이함을 고려하여 각 종목들에 따른 개별 어휘사전을 구축하여 분석에 활용함으로써 감성분석의 성능을 향상을 도모하였으며, 이를 바탕으로 실제 기업의 주식시장대비 초과수익률의 방향성을 예측하고자 하는 것이 기존 연구들과의 가장 큰 차이점이라고 할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 감성 분석(sentiment analysis)과 주식 시장에서의 예측 분석에 대한 기존 선행 연구들을 검토한다. 3장에서는 온라인상에서 발생된 각각의 기업 주가 관련 뉴스들의 감성 분석을 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.

Ⅱ. 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 및 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(김영민 외, 2014), 주식 관련 뉴스 데이터에 오피니언 반의법 규칙(Opinion Antonym Rule: OAR) 알고리즘을 적용하여 감성 사전 구축 후, 이를 통한 주식 상승·하락 분석 연구(조혜진 외, 2015), 주식 시장에 특화 된 감성 사전 구축을

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

통한 주가 지수의 방향성 예측 연구(유은지 외, 2013) 등이 있다. 이외에도 온라인상에서의 텍스트 데이터 분석을 활용한 주가 예측 분석과 관련된 국내외 대표적 연구들은 다음과 같다(<표 1> 참조).

<표 1> 감성분석을 활용한 주가 예측 분석(Predictive Analytics) 연구

연구자	연구내용
조혜진, 서지훈, 최진탁 (2015)	주식 관련 뉴스 텍스트 데이터의 패턴 분석을 기반으로 오피니언 반의법 규칙(Opinion Antonym Rule: OAR) 알고리즘을 적용하여 감성사전을 구축. 구축된 감성사전을 기반으로 코스피 지수 등락과의 관계를 분석.
김영민, 정석재, 이석준 (2014)	온라인상의 증권 종목 토론실에서 다수 논의되는 기업들을 대상으로 텍스트 데이터를 수집. 구글 응용 프로그램을 사용하여 영문으로 변환 후, 감성분석 기법을 활용한 주가 등락 예측 방안을 제시.
김유신, 김남규, 정승렬 (2012)	온라인상에서의 뉴스 정보에 대한 감성분석을 바탕으로 코스피 지수의 등락과 비교. 주식 시장 개장 전 시황·전망·해외 뉴스의 긍정·부정 비율을 활용한 로지스틱 회귀분석을 통하여 투자의사결정 모형 제시.
Evangelopoulos, N., Magro, M. J., and Sidorova, A., (2012)	기업에 대한 언급이 포함된 트윗을 수집, 수집 된 트윗들의 텍스트 마이닝 분석 기법을 바탕으로 개별 기업에 대한 주식 수익률 예측 모델 제시. 트윗 수와 주제가 주가 예측에 강한 상관관계가 있음을 확인.
Bank, M., Larch, M., Peter, G., (2011)	기업 무역활동과 같이 유동성에 미치는 영향과 투자자에게 관심이 높은 개별 기업의 구글 검색량과 주식 수익률에 대한 관계 분석.
Bollen, J., Mao, H., and Zeng, X. (2011)	트위터 사용자들 트윗에 대한 감성분석을 통하여 주식 시장 변화의 예측 방안에 대한 연구 수행.

기존 온라인상의 텍스트 데이터를 활용한 주가예측 관련 연구들의 검토를 통하여 미래 주가에 대한 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

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

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 차이점은 다음과 같다. 첫째, 기업별 주가 예측 시에 기업별 감성사전 구축을 통하여 기업별 온라인 뉴스 정보에 대한 보다 정확한 확인을 도모하였으며, 둘째, 기업별 주식의 초과수익률을 예측 대상으로 하여 기업 외부의 경제적 요인들을 고려하고자 하였다. 마지막으로 다양한 산업군에 속한 기업들을 연구 대상으로 선정하여 제안하는 연구 방안에 대하여 실증적인 유용성 검증을 꾀하였다.

2.2 감성 분석(Sentiment Analysis)

감성 분석(Sentiment Analysis) 기법은 사람이 사용하는 자연어에 대하여 기계 학습 기법을 바탕으로 문장의 긍정·부정과 같은 문맥 정보 등을 추출하거나 분류하고자 하는 기법을 의미한다. 감성 분석 방안으로는 감성 사전을 활용하여 문서 내 단어의 빈도를 기반으로 한 분석 방안과 문맥 전체의 정보를 해석하여 분석하고자 하는 방안들이 존재한다. 본 연구에서는 기존 다수 관련 연구에서 활용된 감성 사전을 활용한 감성 분석 방안을 수행하고자 한다. 감성 사전을 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 감성 분석 성능에 높은 영향을 미치는 중요한 요소 중 하나이다(김승우, 김남규, 2014). 감성 사전을 구축하는 방안으로는 대표적으로 문장 또는 단어의 극성이 사전에 정의된 감성 사전을 바탕으로 새로운 어휘들의 유사성 및 거리 관계 등의 비교를 통해 감성 사전을 구축하는 방안이 있으며, 실제로 수집된 문장들에 대한 형태소 분석을 바탕으로 개별 단어의 극성을 정의하여 구축하는 방안이 있다(조은경, 2012). 구축 된 감성 사전을 바탕으로 실제 텍스트에 대한 감성 분석 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 등의 통계적 분석 기법을 사용하여 분류하는 방안들이 존대하며, 이 밖에도 언어학적 방법론들을 적용한 연구들이 다양한 형태로 수행되고 있다(송상일 외, 2010; 조은경, 2012; Kim and Kim, 2014).

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

Ⅲ. 연구 방안

3.1 연구 데이터 및 선정

본 연구에서는 국내 대표 포털 사이트인 네이버의 증권 정보 서비스에서 기업별 '종목뉴스' 게시판에서 2013년 1월부터 2015년 4월까지 수집한 데이터를 활용하였다. 2013년 1월부터 2014년 4월까지 총 123,985개의 데이터를 훈련 데이터로 기업별 감성 사전 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 예측성과 확인을 위한 검증 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(<표 2> 참조).

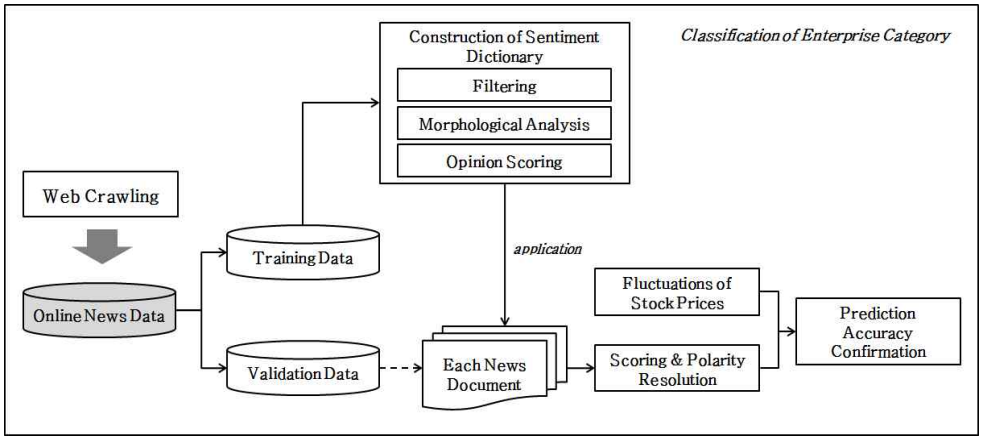
<표 2> 기업별 연구 데이터에 따른 뉴스의 수

종목	훈련데이터	검증데이터	종목	훈련데이터	검증데이터
현대건설	1,284	976	삼성전자	22,224	16,367
두산중공업	827	538	SK하이닉스	5,607	4,747
대림산업	1,084	740	NAVER	4,504	3,888
대우건설	1,293	869	LG전자	6,632	4,594
두산	676	492	LG	1,339	841
현대중공업	3,573	1,978	신한지주	3,673	2,565
삼성중공업	1,222	1,166	삼성생명	2,877	2,120
현대글로비스	403	737	KB금융	3,217	2,468
대우조선해양	1,386	1,176	하나금융지주	1,591	994
현대미포조선	586	610	삼성화재	1,202	992
POSCO	5,382	4,140	한국전력	3,964	3,661
현대제철	1,193	961	SK텔레콤	5,635	4,657
고려아연	430	261	KT&G	780	851
현대하이스코	661	479	LG생활건강	1,010	972
영풍	74	64	KT	4,034	2,716
LG화학	3,411	1,788	현대차	12,171	8,671
SK이노베이션	1,686	1,097	현대모비스	5,143	3,197
SK	1,568	959	기아차	7,984	4,477
S-Oil	683	881	롯데쇼핑	1,850	1,124
롯데케미칼	644	611	강원랜드	482	392

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

3.2 연구 절차

본 연구는 다음과 같은 연구절차를 통하여 수행되었다(<그림 1> 참조).



<그림 1> 기업별 감성사전을 활용한 주가 예측 방안

(1) 데이터 수집 및 활용

연구에 활용된 데이터는 네이버의 '종목뉴스'에서 기업별로 수집하였으며, 2013년 1월부터 2015년 4월까지의 온라인 뉴스이다. 본 연구에서 제시하는 방안을 검증하기 위해 훈련용 데이터와 검증용 데이터 집합으로 나누었으며, 훈련용 데이터를 활용하여 감성사전을 구축하였다. 또한, 기업별 감성사전의 성능을 예측하기 위하여 검증용 데이터 집합을 사용하였다.

(2) 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, 이 외에 뉴스 작성자의 메일주소와 특수기호 등 불용어들을 제거하였다. 최종적으로 추출된 명사의 감성 점수화를 통해 기업별 감성사전을 구축한다. 각 어휘의 감성 점수의 계산 방법은 다음의 식 (1), (2), (3)과 같다.

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

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

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

식 (1)과 (2)에서 TermScore(i)는 어휘 i의 감성 점수이며, 0에서 1의 값을 가진다. TotalNum(i)는 뉴스 전체에서 나온 i의 출현 빈도이며, Num(i)∈PosDocs는 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)∈NegDocs는 부정적 영향을 갖는 뉴스에서 발생한 i의 출현 빈도이다.

$$TermScore(i) = TermScore(i_p) + TermScore(i_n) \quad 식\ (3)$$

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.

$$AR_{jt} = R_{jt} - (\hat{\alpha}_j + \hat{\beta}_j R_{mt}) \quad 식\ (4)$$

AR_{jt}는 t시점에서 기업 j의 초과수익률이며, R_{jt}는 t시점에서 기업 j의 수익률이다. $\alpha_j + \beta_j R_{mt}$는 t시점에서 기업 j의 기대수익률이며, R_{mt}는 t시점의 시장수익률이다.

(3) 감성사전을 활용한 개별 기업의 주가 예측 방안

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

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

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

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

Ⅳ. 연구 결과

본 논문에서 제안한 개별 기업에 대한 감성사전 구축을 통한 주가 등락 예측 방안의 실제 예측 정확도는 다음과 같다(<표 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 16367개로 가장 많은 뉴스가 분석 대상 기간 동안 발생되었으며, 다음으로 '현대차'가 8671개, 'SK하이닉스'가 4747개, 'LG전자' 4594개로 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.

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

<표 3> 기업별 주가 예측 정확도

산업 구분	종목	예측 대상일	발생뉴스	정확도(%)	평균(%)
건설/기계	현대건설	219	976	54.34	55.68
	두산중공업	155	538	60.65	
	대림산업	196	740	55.61	
	대우건설	205	869	59.02	
	두산	166	492	48.80	
조선/운송	현대중공업	242	1978	57.44	53.85
	삼성중공업	209	1166	51.67	
	현대글로비스	172	737	51.16	
	대우조선해양	219	1176	56.62	
	현대미포조선	170	610	52.35	
철강/소재	POSCO	244	4140	52.46	55.17
	현대제철	213	961	58.22	
	고려아연	102	261	63.73	
	현대하이스코	143	479	60.14	
	영풍	46	64	41.30	
에너지/화학	LG화학	241	1788	60.17	60.52
	SK이노베이션	223	1097	66.82	
	SK	226	959	56.19	
	S-Oil	198	881	57.58	
	롯데케미칼	173	611	61.85	
정보기술	삼성전자	244	16367	49.59	52.19
	SK하이닉스	244	4747	61.07	
	NAVER	244	3888	52.05	
	LG전자	244	4594	53.69	
	LG	220	841	44.55	
금융	신한지주	244	2565	58.20	54.56
	삼성생명	244	2120	47.54	
	KB금융	240	2468	54.58	
	하나금융지주	226	994	52.21	
	삼성화재	214	992	60.28	
생활소비재	한국전력	244	3661	59.43	59.48
	SK텔레콤	244	4657	56.56	
	KT&G	193	851	66.84	
	LG생활건강	211	972	59.24	
	KT	244	2716	55.33	
경기소비재	현대차	244	8671	56.15	59.16
	현대모비스	244	3197	62.70	
	기아차	244	4477	54.92	
	롯데쇼핑	225	1124	55.11	
	강원랜드	130	392	66.92	

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

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.

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 기업들의 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.

Ⅴ. 결 론

5.1 연구 결과 정리

본 연구에서는 국내 주식 시장의 개별 기업에 대한 뉴스 정보를 수집 후, 이에 대한 감성분석을 활용한 주가 예측 방안에 대하여 연구하였다. 연구 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', '삼성

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

생명', '두산'과 같이 약 40%대의 예측 정확도를 보이는 기업들도 존재하였다. 특히, 데이터의 수가 가장 적었던 '영풍'은 약 41%의 예측 정확도를 보였으나, 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 경우 또한 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 존재하나, 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 추후 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. 그러나 뉴스 기사에서 다루어지는 국내외를 비롯한 복잡한 기업 경영 정보에 대한 이해와 해석에는 단어가 갖는 의미들의 연결 관계에 대한 이해와 의견 도출이 가능한 감성 분석 방안의 연구가 필요하다.

둘째, 수집 된 뉴스에 대하여 해당 기업에 대하여 직·간접적 연관이 존재하는 뉴스들의 구분이 필요하다. 실제 수집 된 증권 관련 뉴스의 일부는 해당 기업과는 직접

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

적 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. 향후 이러한 연구 한계점을 보완한 연구를 통하여 보다 정확한 개별 기업의 주가 예측 방안 제안이 가능할 것이다.

참 고 문 헌

김승우, 김남규, “오피니언 분류의 감성사전 활용효과에 대한 연구,” 지능정보연구, 제20권, 제1호, 2014, pp.133-148.

김영민, 정석재, 이석준, “소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구,” Entrue Journal of Information Technology, 제13권, 제3호, 2014, pp. 59-70.

김유신, 김남규, 정승렬, "뉴스와 주가: 빅데이터 감성분석을 통한지능형 투자의사결정 모형," 지능정보연구, 제18권, 제2호, 2012, pp. 143-156.

송민정, “빅데이터(Big Data)를 활용한 비즈니스 모델 혁신,” 과학기술정책, 제192호, 2013, pp. 86-97.

송상일, 이동주, 이상구, "PMI 를 이용한 우리말 어휘의 의미 극성 판단," 한국컴퓨터종합학술대회 논문집, 2010, pp. 260-26.

유은지, 김유신, 김남규, 정승렬, “주가지수 방향성 예측을 위한 주제지향 감성사전 구축 방안,” 지능정보연구, 제19권, 제1호, 2013, pp. 95-110.

조은경, “감성 분석 연구의 현황과 말뭉치에 기반한 사례 분석,” 언어과학연구, 제61권, 2012, pp. 259-282.

조혜진, 서지훈, 최진탁, “주식 뉴스 콘텐츠를 활용한 오피니언마이닝 기반의 OAR 감성사전 알고리즘 기법,” 한국정보기술학회논문지, 제13권, 제3호, 2015, pp. 111-119.

Bank, M., Larch, M., and Peter, G., “Google search volume and its influence on liquidity and returns of German stocks,” *Financial markets and portfolio management*, Vol. 25, No. 3, 2011, pp. 239-264.

Bollen, J., Mao, H., and Zeng, X., “Twitter mood predicts the stock market,” *Journal of Computational Science,* Vol. 2, No. 1, 2011, pp. 1-8.

de Fortuny, E. J., De Smedt, T., Martens, D., and Daelemans, W., “Evaluating and understanding text-based stock price prediction models,” *Information Processing & Management*, Vol. 50, No. 2, 2014, pp. 426-441.

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)

Evangelopoulos, N., Magro, M. J., and Sidorova, A., "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*, 2012, Vol. 15, pp. 247-268.

Kim, D. S., and Kim, J. W., "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, pp. 373-384.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N., "Big data, analytics and the path from insights to value," *MIT Sloan Management Review*, Vol. 52, No. 2, 2013, pp. 21-31.

Lee, J., Lapira, E., Bagheri, B., and Kao, H. A., "Recent advances and trends in predictive manufacturing systems in big data environment," *Manufacturing Letters*, Vol. 1, No. 1, 2013, pp. 38-41.

Schumaker, R. P., and Chen, H., "A quantitative stock prediction system based on financial news," *Information Processing&Management,* Vol. 45, No. 5, 2009, pp. 571-583.

Schumaker, R. P., Zhang, Y., Huang, C. N., and Chen, H., "Evaluating sentiment in financial news articles. *Decision Support Systems*," Vol. 53, No. 3, 2012, pp. 458-464.

가천대학교 | IP:203.249.***.201 | Accessed 2019/09/23 10:44(KST)