J. Soc. Korea Ind. Syst. Eng Vol. 41, No. 3 : 154-161, September 2018
https://doi.org/10.11627/jkise.2018.41.3.154
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

A Prediction of Stock Price Through the Big-data Analysis

Ji Don Yu* · Ik Sun Lee**†

*Entrepreneurship Education Center, Gwangju Institute of Science and Technology
**Dept. of Business Administration, Dong-A University

인터넷 뉴스 빅데이터를 활용한 기업 주가지수 예측

유지돈* · 이익선**†

*광주과학기술원 혁신기업가교육센터
**동아대학교 경영학과

This study conducted to predict the stock market prices based on the assumption that internet news articles might have an impact and effect on the rise and fall of stock market prices. The internet news articles were tested to evaluate the accuracy by comparing predicted values of the actual stock index and the forecasting models of the companies. This paper collected stock news from the internet, and analyzed and identified the relationship with the stock price index. Since the internet news contents consist mainly of unstructured texts, this study used text mining technique and multiple regression analysis technique to analyze news articles. A company H as a representative automobile manufacturing company was selected, and prediction models for the stock price index of company H was presented. Thus two prediction models for forecasting the upturn and decline of H stock index is derived and presented. Among the two prediction models, the error value of the prediction model ① is low, and so the prediction performance of the model ① is relatively better than that of the prediction model ②. As the further research, if the contents of this study are supplemented by real artificial intelligent investment decision system and applied to real investment, more practical research results will be able to be developed.

Keywords : Big-Data, Stock Price, Multiple Linear Regression

1. 서 론

글로벌적인 경제위기 속에서 주식시장의 상황에 대한 관심은 지속적으로 증가하고, 주식시장으로 유동성 자금이 모여드는 분위기가 지속되고 있다. 이러한 국제적인 상황 속에서 대한민국 정부도 해외 투자로 빠져나갔던 자금을 국내로 돌아오게 만드는 계기를 만들려는 의도로 2016년 8월부터 30분 거래시간을 연장하게끔 하는 법안을 통과시켜 실행하고 있다. 이와 같이 세계적으로 주식시장에 대한 관심이 상승하고 있는 상황들 속에서 다수의 개인 투자자들은 각종 인터넷 매체들을 통해 노출되는 주식 뉴스기사들을 판단의 근거들 중의 하나로 활용하고 있다. 이러한 정보를 참조하여 투자자들은 투자의 시기를 조율하며, 기업의 현재와 미래 가치를 평가하는 등으로 자신들의 주식투자 전략을 수립하는 것에 참조하고 있다. 투자자들은 인터넷에 뉴스들을 전문가들의 의견이 반영되어 있다고 생각하고 있으며, 즉 어느 정도의 공신력이 있다고 믿고 있으며, 이를 참조하는 투자자들이 많을 것으로 기대하기 때문에 투자의 분위기를 조성하는 것에 어느 정도 기여한다고 생각한다. 그러므로 투자자들은 인터넷 뉴스기사들의 기업의 주가지수에 제한적이나마 영향을 미치는 것으로 기대하고 있으며, 이를

Received 15 May 2018; Finally Revised 13 September 2018; Accepted 14 September 2018
† Corresponding Author : lis1007@dau.ac.kr

적극적으로 자신들의 투자전략에 반영하기도 한다. 그렇지만 인터넷 뉴스기사들은 매우 방대하게 쏟아지기 때문에, 그 중에서 가치 있는 정보를 선별하는 것은 쉽지 않다. 추가적으로 주식시장은 물가, 금리, 환율 등의 외부적 요인을 비롯하여, 기업의 실적, 자산, 재무제표 등의 내부적 요인에 적지 않은 영향을 받기 때문에 기업의 주가지수를 정확하게 예측하기는 생각보다 쉬운 일이 아니다.

이코노미스트(Economist)는 대략적으로 전 세계 600개 기업에게 빅데이터 활용 사례에 관해 조사를 실시하였다. 조사 대상자의 10%는 빅데이터가 지금까지의 비즈니스 모델을 전혀 새롭게 혁신할 것이며, 46%의 응답자는 기업에서 중요한 의사결정요인이 될 것이라고 예상하였다. 하지만 25%의 응답자는 기업 내부에 데이터가 많아도 이러한 데이터를 잘 써먹지 못하고 있으며, 53%의 응답자는 극히 일부의 데이터만을 사용하고 있다고 답변했다. 이러한 조사결과는 보다 나은 기업 의사결정을 위해서 대용량의 데이터 활용이나 분석 등에 많은 관심과 투자가 동반되어야 함을 나타내고 있다.

구글은 엄청나게 방대한 데이터로부터 가치있는 정보를 추출하여 이를 다방면의 비즈니스에 활용하고 있으며 인터넷 검색분야에서 선두를 달리고 있는 세계적인 기업이다. 구글은 사용자들이 접속하는 모든 웹페이지를 탐색하여 검색 단어와 웹페이지의 제목과 내용이 얼마나 관련성이 있는지를 표현하는 지수를 계산한다. 구글은 이러한 방대한 자료처리와 계산과정을 짧은 시간안에 처리해내기 위하여 분산파일 시스템을 개발했으며, 맵리듀스 프로그램을 개발하여 활용하고 있다. 구글이 개발한 언어 번역 시스템을 구글은 통계적 번역이라고 설명하고 있다. 구글 번역 시스템은 문법을 컴퓨터에게 가르치는 대신에, 사람들이 번역한 수억 건의 문서들의 패턴을 분류하여 언어번역의 규칙을 만들도록 하는 방법이다. 문법을 컴퓨터에게 가르치기는 어려움이 많기 때문에, 이러한 문서들을 활용하여 컴퓨터를 학습시키는 접근법으로서, 문서의 양이 많아질수록 번역시스템의 정확도는 높아질 것이다.

IBM은 슈퍼컴퓨터 '왓슨'을 개발하였는데, IBM은 인간의 언어자료에 대한 분석 및 이해를 토대로 대량의 정보를 짧은 시간 안에 탐색하는 기술의 가능성을 새롭게 입증한 사례라고 말할 수 있다. 2011년 2월 왓슨은 미국의 인기 있는 퀴즈쇼에 출연하여 인간 챔피언과 대결하여 승리를 얻었다. 퀴즈쇼에서 다룬 질문들은 분야가 매우 다양했으며 시적이고 은유적인 표현들이 다수 포함되었기 때문에 퀴즈쇼 시청자들도 의미를 이해하기가 난해했다. 왓슨은 4테라바이트 하드디스크 공간에 축척된 2억 페이지를 넘어서는 콘텐츠를 활용할 수 있었다. 현재 왓슨은 의료보험에 관한 데이터 분석과 병원에서 종양의 진단과 처방 등에 활용될 예정이며, 더 나아가서 씨티그룹 등을 비롯한 금융분야의 활용될 수 있는 방안을 모색하는 중이다. 미국 온라인 쇼핑몰 아마존도 빅데이터를 전략적으로 활용하는 기업 중의 하나이다. 아마존은 소비자의 구매 데이터들을 분석하여 소비자의 취향과 흥미에 부합하는 도서를 추천하는 프로그램을 개발해냈다. 이러한 접근방법은 빅데이터를 활용한 전형적인 마케팅 분석 기법이라고 말할 수 있다.

우리의 언어는 문법적으로 어휘적으로 다양한 특성들이 있고, 동일한 의미를 전달할지라도 표현의 형태는 매우 복잡하고 다양해질 수 있다. 그래서 정형화된 방식으로 일괄규정하기 어려운 경우가 많고, 또한 언어는 사회적으로 활용되는 방식과 유형에 따라 지속적으로 변화하고 발전하는 특성이 있다. 문자로 표현되어진 언어를 컴퓨터를 이용하여 수집하고, 가공 및 처리하여 구조와 그 의미를 파악하는 기술을 자연언어처리 기술이라고 부른다. 텍스트마이닝을 활용하여 대량으로 쏟아져나오는 정보들 속에서 가치 있는 의미를 도출하고, 다른 요인과 상관성을 파악하는 등으로 접근법을 시도하는 것이 최근의 추세이다. 텍스트마이닝 기법은 현재까지 다양하게 연구되어왔지만, 컴퓨터 및 인공지능, 빅데이터의 발전과 더불어서 앞으로 더 많은 연구들이 진행되고 발전될 가능성이 충분한 연구분야이다.

오피니언 마이닝 기술은 빅데이터에 포함되는 사건이나 이슈, 인물 등에서 발생되는 사람들의 감정이나 생각, 의견 등을 분석하는 기술이다. 오피니언은 제품을 구매할 때 상품평이나 영화를 본 감상평과 같이 특정한 주제에 대해서 사람들의 주관적인 생각이나 감정 등을 일컫는다. 오피니언 마이닝은 이러한 사람의 개인적 의견이나 감정이 담긴 메시지를 통해 패턴을 밝히고 의견이 긍정인지 부정인지를 찾아낸다.

인터넷 포털사이트에서 거의 실시간으로 출현하는 수십만 건의 기사들을 비정형 빅데이터으로 간주하여 가공, 처리 및 분석해서 동향을 파악하고자 한다. 인터넷에서 넘쳐나는 다양한 유형의 비정형 빅데이터는 동시다발적으로 대량의 정보를 생성함으로써 사람들에게 정보를 제공하고 있다. 미디어 홍수속의 비정형 텍스트 데이터는 단순하게 문장들의 전체적 의미보다는 어휘와 어휘들 간의 의미가 더 중요하다. 어휘들마다 내포되어져 있는 뜻이 글을 작성한 사람의 긍정적, 부정적 사고나 감정, 의견 등을 충분히 담고 있기 때문에 이를 가공하여 빠른 시간 안에 다량의 텍스트를 파악할 수 있다는 것이 빅데이터 연구의 장점이라고 할 수 있다. 하지만 정형·비정형・반정형으로 구성되어지는 빅데이터는 종류와 형태에 따라서 다양한 분석 방법론이 존재하기는 하지만, 매우 복잡한 구조를 가지고 있고, 해석은 연구자의 주관에 의존하여 달라질 가능성이 있기에 어휘들 속에서 정보를

분석하는 것은 쉽지 않은 실정이다.

본 논문은 이러한 한계들 속에서도 인터넷 뉴스기사를 통해 추출되어진 어휘를 통해 주가와 상관관계를 분석하는 연구를 수행하였다. 인터넷의 포털 '다음'에서 뉴스기사들이 수집되었는데, 특정 날짜에 따라 검색되어진 인터넷 뉴스기사를 수집하였다. 본 논문은 수집된 데이터들을 긍정 또는 부정적인 의미에 따른 빈도로 측정하고, 주가지수에 관련이 큰 단어들만을 선별하였다. 이를 통해 추출되어진 어휘와 주가와의 상관관계를 입증하고, 예측 정확도 비교 테스트를 통해 도출되어진 예측모델들의 평균 오차를 파악하여 가장 우수한 투자 예측모델을 제시하고자 한다.

주식시장에서 기업의 주가지수를 예측하는 것은 많은 연구자들의 관심을 받아왔다. 경제학, 통계 및 전산, 수학 등에 이르는 많은 분야에 걸쳐 오랜 기간 동안 주요한 연구 주제로 다루어지고 있다. 지금까지 주가지수를 예측하기 위한 방법으로서 전문회사들 혹은 경제학자들이 주로 활용해온 방법론으로는 계량적 분석방식이라고 불리는데, 이러한 방법은 수학적인 접근방법으로서 미래가치를 수치적으로 예측하고, 이를 활용하여 투자분석 포트폴리오를 제안하여 투자전략을 위한 의사를 지원하는 기법으로서 사용되어져 있다. 가장 대표적인 수학적 모델로 말하자는 거래의 가격범위를 제한하고 매매주문의 가격변동을 예측하는 여과방법과 시계열적인 데이터 움직임을 분석하고 데이터간의 연관성과 미래 변동을 예측하는 웨이블렛 변환 등의 방법론들이 있다[1].

최근에는 컴퓨터에 기초한 연구의 발전과 더불어 과학적 통계방법론들을 이용한 주가지수예측과 그 활용 전략에 대한 연구가 활발히 진행되고 있다. 종합주가지수 KOSPI를 예측하는 논문들을 소개하자면, 김선웅 및 안현철[7]은 Genetic 알고리즘을 활용하여 지능적인 투자방안을 고안하였고, 박종엽 및 한인구[13]는 Neural Network 알고리즘에 근거하여 종합주가지수 KOSPI를 예측하는 기법을 제안하였다. 허양민[3]은 인터넷의 검색어들을 수집하고, 블로그 및 SNS의 텍스트들의 감정을 긍정 혹은 부정으로 평가하여 종합주가지수의 상승 또는 하락을 예측하는 방법을 연구하였다. 인터넷 검색어에 기반하여, SNS 및 블로그에서 출현되는 관련된 텍스트들을 수집하여, 이러한 텍스트들과 종합주가지수와의 상관성을 분석하였다.

인터넷 뉴스기사들을 활용하여 종합주가지수 KOSPI의 동향을 예측했던 연구들이 존재하는데, 이들은 인터넷 뉴스기사들을 긍정적인 뉴스와 부정적인 뉴스로 분류하고, 이러한 긍정, 부정의 뉴스들의 건수와 비중과 종합주가지수 KOSPI가 얼마나 관련이 있는지를 분석하였다[1, 14]. 천세원[2]은 인터넷에 존재하는 여러 뉴스매체들을 비교분석하는 연구를 수행하였는데, 인터넷 뉴스매체별로 주가지수를 상승 또는 하락할 것이라고 전망하는데, 이러한 전망을 분석하여 어떤 뉴스매체의 전망이 가장 정확한지 분석했다.

종합주가지수 KOSPI의 동향을 예측하기 위한 색다른 연구로서 감성사전을 만들고, 이러한 감성사전에 기초로 뉴스기사들에 대해서 감성점수를 평가한 연구들도 존재한다[4, 6, 8]. 측정된 감성점수를 활용하여 긍정 혹은 부정적인 느낌의 정도를 판단하고 종합주가지수 KOSPI와의 관련성을 분석하고 연구했다.

김동영[5], 문하늘 및 김종우[9], 이예지[8]는 종합주가지수 KOSPI보다 개별기업의 주가지수를 예측하는 연구들을 수행하였다. 이러한 논문들의 공통적인 접근법은 단어들의 감성사전을 만들고, 인터넷 뉴스기사들을 수집하여, 감성사전을 활용하여 뉴스기사들의 긍정 혹은 부정의 정도를 평가한다는 점이다. 긍정 또는 부정의 감성점수에 근거하여 개별 기업의 주가지수의 등락을 규명하는 연구들을 수행했다.

본 논문은 인터넷의 뉴스기사들 속에서 선별 및 추출되어진 단어들을 활용하여 한국의 대표적인 H기업의 주가지수를 예측하는 연구를 수행한다. 뉴스기가들은 인터넷 포털 '다음'에서 수집되었는데, 특정 날짜에 따라 검색되어진 인터넷 뉴스기사를 수집한다. 수집된 데이터는 빈도수로 파악하고, 빈도수가 높은 단어들을 추출한다. 이를 통해 추출되어진 단어와 H기업의 주가지수와의 상관관계를 입증하고, 도출되어진 예측모델들을 비교하여 성능테스트를 실시하여 가장 나은 예측모델을 제시한다.

2. 연구방법

H기업에 대한 뉴스를 수집하기 위해 포털사이트 '다음'을 이용하였다. 2015년 특정 날짜에 검색되어진 주가지수 ±2% 이상 상승한 15일치 뉴스와 하락한 15일치의 뉴스 데이터 약 57,000건을 수집하여 실험을 수행하였다.

수집된 H사의 인터넷 뉴스기사들은 먼저 뉴스기사 텍스트들의 형태소를 분석하는 과정을 거쳤다. 수집된 텍스트들이 워낙 방대하기 때문에, 이를 주요한 단어들로만 추출하도록 정제화(cleaning)하는 작업을 수행하였다. 단어들과 함께 수집된 부수적인 기호나 조사들은 분리하는 작업을 수행하였고, 주가지수에 전혀 영향을 미치지 않는 일반적인 지명이나, 장소, 날짜, 날씨 등의 단어들도 정제하였다.

정제화 작업을 하는 동안에 복합명사는 가능한 분리하지 않았으며, 의미를 가지지 않는 어휘나 부호 등은 제거하였다. 비슷한 뜻을 가지고 있지만 다양하게 표현된 단어들은 대표적인 한 개의 단어들로 통합하였다. 예를

들면, ‘현대차그룹’, ‘현대차’, ‘현차’ 등은 ‘현대자동차’으로 통합하였으며, ‘美國’, ‘미’, ‘美’ 등은 미국으로 통합하였다(<Table 1> 참조).

<Table 1> Examples of Symbols Removed During Refining Procedure

,	'	[	]	...
-	|	▷	=	·
<	>	◇	▲	(
)	&	■	▶	*
:	;	"	..	.
·	/	◆	은	들
없이	으며	하여	해서	되며
의	을	하고	로서	으로
電	+	하며	였고	와
를	에	는	처럼	@

정제화 과정을 거친 후에는 한국어 단어 분석에 활용되는 “Krkwic” 소프트웨어를 활용하여 빈도를 분석하였다[10, 11, 12]. 빈도를 분석한 결과를 바탕으로 빈도가 많은 단어들을 선별하고, 그러한 단어들을 회귀분석의 독립변수들로 활용하였다. 즉 본 연구는 H사의 주가지수를 예측하기 위해 H사와 관련된 인터넷 뉴스들을 수집하고 기초분석하여 12개의 ‘명사’ 변수들(상승, 이익, 매수(순매수) 증가, 강세, 개선, 하락, 손실, 부진, 매도(순매도) 감소, 적자)과 내용을 서술해주는 10개의 ‘동사’ 변수들(반등했다, 유지했다, 오른, 기대된, 회복했다, 매각했다, 우려된, 내린, 급락했다, 둔화되다)로서 분류작업을 실행하여, 8개의 ‘긍정’ 변수들과 8개의 ‘부정’ 변수들로 분류하는 작업을 실행하였다. 이러한 분류작업을 거친 단어들이 H사의 주가지수와 관련성이 있는지를 확인하기 위해 단순 다중회귀분석을 비롯하여, 단계적 다중회귀분석을 SPSS 22.0 소프트웨어를 이용하여 테스트를 수행하였다. <Table 2>는 본 연구의 분류작업 결과로서 도출된 긍정 및 부정 단어들이다.

3. 연구 분석

본 연구는 H사에 관련하여 ‘다음’ 포털사이트에서 검색된 뉴스기사 ±2% 이상 상승·하락한 날의 각 15일치 데이터를 수집하였고, 전처리분석 과정을 통해 텍스트를 정제 및 분류하였다. 정제되어진 단어들은 주가지수와 관련성이 높고, 빈도출현이 많은 단어들을 중심으로 8개의 긍정 변수 및 8개의 부정 변수들로 정리하였다.

추출된 변수들과 H사의 주가지수와의 상관관계분석을 실시해보았는데 그러한 결과는 다음의 <Table 3>에 잘 나타나 있다. H사의 주가지수와 추출되어진 16개 변수와의 상관관계를 살펴보면 투자하다($r = .536$, $p < .10$), 확대하다($r = .399$, $p < .10$), 늘리다($r = .398$, $p < .10$), 판매하다($r = .29$, $p < .10$), 강화($r = .565$, $p < .10$)의 변수들은 정(+)의 상관이 나타났고, 노조($r = -.036$, $p < .10$), 변수만 부(-)의 상관이 있는 것으로 나타났다.

<Table 3> Correlation Analysis

words	correlation	
	R	p-value
투자하다(invest)	.536	.001
확대하다(expand)	.399	.015
생산하다(produce)	.112	.277
점유율(market share)	-.100	.304
늘리다(increasing)	.298	.055
판매하다(sell)	.290	.060
증가(increase)	.182	.168
강화(reinforcement)	.565	.001
노조(union)	-.360	.027
규제(regulation)	-.200	.151
부진(sluggish)	-.240	.103
통상임금(wage)	.016	.467
하락(degradation)	-.150	.221
약세(weakness)	.229	.112
감소(decrease)	-.220	.120
매도(sell)	.031	.435

<Table 2> Positive and Negative Words

positive words	투자하다 (invest)	확대하다 (expand)	생산하다 (produce)	점유율 (market share)	늘린 (increasing)	판매하다 (sell)	증가 (increase)	강화 (reinforcement)
20150108(+4.71)	26	14	7	12	12	68	13	13
20150217(+2.85)	34	18	9	12	16	20	24	25
20150305(+2.45)	27	12	10	17	15	10	14	11

...

negative words	노조 (union)	규제 (regulation)	부진 (sluggish)	임금 (wage)	하락 (degradation)	약세 (weakness)	감소 (decrease)	매도 (sell)
20150116(-2.01)	110	23	12	210	15	13	22	11
20150126(-2.08)	22	12	29	23	24	14	28	18
20150209(-3.67)	41	22	22	20	20	15	14	16

정제과정과 분류과정을 거쳐서 선정된 16개의 단어 변수들을 통합하여 다중회귀분석을 실시하였다. <Table 4>는 다중회귀분석 결과를 나타낸 것으로 16개의 변수들을 모두 독립변수로 고려하여 다중회귀분석을 실시한 결과자료이다.

<Table 4>의 분석결과에서 수정된 결정계수는 .805이며, 이때 F값의 유의확률은 .000으로 추정된 회귀식이 통계적으로 유의미한 것으로 나타났다. 독립변수들이 통계적으로 유의미한 변수인지를 판단하는 t값의 유의확률을 살펴보면 '투자하다', '늘리다', '판매하다', '강화', '노조', '부진', '하락', '약세' 변수가 .10 이하의 좋은 값을 보이고 있다. 이는 결과에서 보여주듯이 16개의 변수들 중에서 통계적으로 8개의 변수가 주가지수와의 상관관계가 있는 것으로 나타내고 있다. 다중회귀분석을 통해 도출되어진 예측모델 ①은 아래와 같다.

※ 예측모델 ①

$$Y(\text{주가지수}) = -0.663+0.604X_1-0.056X_2-0.051X_3-0.285X_4 +1.633X_5+0.613X_6+0.500X_7+0.474X_8 -0.266X_9-0.264X_{10}+0.530X_{11}+0.198X_{12} -0.826X_{13}-1.030X_{14}-0.038X_{15}-0.211X_{16}$$

<Table 4> The Multiple Linear Regression Results

	R	R^2	adjusted R^2	Std. error of the Estimate
Summary	.955	.913	.805	.19672

		SS	df	MS	F	p-value
ANOVA	Regression	5.253	16	.328	8.484	.000
	Residual	0.503	13	.039		
	Total	5.756	29			

		Unstandardized constants		t	p-value
		B	Std. Error		
Coefficient	(constant)	-0.663	0.235	-2.82	0.014
	투자하다(invest)	0.604	0.15	4.027	0.001
	확대하다(expand)	0.056	0.203	0.279	0.785
	생산하다(produce)	-0.05	0.144	-0.36	0.728
	점유율(market share)	-0.29	0.201	-1.42	0.18
	늘리다(increasing)	1.633	0.433	3.77	0.002
	판매하다(sell)	0.613	0.238	2.572	0.023
	증가(increase)	0.5	0.287	1.741	0.105
	강화(reinforcement)	0.474	0.174	2.731	0.017
	노조(union)	-0.27	0.137	-1.93	0.075
	규제(regulation)	-0.26	0.201	-1.31	0.212
	부진(sluggish)	0.53	0.277	1.91	0.078
	통상임금(wage)	0.198	0.154	1.283	0.222
	하락(degradation)	-0.83	0.31	-2.66	0.02
	약세(weakness)	-1.03	0.26	-3.96	0.002
	감소(decrease)	-0.04	0.148	-0.26	0.802
	매(sell)	-0.21	0.235	-0.9	0.385

주가지수를 예측하기 위해 도출되어진 변수들이 상대적으로 적합한지를 확인하기 위해 단계적 다중회귀분석(stepwise regression analysis)을 실시하였다. 즉 보다 나은 예측모델을 제시하기 위해서 16개의 독립변수를 활용하는 단계적 다중회귀분석을 실시하였다. <Table 5>에서 분석의 결과를 확인할 수 있다. 단계적 다중회귀분석의 결과로서 얻어진 수정된 결정계수는 .652인데 주가지수의 65.2%가 선별된 독립변수들로부터 설명되고 있음을 확인할 수 있다. 즉, 총 5개의 '강화', '부진', '약세', '규제', '점유율' 독립변수가 주가지수의 65.2% 설명하고 있는 것으로 나타났다.

<Table 5> The Independent Variables in the Stepwise Multiple Regression Analysis

	R	R^2	adjusted R^2	Std. error of the Estimate
1	.565	.319	.295	.37420
2	.680	.463	.423	.33836
3	.782	.611	.567	.29329
4	.817	.668	.615	.27655
5	.844	.712	.652	.26266

1. predictors : (constant), 강화(reinforcement)
2. predictors : (constant), 강화(reinforcement), 부진(sluggish)
3. predictors : (constant), 강화(reinforcement), 부진(sluggish), 약세(weakness)
4. predictors : (constant), 강화(reinforcement), 부진(sluggish), 약세(weakness), 규제(regulation)
5. predictors : (constant), 강화(reinforcement), 부진(sluggish), 약세(weakness), 규제(regulation), 점유율(market share)

<Table 6>은 구해진 R^2의 유의도 검증을 실시한 결과로서, 표에서 F 값은 11.887이며 유의확률은 .000으로 나타났다. 그러므로 단계적 다중회귀분석 모델은 유의수준 $p < .000$으로서 통계적으로 유의미한 것으로 나타났으며, 주가지수에 '강화', '부진', '약세', '규제', '점유율'의 변수들은 H사의 주가지수에 의미 있는 영향력을 제공한다고 말할 수 있다.

주가지수에 영향을 주는 변수에 대한 각 예측변인들의 직접적 효과의 유의성 검증은 <Table 6>에서 그 결과를 확인할 수 있다. 즉 <Table 6>의 결과에서 확인하듯이 단계적 다중회귀분석에서 주가지수와 상관관계가 있는 변인으로는 '강화', '부진', '약세', '규제', '점유율' 변수가 통계적으로 $p < .10$에서 유의미한 것으로 나타났다. 5개의 변수중 주가지수가 상승에 가장 많은 영향을 미치는 변수는 '강화' 변수(.672)로 나타났고 하락하는 주가지수에 큰 영향을 끼치는 변수는 '부진'(-.410)로 나타났다. 단계적 다중회귀분석에 의해 도출된 예측모델 ②는 아래와 같다.

<Table 6> ANOVA Analysis

		SS	df	MS	F	p-value
1	Regression	1.835	1	1.835	13.107	.001
	Residual	3.921	28	.140		
	Total	5.756	29			
2	Regression	2.665	2	1.333	11.639	.000
	Residual	3.091	27	.114		
	Total	5.756	29			
3	Regression	3.520	3	1.173	13.639	.000
	Residual	2.236	26	.086		
	Total	5.756	29			
4	Regression	3.844	4	.961	12.566	.000
	Residual	1.912	25	.076		
	Total	5.756	29			
5	Regression	4.100	5	.820	11.887	.000
	Residual	1.656	24	.069		
	Total	5.756	29			

<Table 7> Significance of Variables in the Stepwise Multiple Regression Analysis

		Unstandardized constants		t	p-value
		B	Std. Error		
5	(constant)	.110	.105	1.039	.309
	강화(reinforcement)	.672	.117	5.738	.000
	부진(sluggish)	-.410	.130	-3.147	.004
	약세(weakness)	.524	.131	4.010	.001
	규제(regulation)	-.341	.168	-2.034	.053
	점유율(market share)	-.246	.128	-1.927	.066

※ 예측모델 ②

$$Y(\text{주가지수}) = -0.110+0.672X_1-0.410X_2+0.524X_3 -0.341X_4-0.246X_5$$

<Table 8>은 단순 다중회귀분석을 통해 도출되어진 변수 예측값과 H사의 2016년도 주가지수 실제값 정확도 테스트이다. 2016년 상승했던 날의 예측값과 실제값 사이의 오차를 분석한 결과 최소오차와 최대오차는 각각 1.13%와 2.49%였으며, 전체 평균 오차값은 1.77%로 최소, 최대 오차의 편차가 적지 않다고 말할 수 있다. 하지만 실제값과 예측값의 오차율을 제외한 일반적인 상승에 대한 예측 정확도는 상당히 뛰어남을 알 수 있다.

2016년 하락했던 날들의 최소 오차와 최대 오차는 각각 0.83%와 3.46%, 전체 오차 평균은 1.92%로 최소, 최대 오차값의 범위가 상승구간에 비해 오차 범위가 큰 것을 확인할 수 있다. 하지만 하락값에 대한 일반적인 예측 정확도는 특정 하루를 제외한 모든 일자가 하락률을 예측하고 있기에 등락의 정확도는 비교적 정확한 것으로 나타났다.

<Table 9>는 단계적 다중회귀분석을 통해 도출되어진 예측모델 ②에 의한 예측값과 2016년 주가지수의 실제값 정확도 테스트이다. 2016년 상승했던 날들의 실제값과 예측값의 정확도의 차이를 분석한 결과 최소오차와 최대오차는 각각 2.06%와 3.44%였으며, 전체 평균값은 2.76%로 예측모델 ①보다 상대적으로 좋지 않은 예측력을 보였다.

2016년 하락했던 날들의 최소오차와 최대오차는 각각 1.34%와 3.39%, 전체 오차 평균은 2.04%로 최소, 최대 오차값의 범위가 예측모델 ①보다 큰 차이가 없음을 확인할 수 있다. 개별적 데이터 오차값에서도 규칙적인 데이터 검증을 확인할 수 있었던 예측모델 ①보다는 예측 정확도가 다소 좋지 않음을 확인할 수 있다.

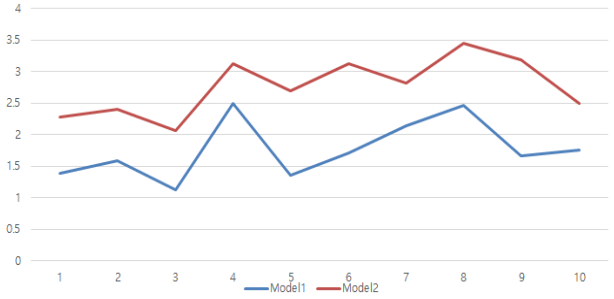
<Figure 1>은 2가지 유형의 다중회귀모형에 따른 상승했던 날들의 오차의 추세를 표현한 그래프로서 예측모형들의 예측의 성능을 나타내고 있다. X축은 주가지수, Y축은 평균 오차를 표현하고 있으며 2가지 예측모델 중에서 예측모델 ①, 예측모델 ②와 비슷한 추이를 보여주고 있음을 알 수 있다.

<Table 8> The Prediction Performance of the Prediction Model ①

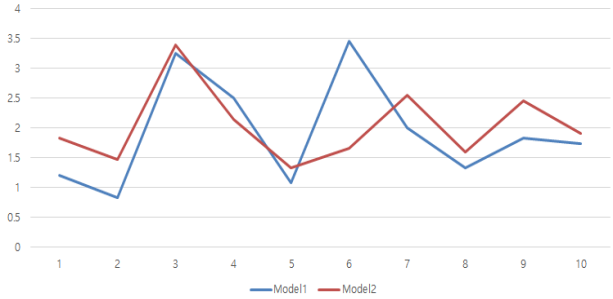
	Rising day in 2016	Actual value	predicting value	gap		Declining day in 2016	Actual value	predicting value	gap
1	1.11	2.71%	1.32%	1.39%	1	1.06	-2.01%	-0.80%	1.21%
2	2.04	2.85%	1.26%	1.59%	2	1.02	-2.08%	-1.25%	0.83%
3	2.12	2.45%	1.32%	1.13%	3	3.25	-3.67%	-0.42%	3.25%
4	2.17	3.70%	1.21%	2.49%	4	4.04	-2.86%	-0.36%	2.50%
5	3.22	2.40%	1.05%	1.35%	5	4.29	-2.20%	-1.12%	1.08%
6	4.14	2.27%	0.56%	1.71%	6	5.09	-2.17%	1.29%	3.46%
7	5.31	3.34%	1.20%	2.14%	7	6.13	-3.31%	-1.31%	2.00%
8	7.13	2.74%	0.27%	2.47%	8	6.03	-2.36%	-1.02%	1.34%
9	8.09	2.69%	1.03%	1.66%	9	7.29	-3.12%	-1.28%	1.84%
10	9.05	2.96%	1.20%	1.76%	10	10.1	-2.2%	-0.46%	1.74%
Average		2.81%	1.04%	1.77%	Average		-2.60%	-0.67%	1.92%

<Table 9> The Prediction Performance of the Prediction Model ②

	Rising day in 2016	Actual value	predicting value	gap		Declining day in 2016	Actual value	predicting value	gap
1	1.11	2.71%	0.43%	2.28%	1	1.06	-2.01%	-0.17%	1.84%
2	2.04	2.85%	0.45%	2.40%	2	1.02	-2.08%	-0.61%	1.47%
3	2.12	2.45%	0.39%	2.06%	3	3.25	-3.67%	-0.28%	3.39%
4	2.17	3.7%	0.57%	3.13%	4	4.04	-2.86%	-0.71%	2.15%
5	3.22	2.4%	-0.29%	2.69%	5	4.29	-2.2%	-0.86%	1.34%
6	4.14	2.27%	-0.86%	3.13%	6	5.09	-2.17%	-0.50%	1.67%
7	5.31	3.34%	0.52%	2.82%	7	6.13	-3.31%	-0.76%	2.55%
8	7.13	2.74%	-0.70%	3.44%	8	6.03	-2.36%	-0.76%	1.60%
9	8.09	2.69%	-0.50%	3.19%	9	7.29	-3.12%	-0.66%	2.46%
10	9.05	2.96%	0.47%	2.49%	10	10.1	-2.2%	-0.28%	1.92%
Average		2.81%	0.05%	2.76%	Average		-2.60%	-0.56%	2.04%



<Figure 1> The Gap Curve of the Rising Days



<Figure 2> The Gap Curve of the Declining Days

하지만 예측모델 ①은 최소, 최대 오차의 범위가 각각 1.13%, 2.49%로 특수한 경우를 배제한다면 평균 오차는 실질적으로 예측모델 ② 보다 더 뛰어난 예측 정확도를 보여준다고 말할 수 있다.

<Figure 2>는 2가지 예측모델에 따른 하락한 날의 오차 추이를 보여주고 있는 그래프이다. 2가지 예측모델 중 예측모델 ①과 예측모델 ②가 모두 비슷한 파형을 보여주고 있으며 단순 다중회귀분석과 단계적 다중회귀분석이 비슷한 정확도를 예측하고 있지만, 예측모델 ①이 예측모델 ②보다 오차의 범위가 더 작게 나타남을 보여주고 있다고 말할 수 있으며, 예측모델 ②가 안정적으로 주가지수의 하락세에 대한 예측 정확도를 보여주고 있다고 말할 수 있다.

3. 연구결과

본 연구는 인터넷의 주식관련 뉴스기사가 기업의 주가지수의 상승 및 하락에 영향력을 가질 것이라는 가정하에서 기업의 주가지수 변동성을 예측하는 연구를 수행하였다. 즉 인터넷 뉴스기사들을 수집하고 정제 및 가공하였으며, 기업의 실제 주가지수와 예측모델들의 예측값을 비교하여 정확도를 평가하는 테스트를 실시하였다.

본 연구는 우리나라의 대표적인 자동차제조기업 H사를 선정하여, H사의 주가지수를 예측하는 예측모델을 제시하였다. 본 연구는 H사 주가지수의 상승·하락 등락률을 예측하기 위해 다중회귀분석에 기반하여 2가지의 예측모델을 제시하였다. 2가지 예측모델 중에서 예측모델 ①의 오차값이 낮게 나타났으며 상대적으로 예측모델 ② 보다 나은 성능을 보임을 확인하였다. 향후에 본 연구의 내용을 기본으로 실제 인공지능형 투자결정시스템으로 보완 및 구현하여 실물투자에 적용해본다면 보다 현실적인 예측모델을 고안할 것으로 기대해본다.

본 연구의 연구내용을 토대로 판단해보면 인터넷에서 실시간으로 꾸준히 출현하는 주식과 관련 뉴스기사들의 긍정적 혹은 부정적인 단어의 발생은 기업의 주가지수 변동성에 충분히 영향을 미치고 있다고 말할 수 있다. 본 연구의 예측모델이 정확한 주가지수 자체를 예측하기에는 다소나마의 한계가 존재하지만 기업의 주가지수의 등락의 경향성은 비교적 잘 예측한다는 점에서는 적지 않은 의미가 있다고 말할 수 있다.

본 연구에서 활용하는 예측방법론은 상승 또는 하락하는 날에 빈번하게 출현하는 단어들에 기반하여 예측했기 때문에, 정부의 정책발표나 대내외의 상황변화 등의

외부적인 돌발 이벤트가 발생하는 경우에는 기업의 주가지수의 변동성을 예측하는 것에는 어렵다고 말할 수 있을 것이다. 실험의 결과에서 확인하듯이 외부적인 돌발 이벤트에 기인하여 큰 변동성을 보이는 특별한 날에는 예측모델의 성능이 낮아지는 경향을 확인할 수 있었다. 이러한 현상으로부터 본 논문의 과정에서 선정된 긍정 및 부정의 단어들이 기업의 주가지수의 변동을 모두 설명한다고 말할 수 없다는 것을 방증한다. 그렇기에 글로벌 정치 경제 동향이나 중국 및 미국의 정책발표 등의 특이한 외부사건들이 발생하는 날의 주가지수 변동성을 예측하는 것에는 다소 어려움이 있다고 말해야 할 것이다.

본 연구를 추가적으로 확장해본다면, 본 연구는 명사 및 동사의 어휘군들을 활용하였는데, 이러한 일부의 어휘만을 사용하기보다는 우리나라 국어의 언어적 성질을 폭넓게 반영하는 추가적인 연구가 수행될 필요가 있을 것이다. 언어는 복잡하고 미묘한 상황에 따라 똑같은 표현일지라도 상황에 따라서 어떤 단어 의미가 긍정으로 해석되기도 하고, 정반대로 부정으로 해석되기도 한다. 이러한 언어적인 특수한 성질을 더욱 면밀히 반영한다면 예측모형의 정확성을 향상시킬 가능성이 있으리라 사료된다.

Acknowledgement

This study was supported by research funds from Dong-A University.

References

[1] Ann, S.W. and Cho, S.B., Stock Prediction Using News Text Mining and Time Series Analysis, *Korea Computer Congress*, 2010, Vol. 37, No. 1, pp. 364-369.

[2] Chun, S.W., Kim, Y.S., and Jung, S.Y., A Comparative Study on the Accuracy of Stock Price Prediction by Medium by Opinion Mining of News Contents, *2013 Spring Conference of the Korea Intelligent Information System Society*, pp. 133-137.

[3] Hur, Y.M., The correlation between big data and stock index using real-time popular search terms [dissertation] : Kunkook University, 2014.

[4] Jung, H. and Park, M.S., A Study of Big data-based Machine Learning Techniques for Wheel and Bearing Fault Diagnosis, *Journal of the Korea Academia-Industrial cooperation Society*, 2018, Vol. 19, No. 1, pp. 75-84.

[5] Kim, D.Y., Park, J.W., and Choi, J.H., A Comparative Study between Stock Price Prediction Models Using Sentiment Analysis and Machine Learning Based on SNS and News Articles, *Korea Society of IT Services*, 2014, Vol. 13, No. 3, pp. 211-233.

[6] Kim, H.S. and Kim, C.S., An Analysis of IT Proposal Evaluation Results using Big Data-based Opinion Mining, *Journal of Society of Korea Industrial and Systems Engineering*, 2018, Vol. 41, No. 1, pp. 1-10.

[7] Kim, S.W. and Ahn, H.C., Development of an Intelligent Trading System Using Support Vector Machines and Genetic Algorithms, *Journal of Intelligence and Information Systems*, 2010, Vol. 16, No. 1, pp. 71-92.

[8] Lee, Y.J., Prediction of individual Stock Price through News Bigdata Keyword Analysis [dissertation] : Chungbook University, 2014.

[9] Moon, H.R. and Kim, J.W., A Study on Prediction Model of Stock Price using Internet News, *2014 Spring Conference of the Korea Intelligent Information System Society*, pp. 387-393.

[10] Park, H.W. and Lee, Y.O., A Mixed Text Analysis of User Comments on a Portal Site, *Journal of the Korean Data Analysis Society*, 2009, Vol. 11, No. 2, pp. 731-744.

[11] Park, H.W. and Leydesdorff, L., Understanding the KrKwic : A computer program for the analysis of Korean text, *Journal of The Korean Data Analysis Society*, 2004, Vol. 6, No. 5, pp. 1377-1388.

[12] Park, H.W., Humanities and Sociology in the E-Science Age, *Social Science Studies*, 2010, Vol. 30, No. 2, pp. 195-211.

[13] Park, J.Y. and Han, I.K., Predicting Korea Composite Stock Price Index Movement Using Artificial Neural Network, *Journal of Intelligence and Information Systems*, 1995, Vol. 1, No. 2, pp. 103-121.

[14] Song, C.Y., A Study on the Influence of News on Financial Market, *International Economic Journal*, 2002, Vol. 8, No. 3, pp. 1-34.

ORCID

Ji Don Yu | https://orcid.org/0000-0002-7979-5162

Ik Sun Lee | https://orcid.org/0000-0001-9513-6709