인터넷 뉴스 콘텐츠 빅데이터를 활용한 종합주가지수 예측

A Prediction of KOSPI Based on News Big-Data Analysis

저자 (Authors)
유지돈, 이익선
Ji Don Yu, Ik Sun Lee

출처 (Source)
경영과학 35(4), 2018.12, 1-14(14 pages)
KOREAN MANAGEMENT SCIENCE REVIEW 35(4), 2018.12, 1-14(14 pages)

발행처 (Publisher)
한국경영과학회
The Korean Operations Research and Management Science Society

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07590884

APA Style
유지돈, 이익선 (2018). 인터넷 뉴스 콘텐츠 빅데이터를 활용한 종합주가지수 예측. 경영과학, 35(4), 1-14

經營科學
第35卷 第4號
2018年 12月, pp.1-14
https://doi.org/10.7737/KMSR.2018.35.4.001

인터넷 뉴스 콘텐츠 빅데이터를 활용한 종합주가지수 예측

유지돈[1] · 이익선[2†]
[1]광주과학기술원 혁신기업가교육센터, [2]동아대학교 경영학과

A Prediction of KOSPI Based on News Big-Data Analysis

Ji Don Yu[1] · Ik Sun Lee[2†]
[1]Entrepreneurship Education Center, Gwangju Institute of Science and Technology
[2]Department of Business Administration, Dong-A University

■ Abstract ■

Are news on the stock market and stock price index related? Generally, investors acquire many information from a lot of media outlets. However, numerous news articles are produced every day in real time. This research collected, processed and analyzed news articles related to stocks to analyze the relationship with stock price fluctuations. To analyze stock-related non-formal articles, we carried out text mining and multiple regression analysis. Six types of investment decision models was suggested. Specifically, variables in the form of "noun", "verb", and "noun & verb" were utilized to suggest analysis methods.

Keywords : Bigdata, KOSPI, Text Mining, Multiple Regression Analysis

논문접수일 : 2018년 04월 27일 논문게재확정일 : 2018년 11월 20일
논문수정일 : 2018년 10월 04일
* This study was supported by research funds from Dong-A University.
† 교신저자, lis1007@dau.ac.kr

1. 서 론

글로벌 경제 불황속에서도 주식시장에 대한 관심이 증가하고, 유동성 자금이 주식시장으로 모여드는 추세가 지속되고 있다. 이러한 분위기 속에서 한국에서도 해외 직접 투자로 빠져나간 자금을 국내 시장으로 유인하는 계기를 만들고자 2016년 8월부터 거래시간을 30분 연장하는 법안을 통과하여 실시하고 있다. 이처럼 전 세계적으로 주식시장에 대한 기대감이 상승하고 많은 개인 투자자들은 각종 매체를 통해 생성되는 인터넷 주식 뉴스를 정보로 활용하고 있다. 이런 정보를 통해 개인 투자자들은 기업의 가치를 분석하여 현재 주식가격의 동향을 판단하고 기업의 미래 가치를 따져 투자를 하게 된다. 다수의 투자자들이 전문가 집단을 통해 생성되는 뉴스가 종목 주가지수에 어느 정도 연관성이 있을 것으로 기대하고, 기업 가치에 대한 분석에 참조하고 있다. 하지만 방대하게 쏟아지는 각종 뉴스 기사들 중에서 의미 있는 정보만을 추출하기에는 쉽지 않다. 또한 주식시장의 자체가 환율, 금리, 물가 등을 나타내는 외부적 요인과 기업의 자산 가치, 실적, 재무제표 등을 나타내는 내부적 요인에 따라서 주가의 흐름이 달라지기 때문에 기업의 주가를 실제적으로 예측하기는 많은 어려움이 있다.

이코노미스트(Economist)가 약 600개 글로벌 기업을 대상으로 실시한 빅데이터 활용 사례에 관한 조사에서 조사 응답자의 10%는 빅데이터가 기존에 사용되는 비즈니스 모델을 새롭게 바꿀 것이며, 46%는 기업 의사결정에 중요한 요인으로 작용할 것으로 응답했다. 그러나 응답자 중 25%는 기업 내부에 활용 가능한 데이터는 많지만 대부분의 데이터를 활용하지 못하고 있으며, 53%는 극소수의 데이터만 활용하고 있다고 응답했다. 이는 부가가치 창출을 위해서 데이터 활용 및 분석에 더 많은 노력과 투자가 필요함을 시사하고 있다.

텍스트 마이닝은 비정형 텍스트 데이터로 구성된 빅데이터에서 자연어 처리 기술에 기반을 두어 의미가 있는 정보로 가공하는 기술이다. 데이터 마이닝 방식과 비슷하지만 데이터 마이닝의 분석은 XML 문서와 관계형 데이터베이스와 같은 구조화된 데이터이고, 텍스트 마이닝 분석은 이메일, 텍스트 문서, HTML 파일 등과 같은 비정형 텍스트 데이터라는 차이가 있다.

우리 인간의 언어는 문법적, 어휘적 특성이 있으며, 그 표현의 형태가 매우 복잡 · 다양하기 때문에 일괄된 방식으로 규정하기 힘든 경우가 많고, 언어가 사용되는 방식에 따라서 계속 변화하는 특징을 지니고 있다. 세계에 공존하는 언어 중 문자로 표현되는 언어를 컴퓨터로 수집 · 가공 처리하고 구조와 의미를 파악하고자 하는 기술이 바로 자연언어처리 기술이다. 텍스트 마이닝 기술을 통해 대량의 정보 흐름 속에서 의미 있는 정보를 도출해 내고, 다른 정보와의 관련성을 파악하는 등, 단순한 정보를 찾아내는 그 이상의 의미를 찾아낼 수 있다. 따라서 텍스트 마이닝 기술은 컴퓨터가 발전하면서부터 끊임없이 연구되어지는 기술 연구 분야이지만, 언어의 복잡성으로 인해 여전히 많은 연구자들에 의해 진행 · 발전되고 있는 연구 분야이다.

주식 시장의 분석을 통한 예측은 경제 뿐 아니라 전산 및 통계, 수학분야에 이르기까지 모든 분야에 걸쳐 오랫동안 중요 연구 과제로 여겨져 왔다. 현재는 금융공학의 발전과 과학적 방법을 활용한 주가예측 및 그 활용 방안에 대한 연구가 꾸준히 진행되고 있다. 수많은 주식투자 회사들이나 경제학자들이 사용하는 계량적 투자 전략 방식은 수학적 접근 방법에 의하여 수치화 된 미래 가치를 예측하고 이를 통해 포트폴리오를 생성하여 투자에 대한 의사를 결정하는 기법이다. 가장 대표적이라고 할 수 있는 수학적 모델로는 거래 가격의 범위를 제한한 채 매매주문에 대한 가격 변동을 연구하는 여과방법과 시계열 데이터의 움직임을 분석하여 데이터간 연관성과 미래의 변동 예측에 활용하는 웨이블렛 변환 방법 등이 있다[3, 13, 15].

종합주가지수를 예측한 연구들을 소개하면, 김선웅, 안현철[4]은 유전자 알고리즘에 기반한 지능적인 트레이딩 시스템을 제안하였고, 박종엽, 한인구[8]는

인공신경망 알고리즘을 활용하여 종합주가지수를 예측하는 방법론을 제안한 바 있다. 허양민[18]은 실시간으로 생성되는 인기 검색어를 수집하여 블로그, SNS의 텍스트에서 발생되는 감정을 긍・부정으로 분석하여 주가지수의 상승・하락을 예측하는 연구를 수행하였다. 실시간 인기 검색어를 통해 SNS, 웹사이트, 블로그에서 발생되는 관련 텍스트들을 수집하고 분석하여 본문, 꼬리말, 머리말, 날씨 등의 변수와 종합주가지수와의 관련성을 제시하였다.

뉴스 기사를 통해 종합주가지수의 동향을 예측한 연구로서 송치영[12], 안성원 외[13]은 긍정적인 뉴스기사와 부정적인 뉴스기사를 분류하여 이러한 뉴스기사들의 건수와 비중에 따라서 다음 날의 종합주가지수가 얼마나 영향을 받는지를 연구하였다. 천세원[17]은 인터넷 뉴스매체별로 다양하게 주가지수의 상승 또는 하락에 대한 전망을 내어놓는데, 이를 분석하여 어떤 인터넷 뉴스매체의 예측의 더 정확한지를 분석하였다.

주가지수를 예측하기 위한 보다 최근의 연구로서 김민수, 구평회[2], 김동영[1], 문하늘, 김종우[7], 이예지[19], 각 인터넷포털의 뉴스섹션에서 수집한 뉴스데이터를 활용하였는데, 이러한 연구들의 공통점은 감성사전을 구축하여 수집한 뉴스의 긍정/부정에 관한 감성점수를 매긴다는 점이다. 긍정 혹은 부정의 감성점수를 활용하여 개별 주가지수 등락과의 관련성을 규명하는 연구들을 수행하였다.

위의 연구들은 개별기업들의 주가를 예측한 연구들이었지만, 김유신 외[5], 유은지[14], 이예지[16]는 종합주가지수 KOSPI의 등락을 예측하는 다양한 연구를 수행하였다. 이 연구들도 공통적으로 감성사전을 구축하여 이러한 사전에 근거한 감성점수를 인터넷 뉴스별로 스코어링하였다. 이러한 감성점수를 근간으로 하여 긍정 또는 부정의 강함의 정도를 파악하고 이를 종합주가지수와의 상관성을 분석하려고 연구하였다.

본 논문은 인터넷 환경에서 실시간으로 생성되는 수천만 건의 뉴스 기사를 비정형 빅데이터로 간주하고 가공・처리・분석을 실시한다. 인터넷에서 쏟아지는 여러 유형의 비정형 텍스트 데이터는 대량의 정보를 동시다발적으로 생성하여 사람들에게 정보를 전달해 주고 있는데, 본 연구는 인터넷 뉴스 기사에서 추출되어진 단어들을 통해 종합주가지수를 예측하는 연구를 수행한다. 또한, 오피니언 마이닝을 통해 분석을 했던 연구들은 반어법이나 의미가 불분명한 단어 등에도 감정・감성 지수를 주어 감정 표현의 정도를 정확하게 계량화 할 수 있기에는 정확성이 떨어진다는 점이 문제로 지적되었기에 본 연구에서는 뉴스기사 수집은 인터넷 포털사이트 '다음'의 주식 섹션에서 2014년 종합주가지수 ±1% 상승한 20일, 하락한 20일치의 뉴스 데이터 약 224,000건을뉴스 기사들을 수집하였고, 인터넷 포털사이트에서 수집한 데이터는 긍정 및 부정에 의거한 빈도수로 처리하여, 주가지수에 관련이 높아 보이는 단어들만을 선별하였다. 이를 통해 추출되어진 어휘와 2015년 특정 날짜에 검색되어진 단일종목 ±2% 이상 상승한 15일치 뉴스와 하락한 15일치 뉴스 데이터 57,000건을 수집해 추출되어진 어휘를 통해 종합주가지수(변화율)와의 상관관계를 입증하고, 도출되어진 모델들의 비교 테스트를 통해서 가장 우수한 예측모델을 제시하고자 한다.

2. 연구방법

본 연구는 포털사이트 '다음' 뉴스의 주식 섹션에서 2014년 종합주가지수 1% 이상 상승한 20일, 1% 이상 하락한 20일치의 뉴스 데이터 약 224,000건을 수집하였다.

수집된 인터넷 뉴스 기사는 가장 먼저 텍스트의 형태소 분석 과정을 실시한다. 다양하고 방대한 텍스트에서 추출된 키워드들은 다양한 정제화(cleaning) 작업을 통해 주요 어휘들로만 구성하는 총 5단계 과정을 거치게 된다. 어휘와 함께 쓰인 불필요한 기호와 조사에 대한 분리작업을 실시하였고, 종합주가지수에 영향을 주지 않는 어휘들을 정제하였다.

정제화 작업시 복합명사의 경우에는 되도록 분리되지 않도록 하였고, 독립적으로 의미가 없는 어휘

나 문장부호 등은 제거하였다. 또한 유사한 의미를 내포하고 있지만 다르게 표현된 어휘들은 한 개의 대표적 어휘로 통합하였다. 예를 들어, '현대차', '현대차그룹', '현차' 등은 '현대자동차'로 '미', '美', '美國' 등은 미국으로 통일하였다(<표 1> 참조).

〈표 1〉 정제 작업을 통해 제거된 기호와 조사의 일부

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

두 번째 단계에서는 한국어 형태소 분석에 활용되는 소프트웨어 "Krkwic"[9, 10, 11]를 이용하여 빈도분석을 실시하였다. 빈도분석 결과를 바탕으로 본 연구에서는 회귀분석에 사용할 상위 빈도의 어휘들을 추출하였고, 형태소로 분리된 어휘들은 긍정 또는 부정의 의견을 판단할 수 있는 어휘들을 추출해내는 과정을 거쳤다. 이는 모든 어휘들에 대해서 똑같은 과정을 통해 명사와 동사의 두 가지 품사로 분류작업을 수행하였다.

본 연구에서는 종합주가지수에 관련된 뉴스들을 수집하여 기초분석을 통해 정제된 어휘들을 변수들로 사용하기 위해서 핵심 개념을 나타내는 '명사' 변수 각 12개(상승, 이익, 매수(순매수) 증가, 강세, 개선, 하락, 손실, 부진, 매도(순매도) 감소, 적자)와 내용을 서술해주는 '동사' 변수 각 10개(반등했다, 유지했다, 오른, 기대된, 회복했다, 매각했다, 우려된, 내린, 급락했다, 둔화되다)의 어휘들로 분류작업을 실시하였다.

분류작업을 거쳐 수집된 어휘들은 종합주가지수에 영향을 미치는지를 알아보기 위해 입력 다중회귀분석과 상대적으로 설명력이 높은 변수들을 알아볼 수 있는 단계적 다중회귀분석을 SPSS 22.0 통계 패키지를 활용하여 실험을 수행하였다.

어휘 선정과 관련해서 품사들 전체를 대상으로 독립변수 데이터를 선정한 것이 아니기 때문에 뉴스기사의 전체를 대변하지는 않는다. 하지만 뉴스기사의 핵심을 파악할 수 있도록 분석 대상의 전체적 구조를 명확하게 보여주고 핵심개념을 나타내는 '명사' 품사와 문장의 주체가 되고 단어 간의 관계성을 나타내는 '동사' 품사로 나누어 정리하였다. 따라서 독립변수 데이터는 상승 · 하락한 20일치 뉴스기사에서 주가지수에 영향을 줄 수 있고, 가급적 많이 언급되는 명사, 동사 어휘를 기준으로 선정하여 도출하였다.

3. 연구분석

종합주가지수 분석은 '다음' 포털사이트 주식섹션 뉴스 기사에서 상승 · 하락한 날의 각 20일치 데이터를 수집하였고, 전처리 과정을 통해 텍스트를 정제 · 분석하였다. 정제되어진 어휘들은 종합주가지수와 관련성이 높고, 빈도출현이 많은 어휘를 중심으로 '명사' 변수 12개와 '동사' 변수 10개로 정리하였다. 추출된 변수들과 종합주가지수의 상관분석에 대한 진단결과는 아래의 내용과 같다.

3.1 종합주가지수에 영향을 미치는 요인; 명사

추출되어진 12개의 명사 변수와 종합주가지수와의 상관관계가 <표 2>에 제시되어 있다. 상승($r = .416$, $p < .10$), 이익($r = .267$, $p < .10$), 매수($r = .418$, $p < .10$), 증가($r = .395$, $p < .10$)의 변수들은 정(+)의 상관이 나타났고, 하락($r = -.330$, $p < .10$), 감소($r = -.396$, $p < .10$), 적자($r = -.457$, $p < .10$) 변수들은 부(-)의 상관이 있는 것으로 나타났다.

<표 3>은 입력 방식의 다중회귀분석을 통해 결과를 도출한 통계자료이다. 분석결과에서 수정된 결정계수는 .365이며, 이때 F값의 유의확률은 .011로 추

〈표 2〉 '명사' 변수 상관관계 분석

구 분	상관분석	
	R	p-value
상승	.416	.004
이익	.267	.048
매수	.418	.004
증가	.395	.006
강세	.158	.165
개선	.065	.345
하락	-.330	.007
손실	-.153	.174
부진	-.117	.235
매도	-.018	.456
감소	-.396	.006
적자	-.457	.098

정된 회귀식이 통계적으로 유의미한 것으로 나타났다. 독립변수들이 통계적으로 유의미한 변수인지를 판단하는 t값의 유의확률을 살펴보면 '증가', '감소', '적자'의 변수만 .10 이하의 좋은 값을 보이고 있다. 이는 결과에서 보여주듯이 12개의 '명사' 변수들 중에서 통계적으로 '증가', '감소', '적자' 3개의 변수만이 종합주가지수와 상관관계가 있음을 나타내고 있다. 입력 다중회귀분석을 통해 도출되어진 '명사' 변수 다중회귀모형 모델 ①은 아래와 같다.

다중 회귀모형은 독립변수들 간의 단위가 서로 다를 때 상대적 크기만을 알려주는 표준화계수 베타값으로도 회귀식을 도출할 수 있다. 하지만 본 연구에서는 독립변수들의 단위가 모두 동일한 관계로 절대적 영향력의 크기를 상호 비교해보고자 비표준화 계수의 β값을 사용하여 회귀식을 도출하였다.

※ 모델 ①

$$Y(\text{종합주가지수}) = -4.550+0.726X_1+0.041X_2+2.194X_3+0.962X_4+0.331X_5-0.457X_6+0.786X_7+0.246X_8-0.073X_9-0.306X_{10}-1.149X_{11}-0.733X_{12}$$

〈표 3〉 '명사' 변수 입력 다중회귀분석 결과

구 분	R	R^2	수정된 R^2	추정값의 표준오차
모형요약	.749	.561	.365	1.02023

구 분		제곱합	자유도	평균 제곱	F	유의확률
분산분석	선형회귀분석	35.852	12	2.988	2.870	.011
	잔차	28.103	27	1.041		
	합계	63.955	39			

구 분		비표준화 계수		t	유의 확률
		B	표준오차		
계수 분석	(상수)	-4.550	2.913	-1.562	.130
	상승	.726	.453	1.601	.121
	이익	.041	.316	.129	.898
	매수	2.194	1.380	1.590	.123
	증가	.962	.381	2.527	.018
	강세	.331	.331	.999	.327
	개선	-.457	.491	-.931	.360
	하락	.786	.934	.842	.407
	손실	.246	.413	.595	.557
	부진	-.073	.345	-.213	.833
	매도	-.306	1.006	-.304	.763
	감소	-1.149	.485	-2.370	.025
	적자	-.733	.398	-1.842	.077

한편 명사 변수인 '상승', '이익', '매수' '증가', '강세', '개선', '하락', '손실', '부진', '매도', '감소', '적자' 12개의 변수 중 어느 변수들이 종합주가지수를 예측하는데 상대적으로 설명력이 높은지 알아보기 위해서 단계적 다중회귀분석(stepwise regression analysis)을 실시했다. 단계적 다중회귀분석은 보통 많은 변수들이 존재할 때 쓰는 방식으로 전진과 후진방법을 결합한 방식이다. 독립변수의 기여도를 평가한 후 주요하게 영향력이 높은 변수를 먼저 투입한 후 단계별로 변수를 제거하고 가장 적합도가 큰 변수들을 찾을 수 있다.

'명사' 변수가 종합주가지수에 끼치는 영향력을 분석하기 위해서 단계적 다중회귀분석을 실시하였다. 아래의 분석 결과를 살펴보면 단계적 다중회귀분석을 통해 도출되어진 수정된 결정계수는 .422로서 종합주가지수 분산의 42.2%가 독립변수에 의해서 설명됨을 알 수 있었다. 즉, 총 5개의 '매수', '상승', '증가', '적자', '감소' 독립변수가 종합주가지수를 42.2%만큼 설명하고 있는 것으로 나타났다.

R^2의 유의도 검증을 실시한 결과 F 값은 6.689이고, 유의확률은 .000으로 나타나 유의수준 $p < .10$에서 5개의 변수에 의한 다중회귀분석은 통계적으로 유의한 것으로 나타났으며, 종합주가지수와 '매수', '상승', '증가', '적자', '감소' 간에 전반적으로 유의미한 비례관계가 있음을 알 수 있다.

종합주가지수에 영향을 주는 '명사' 변수에 대한 직접적 효과의 유의성 검증을 실시하였는데, 단계적 다중회귀분석에서 종합주가지수와 명사 변수간의 상관관계가 있는 변인으로 '매수', '상승', '증가', '적자', '감소' 변수가 통계적으로 $p < .10$에서 유의한 것으로 나타났다. 5개의 변수 중 종합주가지수가 상승에 가장 많은 영향을 미치는 변수는 '증가' 변수(.309)로 나타났고 하락에 가장 많은 영향을 미치는 변수는 '적자'(-.268)로 나타났다. 단계적 다중회귀분석을 통해 도출되어진 '명사' 변수 모델 ②는 아래와 같다.

※ 모델 ②

$$Y(\text{종합주가지수}) = -4.282+2.278X_1-0.635X_2+0.820X_3-0.599X_4-0.759X_5$$

3.2 종합주가지수에 영향을 미치는 요인; 동사

추출되어진 10개의 동사 변수와 종합주가지수와의 상관관계를 <표 4>에서 살펴보면, 반등했다($r = .312$, $p < .10$), 내린($r = .349$, $p < .10$), 둔화되다($r = .395$, $p < .10$)의 변수들은 정(+)의 상관이 나타났고, 우려된($r = -.503$, $p < .10$) 변수만 부(-)의 상관이 있는 것으로 나타났다.

〈표 4〉 '동사' 변수 상관관계 분석

구 분	상관분석	
	R	p-value
반등했다	.312	.025
유지했다	-.076	.320
오른	.026	.437
기대되다	.155	.170
회복했다	-.131	.211
매각했다	.069	.335
내린	.349	.014
우려된	-.503	.000
급락했다	-.050	.380
둔화되다	-.236	.071

독립변수들이 통계적으로 유의미한 변수인지를 판단하는 t값의 유의확률을 살펴보면 '반등했다', '기대된다', '회복했다', '내린', '우려된', '둔화되다'의 변수가 .10 이하의 좋은 값을 보이고 있다. 이는 결과에서 보여주듯이 10개의 '동사' 변수들 중에서 통계적으로 '반등했다', '기대된다', '회복했다', '내린', '우려된', '둔화되다' 6개의 변수가 종합주가지수에 상관관계가 있는 것으로 나타나고 있다. 입력 다중회귀분석을 통해 도출되어진 '동사' 변수 모델 ③은 아래와 같다.

※ 모델 ③

$$Y(\text{종합주가지수}) = -0.258+0.732X_1-0.477X_2-0.029X_3+0.822X_4-0.585X_5-0.161X_6+0.872X_7-1.197X_8+0.370X_9-1.330X_{10}$$

'명사' 변수와 동일하게 '동사' 변수들이 종합주가지수를 예측하는데 상대적으로 설명력이 높은지 알아보기 위해서 총 10개(반등했다, 유지했다, 오른, 기대된다, 회복했다, 매각했다, 우려된, 내린, 급락했다, 둔화되다)의 독립변수에 대한 단계적 다중회귀분석(stepwise regression analysis)을 실시했다. <표 5>의 분석 결과를 살펴보면 단계적 다중회귀분석을 통해 도출되어진 수정된 결정계수는 .644로서 종합주가지수 분산의 64.4%가 독립변수에 의해서 설명됨을 알 수 있었다. 즉, 총 7개의 '반등했다', '기대된다', '회복했다', '우려된', '내린', '급락했다', '둔화되다' 독립변수가 종합주가지수를 64.4%를 설명하고 있는 것으로 나타났다.

<표 5>에서 R^2의 유의도 검증을 실시한 결과 F값은 11.095이고, 유의확률은 .000으로 나타나 유의수준 $p < .001$에서 7개의 변수에 의한 다중회귀분석은 통계적으로 유의한 것으로 나타났으며, 종합주가지수와 '반등했다', '기대된다', '회복했다', '우려된', '내린', '급락했다', '둔화되다' 간에 전반적으로 유의미한 선형관계가 있음을 알 수 있다.

단계적 다중회귀분석에서 종합주가지수와 동사 변수간의 상관관계가 있는 변인으로 '반등했다', '기대된다', '회복했다', '우려된', '내린', '급락했다', '둔화되다' 변수가 통계적으로 $p < .10$에서 유의한 것으로 나타났다. 7개의 변수중 종합주가지수가 상승에 가장 많은 영향을 미치는 변수는 '기대된다' 변수(.434)로 나타났고 하락에 가장 많은 영향을 미치는 변수는 '우려된'(-.622)로 나타났다. 위의 도출된 결과를 수식으로 표현하면 다음과 같다. 단계적 다중회귀분석을 통해 도출되어진 '동사' 변수 모델 ④는 아래와 같다.

※ 모델 ④

$$Y(\text{종합주가지수}) = -0.331 - 1.259X_1 + 0.876X_2 + 0.828X_3 - 1.342X_4 + 0.541X_5 - 0.621X_6 + 0.455X_7$$

〈표 5〉 '동사' 변수 입력 다중회귀분석 결과

구 분	R	R^2	수정된 R^2	추정값의 표준오차
모형요약	.853	.728	.635	.77400

구 분		제곱합	자유도	평균 제곱	F	유의확률
분산분석	선형회귀분석	46.582	10	4.658	7.776	.000
	잔차	17.373	29	.599		
	합계	63.955	39			

구 분		비표준화 계수		t	유의확률
		B	표준오차		
계수분석	(상수)	-.258	.509	-.507	.616
	반등했다	.732	.290	2.524	.017
	유지했다	-.477	.342	-1.395	.174
	오른	-.029	.197	-.148	.883
	기대된다	.822	.223	3.689	.001
	회복했다	-.585	.243	-2.406	.023
	매각하다	-.161	.249	-.648	.522
	내린	.872	.233	3.750	.001
	우려된	-1.197	.261	-4.591	.000
	급락했다	.370	.257	1.440	.161
	둔화되다	-1.330	.344	-3.862	.001

3.3 예측 정확도 비교 분석

2016년 종합주가지수의 실제값과 입력 방식의 다중회귀분석을 통해 도출되어진 '명사' 변수 예측값의 정확도 테스트 결과는 <표 6>과 같으며, 실제값과 예측값의 오차를 구하는 공식은 다음과 같이 구할 수 있다.

$$Gap(\%) = |실제값 - 예측값|$$

평균 Gap(%)은 각 실제값과 예측값 오차의 평균값을 의미한다.

2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도 테스트 결과 최소 오차값과 최대 오차값은 각 0.10%와 0.94%였으며, 전체 오차 평균값은 0.72%로 나타났다. 이에 반해 하락한 날의 최소 오차값과 최대 오차값은 각 3.06%와 4.52%, 전체 오차 평균값은 3.60%로 상승일에 비해 다소 부정확한 정확도를 보였다.

<표 7>은 단계적 다중회귀분석을 통해 도출되어진 '명사' 변수 예측값과 2016년 종합주가지수의 실제값과의 정확도 테스트 결과이다. 2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도를 비교한 결과 최소 오차값과 최대 오차값은 각 0.33%와 0.69%였으며, 전체 평균값은 0.50%로 모델 ①의 정확도 결과값보다 최소, 최대 오차값의 범위와 전체 오차 평균값이 더 좋은 결과값을 보였다.

〈표 6〉 모델 ① 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	2.15%	0.81%	1	01.20	-2.34%	1.96%	4.30%
2	01.22	2.11%	2.01%	0.10%	2	01.26	-1.15%	1.99%	3.14%
3	01.27	1.40%	2.18%	0.78%	3	02.11	-2.93%	1.59%	4.52%
4	02.04	1.35%	2.22%	0.87%	4	02.12	-1.41%	1.97%	3.38%
5	02.15	1.47%	2.24%	0.77%	5	04.01	-1.12%	1.74%	2.86%
6	02.16	1.40%	2.34%	0.94%	6	06.13	-1.91%	1.82%	3.73%
7	02.18	1.32%	2.25%	0.93%	7	07.06	-1.85%	2.02%	3.87%
8	03.02	1.60%	2.25%	0.65%	8	08.03	-1.20%	1.87%	3.07%
9	04.14	1.75%	2.35%	0.60%	9	09.09	-1.25%	1.81%	3.06%
10	05.25	1.18%	2.09%	0.91%	10	09.12	-2.28%	1.77%	4.05%
Average		1.49%	2.21%	0.72%	Average		-1.74%	1.86%	3.60%

〈표 7〉 모델 ② 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	2.00%	0.66%	1	01.20	-2.34%	1.23%	3.57%
2	01.22	2.11%	1.78%	0.33%	2	01.26	-1.15%	0.53%	1.68%
3	01.27	1.40%	2.04%	0.64%	3	02.11	-2.93%	-1.34%	1.59%
4	02.04	1.35%	2.02%	0.67%	4	02.12	-1.41%	0.94%	2.35%
5	02.15	1.47%	2.09%	0.62%	5	04.01	-1.12%	-0.21%	0.91%
6	02.16	1.40%	2.05%	0.65%	6	06.13	-1.91%	-0.01%	1.90%
7	02.18	1.32%	2.01%	0.69%	7	07.06	-1.85%	-1.09%	0.76%
8	03.02	1.60%	1.96%	0.36%	8	08.03	-1.20%	-0.15%	1.05%
9	04.14	1.75%	2.12%	0.37%	9	09.09	-1.25%	-0.96%	0.29%
10	05.25	1.18%	1.82%	0.64%	10	09.12	-2.28%	-0.48%	1.80%
Average		1.49%	1.99%	0.50%	Average		-1.74%	-0.15%	1.59%

〈표 8〉 모델 ③ 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	0.94%	0.40%	1	01.20	-2.34%	0.45%	2.79%
2	01.22	2.11%	1.26%	0.85%	2	01.26	-1.15%	-0.63%	0.52%
3	01.27	1.40%	1.40%	0.00%	3	02.11	-2.93%	0.21%	3.14%
4	02.04	1.35%	1.50%	0.15%	4	02.12	-1.41%	-0.37%	1.04%
5	02.15	1.47%	0.81%	0.66%	5	04.01	-1.12%	-1.49%	0.37%
6	02.16	1.40%	1.68%	0.28%	6	06.13	-1.91%	-0.97%	0.94%
7	02.18	1.32%	-0.01%	1.33%	7	07.06	-1.85%	1.13%	2.98%
8	03.02	1.60%	1.24%	0.36%	8	08.03	-1.20%	-1.41%	0.21%
9	04.14	1.75%	1.59%	0.16%	9	09.09	-1.25%	-1.18%	0.07%
10	05.25	1.18%	1.17%	0.01%	10	09.12	-2.28%	1.33%	3.61%
Average		1.49%	1.16%	0.33%	Average		-1.74%	-0.29%	1.45%

2016년 하락한 날의 최소 오차값과 최대 오차값은 각 0.29%와 3.57%, 전체 오차 평균은 1.59%로 최소, 최대 오차값의 범위가 크게 나타남을 확인할 수 있지만 예측값이 실제값에 비해 크게 벗어나는 특정한 날을 제외한 개별적 오차값을 확인해보면 모델 ①의 성능보다 뛰어남을 보였다.

2016년 종합주가지수의 실제값과 입력 다중회귀분석을 통해 도출되어진 '동사' 변수 예측값의 정확도 테스트 결과는 <표 8>과 같다. 2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도를 비교한 결과 최소 오차값과 최대 오차값은 각 0%와 1.33%였으며, 전체 평균값은 0.33%로 정확도 테스트에서 좋은 결과값을 보였다.

2016년 하락한 날의 최소 오차값과 최대 오차값은 각 0.07%와 3.61%, 전체 오차 평균은 1.45%로 최소, 최대 오차값의 범위가 크게 나타남을 확인할 수 있다. 하지만 예측값이 실제값에 비해 크게 벗어나는 특정한 날을 제외한 개별적 데이터 오차값을 확인해보면 더 좋은 성능을 보여주고 있다.

<표 9>는 단계적 다중회귀분석을 통해 도출되어진 '동사' 변수 예측값과 2016년 종합주가지수의 실제값 정확도 테스트 결과이다. 2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도를 비교 분석한 결과 최소 오차값과 최대 오차값은 각 0.12%와 0.77%였으며, 전체 평균값은 0.09%으로 모델 ③보다 모든 결과에서 뛰어난 값을 보였다.

〈표 9〉 모델 ④ 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	0.57%	0.77%	1	01.20	-2.34%	-0.09%	2.25%
2	01.22	2.11%	1.62%	0.49%	2	01.26	-1.15%	-0.94%	0.21%
3	01.27	1.40%	1.73%	0.33%	3	02.11	-2.93%	-1.07%	1.86%
4	02.04	1.35%	1.83%	0.48%	4	02.12	-1.41%	-1.14%	0.27%
5	02.15	1.47%	1.88%	0.41%	5	04.01	-1.12%	0.58%	1.70%
6	02.16	1.40%	1.75%	0.35%	6	06.13	-1.91%	-0.80%	1.11%
7	02.18	1.32%	1.55%	0.23%	7	07.06	-1.85%	-1.42%	0.43%
8	03.02	1.60%	1.74%	0.14%	8	08.03	-1.20%	-1.03%	0.17%
9	04.14	1.75%	1.63%	0.12%	9	09.09	-1.25%	-0.13%	1.12%
10	05.25	1.18%	1.51%	0.33%	10	09.12	-2.28%	1.16%	3.44%
Average		1.49%	1.58%	0.09%	Average		-1.74%	-0.49%	1.25%

2016년 하락한 날의 최소 오차값과 최대 오차값은 0.17%와 3.34%, 전체 오차 평균은 1.25%로 최소, 최대 오차값의 범위가 크게 나타남을 확인할 수 있다. 하지만 예측값과 실제값이 크게 벗어나는 특정한 날을 제외 하면 모델 ③보다 좋은 결과를 보여주고 있음을 확인할 수 있다.

품사들로 구분지은 어휘 '명사' 12개의 변수와 '동사' 10개의 변수들을 통합하여 다중회귀분석을 실시하였다. <표 10>은 다중회귀분석 결과를 나타낸 것이다.

수정된 결정계수는 .694이며, 이때 F값의 유의확률은 .001로 제안된 회귀모델이 통계적으로 유의한 것으로 나타났다. 독립변수들이 통계적으로 유의미한 변수인지를 판단하는 t값의 유의확률을 살펴보면 '오른', '기대된다', '회복했다', '내린', '우려된', '급락했다', '둔화되다', '증가', '하락' 변수가 .10 이하의 좋은 값을 보이고 있다. 이는 결과에서 보여주듯이 22개의 '명사+동사' 변수들 중에서 통계적으로 9개의 '명사+동사' 변수가 종합주가지수에 영향을 미치는 것으로 설명됨을 나타내고 있다. 입력 방식의 다중회귀분석을 통해 도출되어진 '명사+동사' 변수 모델 ⑤는 아래와 같다.

〈표 10〉 '명사+동사' 변수 입력 다중회귀분석 결과

구 분	R	R^2	수정된 R^2	추정값의 표준오차
모형요약	.931	.867	.694	.70802

구 분		제곱합	자유도	평균 제곱	F	유의확률
분산분석	선형회귀분석	55.433	22	2.520	5.026	.001
	잔차	8.522	17	.501		
	합계	63.955	39			

구 분		비표준화 계수		t	유의확률
		B	표준오차		
계수 분석	(상수)	-4.293	2.251	-1.907	.074
	반등했다	.438	.338	1.297	.212
	유지했다	-.486	.393	-1.238	.232
	오른	.561	.254	2.211	.041
	기대된다	.789	.286	2.755	.014
	회복했다	-.591	.244	-2.418	.027
	매각하다	.335	.304	1.102	.286
	내린	.596	.251	2.379	.029
	우려된	-1.249	.372	-3.354	.004
	급락했다	.556	.280	1.988	.063
	둔화되다	-.978	.496	-1.971	.065
	상승	-.009	.443	-.021	.983
	이익	.355	.249	1.427	.172
	매수	.092	1.284	.072	.944
	증가	1.068	.359	2.978	.008
	강세	.324	.255	1.269	.222
	개선	-.251	.445	-.564	.580
	하락	2.095	.785	2.669	.016
	손실	-.481	.360	-1.335	.200
	부진	.140	.302	.465	.648
	매도	-.412	.732	-.563	.581
	감소	-.482	.431	-1.119	.279
	적자	.117	.346	.337	.740

※ 모델 ⑤

$$Y(\text{종합주가지수}) = -0.486+0.438X_1-0.486X_2+0.561X_3+0.789X_4-0.591X_5+0.335X_6+0.596X_7-1.249X_8+0.556X_9-0.978X_{10}-0.009X_{11}+0.355X_{12}+0.092X_{13}+1.068X_{14}+0.324X_{15}-0.251X_{16}+2.095X_{17}-0.481X_{18}+0.14X_{19}-0.412X_{20}-0.482X_{21}+0.117X_{22}$$

'명사+동사' 변수들이 종합주가지수를 예측하는데 상대적으로 설명력이 높은지 알아보기 위해서 단계적 다중회귀분석(stepwise regression analysis)을 실시했다. <표 11>을 살펴보면 단계적 다중회귀분석을 통해 도출되어진 수정된 결정계수는 .701로서 종합주가지수 분산의 70.1%가 독립변수에 의해서 설명됨을 알 수 있었다. 즉, 총 8개의 '우려된', '회복했다', '증가', '내린', '기대된다', '급락했다', '둔화되다' '반등했다' 독립변수가 종합주가지수를 70.1%를 설명하고 있는 것으로 나타났다.

〈표 11〉 단계적 다중회귀분석 '명사+동사' 변수 수정된 결정계수

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.500	.250	.230	1.12132
2	.610	.370	.330	1.04623
3	.680	.460	.420	.96978
4	.730	.530	.480	.87962
5	.780	.600	.540	.84250
6	.812	.659	.597	.79253
7	.840	.706	.642	.76367
8	.834	.696	.640	.73976
9	.859	.739	.681	.69234
10	.873	.762	.701	.85256

예측값 : (상수), 우려된
예측값 : (상수), 우려된, 매수
예측값 : (상수), 우려된, 매수, 회복했다
예측값 : (상수), 우려된, 매수, 회복했다, 증가
예측값 : (상수), 우려된, 매수, 회복했다, 증가, 내린
예측값 : (상수), 우려된, 매수, 회복했다, 증가, 내린, 기대된다
예측값 : (상수), 우려된, 매수, 회복했다, 증가, 내린, 기대된다, 급락했다
예측값 : (상수), 우려된, 회복했다, 증가, 내린, 기대된다, 급락했다
예측값 : (상수), 우려된, 회복했다, 증가, 내린, 기대된다, 급락했다, 둔화되다
예측값 : (상수), 우려된, 회복했다, 증가, 내린, 기대된다, 급락했다, 둔화되다, 반등했다

R^2의 유의도 검증을 실시한 결과에서 F 값은 12.44이고 유의확률은 .000으로 나타나 유의수준 $p < .000$에서 8개의 변수에 의한 다중회귀분석은 통계적으로 유의한 것으로 나타났다. 종합주가지수와 '우려된', '회복했다', '증가', '내린', '기대된다', '급락했다', '둔화되다' '반등했다' 간에 전반적으로 유의미한 비례관계가 있음을 알 수 있다.

단계적 다중회귀분석에서 '명사+동사' 변수와 종합주가지수와의 상관관계가 있는 변인으로 '우려된', '회복했다', '증가', '내린', '기대된다', '급락했다', '둔화되다', '반등했다' 변수가 통계적으로 $p < .10$에서 유의한 것으로 나타났다. 8개의 변수중 종합주가지수가 상승에 가장 많은 영향을 미치는 변수는 '기대된다' 변수(.425)로 나타났고 하락에 가장 많은 영향을 미치는 변수는 '우려된'(-.640)로 나타났다. 단계적 다중회귀분석을 통해 도출되어진 '명사+동사' 변수 모델 ⑥은 아래와 같다.

※ 모델 ⑥

$$Y(\text{종합주가지수}) = -1.102-1.285X_1-0.605X_2+0.748X_3+0.779X_4+0.857X_5+0.635X_6-0.947X_7+0.395X_8$$

<표 12>는 입력 다중회귀분석을 통해 도출되어진 '명사+동사' 변수 예측값과 2016년 종합주가지수의 실제값의 정확도 테스트이다. 2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도를 비교한 결과 최소 오차값과 최대 오차값은 각 0.03%와 0.97%였으며, 전체 평균값은 0.68%로 최소, 최대 오차값 범위와 전체 오차 평균값 모두 좋은 결과값을 보였다.

2016년 하락한 날의 최소 오차값과 최대 오차값은 각 0.02%와 4.57%, 전체 오차 평균은 2.71%로 최소,

〈표 12〉 모델 ⑤ 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	1.91%	0.57%	1	01.20	-2.34%	2.23%	4.57%
2	01.22	2.11%	2.08%	0.03%	2	01.26	-1.15%	2.27%	3.42%
3	01.27	1.40%	2.25%	0.85%	3	02.11	-2.93%	-1.90%	1.03%
4	02.04	1.35%	2.32%	0.97%	4	02.12	-1.41%	-1.96%	0.55%
5	02.15	1.47%	2.27%	0.80%	5	04.01	-1.12%	2.19%	3.31%
6	02.16	1.40%	2.28%	0.88%	6	06.13	-1.91%	2.20%	4.11%
7	02.18	1.32%	2.19%	0.87%	7	07.06	-1.85%	2.44%	4.29%
8	03.02	1.60%	2.19%	0.59%	8	08.03	-1.20%	2.30%	3.50%
9	04.14	1.75%	2.20%	0.45%	9	09.09	-1.25%	2.18%	3.43%
10	05.25	1.18%	2.06%	0.88%	10	09.12	-2.28%	-2.26%	0.02%
Average		1.49%	2.17%	0.68%	Average		-1.74%	0.97%	2.71%

최대 오차값의 범위가 비교적 상승한 날에 비해 크게 나타남을 확인할 수 있다. 뿐만 아니라 위에서 제시한 모델 ①~모델 ④의 개별적 평균 오차값을 비교하더라도 결과에서 좋지 않은 값들을 보이고 있다.

<표 13>은 단계적 다중회귀분석을 통해 도출되어진 '명사+동사' 변수 예측값과 2016년 종합주가지수의 실제값 정확도 테스트이다. 2016년 상승한 날의 실제 상승값과 예측 상승값의 정확도를 비교한 결과 최소 오차값과 최대 오차값은 각 0.30%와 1.45%였으며, 전체 평균값은 0.12%로 모델 ⑤보다 뛰어난 결과값을 보였다.

2016년 하락한 날의 최소 오차값과 최대 오차값은 2.88%와 0.74%, 전체 오차 평균은 1.82%로 최소, 최대 오차값의 범위가 크게 나타남을 확인할 수 있다. 하지만 예측값과 실제값의 오차값의 범위가 크게 벗어나는 특정한 날을 제외한 개별적 데이터 오차값을 확인해보면 모델 ⑤보다 뛰어난 결과값을 나타내고 있음을 확인할 수 있다.

언어 분석의 핵심은 그 언어가 명사를 중심으로 하는가, 동사를 중심으로 하는가의 차이라고 말할 수 있는데 명사는 사물, 개체의 대상화를 중요시하는 반면에, 동사는 개체들 간의 관계를 중요시한다.

〈표 13〉 모델 ⑥ 예측 정확도 비교

구 분	2016년 상승일	실제 상승값	예측 상승값	gap	구 분	2016년 하락일	실제 하락값	예측 하락값	gap
1	01.13	1.34%	0.19%	1.15%	1	01.20	-2.34%	-0.25%	2.08%
2	01.22	2.11%	1.35%	0.76%	2	01.26	-1.15%	-0.02%	1.12%
3	01.27	1.40%	1.65%	0.26%	3	02.11	-2.93%	-0.19%	2.74%
4	02.04	1.35%	1.23%	0.12%	4	02.12	-1.41%	0.29%	1.70%
5	02.15	1.47%	1.63%	0.16%	5	04.01	-1.12%	0.70%	1.82%
6	02.16	1.40%	1.83%	0.43%	6	06.13	-1.91%	0.41%	2.32%
7	02.18	1.32%	0.61%	0.71%	7	07.06	-1.85%	-0.72%	1.13%
8	03.02	1.6%	1.31%	0.29%	8	08.03	-1.20%	0.43%	1.63%
9	04.14	1.75%	0.40%	1.35%	9	09.09	-1.25%	1.63%	2.88%
10	05.25	1.18%	0.48%	0.69%	10	09.12	-2.28%	-1.54%	0.74%
Average		1.49%	1.07%	0.42%	Average		-1.74%	0.07%	1.82%

따라서, 문장의 핵심개념을 나타내는 명사보다는 문장 속에 담겨있는 대부분의 관계를 표현하고, 핵심내용을 서술해주는 동사가 문장의 의미를 파악하는데 더 중요한 역할은 한다. 따라서 변수의 영향도가 큰 순서대로 나열하면 '동사' 모델 > '명사' 모델 ≥ '동사+명사모델'이 된다.

본 연구의 모델별 성능 원인을 규명해보자면 아래와 같다.

ex) ① 손실(부정)+증가했다(긍정) = 부정
② 손실(부정)+감소했다(부정) = 긍정
③ 상승(긍정)+기대된다(긍정) = 긍정
④ 상승(긍정)+둔화된다(부정) = 부정

위의 예시와 같이 명사의 의미가 긍정, 부정적 의미를 가진다 하더라도 결국 따라오는 동사의 관계성으로 인해 긍정, 부정의 의미가 달라질 수 있다. 따라서 명사만을 가지고 의미를 파악하는 것은 한계가 있으며 동사보다는 긍정, 부정에 대한 영향도가 더 작을 수 있다.

4. 연구결과

본 연구에서는 비정형 텍스트 데이터인 뉴스 기사가 주가의 상승 또는 하락에 어떠한 영향을 미친다는 가정에서 출발하여 종합주가지수의 상승・하락에 대한 변동성을 예측하고자 하였다. 뉴스 기사를 분석하기 위해 글쓴이의 생각이나 의견, 감정 등으로 이루어진 비정형 텍스트 뉴스 기사를 수집・가공하였고, 실제 상승된 주가와 모델들을 통해 예측 되어진 주가와의 정확도를 비교하는 실험을 실시하였다.

종합주가지수가 상승한 날의 6가지 모델 중에서는 모델 ③의 오차값이 가장 낮게 나타났으며, '동사' 변수의 영향력이 상대적으로 '명사' 변수보다 더 큰 것으로 나타났다. 하락한 날의 6가지 모델을 비교한 결과에서는 모델 ③과 모델 ④의 오차값이 가장 낮은 것으로 나타났다. 상승한 날의 모델보다 다소 예측 정확도가 떨어지는 것을 보여주고 있지만 하락세에 대한 결과를 비교적으로 정확하게 예측함을 확인하였다.

본 연구의 결과를 통해 일정기간 동안 꾸준히 생성되는 주식관련 뉴스 기사의 긍정・부정 어휘 발생 빈도는 주가 변동성에 영향을 충분이 미치는 것으로 나타났다. 따라서, 정확한 주가의 등락률을 예측하기에는 다소 한계가 있지만 주가의 등락의 방향성을 비교적 정확하게 예측해 주고 있다는 점에서는 의미가 있다.

본 연구에서 도출되어진 모델은 실제 인공지능형 투자결정시스템으로 보완・구현하고 실물투자에 적용해본다면 본 연구의 결과보다 뛰어난 성능의 결과물을 산출할 수 있을 것으로 예상한다. 또한 종합주가지수의 빠르고 정확한 대표성을 확보하기 위해 분석에 사용되는 툴이나 프로그램의 완전 자동화가 필요하며 추후 연구에서는 이러한 실험 환경을 통해 주가의 동향에 대한 예측의 논리적 근거를 보완한다면 한층 더 진보된 투자결정시스템으로 발전할 수 있을 것이다.

참 고 문 헌

[1] 김동영, 박제원, 최재현, "SNS와 뉴스기사의 감성분석과 기계학습을 이용한 주가예측 모형에 관한 연구", 「한국IT서비스학회지」, 제13권, 제3호(2014), pp.211-233.
[2] 김민수, 구평회, "인터넷 검색추세를 활용한 빅데이터 기반의 주식투자전략에 대한 연구", 「한국경영과학회지」, 제38권, 제4호(2013), pp.53-63.
[3] 김범수, "빅데이터를 통한 대형할인매장 촉진활동 전략 분석 : 베이지언 네트워크기법 응용을 중심으로", 「경영과학」, 제34권, 제2호(2017), pp.15-33.
[4] 김선웅, 안현철, "Support Vector Machines와 유전자 알고리즘을 이용한 지능형 트레이딩 시스템 개발", 「지능정보연구」, 제16권, 제1호(2010), pp. 71-92.
[5] 김유신, 김남규, 정승렬, "뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자의사 결정모형", 「지능정보연구」, 제18권, 제2호(2012), pp.143-156.

[6] 김태환, 이상용, “기업의 SNS 노출이 주가에 영향을 미치는가 : 한국의 트위터와 블로그를 중심으로”, 한국경영정보학회 추계학술대회, (2013).

[7] 문하늘, 김종우, “인터넷 뉴스를 활용한 개별 주식 수익률 예측 모델 연구”, 한국지능정보시스템학회 2014년 춘계학술대회, (2004), pp.387-393.

[8] 박종엽, 한인구, “인공신경망을 이용한 한국 종합주가지수의 방향성 예측”, 「한국전문가시스템학회지」, 제1권, 제2호(1995), pp.103-121.

[9] 박한우, “e-사이언스 시대의 인문사회학 연구하기-인터넷 연구방법을 중심으로”, 「사회과학연구」, 제30권, 제2호(2010), pp.195-211.

[10] 박한우, 이연옥, “복합적 텍스트 분석을 이용한 포털 댓글에 관한 연구”, 「Journal of the Korean Data Analysis Society」, 제11권, 제2호(2009), pp.731-744.

[11] 박한우, Loet Leydesdorff, “한국어의 내용분석을 위한 KrKwic 프로그램의 이해와 적용 : Daum.net에서 제공된 지역혁신에 관한 뉴스를 대상으로”, 「Journal of The Korean Data Analysis Society」, 제6권, 제5호(2004), pp.1377-1388.

[12] 송치영, “뉴스가 금융시장에 미치는 영향에 관한 연구”, 「국제경제연구」, 제8권, 제3호(2002), pp.1-34.

[13] 안성원, 조성배, “뉴스 텍스트 마이닝과 시계열 분석을 이용한 주가예측”, 「한국컴퓨터종합학술대회 논문집」, 제37권, 제1호(2010), pp.364-369.

[14] 유은지, “오피니언 마이닝 정확도 향상을 위한 주제지향 감성사전 구축 및 활용 : 주가예측 적용 사례”, 국민대학교 정보미디어경영 석사학위논문, (2013).

[15] 이상열, 원중호, “빅데이터 분석을 위한 슈퍼컴퓨터 환경에서 R의 병렬처리”, 「한국경영과학회지」, 제39권, 제4호(2014), pp.19-31.

[16] 이예지, “뉴스 빅데이터 분석을 통한 종목별 주가예측”, 충북대학교 비즈니스데이터융합 석사학위논문, 2014.

[17] 천세원, 김유신, 정승렬, “뉴스 콘텐츠의 오피니언 마이닝을 통한 매체별 주가상승 예측정확도 비교 연구”, 한국지능정보시스템학회 2013년 춘계학술대회, (2013), pp.133-137.

[18] 허양민, “실시간 인기검색어를 이용한 빅데이터와 주가지수의 상관관계”, 건국대학교 정보통신대학원 석사학위논문, (2014).