金融工學硏究, 제12권 제2호 (2013. 6) 79~96

빅데이터에 나타난 감성 분석

이득환* · 강형구** · 김수현*** · 이창민****

<요 약>

본 연구는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)에 나타난 9가지 감성(Sentiment)들의 특징을 자세히 살펴보았다. 기존에는 감성들의 추출에 대한 어려움으로 인해 감성들이 실제 주식 시장에 끼칠 수 있는 영향력이 등한시 되어 있는 실정이다. 본 연구에서는 Daum-soft에서 제공 받은 감성 자료를 대상으로 자기상관 분석, 주성분 분석, VAR 추정을 실시하여 감성이 가지고 있는 특징을 실증 분석한다.

그 결과 감성들은 일정한 패턴을 가지고 있음을 확인 할 수 있었다. 즉, 자기상관 분석 결과 감성들의 자기회귀성과 주기를 확인 할 수 있었으며 주성분 분석 결과 9가지 감성들이 긍정성, 부정성으로 묶일 수 있음을 보였다. 마지막으로 VAR분석을 통해 음의 자기회귀 계수를 가짐을 알 수 있었으며 상호 다양한 시차에서 영향을 주고받음을 확인 할 수 있었다. 이는 빅데이터(Big-data)에 나타난 주가 정보를 담고 있는 감성들은 무작위적인 정보의 나열이 아니라 주식시장과 흐름을 같이 하고 있으며 과거 값을 통해 예측이 가능함을 시사하고 있다.

주제어 : 자기회귀, 주성분 분석, VAR, 트위터, 빅데이터(Big-data)

Ⅰ. 서 론

금융 시장은 수많은 요인들에 의해 영향을 받으며 그 결과 주가에는 시장 참여자들의 정보가 포함되어 있다. 최근에는 통신 기술이 발전함에 따라 개별 경제 주체들이 가지고 있는 정보는 손쉽게 확산되고 있으며 가계, 기업, 정부 등 광범위한 분야에 영

논문접수일 : 2013. 4. 10 1수정일 : 2013. 5. 29 게재확정일 : 2013. 5. 31
논문심사과정에서 유익한 조언을 해 주신 익명의 세 분의 심사위원께 감사드립니다.
* 제1저자, 한양대학교 경영학과 박사과정, E-mail : Yidhs@naver.com
** 공동저자, 한양대학교 경영학과 조교수, E-mail : hyoungkang@hanyang.ac.kr
*** 공동저자, 삼성자산운용 퀀트 애널리스트, E-mail : soo_hyun.kim@samsung.com
**** 교신저자, 한양대학교 경영학과 조교수, E-mail : changmin74@hanyang.ac.kr

향을 끼치는 실정이다. 이러한 인터넷과 사회적 네트워크 등을 통해 무수히 쏟아지는 엄청난 양의 빅데이터(Big-data)를 효율적으로 관리, 분석함으로 인해서 인간의 행동까지 예측을 시도한다. 그렇다면 주식 시장은 어떠한가? 아직까지 국내에서는 빅데이터(Big-data), 특히 투자자의 감성과 관련된 빅데이터(Big-data)와 주식시장의 상관관계를 분석하는 연구는 미흡한 실정이다. 만약 빅데이터(Big-data)의 적절한 가공을 통해서 금융 시장의 흐름을 읽을 수 있다면 또 다른 하나의 지표로써 주가의 예측과 시장 변동 상황을 이해하는데 도움이 될 것이다.

기존의 외국 문헌들 중에 투자자의 감성과 주식시장과의 연관관계를 분석한 것으로는 대표적으로 Baker and Wurgler(2006, 2007), Han(2008) 등이 있는데 이는 인터넷과 네트워크 등의 빅데이터(Big-data)를 이용한 것들이 아니다. 최근에 Bollen, Mao, and Zeng(2010)은 빅데이터(Big-data)를 통해서 주가의 움직임을 예측하려 했다. 그들은 2008년부터 약 1000만개의 트위터 감성을 심리학 도구를 사용하여 Calm, Alert, Sure, Vital, Kind, Happy등 6가지 감성으로 분류한다. 감성을 사용하여 미국 다우존스 지수와 비교한 결과 적중률 86.7%로 다우존스의 등락을 예측하였다. 특히 빅데이터(Big-data)라는 도구를 통해 개인들이 가지고 있는 생각의 흐름이 금융시장의 미래에 어떠한 영향을 미치는지 분석에 성공 했다는 점에서 의의를 가지고 있다. 또 다른 연구로써 Hristidis(2012)는 약 3억 4천만개의 트위터를 분석한 결과 특정 기업에 대해 언급이 많을수록 주가상승확률이 높아짐을 실증분석 하였다. 특히 4개월 동안의 모델 시뮬레이션 결과 빅데이터(Big-data)에 많이 언급된 회사일수록 다우지수와 비교하여 약 2%정도 덜 하락함을 보였다. 또한 기업의 신제품 출시 등의 정보가 해당 기업의 주식 거래량에 영향을 미침을 보였다. 그로써 빅데이터(Big-data)와 주식시장의 상관관계를 파악하려 했다는 점에 의의를 가지고 있다.

국내 연구로는 김유신, 김남규, 정승렬(2012)이 있다. 그들은 뉴스 컨텐츠를 분석하기 위해 오피니언 마이닝이라는 빅데이터(Big-data) 감성분석 기법을 적용하였고, 이를 통해 주가지수의 등락을 예측하는 지능형 투자의사결정 모형을 제시하였다. 뉴스의 감성 분석 결과와 주가지수 등락은 유의미한 관계가 있으므로, 뉴스의 감성분석 결과를 이용하여 주가지수의 변동성 예측이 가능할 것으로 판단되었다. 한편 정정현, 김수경(2009)은 투자심리의 대용치로 거래회전율을 사용하여 주식 수익률에 미치는 영향을 분석하였다. 그 결과 투자심리가 주식시장에 유의한 영향력을 주는 것을 확인할 수 있다. 또한 옥기율, 김지수(2012)는 소비자 심리지수와 KOSPI 수익률에 관한 실증분석을 통해 소비자 심리지수와 같은 공시 정보가 국내 주식시장에 영향을 미칠 수 있음을 보였다. 그 결과 긍정적인 정보에 과소반응을 하며 부정적인 정보에 과대반응을 하며 부정적인 정보로부터 초래된 주식시장의 하락은 긍정적인 정보 보다 더

빠르게 평균으로 회귀하는 것을 확인 할 수 있다. 미국과 한국의 정보의 전이를 검증한 국내 문헌으로는 김규영 · 김영빈(2010), 유일성(2012)의 연구가 있다. 김규영 · 김영빈(2010)은 Granger 인과관계 분석을 통해 미국 시장의 정보가 국내 시장을 선도함을 보였다. 마찬가지로 유일성(2012)은 GARCH모형 추정을 통하여 미국에서 전이되는 정보가 아시아와 유럽 9개국 수익률에 유의한 영향을 준다는 결과를 제시한다. 그러나 이를 해외 정보의 전이를 이용한 기술적 거래전략 실행 결과 거래비용을 감안할 경우 모든 국가에서 시장대비 초과 수익률 창출에 실패하였다고 한다.

하지만 아직까지 국내외를 막론하고 트위터 등에 나타나는 투자자의 감성에 관한 빅데이터(Big-data)와 금융시장간의 상관관계를 실증 분석하는 연구는 많지 않다. 왜냐하면 금융시장과 연관된 감성이나 예측력을 가지는 요인들을 빅데이터(Big-data)에서 추출하는 것이 어렵기 때문이다. 따라서 본 연구에서는 이의 측정 도구로써 국내 Daum-soft에서 제공하는 9가지 감성을 대상으로 실증분석 하기로 한다. 기존의 문헌연구의 경우 빅데이터(Big-data)를 활용하여 종합주가지수의 등락을 살펴본 반면 본 연구는 감성들의 자기회귀성을 파악하고 주성분 분석을 통해 상호 어떠한 관계를 맺고 있는지, 예측할 수 있는 정보를 포함하고 있는지를 실증 분석한다. 구체적으로 첫째, 주가정보를 담고 있는 9가지 감성들의 자기상관 여부를 파악하여 미래를 예측할 수 있는 정보를 포함하는 지를 살펴 볼 것이다. 이때의 주요 변수인 9가지 감성들은 빅데이터(Big-data)에 나타나있는 주식시장에 대한 느낌을 포함하고 있는 단어들의 조합으로 이루어져 있다. 즉, 블로그, 트위터, 페이스북, 게시판 및 뉴스 등으로부터 정보를 수집하여 스팸 및 노이즈 데이터를 제거한다. 이후 자연어 처리, 텍스트 마이닝을 통해 주식 시장에 대한 분노(ANGER), 미움(HATE), 싫음(DISLIKE), 두려움(FEAR), 사랑(LOVE), 수치심(SHAME), 슬픔(SADNESS), 바람(HOPE), 기쁨(JOY)의 9가지 감성들로 분류한다. 감성들 자체에 주가정보가 담겨있으므로 만약 이들이 자기회귀성을 가진다면 이를 통해 실제 주가의 예측이 가능할 것이다. 둘째, 주성분 분석을 통해 9가지 감성들이 상호 어떠한 관계를 맺고 있는지 파악해 볼 것이다. 셋째, VAR 분석을 통해서 각 감성들의 상관관계를 구체적으로 살펴 볼 것이다. 이러한 의미에서 우리의 연구는 Castillo, Mendoza, and Poblete(2011)의 연구와 유사하다고 할 수 있다. 그들은 트위터에 있는 정보들의 신뢰성을 테스트하는 방법을 제시하여, 어떤 정보들이 신뢰성이 있고 어떤 정보들은 아닌지 분류해 내었다. 그러나 우리의 연구는 트위터에 나타난 감성 정보들의 동학(dynamic)을 분석하는 것으로 분석의 초점에서 큰 차이를 보인다.

우리의 주요 발견은 다음과 같다. 우선, 자기상관 분석 결과 감성들은 7일을 주기로 가지며 각 감성의 과거 값들은 미래를 예측하는데 유용한 정보를 포함하고 있음을 알

수 있다. 둘째, 주성분 분석 결과 9가지 감성들이 크게 긍정성과 부정성으로 묶일 수 있음을 확인할 수 있었다. 마지막으로, VAR 추정에서 자기상관 분석 결과와 마찬가지로 감성들의 자기회귀성을 알 수 있었다.

Ⅱ. 표 본

본 연구는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 주식시장의 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)을 대상으로 실증 분석한다. 또한 9가지의 감성 역시 주성분 분석을 통해 3가지 공통 성분으로 나누어 살펴보았다. 연구 자료는 일간 자료(daily data)이며 감성의 1일 변화량을 변수로 한다. 주식시장의 감성은 Daum-soft에서 제공받았다. 특히 Daum-soft에서 제공받은 감성에 관한 자료는 영업상 기밀로 자세히 밝히지 못하지만 한글 트위터 사용자들의 어휘를 분석하여 이를 9개의 카테고리로 분류하였다. 이후 각 분류별로 나타난 감성의 숫자를 세었다. 9가지 감성 및 단어의 분류는 자체적으로 정의된 사전을 이용하였다. 어휘를 분석하는데 사용한 사전은 Daum-soft의 지적재산으로 공개하지 않았다.

9가지 감성을 나타내는 데이터의 개괄적인 수집방법은 다음과 같다. 첫 번째로 블로그, 트위터, 페이스북, 게시판 및 뉴스 등의 정보원으로부터 각 문서에 포함된 스팸 및 노이즈 데이터를 제거한다. 이후 형태소 분석, 구문 분석 등의 자연어 처리를 통해 텍스트 마이닝 및 데이터 마이닝을 하여 필요한 정보를 포함하고 있는 9가지 감성이 구분되게 된다. 본 연구에서는 주식시장의 정보를 포함하고 있는 9가지 감성을 대상으로 실증분석을 하였다.

우리는 Daum-soft의 소셜미디어 분석 시스템 중 트렌드맵을 사용하였는데 이는 <그림 1>에서와 같이 확인할 수 있다. Daum-soft의 트렌드맵은 소셜미디어로부터 사람들과 사회의 생각을 읽고 인사이트를 발굴하기 위해 개발된 시스템이다. 이는 블로그나 SNS, 커뮤니티 등으로부터 실시간으로 문서를 수집하고 고수준의 언어처리를 이용하여 분석하고 그 결과를 저장한 후 비즈니스 활용을 위한 서비스에 제공하는 소셜 빅데이터 분석 플랫폼(Social Big Data Analytics Platform)이다. 트렌드맵에서 처리하는 어휘에는 제품이나 브랜드와 같은 개체명 뿐만 아니라 이들 개체명과 관련된 감성어(sentiment)들이 포함되어 다양한 형태의 감성 분석이 가능하다. Daum-soft에서는 소셜미디어에 나타난 감성의 분석을 위해 감성어의 의미에 내재하는 기본 속성

에 따라 인간의 감성을 9가지로 분류한다.

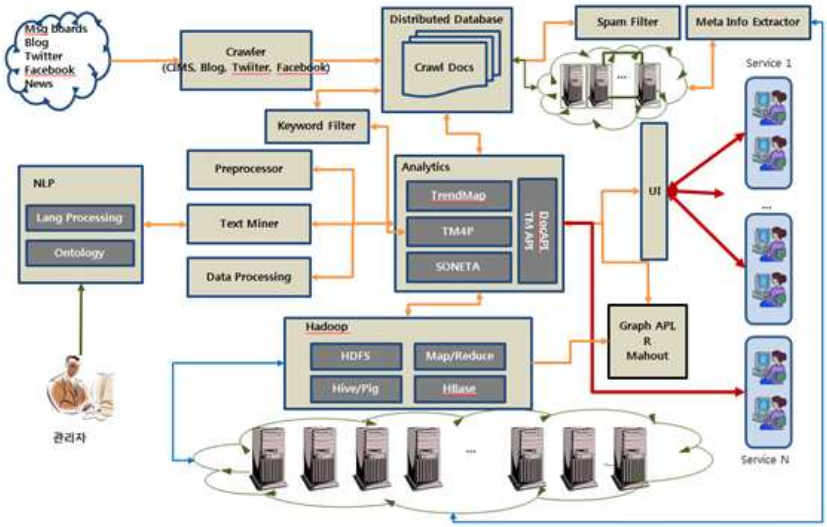
현재 Daum-soft 트렌드맵은 2011년 1월 1일부터 하루 평균 약 500만건의 문서가 수집되어 분석되고 있다. 이에 본 연구에서는 감성분류에 포함된 총 350개의 어휘를 기본으로 주식 시장과 관련된 단어를 결합하여 총 735일간 약 1억9천만개의 빅데이터(Big-data)를 분석하였다.

<표 1>을 통해 살펴보면 시장 부정성을 나타내는 ANGER, HATE, DISLIKE, FEAR의 평균은 양의 값을 가지는 반면 시장 긍정성을 나타내는 HOPE, JOY의 평균은 모두 음의 값을 가지고 있다. 동시에 다른 성향의 감성의 평균이 서로 반대 부호를 가지는 것을 보아 감성들은 잘 구분 되어 있음을 알 수 있다.

<표 1> 기초 통계량

아래의 표는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)들의 기초 통계량임. 9가지 감성들은 Daum-soft에서 제공받았음. 주말과 공휴일을 포함한 총 735일간의 1일 간의 감성 변화량을 보여주고 있음. 즉, 오늘의 감성 변화량(St)은 ln(St/St-1) 으로 측정되었음.

변수	Obs.	최소값	최대값	평균	표준편차
ANGER	735	-.7217	.9098	.001140	.1423444
HATE	735	-.9907	.6657	.001020	.1471391
DISLIKE	735	-.7645	.5346	.000936	.1236879
FEAR	735	-.9664	1.1105	.000885	.1457385
LOVE	735	-.9490	.3889	.000781	.1138936
SHAME	735	-1.0869	1.1430	.001177	.1632425
SADNESS	735	-.8726	.4611	.000377	.1115076
HOPE	735	-.8935	1.3414	-.000584	.1916136
JOY	735	-.9763	.6176	-.000631	.1425756



위의 그림은 Daum-soft의 트렌드맵 시스템을 도식화한 그림임. 이 시스템은 대량의 문서로부터 형태소분석, 품사중의성해소, 구문분석, 개체명인식 등의 언어처리 후 텍스트마이닝을 이용하여 중요 어휘들을 자동으로 추출하고 빈도 및 상호 연관관계를 날짜별로 추출한 후 시각화(visualization) 기술을 이용해 사용자에게 데이터에 대한 이해를 쉽게 도와주는 기능을 가지고 있음.

<그림 1> 트렌드맵 시스템

Ⅲ. 실증 분석

1. 감성들의 자기상관

본 연구는 주식시장에 대한 감성이 예측할 수 있는 정보를 포함하는가를 알아본다. 우선 각 감성들의 변화량이 자기상관 성격이 있는지 실증 분석하도록 한다. 만약 자기상관이 있다면 이는 주식시장에 대한 정보를 포함하는 감성들이 실제로 주식시장에 영향을 줄 수 있음을 암시할 것이다. <표 2-1>, <표 2-2>, <표 2-3>은 14일까지의 자기상관을 분석한 표이며 <그림 1>은 이를 바탕으로 도식화 한 것이다.

<표 2-1>, <표 2-2>, <표 2-3>를 통해 살펴보면 각 감성들 마다 약간의 차이를 가지지만 모두 공통적으로 7시차마다 유의한 양의 자기상관계수를 가짐을 알 수 있다. 마찬가지로 <그림 1>을 통해 보면 감성들은 신뢰한계를 벗어나는 시차가 존재함을 알 수 있다. 이는 현재의 감성들은 미래를 예측할 때 사용될 수 있으며 일주일을 주기로 과거 값과 현재 값 사이에 관계가 유의미하다는 것을 확인 할 수 있다. 즉, 빅

데이터(Big-data)에 나타난 주식 시장에 대한 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)은 무작위적으로 나타나는 무의미한 값이 아니라 예측성과 주기를 갖는 정보(Information)의 성격을 가지고 있음을 암시한다. 특히 그 주기가 7일이라는 것은 주목할만 한 점이다.

<표 2-1> 자기상관계수(1)

아래의 표는 Big-data에 나타난 9가지 감성 중 ANGER, HATE, DISLIKE의 자기상관계수를 정리한 표임. AC는 자기상관계수를 의미하며 Q-Stat.는 Box-Ljung 통계량을 의미함. Value는 Box-Ljung 통계량의 값이며 그에 따른 P-값(Sig.)을 확인 할 수 있음.

Lag	ANGER			HATE			DISLIKE		
	AC	Q-Stat.		AC	Q-Stat.		AC	Q-Stat.	
		Value	Sig.		Value	Sig.		Value	Sig.
1	-.238	41.748	.000	-.212	33.190	.000	-.220	35.665	.000
2	-.127	53.697	.000	-.153	50.554	.000	-.144	50.925	.000
3	-.044	55.155	.000	-.063	53.505	.000	-.056	53.222	.000
4	-.046	56.732	.000	-.052	55.511	.000	-.076	57.449	.000
5	-.038	57.792	.000	-.073	59.510	.000	-.055	59.688	.000
6	-.016	57.991	.000	.059	62.113	.000	-.022	60.033	.000
7	.148	74.209	.000	.141	77.007	.000	.239	102.617	.000
8	-.034	75.052	.000	.039	78.134	.000	-.013	102.735	.000
9	-.038	76.113	.000	-.064	81.180	.000	-.025	103.188	.000
10	-.005	76.128	.000	-.079	85.847	.000	-.081	108.101	.000
11	-.005	76.147	.000	-.025	86.296	.000	-.010	108.183	.000
12	-.068	79.566	.000	-.055	88.578	.000	-.140	122.918	.000
13	.028	80.161	.000	.038	89.659	.000	.067	126.240	.000
14	.086	85.741	.000	.157	108.168	.000	.235	167.699	.000

<표 2-2> 자기상관계수(2)

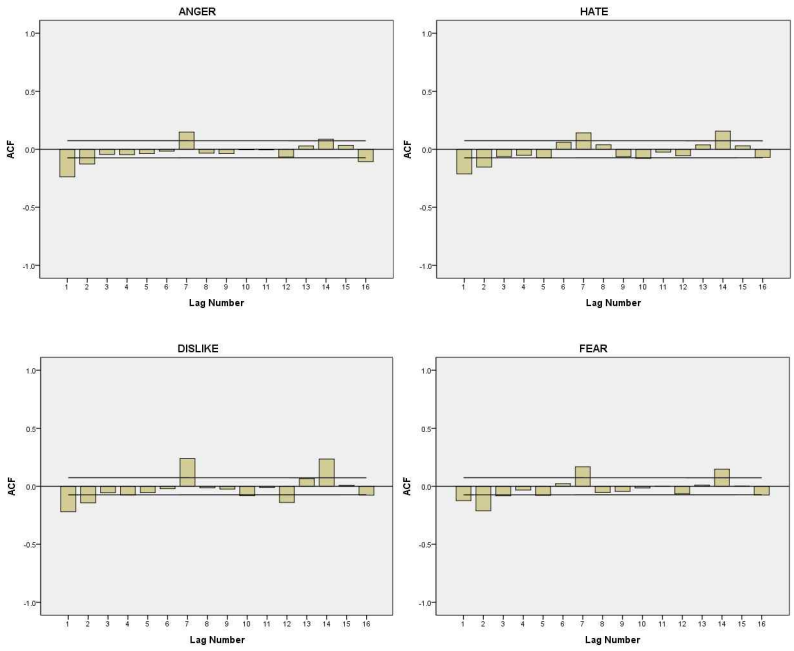
아래의 표는 Big-data에 나타난 9가지 감성 중 FEAR, LOVE, SHAME의 자기상관계수를 정리한 표임. AC는 자기상관계수를 의미하며 Q-Stat.는 Box-Ljung 통계량을 의미함. Value는 Box-Ljung 통계량의 값이며 그에 따른 P-값(Sig.)을 확인 할 수 있음.

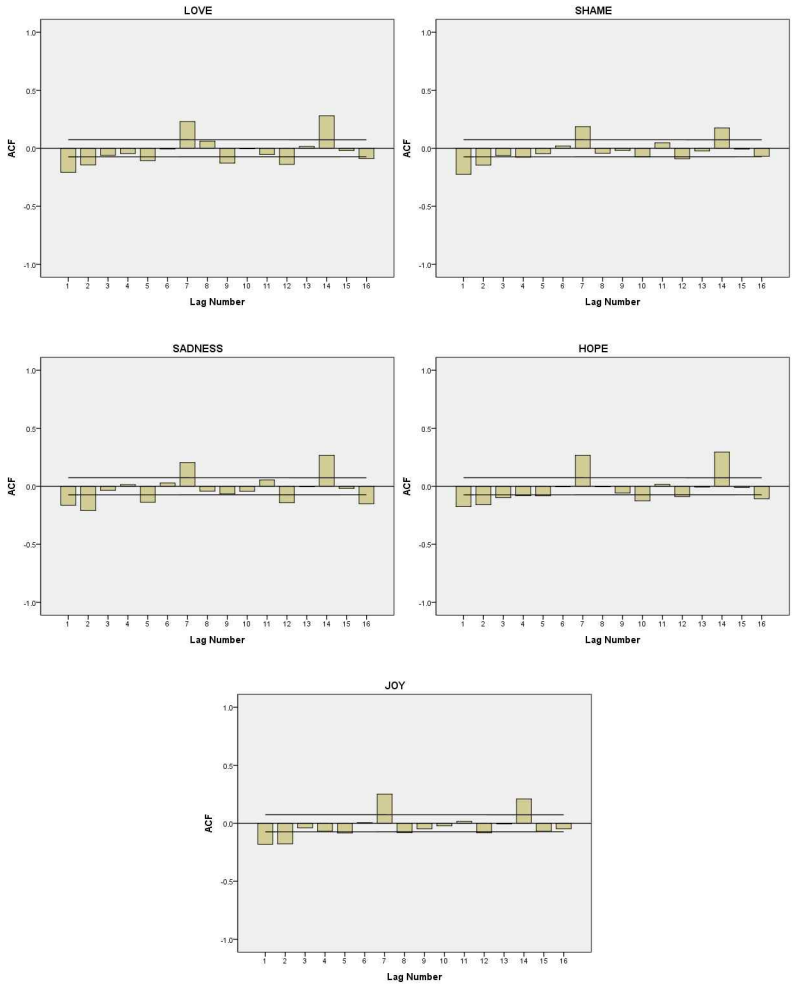
Lag	FEAR			LOVE			SHAME		
	AC	Q-Stat.		AC	Q-Stat.		AC	Q-Stat.	
		Value	Sig.		Value	Sig.		Value	Sig.
1	-.124	11.263	.001	-.207	31.764	.000	-.224	37.103	.000
2	-.211	44.117	.000	-.144	47.048	.000	-.146	52.769	.000
3	-.083	49.186	.000	-.061	49.828	.000	-.063	55.701	.000
4	-.033	49.983	.000	-.045	51.359	.000	-.077	60.144	.000
5	-.079	54.652	.000	-.106	59.727	.000	-.046	61.712	.000
6	.022	55.022	.000	-.006	59.756	.000	.020	62.004	.000
7	.168	75.954	.000	.231	99.319	.000	.187	87.914	.000
8	-.053	78.073	.000	.061	102.107	.000	-.043	89.265	.000
9	-.044	79.488	.000	-.128	114.402	.000	-.019	89.523	.000
10	-.014	79.641	.000	-.003	114.409	.000	-.074	93.586	.000
11	.002	79.643	.000	-.054	116.558	.000	.048	95.277	.000
12	-.065	82.766	.000	-.139	130.960	.000	-.090	101.357	.000
13	.009	82.833	.000	.016	131.152	.000	-.023	101.765	.000
14	.147	99.067	.000	.281	190.472	.000	.176	125.003	.000

<표 2-3> 자기상관계수(3)

아래의 표는 Big-data에 나타난 9가지 감성 중 SADNESS, HOPE, JOY의 자기상관계수를 정리한 표임. AC는 자기상관계수를 의미하며 Q-Stat.는 Box-Ljung 통계량을 의미함. Value는 Box-Ljung 통계량의 값이며 그에 따른 P-값(Sig.)을 확인 할 수 있음.

Lag	SADNESS			HOPE			JOY		
	AC	Q-Stat.		AC	Q-Stat.		AC	Q-Stat.	
		Value	Sig.		Value	Sig.		Value	Sig.
1	-.164	19.744	.000	-.175	22.644	.000	-.181	24.154	.000
2	-.208	51.806	.000	-.160	41.457	.000	-.178	47.572	.000
3	-.036	52.741	.000	-.099	48.763	.000	-.039	48.714	.000
4	.014	52.883	.000	-.082	53.719	.000	-.069	52.199	.000
5	-.138	66.974	.000	-.083	58.780	.000	-.084	57.380	.000
6	.028	67.562	.000	-.003	58.788	.000	.005	57.402	.000
7	.205	98.770	.000	.268	112.180	.000	.252	104.503	.000
8	-.042	100.074	.000	-.004	112.193	.000	-.081	109.387	.000
9	-.066	103.336	.000	-.057	114.657	.000	-.048	111.098	.000
10	-.042	104.669	.000	-.126	126.503	.000	-.022	111.466	.000
11	.055	106.899	.000	.017	126.714	.000	.016	111.667	.000
12	-.142	122.018	.000	-.089	132.584	.000	-.083	116.800	.000
13	-.002	122.021	.000	-.007	132.623	.000	-.006	116.826	.000
14	.268	175.961	.000	.295	198.079	.000	.209	149.700	.000





아래 그림은 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)들의 1시차부터 14시차까지의 자기상관함수(ACF)를 보여줌. x축은 시차, y축은 자기상관계수를 의미하며 0을 기준으로 상하의 참조선은 신뢰한계를 보여주고 있음.

<그림 2> 감성의 자기상관 분석

2. 주성분 분석

변수들을 대표할 수 있는 성분을 추출하기 위해 주성분 분석을 시행한다. 이는 총 표본의 분산 중 대부분을 설명하는 몇 개의 주요인을 추출하는데 의의가 있다. 요인 추출방법에는 요인 적재 값의 제곱 합인 Eigen-value 1이상의 값을 기준으로 삼을 수

있다. 하지만 본 연구에서는 시장 감성을 긍정성과 부정성 및 중립성으로 구분해서 살펴보고자 감성들의 요인 수를 사전에 3으로 고정한다. 이때 요인회전 방법으로는 직각요인회전 방식 중 가장 많이 사용되는 베리맥스(Varimax)방식을 25회 반복 수행하여 주성분 요인 벡터를 추출하였다.

<표 3>는 감성들의 회전된 성분행렬을 나타낸 표이다. ANGER, HATE, DISLIKE, FEAR는 주성분 1로 구분 되었다. LOVE, SHAME, SADNESS는 주성분 2로 구분 되었으며 HOPE, JOY는 주성분 3으로 묶일 수 있음을 보여 준다. 첫 번째 성분들은 ANGER, HATE, DISLIKE, FEAR가 포함되어 있으며 주로 시장 부정성을 대표한다. 두 번째 성분들은 LOVE, SHAME, SADNESS이며 세 번째 성분들은 HOPE, JOY가 포함되어 시장 긍정성을 주로 나타내고 있다. 이렇게 감성의 성격에 따라 비슷한 감성들이 하나의 성분으로 묶일 수 있다는 것은 감성들의 자기상관을 분석한 결과를 뒷받침한다.

<표 3> 회전된 성분행렬

아래의 표는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)들의 주성분 분석(principal component analysis)을 한 결과임. 이를 통해 추출된 3개의 주성분들에 대한 회전된 성분행렬을 확인 할 수 있음. 요인회전 방법으로는 Kaiser 정규화가 있는 베리맥스(Vari max)방식을 사용하였음.

	주성분1	주성분2	주성분3
ANGER	.787	0.329	-0.018
HATE	.769	0.101	0.243
DISLIKE	.658	0.368	0.225
FEAR	.644	0.236	0.343
LOVE	0.197	.798	0.154
SHAME	0.375	.741	0.057
SADNESS	0.563	.594	0.221
HOPE	0.338	0.021	.866
JOY	0.057	0.559	.709

송치영(2002)은 주가가 경제동향, 해외경제, 해외정치·사회 뉴스에 대하여 유리한 뉴스와 불리한 뉴스에 다르게 반응하는 비대칭적인 반응을 한다고 하였다. 본 연구에서 분류된 시장 긍정성과 시장 부정성에 대한 빅데이터(Big-data)는 주가에 각기 다른 영향을 미칠 수 있을 것이다. 즉, 주식 시장에 대한 감성들이 정보(Information)의 역할을 하여 그날의 시장에 대한 긍정성이나 부정성을 통해 주가의 예측이 가능함을 시사한다.

특히 LOVE, SHAME, SADNESS는 긍정성, 부정성 집단 중 어느 집단에도 포함되

지 못하였는데 주식 시장에 대한 사랑(LOVE), 수치심(SHAME), 슬픔(SADNESS)은 주식 시장에 정(+)과 부(-)의 관계가 없는 중립적인 정보(Information) 또는 잡음(Noise)를 포함할 것이다. 기업정보관리를 위한 오픈소스를 제공하는 MIKE2.0에서는 빅데이터(Big-data)를 유의미한 빅데이터(Big-data) 추출이 쉽지 않음을 시사한다.[1] 수많은 정보가 있는 인터넷의 공간에서 단어들의 조합을 통해 정확히 원하는 정보를 뽑아낸다는 것은 기술적으로 매우 힘든 작업이다. 향후 기계적으로 분류된 사랑(LOVE), 수치심(SHAME), 슬픔(SADNESS)의 감성에 대한 추가적인 분석이 필요할 것이다.

<표 4>을 통해 살펴보면 추출된 3개의 주성분 중 첫 번째 성분이 전체의 약 51%를 설명하고 있다. 이는 전체 감성을 가장 크게 대표한다고 할 수 있다. 두 번째와 세 번째 성분들은 각각 전체의 약 10%와 9%를 설명하고 있다.

<표 4> 추출 주성분의 설명된 총 분산

아래의 표는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)들의 주성분 분석을 실시한 결과임. 구분된 3가지 주성분들의 설명된 총 분산을 보여주고 있음. 추출된 요인의 수는 사전에 3으로 고정하였으며 설명력(%분산)은 각 요인들이 전체의 얼마를 설명하는지를 보여줌.

성분	초기 고유값			추출 제곱합 적재값			회전 제곱합 적재값		
	합계	% 분산	% 누적	합계	% 분산	% 누적	합계	% 분산	% 누적
1	4.604	51.156	51.156	4.604	51.156	51.156	2.673	29.701	29.701
2	.932	10.354	61.510	.932	10.354	61.510	2.162	24.024	53.725
3	.856	9.516	71.026	.856	9.516	71.026	1.557	17.301	71.026

3. VAR(Vector Auto-regression) 분석

VAR모형은 특정한 경제이론의 제약 없이 주요 경제, 금융 변수들의 상호의존성이나 내생성을 고려하여 모형화 하는 데 적절한 도구이다. 하지만 시계열이 자기회귀 구조를 가지고 있어야 하므로 기본적으로 해당 시계열의 안정성을 전제로 한다. VAR 분석을 수행하기 앞서 감성성들의 안정성을 검정하였다. 이때 ADF단위근 검정을 통하여 단위근을 검정한다. 옵션은 수준변수에서 상수항과 추세를 고려하지 않고 분석하였다.

<표 5>에 따르면 모든 변수는 1% 수준에서 안정적인 시계열임을 확인 할 수 있다. 다음으로 변수들의 자기상관을 보다 명확히 살펴보고자 VAR를 분석하도록 한다.

1) http://mike2.openmethodology.org/wiki/Big_Data_Definition

<표 5> Augmented Dickey-Fuller 단위근 검정

아래의 표는 2011년 1월 1일부터 2013년 1월 4일까지 빅데이터(Big-data)를 통해 살펴 본 9가지 감성(ANGER, HATE, DISLIKE, FEAR, LOVE, SHAME, SADNESS, HOPE, JOY)들의 ADF 단위근 검정 표임. 수준변수(일일 변화률)에서 감성들에 대한 상수항과 추세를 포함하지 않은 결과임. t-Statistic와 Prob.는 각 감성들의 t-통계량과 그에 대응하는 p-값을 의미함.

변수	t-Statistic	Prob.
ANGER	-27.1959	0.0000
HATE	-20.122	0.0000
DISLIKE	-25.1569	0.0000
FEAR	-23.8793	0.0000
LOVE	-2.3843	0.0000
SHAME	-6.27236	0.0000
SADNESS	-4.15555	0.0000
HOPE	-24.8433	0.0000
JOY	-20.78481	0.0000

<표 6>, <표 7>, <표 8>은 9가지 감성들을 VAR를 분석한 표이다. 이들을 종합하여 살펴보면 모든 감성들은 비교적 긴 시차에서 자기회귀성을 가진다고 할 수 있다. 이의 결과는 자기상관성을 분석한 결과와 일치하고 있음을 알 수 있다. 특히 감성들의 자기회귀 계수는 대부분 음의 값을 가지며 1%수준에서 통계적으로 유의함을 확인할 수 있다. 또한 하루 전의 자기회귀 계수값들이 가장 큼을 알 수 있으며 시차가 증가할수록 점점 약해짐을 알 수 있다. 각 변수들은 다양한 시차에서 양과 음의 회귀계수 값을 유의적으로 가지는 것으로 보아 상호 유기적인 움직임을 확인 할 수 있다. 공통적으로 부정성을 나타내는 ANGER, HATE, DISLIKE, FEAR의 과거값은 다른 감성들에 주로 양의 영향력을 불규칙하게 나타내는 반면 긍정성을 나타내는 HOPE의 과거값은 다른 감성들에 주로 음의 영향력을 불규칙하게 보임을 알 수 있다. 이는 빅데이터(Big-data)의 특성을 나타나는 것으로 1일간의 긍정(부정)이 증가(감소)하였다고 할지라도 1일간의 전체 값이 증가(감소)하여 나타난 것인지 아닌지를 알 수 없기 때문에 표준화된 방법의 사용이 필요함을 시사한다.

한편 옥기율·김지수(2012)는 우리나라 주식시장에서 거시경제정보가 사전에 다른 정보원천에 의해 예측되고 그 내용이 주가에 반영된다고 하였다. VAR분석의 결과는 빅데이터(Big-data)를 통해 추출된 대량의 정보(Information) 혹은 감성(Sentiment)은 새로운 정보원천으로써 실시간으로 보다 정확하게 파악하여 주가를 예측할 수 있음을 알 수 있다.

<표 6> VAR 추정 결과(1)

아래의 표는 주성분 1에 해당하는 변수(ANGER, HATE, DISLIKE, FEAR)들을 대상으로 그들의 일일 변화량에 대한 VAR(Vector Auto Regression) 분석 결과임. 분석한 VAR모형에는 9가지 감성이 모두 포함되어 있으나 지면상 <표 6>, <표 7>, <표 8>로 나누어 살펴봄. *와 **은 각각 5%와 1%수준에서 유의함을 의미함.

	ANGER	HATE	DISLIKE	FEAR	LOVE	SHAME	SADNESS	HOPE	JOY
ANGER(-1)	-0.534**	0.047	0.062	0.021	0.067	0.137**	0.043	0.089	0.158**
ANGER(-2)	-0.379**	0.065	0.083	0.029	0.080	0.036	0.042	0.090	0.220**
ANGER(-3)	-0.375**	0.039	0.042	-0.107	0.047	0.046	0.002	0.171**	0.144**
ANGER(-4)	-0.294**	0.072	0.053	-0.014	0.101*	0.118	0.040	0.127	0.147**
ANGER(-5)	-0.188**	0.127	0.065	0.024	0.093*	0.083	0.094*	0.230**	0.169**
ANGER(-6)	-0.185**	0.082	0.060	0.051	0.013	-0.00	0.006	0.123	0.022
ANGER(-7)	-0.105*	0.015	0.058	0.002	0.025	0.042	-0.035	-0.041	-0.091*
HATE(-1)	0.038	-0.610**	0.006	-0.034	-0.017	-0.012	-0.029	-0.018	-0.084*
HATE(-2)	0.024	-0.494**	0.062	-0.014	-0.020	0.021	-0.011	0.003	-0.007
HATE(-3)	-0.050	-0.401**	-0.013	-0.037	-0.037	-0.015	-0.062	-0.129	0.022
HATE(-4)	0.049	-0.250**	0.051	0.022	0.010	0.097	0.050	-0.015	0.040
HATE(-5)	0.085	-0.197**	0.089*	0.107	0.068	0.075	0.026	-0.055	0.003
HATE(-6)	0.179*	-0.060	0.093**	0.119**	0.030	0.031	0.070	0.102	0.017
HATE(-7)	0.128**	-0.023	0.107**	0.110**	0.020	0.004	0.055	0.130**	0.076
DISLIKE(-1)	0.065	0.046	-0.598**	0.021	0.042	0.110	0.089*	0.017	0.031
DISLIKE(-2)	-0.024	-0.019	-0.544**	0.025	-0.026	0.150**	0.016	-0.026	-0.015
DISLIKE(-3)	0.112	-0.010	-0.416**	0.001	0.021	0.205**	0.060	0.005	-0.044
DISLIKE(-4)	-0.008	-0.077	-0.433**	-0.161	-0.028	0.059	-0.003	-0.174	-0.068
DISLIKE(-5)	-0.09	-0.138	-0.348**	-0.199**	0.012	-0.075	-0.034	-0.140	-0.033
DISLIKE(-6)	0.0067	-0.071	-0.248**	-0.189**	0.0949	-0.021	-0.0001	-0.253**	-0.040
DISLIKE(-7)	0.038	0.125*	-0.039	0.0374	0.103*	0.101	0.096**	-0.119	-0.016
FEAR(-1)	0.103**	0.165**	0.077**	-0.322**	0.060	0.123**	0.139**	0.257**	0.085*
FEAR(-2)	0.083	0.097*	0.024	-0.356**	0.053	0.118	0.084	0.149	0.068
FEAR(-3)	0.184**	0.251**	0.107**	-0.195**	0.042	0.167**	0.158**	0.219**	0.063
FEAR(-4)	0.119**	0.148**	0.123**	-0.119*	0.063	0.143**	0.108**	0.151**	0.071
FEAR(-5)	0.041	0.089	0.029	-0.153**	-0.037	0.027	0.018	0.048	-0.007
FEAR(-6)	0.062	0.091	0.128	-0.053**	-0.020	0.055	0.068	0.109	0.051
FEAR(-7)	0.047	0.008	0.061*	-0.008	0.017	-0.014	0.023	0.020	-0.021

<표 7> VAR 추정결과(2)

아래의 표는 주성분 2에 해당하는 변수(LOVE, SHAME, SADNESS)들을 대상으로 그들의 일일 변화량에 대한 VAR(Vector Auto Regression) 분석 결과임. 분석한 VAR모형에는 9가지 감성이 모두 포함되어 있으나 지면상 <표 6>, <표 7>, <표 8>로 나누어 살펴봄. *와 **은 각각 5%와 1%수준에서 유의함을 의미함.

	ANGER	HATE	DISLIKE	FEAR	LOVE	SHAME	SADNESS	HOPE	JOY
LOVE(-1)	0.012	-0.010	0.063	0.075	-0.530**	-0.058	0.032	-0.031	0.005
LOVE(-2)	-0.034	-0.068	-0.027	-0.059	-0.468**	-0.129	-0.101*	-0.044	-0.092
LOVE(-3)	-0.117	-0.102	-0.096	0.007	-0.477**	-0.306**	-0.087	-0.072	-0.079
LOVE(-4)	-0.055	-0.086	-0.072	-0.025	-0.401**	-0.218**	-0.062	-0.157	-0.046
LOVE(-5)	0.048	0.007	-0.008	0.040	-0.344**	-0.172**	-0.046	0.098	-0.072
LOVE(-6)	-0.029	-0.050	-0.036	0.042	-0.254**	-0.115	-0.025	0.047	-0.103
LOVE(-7)	0.087	0.009	0.032	0.126*	-0.082	0.055	0.073	0.009	0.045
SHAME(-1)	0.012	0.009	-0.008	0.076	0.066*	-0.638**	0.006	-0.044	-0.027
SHAME(-2)	-0.058	-0.027	-0.035	0.084	-0.021	-0.522**	-0.052	-0.005	-0.001
SHAME(-3)	-0.013	-0.060	0.002	0.095	0.001	-0.511**	-0.026	0.031	-0.027
SHAME(-4)	-0.001	-0.098	0.022	0.013	-0.068	-0.525**	-0.102**	0.036	-0.142**
SHAME(-5)	-0.087	-0.120*	-0.010	-0.023	-0.073	-0.385**	-0.103**	-0.066	-0.134**
SHAME(-6)	-0.061	-0.088	-0.129**	-0.095	-0.089**	-0.280**	-0.095**	-0.121	-0.086
SHAME(-7)	0.033	0.024	-0.065	-0.044	0.012	-0.114**	-0.015	-0.017	0.075
SADNESS(-1)	0.015	0.117	0.019	-0.152*	-0.095	0.066	-0.520**	0.032	0.008
SADNESS(-2)	-0.020	0.024	0.019	-0.225**	-0.077	-0.092	-0.379**	-0.127	-0.102
SADNESS(-3)	0.082	0.066	0.085	-0.015	0.045	0.105	-0.275**	-0.032	0.033
SADNESS(-4)	-0.006	0.099	0.011	-0.003	0.074	0.080	-0.142**	0.040	0.120
SADNESS(-5)	0.038	0.032	-0.094	0.030	-0.010	0.187*	-0.138**	-0.025	0.128
SADNESS(-6)	-0.072	-0.068	-0.091	-0.019	0.013	0.257**	-0.112	-0.095	0.089
SADNESS(-7)	-0.039	-0.109	-0.114*	-0.082	-0.065	0.076	-0.092	-0.016	0.109

<표 8> VAR 추정결과(3)

아래의 표는 주성분 3에 해당하는 변수(HOPE, JOY)들을 대상으로 그들의 일일 변화량에 대한 VAR(Vector Auto Regression) 분석 결과임. 분석한 VAR모형에는 9가지 감성이 모두 포함되어 있으나 지면상 <표 6>, <표 7>, <표 8>로 나누어 살펴봄. *와 **은 각각 5%와 1%수준에서 유의함을 의미함.

	ANGER	HATE	DISLIKE	FEAR	LOVE	SHAME	SADNESS	HOPE	JOY
HOPE(-1)	-0.023	0.057	-0.026	0.022	-0.059*	-0.098**	-0.060**	-0.547**	0.007
HOPE(-2)	-0.013	-0.011	-0.029	0.024	-0.050	-0.051	-0.029	-0.500**	-0.013
HOPE(-3)	-0.116**	-0.100**	-0.106**	-0.033	-0.078**	-0.120**	-0.064*	-0.499**	-0.074
HOPE(-4)	-0.073	-0.128**	-0.069	-0.037	-0.104**	-0.102*	-0.096**	-0.397**	-0.047
HOPE(-5)	-0.080	-0.103**	-0.024	-0.015	-0.083**	-0.024	-0.074**	-0.349**	-0.030
HOPE(-6)	-0.028	-0.023	-0.013	0.004	-0.056	0.094*	0.012	-0.215**	0.001
HOPE(-7)	-0.075*	0.035	0.004	-0.030	-0.068**	-0.034	0.000	0.007	-0.012
JOY(-1)	-0.020	-0.063	-0.039	-0.027	-0.051	0.094	0.022	0.070	-0.513**
JOY(-2)	0.000	0.056	0.015	0.001	0.018	0.067	-0.007	0.026	-0.505**
JOY(-3)	-0.010	0.099	-0.024	-0.079	-0.008	0.131*	-0.010	0.047	-0.415**
JOY(-4)	0.004	0.074	-0.021	-0.007	-0.044	0.122	-0.017	0.017	-0.397**
JOY(-5)	0.046	0.086	-0.021	-0.067	-0.019	0.061	0.027	-0.004	-0.297**
JOY(-6)	0.002	0.001	0.020	-0.080	0.033	-0.076	-0.013	0.022	-0.179**
JOY(-7)	0.063	0.041	0.054	0.035	0.101**	0.157**	0.085**	0.067	0.025

Ⅳ. 결론 및 한계점

본 연구는 빅데이터(Big-data)의 감성들의 성격을 자기상관, 주성분 분석, VAR 모형을 통해 자세히 살펴보았다. 2011년 1월 1일부터 2013년 1월 4일까지를 표본 기간으로 하여 감성들의 일일 변화량을 변수로 하여 자기상관 여부를 살펴보았다. 이후 상호간의 영향력을 구체적으로 살펴보고자 빅데이터(Big-data)의 분노(ANGER), 미움(HATE), 싫음(DISLIKE), 두려움(FEAR), 사랑(LOVE), 수치심(SHAME), 슬픔(SADNESS), 바람(HOPE), 기쁨(JOY)의 9가지 감성을 주성분 분석을 하였으며 VAR 분석을 통해 상호 상관관계를 살펴보았다. 그 결과를 요약하면 아래와 같다.

첫째, 감성들의 자기상관 분석 결과 감성들은 7일을 주기로 가지며 각 감성의 과거 값들은 미래를 예측하는데 유용한 정보를 포함하고 있음을 알 수 있다. 특히 주가에 대한 정보가 감성들에 포함 되어 있으므로 이를 통해 예측을 시도할 수 있을 것이다. 둘째, 주성분 분석 결과 9가지 감성들이 크게 긍정성과 부정성으로 묶일 수 있음을 확인할 수 있었다. 향후 이들의 성분점수는 주가 수익률에 유의미한 영향을 가져올 수 있음을 시사한다. 예를 들어 긍정성은 주가 수익률에 양(+)의 영향을 미칠 것이며 부정성은 음(-)의 영향을 미칠 것이다. 셋째, VAR 추정에서 자기상관 분석 결과와 마찬가지로 감성들의 자기회귀성을 알 수 있었으며 감성들이 시차가 변함에 따라 상호 영향을 주고받음을 보였다.

French, Schwert,and Stambaugh(1987)은 포트폴리오 수익률이 자기상관을 가짐을 고려하여 주가 수익률과 변동성 예측에 활용하였으며 Engle(1982)은 인플레이션의 자기상관 관계를 활용하여 p-차 자기회귀 조건부 이분산(Autoregressive conditional heteroskedasticity)모형을 제시하였다. 한편 심홍진, 황유선(2010)은 트위터 이용 동기에 대해 사회참여 및 여론 형성, 팔로어 그룹 형성 등을 가짐을 통해 트위터가 공적이고 사적인 미디어의 영역에 활용될 수 있음을 실증하였다. 이러한 빅데이터(Big-data)의 활용성을 금융 시장으로 확대하여 볼 때 빅데이터(Big-data)를 통해 나타난 주식시장에 관한 정보(Information) 혹은 감성(Sentiment)이 자기상관성을 가짐을 활용한다면 주가 및 변동성 등 금융 시장 전반의 연구 및 실무에 시사하는 바가 클 것이다.

하지만 본 연구는 다음과 같은 한계점을 가지고 있다. 첫째, 주성분 분석에서 LOVE, SHAME, SADNESS의 특징을 명확히 정의 하지 못하였다. 즉, LOVE, SHAME, SADNESS는 긍정성 집단과 부정성 집단 중 어느 집단에도 소속되지 못함을 확인할 수 있다. 이는 시장에 잡음(Noise)으로 작용할 가능성이 높다. 빅데이터

(Big-data)에서 감성들을 추출할 때 일정한 단어의 조합을 통해 각 감성이 분류가 되는데 기술적으로 충분한 분류가 되었더라도 정확한 분류에는 한계가 존재한다. 이들의 명확한 정의는 향후 연구과제로 남긴다. 둘째, VAR 모형 추정에서 긍정성에 속하는 감성과 부정성에 속하는 감성들의 상호 회귀계수들이 시차에 따라 양과 음의 관계가 있었다. 이론적으로 1일간의 긍정성 증가는 부정성의 감소를 가져와야 하지만 일부 시차에서 함께 증가 또는 감소함으로써 감성의 명확한 추출의 어려움이 있음을 알 수 있다.

참고문헌

김규영·김영빈 (2010), "한·미 주식시장 간의 정보 전달에 관한 실증연구 : 투자집단별 순매수 의사결정을 중심으로," 금융공학연구, 제9권 제2호, 53-75.

김유신·김남규·정승렬 (2012), "뉴스와 주가 : 빅데이터 감성분석을 통한 지능형 투자 의사결정모형," 지능정보연구, 제18권 제2호, 143-156.

송치영 (2002), "뉴스가 금융시장에 미치는 영향에 관한 연구," 국제경제연구, 제8권 제3호, 1-34.

심홍진·황유선 (2012), "마이크로블로깅(micro-blogging) 이용동기에 관한 연구 : 트위터(twitter)를 중심으로," 한국방송학보, 제24권 제2호, 192-234.

옥기율·김지수 (2012), "소비자 심리지수가 KOSPI 수익률에 미치는 비대칭적 영향에 대한 연구," 금융공학연구, 제11권 제1호, 17-37.

유일성 (2012), "해외정보전이와 국가별 초과이익 창출기회," 금융공학연구, 제11권 제3호, 31-57.

정정현·김수경 (2009), "투자자 심리의 척도로서의 시장유동성이 주식수익률에 미치는 영향," 금융공학연구, 제8권 제4호, 65-90.

Bing, H. (2007), "Investor Sentiment and Option Prices," *Review of Financial Studies*, 21, 387-414.

Bollen, J., H. Mao, and X. J. Zeng (2010), "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2(1), 1-8.

Castillo, C., M. Mendoza, and B. Poblete (2011), "Information Credibility on Twitter," In Proceedings of World Wide Web Conference, 675-684.

French, K. R., G. W. Schwert, and R. F. Stambaugh (1987), "Expected Stock

Returns and Volatility," *Journal of Financial Economics*, 19, 3-29.

Malcolm, B. and J. Wurgler (2006), "Investor Sentiment and the Cross-Section of Stock Returns," *Journal of Finance*, 61(4), 1645-1680.

Malcolm, B. and J. Wurgler (2007), "Investor Sentiment in the Stock Market," *Journal of Economic Perspectives,* 21, 129-151.

Robert, F. E. (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987-1007.

Vagelis, H. (2012), "Correlating Financial Time Series With Micro-blogging Activity," *WSDM*, 513-522.

Abstract

Autocorrelation Analysis of the Sentiment with Stock Information Appearing on Big-Data

Deuk Hwan Lee[*], *Hyoung Goo Kang*[**], *Soo Hyun Kim*[***], *and Chang Min Lee*[****]

We study thoroughly by looking into nine different sentiments found in approximately 190 million pieces of Big-data gained from January 1st, 2011 to January 4th, 2013. In the past, it was not easy to extract the sentiments and because of that, until now, any influences that the sentiments could actually have on the stock market have been neglecting. In the study, with the sentiment references provided by Daum-soft, features of the sentiments were examined by autocorrelation analysis, principle component analysis and VAR. According to the results, we find that the sentiments are observed to have some regular patterns. In other words, the findings from the autocorrelation analysis prove autocorrelation and period of the sentiments while the results from the principle component analysis report that the nine sentiments could be connected with positivity and negativity. Lastly, via VAR, the sentiments appeare to have negative autoregressive parameters as they would be affected by each other at various lag-times. Those results from the analyses indicate that the sentiments with stock information appearing on Big-data would integrate with changes in the stock market as they can be possibly estimated based on values from the past.

Key Words : Autocorrelation, Principle Component Analysis, VAR, Twitter, Big-data

[*] Graduate Student, Department of Business Adminstration, Hanyang University, Yidhs@naver.com
[**] Assistant Professor, Department of Business Adminstration, Hanyang University, hyoungkang@hanyang.ac.kr
[***] Quantitative Research Analyst, Quant Investment Team, Samsung Asset Management, soo_hyun.kim@samsung.com
[****] Corresponding Author, Assistant Professor, Department of Business Adminstration, Hanyang University, changmin74@hanyang.ac.kr