creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조선대학교
CHOSUN UNIVERSITY

[UCI]I804:24011-200000334342

2020년 8월
석사학위 논문

온라인 주식게시판 정보를 활용한 주가분석

조선대학교 대학원

경영학과

박 단 비

조선대학교
CHOSUN UNIVERSITY

온라인 주식게시판 정보를 활용한 주가분석

Quantifying Stock Price
using Social Media

2020년 8월 28일

조선대학교 대학원

경영학과

박 단 비

조선대학교
CHOSUN UNIVERSITY

온라인 주식게시판 정보를 활용한 주가분석

지도교수 오 갑 진

이 논문을 경영학 석사학위 신청 논문으로 제출함

2020년 5월

조선대학교 대학원

경영학과

박 단 비

조선대학교
CHOSUN UNIVERSITY

박단비의 석사학위논문을 인준함

위원장 조선대학교 교 수 이 계 원 (인)

위 원 조선대학교 교 수 이 현 철 (인)

위 원 조선대학교 교 수 오 갑 진 (인)

2020년 6월

조선대학교 대학원

조선대학교
CHOSUN UNIVERSITY

<목 차>

제1장 서 론 ---------- 1
제1절 연구동기 ---------- 1

제2장 선행연구 ---------- 11

제3장 연구모형 ---------- 17
제1절 회귀분석 ---------- 17
제2절 포트폴리오 분석 ---------- 18

제4장 자료 및 변수의 정의 ---------- 22
제1절 자료 ---------- 22
1. 온라인 주식게시판 자료 ---------- 22
2. 주식시장 관련 자료 ---------- 23
제2절 텍스트 데이터 분석 ---------- 24
1. 게시글의 센티멘트 분류 ---------- 24
2. 투자의견(Bullish) 변수 및 의견 불일치(Disagree)변수 ---------- 27

제5장 실증분석결과 ---------- 29
제1절 기술통계 ---------- 29
제2절 투자심리지표의 정보력 검증 ---------- 30
1. 전체기업에서 투자심리지표와 주식수익률, 변동성 간의 회귀분석 ---------- 30
2. 기업특성에 따른 투자심리지표와 주식수익률, 변동성 간의 회귀분석 ---------- 32
제3절 매수의견점수의 3분위 포트폴리오 분석 ---------- 33
1. 매수의견점수에 따른 단일정렬 포트폴리오 ---------- 33
2. 매수의견점수와 개인투자자 거래량에 따른 이중정렬

조선대학교
CHOSUN UNIVERSITY

포트폴리오 ---------- 36
3. 보유기간에 따른 정보력 검증 ---------- 38

제6장 결론 ---------- 40

[참고문헌] ---------- 44

[부록] ---------- 62

조선대학교
CHOSUN UNIVERSITY

<표 목 차>

<표 1> 매수 및 매도로 분류된 게시글 예시

<표 2> 회귀변수 기초통계량

<표 3> 위험요인 기초통계량

<표 4> 회귀분석 결과

<표 5> 표본 외 회귀분석 결과

<표 6> 개인투자자 그룹별 회귀분석 결과

<표 7> 개인투자자 그룹별 표본 외 회귀분석 결과

<표 8> 매수의견점수에 따른 3분위 포트폴리오

<표 9> 개인투자자 그룹별 매수의견점수 포트폴리오

<표 10> 보유기간에 따른 정보력 검증

조선대학교
CHOSUN UNIVERSITY

[그림 목 차]

[그림 1] 네이버 금융 종목토론실 게시글

[그림 2] 요일별 게시글 분포, 표본기간동안 시간대별 게시글 분포

[그림 3] 단일정렬 포트폴리오의 누적수익률

[그림 4] 개인투자자의 거래량과 매수의견점수에 따른 이중정렬 포트폴리오의 누적수익률

조선대학교
CHOSUN UNIVERSITY

ABSTRACT

Quantifying Stock Price using Social Media

Danbi Park
Advisor : Prof. Gabjin Oh, Ph.D.
Division of Business Administration,
Graduate School of Chosun University

In this paper, we analyse whether investor sentiment measured, bullishness and disagreement, from text data on social media has information power and predictive power on stock returns, and examine how the impact of sentiment differs according to the trading volume of individual investors. Through empirical studies, we realized that the statistically significant positive relationship between the stock return and bullishness, and volatility and disagreement. In addition, as a result of portfolio analysis constructed by according to bullishness score, all of portfolio have higher sharpe ratio than the market portfolio. These results have been more influential on companies with high trading volumes of individual investors. Our findings suggest that individual investors, who are under information asymmetry, use social media to obtain information, and the individual investors, known as irrational investors, will have information by sharing their opinion, which is influential as a market participant.

조선대학교
CHOSUN UNIVERSITY

제1장 서 론

제1절 연구동기

Fama(1970)의 효율적시장가설(efficient market hypothesis)에 따르면 마찰 없는 시장에서 정보 측면의 효율성에 의해 정보는 공시될 때 주가에 즉시 반영되기 때문에 투자자는 공개적으로 이용 가능한 정보로 주식시장을 예측하여 초과수익을 얻는 것이 불가능하다. Malkiel(1999)의 랜덤워크이론(random walk theory)은 주가의 변화는 과거의 패턴에 관계없이 독립적으로 움직이기 때문에 과거의 정보를 통해서 미래의 가격 변동을 예측 할 수 없다고 주장한다. 그러나 행동재무학(behavioral finance) 연구에 의하면 제한된 합리성을 가진 투자자(Hirshleifer et al, 2011)가 존재하고, 차익거래자의 활동이 제한되어 있는 상황이라면 비이성적 투자자의 투자심리가 자산가격에 영향을 미칠 수 있기 때문에(Shleifer and Summers, 1990; Kumar and Lee, 2006) 자산가격이 과대 혹은 과소평가 되더라도 즉시 진정한 가치로 수렴하지 않고 가격왜곡현상이 지속되어 이를 통해 초과수익을 얻을 수 있는 유인이 존재한다(Shiller et al, 1984; Shleifer, 1997). 현실에서는 공매도의 제약 등의 거래비용으로 인해 투자자의 차익거래 유인이 감소(De Long et al., 1990)되기 때문에 차익거래 활동에 제한이 있고 따라서 자산가격에 영향을 미칠 수 있는 투자심리(investor sentiment)를 어떻게 측정하는지가 중요한 연구주제 중 하나이다.

투자심리가 주식수익률에 미치는 영향을 분석한 기존의 연구들은 일반적으로 주식시장의 과거 데이터를 이용하여 투자심리의 대용치를 만들어 분석에 사용하여왔다. 예를 들어, Baker and Stein(2004)은 공매도 제약이 있고, 비이성적 투자자가 비공개적 정보를 과신(overconfidence)한다는 가정하에 거래회전율이 증가할 때 다음 연도의 인덱스 예상수익률이 감소한다고 제시하였다. Kumar and Lee(2006)은

조선대학교
CHOSUN UNIVERSITY

개인투자자들이 특정 주식을 매수 할 때 다른 주식들도 동시에 매수하는 경향이 나타남으로써 개인투자자의 매수-매도의 불균형(bull-sell imbalance)의 상관관계가 매우 높다는 사실을 발견하였고, 이러한 거래 패턴은 주식수익률에서 영향을 미친다는 것으로 나타난다. Baker and Wurgler(2006)은 기존에 투자심리 대용치로 사용되었던 변수 중 폐쇄형 펀드 할인율, 거래회전율, IPO의 첫 거래일의 평균 수익률, 배당금 프리미엄, 채권발행 대비 주식발행 비율을 이용하여 주성분분석(principal component analysis, PCA)을 통해 투자 심리지수(investor sentiment index)를 추정하였고, 투자심리지수와 향후 시장 수익률 간에 음의 관계가 존재한다는 결과를 얻었다. 이렇듯 주식시장의 과거 데이터를 이용한 투자심리 측정치가 주식수익률에 영향을 미친다는 결과가 기존의 연구들을 통해 잘 알려져 있지만 투자자의 투자 심리를 간접적으로 측정하였다는 한계가 존재한다.

이에 본 연구는 실시간으로 투자자들이 정보를 얻고 의견을 공유함으로써 직접적으로 투자 심리를 나타낼 수 있는 소셜 네트워크 서비스(social network service, 이하 'SNS'라 한다)의 텍스트 데이터를 통해 Antweiler and Frank(2004)가 제시한 매수의견점수와 의견불일치지수를 산출함으로써 투자자의 투자심리를 직접적으로 추정하고 주식시장에 미치는 영향을 살펴보고자 한다.

2000년대에 들어서면서 인터넷의 보급으로 그에 대한 접근성이 높아짐에 따라 사용자가 직접 정보를 생성하고 공유할 수 있게 되었고, 더불어 스마트 기기 대중화와 소셜미디어의 등장으로 데이터가 급증하면서 이른바 '빅 데이터' 시대에 도래함으로써 연구자들은 대량의 데이터에 직면하게 되었다. 특히, 글쓴이의 의견 및 감정을 직접적으로 나타낼 수 있는 텍스트 데이터의 양 또한 급증하게 되었고, 컴퓨터 성능 및 정보 분석 기술의 발전으로 비정형 데이터인 텍스트 데이터의 분석이 용이해지게 됨에 따라 온라인 상의 SNS 메시지나 구글검색, 뉴스 등의 비정형 데

조선대학교
CHOSUN UNIVERSITY

이터를 통해 주식시장을 설명하고자 하는 시도들이 활발하게 진행되고 있다. Preis et al.(2013)의 구글 트렌드(Google trend)의 비정상적 검색량이 주식시장이 불안정할 때 더욱 커진다는 연구결과에서 볼 수 있듯이 투자자는 주식을 매수 또는 매도하기 이전에 정보를 얻기 위해 인터넷을 사용한다는 것을 알 수 있다. Antweiler and Frank(2004)는 온라인 상에서 투자자들이 쓴 글을 이용해 주식시장에 미치는 영향을 분석한 연구에서 투자자들이 정보를 얻기 위해서 뿐만 아니라 정보를 공유하고, 이를 통해 형성된 투자심리가 주식시장에 영향을 미친다고 하였다. 이러한 연구 결과에서 알 수 있듯 투자자들은 투자를 하기 전 인터넷을 통해 정보를 수집하고, 자신의 생각이 맞는지 확인하기 위해 다른 사람들과 의견을 공유하여 형성된 투자심리가 주식시장에 영향을 미칠 수 있기 때문에 인터넷에서 생성된 글쓴이의 의견을 직접적으로 관찰 할 수 있는 텍스트 데이터의 분석이 중요하고 이러한 텍스트 데이터를 통한 투자심리 측정이 높은 관심을 받고 있다(Gentzkow and Taddy, 2019; Ke et al., 2019).

실제로 텍스트 데이터를 통한 투자심리 측정치가 이전의 시장데이터를 통한 측정치보다 주식수익률을 더 잘 설명한다고 최근 연구들에서 밝혀지고 있다. Jiang et al.(2018)은 기업의 재무보고서를 통해 경영자의 심리지수를 측정하였다. 보고서를 통해 측정된 심리지수와 기존의 연구에서 제시되었던 5가지 투자심리지수[1] 및 거시경제변수의 주식수익률에 대한 예측력을 비교분석한 결과 텍스트 데이터인 재무보고서를 통해 측정된 경영자의 심리지수가 가장 높은 예측력을 갖는다는 것을

[1] Jiang et al.(2018)의 재무보고서를 통해 측정된 경영자 투자심리지수와 비교분석하기 위해 사용된 기존의 투자심리측정치는 시장데이터를 통해 측정된 Baker and Wurgler(2006)의 투자심리지수((investor sentiment index), Huang et al.(2015)에 의해 조정된 투자심리지수(aligned investor sentiment index)와 설문에 기반하여 측정된 미시간대학교의 소비자심리지수(University of Michigan consumer sentiment index), 미국의 컨퍼런스 보드(Conference Board, CB) 소비자신뢰지수, 그리고 Da et al.(2015)의 검색량에 기반을 둔 투자심리지수(Financial and economic attitudes reveals by search(FEARS)이다.

조선대학교
CHOSUN UNIVERSITY

발견하였고, 이는 기존의 시장데이터를 통해 측정된 투자심리지표가 가지고 있지 않는 정보력을 텍스트데이터를 통해 측정된 투자심리지수가 추가적으로 더 가지고 있기 때문이라고 설명하고 있다. Cookson and Niessner(2020)은 주식수익률의 변동성과 같은 시장데이터로부터 측정된 투자자의 투자심리 대용치는 간접적인 측정치임을 지적하면서 투자자의 의견 불일치 정도가 거래량에 미치는 영향을 보기 위해 소셜 미디어 투자 플랫폼인 StockTwits의 게시글에서 측정된 투자심리를 매수의견과 매도의견으로 구분하여 투자자의 의견 불일치정도를 Antweiler and Frank(2004)가 제시한 의견불일치지수를 이용하여 측정하였다. 게시글 글쓴이가 제시하는 정보를 이용하여 글쓴이를 기술적분석 그룹과 기본적 그룹으로 구분하여 그룹 내 의견 불일치정도와 그룹 간 의견 불일치정도가 거래량과 갖는 관계를 분석한 결과 주식수익률의 변동성을 통제한 이후에도 의견 불일치정도가 거래량에 영향을 미치며 그룹 내 의견 불일치정도의 거래량에 대한 영향력이 더 높다는 것을 발견하였다.

이처럼 최근의 투자심리가 주식수익률에 미치는 영향을 분석한 선행연구들에 따르면 텍스트 데이터를 통해 직접적인 투자심리를 측정하는 것이 중요하다는 것을 알 수 있다. 특히 한국주식시장은 매매세력 중 개인 투자자의 비중이 높고(구맹회와 이윤선, 2001), 기관 투자자에 비해 상대적으로 정보 열위에 놓여있는 개인 투자자들은 정보를 얻기 위해 온라인상의 커뮤니케이션을 활발하게 이용할 수 있다. 실제로 한국에서 가장 높은 점유율을 가진 검색 엔진인 네이버[2]의 네이버 금융에서 제공하는 종목토론실의 2018년도부터 2019년까지의 하루 평균 게시글 수는 22,128개이고, 글쓴이 수[3]는 13,976명, 게시글의 조회 수는 12,747,945로 온라인

---

2) 웹사이트 분석사이트인 InternetTrend(http://www.internettrend.co.kr)에 의하면 2018년도부터 2019년도까지의 기간을 대상으로 조회한 국내 검색엔진의 점유비율은 네이버가 64.46%로 가장 높은 비중을 갖는 것으로 나타났다.

3) 네이버 금융의 종목토론실에서 제공하는 글쓴이에 대한 정보는 아이피 일부분과 아

조선대학교
CHOSUN UNIVERSITY

상의 소셜 네트워크를 통해 투자자들의 의견 공유가 활발하게 이루어 지는 것으로 나타났다. 이에 본 연구는 한국주식시장에서 개인 투자자들이 활발하게 이용하는 네이버 금융의 종목토론실 게시글을 통해 실시간으로 개인 투자자들이 상호작용함으로써 형성되는 투자심리를 측정하여 주식수익률과 변동성에 대한 정보력과 예측력을 가지고 있는지 분석하고, 온라인 주식게시판(이하 '게시판'이라 한다)에서 형성된 투자심리가 주식수익률에 미치는 영향력이 개인투자자의 거래량 크기에 따라 달라지는지를 알아보고자 한다.

텍스트 데이터에서 나타난 정보와 주식시장 간의 관계를 살펴보기 위해 가장 많이 사용되는 방법으로는 텍스트 데이터에서 형성되는 의견을 분석하는 오피니언 마이닝(opinion mining), 조금 더 세부적으로 글쓴이의 감정을 분석하는 센티멘트 분석(sentiment analysis)이 있다. 텍스트 데이터로부터 투자심리를 측정하기 위해 일반적으로 센티멘트 분석방법이 주로 이용되며, 이와 관련된 선행연구들은 크게 두 가지로 구분할 수 있다. 첫 번째는 기존에 있는 감정단어사전을 이용하여 텍스트 데이터의 센티멘트를 분류하는 경우이다(Tetlock, 2007; Loughran and McDonald, 2011). 이는 텍스트에 있는 단어가 감정사전에 존재하면 해당하는 감정을 부여하는 방식으로 가장 활발하게 사용되는 감정사전은 Loughran and McDonald(2011)가 개발한 주식시장 관련 긍정·부정 단어사전이 있다(Chen et al, 2014; Hillert and Muller, 2014; Calomiris and Mamaysky, 2019). 그러나 기존의 감정사전을 이용하여 현재 시점의 텍스트의 감정을 분석하는 것은 지난 연구로부터 추정되어진 모델을 현재 시점의 새로운 데이터에 적용시키는 것과 유사하다는 한계가 있다(Ke et al., 2019). 더불어 기존의 감정사전을 이용하여 센티멘트를 분석하는 경우 일반적인 감정사전보다 각 분야별로 특화된 감정사전이 텍스트의

이디 일부분으로 본 연구에서는 위 두 가지가 모두 일치하는 경우 동일한 글쓴이로 가정한다.

조선대학교
CHOSUN UNIVERSITY

센티멘트를 비교적 더 잘 분류한다(Li, F, 2010; Henry and Leone, 2016). 영어권 국가는 활발하게 사용되는 주식시장 관련 감정사전이 있는데 반해, 한국어의 경우 주식시장에 기초한 감정사전의 부재로 주식시장 관련 텍스트를 분석하는데 한계가 있다.

두 번째로 텍스트의 센티멘트가 주식시장에 미치는 영향을 분석하기 위해 텍스트의 센티멘트를 측정하는 방법으로는 머신러닝(machine learning)이 있다. 머신러닝은 감정사전의 존재여부에 의존하지 않고 텍스트의 센티멘트를 분석할 수 있는 방법으로 데이터의 양이 많아지고 컴퓨터 기술발전으로 보다 적은 계산으로 많은 양의 데이터를 다룰 수 있게 됨에 따라 머신러닝 방법은 더욱 각광받고 있다(Ke et al., 2019). 이에 외국에서는 게시글, 뉴스 등 다양한 텍스트 데이터의 센티멘트를 머신러닝 방법을 통해 분석함으로써 텍스트 데이터에서 관찰되는 투자심리가 주식시장과 갖는 관계를 분석하는 연구가 증가하고 있다(Das and Chen, 2007; Frank and Sanati, 2018; Cookson and Niessner, 2020). 본 연구에서도 금융시장에 특화된 감정사전이 없다는 한국 금융시장 텍스트 분석의 한계점을 머신러닝의 분류방법을 이용함으로써 극복하고자 한다. 더불어 감정사전에 기반하여 텍스트의 센티멘트를 분류하는 방법은 감정사전에 존재하지 않는 단어에 대해서는 분류하지 못함으로(Kraus and Feuerriegel, 2017) 인터넷 상의 용어가 주로 사용되는 글을 분석할 때에는 감정사전에 존재하지 않는 단어에 대한 정보를 잃게 되어 센티멘트 분류를 정확하게 하지 못할 수 있다. 반면, 머신러닝 방법을 통한 텍스트의 센티멘트 분류는 특정 단어에 집중하기 보다는 텍스트의 전반적인 맥락을 파악하여 텍스트에 사용된 단어들의 정보를 잃지 않고 분석에 사용할 수 있기 때문에(Kearney and Liu, 2014) 본 연구에서 분석에 사용할 온라인 상의 게시글의 센티멘트를 분류하기 위해 머신러닝을 통한 분류방법을 사용하고자 한다.

조선대학교
CHOSUN UNIVERSITY

따라서 본 연구에서는 대용량 데이터에 대한 처리 속도와 학습 능력이 높다고 평가되며 텍스트 분류 뿐만 아니라 다양한 분류 문제를 잘 해결하는 것으로 알려져 있는 서포트 벡터 머신(support vector machine, SVM)[4]을 분류모델로 이용하여 게시글의 센티멘트를 분류하고, Antweiler and Frank(2004)의 논문에서 제시하고 있는 매수의견점수와 의견불일치지수를 투자심리지표로써 사용한다. 텍스트의 센티멘트를 분류하는 과정은 다음과 같이 요약된다. 첫째로 분석대상인 텍스트를 분석을 위해 단어벡터로 변환하는 과정을 거친다. 두번째로 게시글 중 개별 기업의 일별 수익률의 상·하위 극단치에 해당하는 수익률이 발생한 날에 쓰인 게시글을 추출하여 학습데이터를 구성한 후 분류모델을 학습시킨다. 학습데이터를 통해 학습된 분류모델을 이용하여 모든 게시글의 센티멘트를 분류한다. 마지막으로 각 기

---

4) 서포트 벡터 머신은 데이터 샘플들을 n차원의 벡터 $x=(x_{1,\cdots,}x_n)$로 보았을 때 데이터 샘플을 두 범주로 분리하는 초평면을 찾는 방법으로, 초평면과 가장 가까운 데이터 샘플과의 거리를 나타내는 초평면의 마진(margin)이 최대화되는 초평면을 찾도록 학습되는 이진 분류기이다. 초평면을 식으로 나타내면 다음과 같다.

$$f(x)=w \cdot x + b=0$$

여기서 w는 초평면의 법선 벡터이고, b는 초평면과 원점의 거리로 데이터 샘플x를 위 식에 넣은 결과의 부호가 $f(x)>0$ 또는 $f(x)<0$값을 가짐으로 이진분류가 수행된다. 부호가 음수가 나오는 데이터 샘플 범주를 -1로 정의하고 반대로 부호가 양수가 나오는 데이터 샘플의 범주를 1로 정의할 때 각 범주에서 주어진 초평면과 가장 가까운 데이터 샘플을 나타내는 서포트 벡터는 다음과 같이 정의된다.

X^+:+1 범주에 속하는 데이터 중에서 초평면과 가장 가까운 데이터
X^-:−1 범주에 속하는 데이터 중에서 초평면과 가장 가까운 데이터

마진이 최대화되는 초평면을 찾기 위해서는 각 범주의 서포트 벡터를 지나는 초평면 사이의 거리를 최대화 시켜야 한다. 주어진 초평면과 같은 법선 벡터를 지니면서 서포트 벡터를 지나는 초평면은

$$w \cdot x^- + b=1$$
$$w \cdot x^+ + b=-1$$

으로 나타낼 수 있고, 이를 이용하여 두 초평면 사이의 거리인 마진 d는 다음과 같이 계산할 수 있으며 학습 데이터를 통해 이를 최대화시키는 초평면을 찾는 것이 서포트 벡터 머신의 목표이다.

$$d=\frac{2}{||w||}$$

조선대학교
CHOSUN UNIVERSITY

업의 센티멘트가 분류된 게시글을 이용하여 Antweiler and Frank(2004)의 투자심리지표를 일별로 측정한다. 자세한 분류과정은 부록에 제시하였다.

본 연구는 네이버 금융에서 제공하는 종목토론실의 2018년 1월 2일부터 2019년 12월 30일까지의 각 기업의 게시글을 중심으로 표본기간 동안 유가증권시장과 코스닥시장에서 지속적으로 거래되고 있는 상장기업 중 금융업을 제외한 기업들을 대상으로 거래일 평균 게시글 수가 10개 이상인 350개 기업을 표본기업으로 선정하여 실증분석을 실시하였다. 주요 분석 초점은 다음과 같다. 먼저 게시판에서 형성된 투자심리가 국내 주식시장에 유의한 영향을 미치는 투자행동의 채널이 될수 있는지 알아보기 위해 주식수익률과 변동성을 종속변수로, 투자심리지표를 독립변수로 놓고 회귀분석을 실시한다. 그리고 기관투자자에 비해 상대적으로 제한된 정보수집능력을 가지고 있기 때문에 게시판의 주된 이용자로 예상되는 개인투자자의 거래량이 많은 주식일수록 온라인에서 형성된 투자심리가 수익률과 변동성을 더 잘 설명하는지 조사한다. 더불어 게시글을 통해 측정된 매수의견점수가 주식수익률에 대해 예측력을 갖는지 확인하기 위해 동일가중치와 가치가중치를 이용하여 3분위 포트폴리오를 구축한 후 샤프비율(Sharpe ratio)과 Capital Asset Pricing Model(CAPM), Fama-French 3요인 모형(1993, 이하 'FF3 모형'이라 한다), Fama-French 5요인 모형(2015, 이하 'FF5 모형'이라 한다)으로 위험을 조정한 초과수익률을 통해 투자성과를 살펴보고, 개인투자자의 거래량이 많은 기업으로 구축된 포트폴리오일수록 수익률이 증가하는지 분석한다. 마지막으로 매수의견점수로 측정된 게시판의 투자의견이 갖는 정보력이 어느 정도 지속되는지 알아보기 위해 개별 기업의 매수의견점수를 측정하는 기간과 포트폴리오를 보유하는 기간을 1일, 2일, 3일, 1주일, 2주일, 그리고 1개월로 늘려가며 포트폴리오 수익률이 어떻게 달라지는지 확인한다.

조선대학교
CHOSUN UNIVERSITY

주요 실증 분석결과는 다음과 같다. 첫째, 게시판에서 형성된 센티멘트를 이용하여 측정한 투자심리지표인 매수의견점수와 의견 불일치정도는 주식수익률 및 변동성과 통계적으로 유의한 양(+)의 관계를 갖는 것을 확인하였다. 즉, 게시글에 매수의견이 많을수록 당일과 그 다음날까지의 수익률이 통계적으로 유의하게 상승하고, 의견 불일치정도가 커질수록 당일과 다음날의 변동성이 유의적으로 증가한다는 것으로 나타났다. 이는 게시판의 글이 주식시장에 대해서 정보력을 가지고 있음을 의미한다. 둘째, 개인투자자의 거래량이 많은 기업에서 게시판에서 측정된 투자심리지표의 주식수익률과 변동성에 대한 설명력이 더 높아짐을 확인하였다. 이는 기관투자자에 비해 상대적으로 정보적 열위에 놓여있는 개인투자자가 게시판의 정보를 투자의사결정에 사용할 가능성이 높음을 의미한다. 더불어 개인투자자가 많은 기업일수록 일반적으로 비합리적 투자자로 알려진 개인투자자들의 투자에 대한 반응이 게시판에서의 상호작용을 통해 집단지성의 한 형태를 띄며 개인투자자 집단이 시장 참여자로써 주식가격 형성에 영향을 줄 수 있음을 시사한다. 셋째, 게시글을 통해 측정된 매수의견점수를 이용하여 구축한 포트폴리오는 위험요인을 조정한 경우 초과수익률의 통계적으로 유의한 양(+)의 값을 갖지만 위험요인에 의한 설명력은 매수의견점수가 높은 기업들로 구성된 포트폴리오일수록 낮아졌고, 샤프비율의 경우 시장포트폴리오 대비 다음 기간에 더 높은 샤프비율을 가졌다. 이는 게시판에서 형성된 매수의견의 투자심리가 위험요인에 의해서는 설명되기 어려우나 주식수익률에 대해 예측력은 존재함을 나타낸다. 넷째, 샤프비율로 측정한 주식수익률에 대한 게시글의 매수의견점수 포트폴리오의 투자성과는 개인투자자의 거래량이 많은 기업에서 더 높게 나타났다. 즉, 게시글에서 형성된 투자의견은 개인투자자가 많은 기업집단과 적은 기업집단 각각에서 다른 영향력을 가지고 있고, 개인투자자가 많은 기업집단에서 그 영향력이 증가함을 알 수 있다. 마지막으로

조선대학교
CHOSUN UNIVERSITY

포트폴리오의 보유기간이 짧을수록, 매수의견점수를 측정하는 기간은 길수록 포트폴리오의 시장포트폴리오 대비 샤프비율이 증가한다. 이를 통해 게시판에서 형성되는 투자의견은 추세가 형성되는 경향이 있고 이를 이용한 주식수익률에 대한 예측력은 단기간일수록 높아진다는 것을 의미한다.

본 연구는 다음과 같은 측면에서 기존연구에 추가적인 기여를 할 것으로 판단된다. 첫째, 한국 금융시장에서 통용되는 감정사전이 존재하지 않아 일반적인 텍스트의 센티멘트 분석에서 사용하는 사전 기반 방법을 사용할 수 없는 경우 데이터 특성에 맞는 학습 데이터를 구축하는 방법을 제안한다. 둘째, 개인투자자가 큰 비중을 차지하고 있는 한국 주식시장에서 비합리적 투자자로 알려진 개인투자자의 의견이 게시판을 통해 집단지성으로써 역할을 발견했다는데 의의가 있다. 마지막으로 게시판에서 형성되는 투자심리에 영향을 받는 기업의 특성을 제시함으로써 한국 주식시장에서 기업의 정보 환경에 대한 이해를 높였다는 측면에서 공헌점이 있다.

논문은 다음과 같이 구성되어 있다. 제 2장에서는 본 연구와 관련된 선행연구를 검토하고, 제 3장은 연구방법론과 실증분석에 사용된 자료와 변수 구축 방법에 대해 서술한다. 제 4장에서는 실증분석결과를 구체적으로 제시하고, 마지막으로 제 5장에서는 본 논문을 요약하고 결론 및 시사점을 언급한다.

조선대학교
CHOSUN UNIVERSITY

제2장 선행연구

현실에서는 거래비용 등으로 인해 차익거래가 제한될 수 있고, 비합리적 투자자가 존재하기 때문에 투자자들의 투자심리가 자산가격에 영향을 미칠 수 있다고 알려져 있다(Shleifer and Summers, 1990). 이에 투자자의 심리를 어떻게 측정하는지가 중요한 연구 주제 중 하나이고, 투자자의 투자심리지표를 구축하는 방법에 따른 연구동향은 다음과 같다.

첫째, 주식시장의 과거 데이터에 기반을 둔 방법으로 거래량, 폐쇄형 펀드 할인율, 거래회전율, IPO의 첫 거래일의 평균 수익률 등을 개별적으로 이용하거나 Baker and Wurgler(2007)가 제시한 것처럼 변수들로부터 주성분을 분석하여 투자자의 투자심리를 측정하는 방법이 있다. Baker and Stein(2004)은 공매도에 제약이 있고, 비합리적인 투자자가 비공개적 정보를 과신하는 환경에서 거래회전율이 증가함에 따라 다음 연도의 인덱스 예상수익률이 감소한다고 보고하였다. Kumar and Lee(2006)은 미시 거래 자료인 개인투자자의 매수, 매도 거래량을 이용하여 투자자의 투자심리를 측정하였는데, 개인투자자의 매수-매도 거래량은 불균형적으로 발생하며 이러한 거래 패턴은 주식수익률에서 영향을 미친다는 것을 제시하였다. 더불어 설문조사를 통해 구축된 투자심리지수를 이용해서도 주식시장과의 관계를 분석한 선행연구들이 존재하는데, Lemmon and Portniaguina(2006)은 미시간대학의 소비자심리지수(University of Michigan Consumer Sentiment Index)와 미국의 컨퍼런스 보드(conference board, CB) 소비자신뢰지수를 이용하여 분석한 결과 투자심리지수는 개인투자자의 거래량이 많은 기업의 주식수익률을 예측함을 보였고, 소비자신뢰지수를 투자심리지수로 사용한 Schmeling(2009)은 Hofstede(2003)이 제안한 불확실성회피지수(uncertainty avoidance index; UAI)가 높은 국가일수록, 그리고 집단주의가 강한 국가일수록 투자자의 투자심리가 미래 주식수익률에

조선대학교
CHOSUN UNIVERSITY

미치는 영향이 커진다고 주장하였다. 이렇듯 주식시장 데이터를 이용한 투자심리 측정치가 주식수익률에 영향을 미친다는 결과가 기존 연구들을 통해 잘 알려져 있지만, 이는 투자자의 투자심리를 시장 데이터를 이용해 간접적으로 측정한 측정치이기 때문에 많은 경제 요인에 영향을 받을 수 있다는 한계가 존재한다.

둘째, 글쓴이의 심리를 직접적으로 나타낼 수 있는 텍스트 데이터 양의 급증과 컴퓨터 성능 개선 및 분석 기술 발전에 따라 비정형 데이터인 텍스트 데이터를 보다 적은 계산 비용으로 많은 양을 다룰 수 있게 되었고, 이에 텍스트 데이터를 통해 투자심리지표를 구축하고자 하는 접근이 증가하였다. 이러한 접근법은 두 가지로 구분될 수 있다.

하나는 텍스트에서 나타나는 의견을 분석 할 때 긍정어와 부정어 등의 감정 단어 목록을 먼저 구축한 뒤 이를 통해 텍스트의 센티멘트를 측정하고 주식시장과의 관계를 살펴보는 연구들이 존재한다. 초기에는 모든 분야에서 사용될 수 있는 공용감정사전인 하버드의 감정사전(Harvard IV-4 dictionaries)을 이용하여 텍스트 데이터의 센티멘트를 측정하고 주식수익률과의 관계를 분석하였다(Tetlock,2007; Tetlock 2008; Kothari and Short, 2009). 그러나 이러한 공용감정사전을 이용한 텍스트의 센티멘트 분류는 충분한 정확도를 주지 못한다는 한계가 있었고(Li, 2010), 이에 Loughran and McDonal(2011)는 주식시장에 특화된 감정사전을 구축하여 제시하였다. Loughran and McDonal의 감정사전과 공용감정사전인 하버드 감정사전으로 센티멘트를 분류하여 투자심리를 측정하고 주식수익률에 대한 예측력을 비교분석을 실시한 결과, 주식시장에 특화된 감정사전으로 센티멘트를 분류한 경우에 예측력이 높다는 것이 확인되었고(Henry and Leone, 2016), 현재 Loughrna and McDonal의 감정사전은 영어권 국가에서 연구에 활발하게 사용되고 있다. Loughran and McDonal의 감정사전을 텍스트 데이터의 센티멘트 분류에 사용한 연

조선대학교
CHOSUN UNIVERSITY

구로 Chen and Hwang(2014)는 소셜미디어 기사의 부정적 단어는 미래의 주가 수익률과 어닝서프라이즈(earning surprise)를 예측한다는 사실을 발견하였다. Garcia(2013)는 뉴욕타임즈의 금융분야 뉴스 기사의 센티멘트를 측정하여 뉴스 기사의 긍정 및 부정적인 단어들의 주식수익률에 대한 예측력은 불황에 더 높아진다는 것을 발견하였다. Hillert and Müller(2014)는 뉴스의 센티멘트는 모멘텀 효과의 크기에 상당한 역할을 하고 투자자의 의견과 상호관계를 가지고 있음을 제시하였고, Hillert and Müller(2018)은 미국 지역신문들의 기사들의 센티멘트를 측정하여 기자들의 의견 불일치가 주식시장과 갖는 관계를 살펴본 결과 의견 불일치는 시장수익률을 부정적으로 예측하고 이러한 효과는 불황에서 더 커진다는 것을 확인하였다. 그러나 기존의 감정사전을 이용하여 현재 시점의 텍스트의 감정을 분석하는 경우 과거에 존재하지 않았던 단어나 시기에 따라 다른 감정으로 사용될 수 있는 단어에 대해서는 정확한 분류를 할 수 없기 때문에(Kraus and Feuerriegel, 2017; Ke et al., 2019) 인터넷 상에서 주로 사용되는 단어가 포함된 텍스트를 분석하는 데는 어려움이 있을 수 있고, 한국어의 경우 주식시장에 기초해 통용되는 감정사전이 존재하지 않기 때문에 감정사전을 기반으로 텍스트 데이터의 투자심리를 분석하는데 한계가 있다.

다음으로 기존의 감정사전이 존재하지 않는 경우에도 텍스트의 센티멘트를 분류할 수 있는 방법으로 머신러닝이 있다. 머신러닝은 감정사전의 존재여부에 의존하지 않을 뿐 아니라, 컴퓨터 기술발전으로 인해 많은 양의 비정형 데이터를 용이하게 다룰 수 있게 됨에 따라 더욱 각광받고 있다(Ke et al., 2019). 이에 따라 텍스트 데이터의 센티멘트를 머신러닝의 분류모델을 통해 분류함으로써 투자심리를 측정하고 주식수익률과 갖는 관계를 분석하는 연구가 증가하고 있다. 온라인 상의 게시글을 분석한 초기 연구인 Antweiler and Frank(2004)는 게시글의 센티멘트를

조선대학교
CHOSUN UNIVERSITY

분류한 후 매수의견점수와 의견 불일치정도를 측정하여 주식수익률과 변동성에 미치는 영향을 살펴 본 결과 게시글의 매수의견점수가 높을수록 다음날의 수익률이 하락하는 것을 발견하였고, Das and Chen(2007)은 게시판의 개별 게시글의 센티멘트보다 게시글을 종합하여 측정한 센티멘트가 더 정확하며, 주식시장의 움직임은 게시글의 주 이용자인 개인 투자자의 센티멘트와 게시판이 활발한 정도와 관련이 있음을 보였다. 뉴스기사의 센티멘트를 분류하여 주식시장과의 관계를 분석한 Frank and Sanati(2018)는 주식시장은 긍정적인 뉴스에 대해서는 과대반응을 보이고, 부정적인 뉴스에 대해서는 과소반응을 보이는 경향이 있음을 발견하였다. StockTwits 메시지를 투자심리 측정에 사용한 Giannini and Shu(2018)는 트위터 글에서 형성된 센티멘트는 주가수익률에 대해 음(-)의 예측력을 확인하였다. 마찬가지로 StockTwits의 메시지를 투자자들의 투자심리를 측정하는데 사용한 Cookson and Niessner(2020)은 메시지의 센티멘트를 분류한 후, Antweiler and Frank(2004)가 제시한 투자심리지표 중 하나인 의견 불일치정도를 사용하여 거래량과의 관계를 살펴보았다. StockTwits의 이용자를 기술적 투자자 집단과 기본적 투자집단으로 구분한 후 집단 내 의견 불일치 정도와 집단 간 의견 불일치정도가 거래량과 갖는 관계를 분석한 결과 같은 투자성향을 갖는 투자자 집단 내에서의 의견 불일치가 다른 투자성향을 갖는 투자자 집단 간 의견 불일치보다 거래량과 큰 관계가 있다는 것을 발견하였다. 마지막으로 Giannini and Shu(2019)는 StockTwits에 뉴스데이터까지 추가하였는데, 각각의 데이터 그룹에서 센티멘트를 측정한 후 두 그룹의 센티멘트를 비교하여 투자자들의 의견일치정도를 측정하였다. 투자자들의 의견일치정도가 이익공시일의 수익률에 미치는 영향을 분석한 결과 이익공시일 전 투자자들의 의견 불일치는 이익공시일에 높은 거래량과 이익공시일 후 낮은 수익률을 야기하고, 동질적이던 의견이 이익공시일에 의견 불일치로 바뀌면 양(+)의

조선대학교
CHOSUN UNIVERSITY

초과수익률을, 반대로 이질적이던 의견이 동질적 의견으로 바뀌면 음(-)의 초과수익률을 유발함을 확인하였다.

주식시장에 특화된 감정사전이 존재하는 외국과 달리 한국은 주식시장에 특화된 감정사전이 부재하여 텍스트 데이터를 통해 주식시장을 분석하는 연구가 활발하게 이루어지고 있지는 않다. 텍스트 데이터를 주식시장 분석에 사용한 국내 연구인 남달우 외 4인(2012)은 네이버 금융의 종목토론실 게시글을 이용하여 투자심리를 측정하였는데, 글쓴이들이 투자의견을 선택한 글만을 분석대상으로 하여 투자의견별로 게시글 수를 투자심리로 보고 주식수익률과의 관계를 분석하였다. 분석결과 강력매도의견과 매도의견은 수익률에 음(-)의 영향을 미치고 강력매수의견과 매수의견은 수익률에 양(+)의 영향을 미치며, 강력매도의견과 강력매수의견이 매도의견과 매수의견보다 영향력이 크다는 것을 발견하였다. 김영민 외 2인(2014)은 온라인 상의 게시글을 구글번역기를 사용하여 영어로 번역한 후 Loughran and McDonal의 감정사전을 사용하여 센티멘트를 분류 후 다음날의 주식가격 등락을 예측하는지 분석하였고, 정지선 외 2인(2015)은 뉴스데이터를 분석대상으로 하여 주식가격이 상승한 날의 뉴스는 긍정적 의미가 있는 것으로 간주하고, 반대로 주식가격이 하락한 날의 뉴스는 부정적인 의미가 있는 것으로 간주하여 각 뉴스기사에서 나온 단어의 출현 빈도를 이용하여 긍정어·부정어 단어목록을 연구자가 직접 구축한 후 뉴스의 센티멘트가 해당일의 주가등락과 일치하는지 확인하였다. 최근에는 머신러닝 방법이 각광을 받으며 주식시장에서 통용되는 감정사전이 없는 한국 주식시장에 대해 머신러닝 분류모델을 사용한 시도로써 김용석과 조성욱(2019)은 기업공개(IPO)를 하는 기업이 제출한 증권발행신고서의 센티멘트를 분석하기 위해 학습데이터로 사용될 텍스트를 임의로 추출한 뒤 연구자가 직접 센티멘트를 분류하여 학습데이터를 구축한 후 머신러닝 분류모델을 학습시켜 모든 텍스트의

조선대학교
CHOSUN UNIVERSITY

센티멘트를 분류하는데 사용하였다. 그 결과 증권신고서에서 측정된 센티멘트가 부정적인 회사일수록 공모가격 변화율(offer price revision)이 크게 나타나지만, 증권신고서의 센티멘트가 상장 직후 수익률과는 통계적으로 유의한 관계가 나타나지 않음을 확인하였다.

조선대학교
CHOSUN UNIVERSITY

제3장 연구모형

제1절 회귀분석

본 연구에서는 게시판에서 형성된 투자심리를 나타내는 매수의견점수($Bullish$)와 의견 불일치정도($Disagree$)가 주식수익률과 변동성에 대해 정보력이 있는지 확인하기 위해 회귀분석을 실시한다.

$$Return_{i,t} = a + \beta_1 Bullish_{i,t} + \beta_2 Disagree_{i,t} + \beta_3 Message_{i,t} + \beta_4 MR + \beta_5 Size_{i,t} + \beta_6 MB_{i,t} + u_{i,t} \quad (1)$$

$$Volatility_{i,t} = a + \beta_1 Bullish_{i,t} + \beta_2 Disagree_{i,t} + \beta_3 Message_{i,t} + \beta_4 Volume_{i,t} + \beta_5 Size_{i,t} + \beta_6 MB_{i,t} + u_{i,t} \quad (2)$$

$Return_{i,t}$: 일별 수정주가를 이용하여 계산된 기업 i의 일별수익률
$Volatility_{i,t}$: 기업 i의 과거 17일간의 일별 수정주가를 이용하여 계산된 일별 변동성
$Bullish_{i,t}$: 게시글의 센티멘트를 이용하여 계산된 기업i의 일별 매수의견점수
$Disagree_{i,t}$: 게시글의 센티멘트를 이용하여 계산된 기업i의 일별 의견불일치도
$Message_{i,t}$: 기업i의 일별 게시글 수에 log를 취한 값
MR_t : $KOSPI$ 인덱스의 일별수익률
$Size_{i,t}$: 기업 i에 대한 전년도 6월말 시점의 보통주 시가총액에 log를 취한 값으로 계산된 기업규모
$MB_{i,t}$: 기업 i에대한 시장가치 대 장부가치 비율로써 시장가치는 전년도 6월말 시점의 시가총액 장부가치는 전년도 말의 총 자본에서 우선주 자본금을 차감한 값
$Volume_{i,t}$: 기업 i의 일별 거래

개인투자자는 기관투자자에 비해 상대적으로 정보적 열위에 놓여있기 때문에 정보를 얻고 의견을 공유하기 위해 게시판을 사용할 것이고, 이에 게시판의 주된 이용자는 개인투자자일 것이다. 이에 본 연구는 개인투자자의 거래량이 많은 주식일수록 온라인에서 형성되는 투자심리지표가 수익률과 변동성에 미치는 영향력이 증가하는지 확인하기 위해 분석대상기업을 개인투자자 거래량 크기에 따라 3분위로 구분한 후, 각 집단별로 위 식(1)과 식(2)의 회귀분석을 수행한다. 본 연구에서는

조선대학교
CHOSUN UNIVERSITY

기업에 따른 거래량 크기를 비교하기 위해 개별 기업의 전체 거래량 수 대비 개인 투자자의 거래량 수로 나타낸 비율을 사용하였다.

본 연구는 산업에 따른 효과를 통제하고자 산업더미변수를 연구모형에 고려하였고, 표본기간 동안 월별로 전체 기업에 미칠 수 있는 영향을 통제하기 위해 월별 더미변수를 연구모형에 고려하였다. 오차항의 이분산성과 자기상관성을 해결하기 위해 Newey and West(1987)방법으로 표준오차를 조정하였다.

제2절 포트폴리오 분석

온라인 게시글의 투자심리지표와 주식수익률 및 변동성과의 관련성을 분석하여 정보력을 확인하고, 온라인 게시글에서 측정된 투자심리지표 중 매수의견점수가 주식수익률에 대해 예측력이 있는지를 조사하기 위해 매수의견점수를 기반으로 3분위 포트폴리오를 구성하여 각 포트폴리오별 표본 기업의 시장가치로 가중평균한 일별 수익률과 동일가중치로 평균한 일별 수익률을 살펴본다. 시장포트폴리오 대비 투자성과를 측정하기 위해 시장포트폴리오와 매수의견점수 포트폴리오의 샤프 비율을 비교분석하고, 각 포트폴리오의 일별 수익률이 위험 요인을 통제한 후에도 초과수익률을 얻는지 확인하기 위하여 CAPM과 FF3 모형 및 FF5 모형으로부터 추정된 알파를 확인한다.

$$R_{p,t} - R_{f,t} = \alpha_p + b_1(r_{m,t} - r_{f,t}) + \epsilon_{p,t} \quad (3)$$

$$R_{p,t} - R_{f,t} = \alpha_p + b_1 MKT_t + b_2 SMB_t + b_3 HML_t + \epsilon_{p,t} \quad (4)$$

$$R_{p,t} - R_{f,t} = \alpha_p + b_1 MKT_t + b_2 SMB_t + b_3 HML_t + b_4 RMW_t + b_5 CMA_t + \epsilon_{p,t} \quad (5)$$

조선대학교
CHOSUN UNIVERSITY

식(3)~(5)의 포트폴리오 모형에서 위험요인 MKT, SMB, HML, RMW, CMA은 다음과 같이 구성된다. 먼저 시장요인인 MKT는 시장포트폴리오의 위험프리미엄을 나타내는 것으로 KOSPI 200 인덱스의 수익률에서 무위험 수익률을 차감하여 계산되었고, 무위험 수익률은 354일물 통화안정증권의 금리를 사용하였다. FF3 모형에서 제시된 기업규모 요인인 SMB와 장부가-시가요인 HML은 다음의 방법으로 생성되었다. 먼저, 전년도 6월 말 시장가치를 기준으로 전체기업(KOSPI와 KOSDAQ)을 정렬한 후 시장가치의 중위수를 기준으로 2개의 포트폴리오(B, S)를 구성한다. 또한 전년도 말의 장부가치와 전년도 6월 말의 시장가치를 이용하여 계산한 장부가/시가 비율(BM)을 기준으로 전체기업을 상위30%, 중위 40%, 그리고 하위 30%에 속하는 기업들로 3개의 포트폴리오를 구성한다(H, M, L). 이후 기업규모요인 SMB와 장부가-시가요인 HML을 구성하기 위해 기업규모 기준 2개의 포트폴리오와 장부가/시가 비율 기준 3개의 포트폴리오를 서로 교차하여 6개의 기업규모-장부가/시가 포트폴리오(BH, BM, BL, SH, SM, SL)를 구성한 후, SMB는 기업규모가 작은 3개의 포트폴리오의 수익률을 단순 평균한 값에서 기업규모가 큰 3개의 포트폴리오의 수익률을 단순 평균한 값을 차감함으로써 계산하고((SL+SM+SH)/3-(BL+BM+BH)/3), HML은 BM을 기준으로 BM이 상위 30%인 포트폴리오와 하위 30인 포트폴리오를 사용하여 SMB를 계산하는 방식과 동일하게 BM이 상위 30%인 포트폴리오에 해당하는 포트폴리오의 수익률을 단순 평균한 값에서 BM이 하위 30%인 포트폴리오에 해당하는 포트폴리오의 수익률을 단순 평균한 값을 차감하여 구한다((SH+BH)/2-(SW+BW)/2). FF5 모형에서 제시된 수익성 위험요인 RMW와 자본투자 위험요인 CMA도 위와 유사한 방식으로 구축된다. 먼저 장부가/시가 비율과 동일한 방법으로 수익성((영업이익-비금융이자비용)/총자본)을 기준으로 3개의 포트폴리오(R, M, W)를 구축하고, 자본투자를 나타내는 총자산증가율을 기준으로 3개의 포트폴리오(A, M, C)를 구성한다.

조선대학교
CHOSUN UNIVERSITY

그리고 이를 이용하여 기업규모와 수익성에 따른 각각의 포트폴리오를 서로 교차하여 6개의 포트폴리오(BR, BM, BH, SR, SM, SW)를 만들고, 기업규모와 자본투자에 따른 각각의 포트폴리오를 서로 교차하여 6개의 포트폴리오(BA, BM, BC, SA, SM, SC)를 구성한다. 이 후 RMW는 수익성을 기준으로 수익성이 상위 30%인 포트폴리오의 수익률의 단순 평균값에서 수익성 하위 30%에 해당하는 포트폴리오의 수익률을 단순 평균한 값을 차감하여 구하고((SR+BR)/2-(SW+BW)/2), CMA는 자본투자가 가장 낮은 하위 30%의 포트폴리오의 수익률을 단순 평균한 값에서 자본투자가 가장 높은 상위 30%의 포트폴리오의 수익률의 단순 평균값을 차감하여 계산한다((SC+BC)/2-(SA+BA)/2).

한편 게시판의 주된 이용자로 예상되는 개인투자자의 거래량이 많은 기업으로 구축된 포트폴리오일수록 매수의견점수가 수익률에 미치는 영향력이 증가하여 예측력이 커지는지 분석하기 위해 개인투자자의 거래량 크기를 기준으로 분석대상기업을 3개 집단으로 분류한 후 각 집단을 다시 매수의견점수로 3등분하여 구축한 이중정렬 포트폴리오가 시장포트폴리오 대비 성과가 나타나는지 샤프비율을 통해 확인한다. 또, 각각의 집단 별로 CAPM, FF3 모형, FF5 모형으로 위험 요인을 조정한 이후에도 초과수익률이 나타나는지를 살펴본다. 식(3)~(5)를 통한 분석을 실시함에 있어서도 마찬가지로 오차항의 이분산성과 자기상관성을 해결하기 위해 Newey and West(1987)방법으로 표준오차를 조정하였다.

마지막으로 게시판에서 매수의견점수로 측정된 투자심리지표의 정보력이 어느 정도 지속되는지 알아보고자 한다. 이를 위해 개별 기업별로 매수의견점수를 측정하는 기간과 매수의견점수에 따라 구축된 포트폴리오를 보유하는 기간을 1일, 2일, 3일, 1주일, 2주일, 그리고 1개월로 하여 각각의 기간을 교차해가며 포트폴리오 수익률을 계산하여 각 포토폴리오별로 샤프비율이 어떻게 달라지는지를 확인한다.

조선대학교
CHOSUN UNIVERSITY

이 때 각 포트폴리오의 시장 포트폴리오 대비 얻는 투자성과를 비교하기 위해 각 포트폴리오의 샤프비율에서 시장포트폴리오의 샤프비율을 차감한 값을 분석에 사용한다. 이를 식으로 나타내면 다음과 같다.

$$ExtraSR_{b,h} = BullishPf_SR_{b,h} - MarketPf_SR_{h} \quad (6)$$

b = 평균매수의견점수를 측정하는 기간, b∈{1day, 2day, 3day, 1week, 2week, 1month}
h = 포트폴리오 보유 기간, h∈{1day, 2day, 3day, 1week, 2week, 1month}

표본 데이터의 양이 충분하지 않기 때문에 이러한 거래전략을 사용할 때 하루씩 이동해 가며 매수의견점수를 평균하여 계산하였고, 포트폴리오의 보유기간 또한 하루씩 이동해 가며 수익률을 계산하여 분석에 사용하였다.

조선대학교
CHOSUN UNIVERSITY

제4장 자료 및 변수의 정의

제1절 자료

1. 온라인 주식게시판 자료

대한민국의 포털 사이트 1위인 네이버는 네이버 금융(http://finance.naver.com)을 통해 한국 금융시장에 관한 정보를 제공하고 있는데, 그 중 주식 투자자들이 의견과 정보를 공유할 수 있는 종목토론실 서비스를 2014년부터 제공하고 있다. 웹사이트 분석사이트인 SimilarWeb (http://www.similarweb.com)에 의하면 네이버 금융은 2020년도에 대한민국에서 31위로 많이 이용되는 웹사이트로 금융 관련 웹사이트 중에서는 가장 많이 이용되는 웹사이트이다. 네이버 금융이 제공하는 서비스인 종목토론실은 이용자들에 의해 활발하게 이용되는 대한민국에서 가장 큰 주식관련 온라인 토론 사이트로 대한민국의 주식관련 소셜 미디어 중 가장 많은 글을 보유하고 있다.

본 연구는 네이버 금융에서 제공하는 종목토론실에서 2018 년 1 월 2 일부터 2019 년 12 월 30 일까지의 각 기업의 게시글을 웹 크롤링(web-crawling)방법을 이용해 내려받았다. 다운로드된 게시글은 기업마다 일별로 저장되며 저장되는 웹사이트의 게시글은 [그림 1]와 같다. 각각의 게시글은 글쓴이의 아이디와 아이피 일부분, 글을 쓴 날짜와 시간, 다른 사람들이 그 글에 대해 평가하는 지표인 공감 및 비공감 수, 조회 수 등으로 구성되어 있다. 사이트에서 제공하는 글쓴이의 아이디와 아이피를 합쳐서 한 명의 글쓴이 정보로 보고 아이디와 아이피가 동일하다면 동일한 글쓴이로 가정하고 분석에 사용한다. 게시글은 게시글 이용자들의 최소한의 반응이 있는 글들을 사용하기 위해 공감 수와 비공감 수의 합이 영(0)보다 크고, 개별 기업별로 표준화한 조회수 값을 사용하여 조회수 표준화 값이 -3 시그마를 초과

조선대학교
CHOSUN UNIVERSITY

하는 게시글을 분석대상으로 하였다. 그 결과 분석을 위한 게시글 표본으로 총 16,153,123 개의 글과 2,183 기업, 3,881,542 명의 글쓴이가 선정되었다.

여기에 [그림 1] 삽입

본 연구는 표본기간 동안 유가증권시장과 코스닥시장에서 지속적으로 거래되고 있는 상장기업 중 금융업에 속하지 않은 기업으로서 자본잠식이 아닌 기업을 분석 과정에 포함하였다. 마지막으로 본 연구는 각 기업별로 일별 게시글에서 나타나는 투자자들의 센티멘트를 이용하여 각 기업의 투자심리지표인 매수의견점수와 의견 불일치정도의 일별 관측치를 산출하기 때문에 표본기간 동안 많은 양의 게시글을 가지고 있는 기업을 분석대상으로 하기 위해 평균 일일 게시글이 10 회 미만인 기업을 제외하였다. 그 결과 최종적으로 5,552,485 개의 게시글과 350 개 기업이 분석 대상 표본으로 선정되었다.

[그림 2]는 요일별(패널 A) 및 시간대별(패널 B)로 게시판에 쓰인 일일 평균 게시글 수의 분포를 나타낸다. 게시글은 일반적으로 주식 거래가 이루어지는 평일, 주식시장 개장 이후(am 9:00)부터 폐장 직전(pm 3:30)에 게시되는 것을 알 수 있다. 그러므로 본 연구에서는 투자심리지표를 측정할 때 거래일 동안 주식시장 개장 중에 게시된 게시글만을 사용하도록 한다.

여기에 [그림 2] 삽입

2. 주식시장 관련 자료

본 연구의 표본은 2018년 1월 2일부터 2019년 12월 30일까지 유가증권시장과 코스닥시장에서 지속적으로 거래되고 있는 상장기업 중 금융업에 속하지 않은 기업을 대상으로 하였다. 이 중 음(-)의 장부가치를 갖는 기업과 거래일의 평균 일일 게시글이 10회 미만인 기업은 제외하여 최종적으로 350개의 표본기업이 선정되었

조선대학교
CHOSUN UNIVERSITY

다. 분석에 사용된 주가자료 및 재무제표자료는 FnGuide에서 제공하는 DataGuide를 통해 수집하였고, 무위험이자율에 대한 대용치로는 한국은행 경제통계시스템에서 제공하는 통화안정증권 364일물 금리를 사용하였다.

본 논문에서 실증분석을 위해 사용된 주식시장 관련 자료는 FnGuide로부터 얻었고, 종가와 보통주 발행주식수를 이용하여 시가총액을 계산하고, 장부가치는 자본총계에서 우선주 자본금을 차감 한 값으로 하였다. 수익률은 개별 기업의 수정주가로부터 산출하였고 시장수익률은 한국종합주가지수(KOSPI 200)를 이용하여 개별 기업의 주식수익률과 동일한 기간 및 수익률 산출방법으로 구하였다.

제2절 텍스트 데이터 분석

본 논문은 게시글에 나타난 매수, 매도 의견의 정도를 측정하여 게시판에서 형성된 투자심리가 주가 수익률과 변동성과 어떤 관계를 갖는지 분석하고자 한다. 본 연구에서는 온라인에서 형성되는 투자심리를 파악하기 위해 네이버 금융의 종목토론실을 대상으로 텍스트 데이터를 수집하였다. 게시글이 긍정 센티멘트를 가질수록 매수의견, 부정 센티멘트를 가질수록 매도의견으로 보고 게시글을 센티멘트에 따라 매수, 매도 의견으로 분류하기 위해 게시글에 대해 머신러닝 방법 중 하나인 분류모델을 사용하였다. 머신러닝의 분류모델에 의해 매수, 매도 의견으로 분류된 각 기업의 게시글들을 일별로 종합하여 각 기업의 일별 투자심리지표인 매수의견점수와 의견 불일치 정도를 파악하고자 한다. 게시글 수집 및 데이터 처리 과정은 모두 파이썬(python)을 이용하여 수행하였다.

1. 게시글의 센티멘트 분류

게시글의 센티멘트를 분류하기 위해 비정형 데이터인 텍스트 데이터를 정형화된

조선대학교
CHOSUN UNIVERSITY

수치화 데이터로 바꿔줄 필요가 있다. 이를 위해 다음과 같은 전처리 과정을 거친다. 첫 번째로 분석 대상 글을 선택하고 이를 정제하고, 두 번째로 문서 단어 행렬을 구축한다. 자세한 전처리 과정은 부록 1에서 소개하도록 한다.

문서 단어 행렬을 통해 수치화 데이터로 바뀐 게시글 데이터를 이용하여 각 게시글의 센티멘트를 분류한다. 텍스트 데이터를 분류하는데 있어서 일반적으로 사용되는 방법은 단어의 감정이 분류되어 있는 감정사전이 존재하는 경우 감정사전을 이용하여 각 단어의 감정 점수를 부여하고 이를 합하여 글의 감정을 측정하는 방법과 감정사전이 존재하지 않고 충분한 양의 텍스트 데이터가 있는 경우 머신러닝 방법을 통해 텍스트 데이터를 특징에 따라 분류하는 방법이 있다. 우리나라의 경우 금융시장에 특화된 감정사전이 존재하지 않고, 금융시장에서 사용되는 단어들은 일반적으로 사용되는 단어들의 뜻과 다르게 해석되기 때문에 일반적인 단어들로 구축된 감정사전을 이용하는 것도 어렵다. 또한, 텍스트의 원천(예를 들어 뉴스, 사업보고서, 소셜 미디어 등)에 따라 텍스트를 만들어내는 이용자가 다르고, 이로 인해 텍스트의 원천에 따라 사용되는 단어들도 달라지기 때문에 하나의 공통된 감정사전을 이용하여 텍스트의 감정을 분류하는 데에는 한계가 있다. 더불어 금융시장은 경제 상황에 민감하여 텍스트가 창출된 시기에 따라 그 글에 쓰인 단어가 나타내는 감정이 달라질 수 있으므로 기존에 만들어진 감정사전을 이용하여 분석에 사용하는데 한계가 있다.

본 연구는 게시글의 센티멘트에 따라 긍정적인 센티멘트를 매수의견으로, 부정적인 센티멘트를 매도의견으로 분류하기 위해 머신러닝의 지도학습 중 분류모델을 이용하여 감정사전을 이용하지 않고 게시글을 매수의견(긍정 센티멘트)과 매도의견(부정 센티멘트)으로 구분하고자 한다. 분류모델을 적용하기 위한 학습데이터는 개별 기업의 종가를 이용해 계산된 일별 수익률을 이용하여 구축한다. 일별 수익

조선대학교
CHOSUN UNIVERSITY

률을 수익률 크기에 따라 일별로 정렬한 후 가장 상위의 수익률을 가지고 있는 기업의 게시글들을 매수의견으로 보고 반대로 가장 하위의 수익률을 가지고 있는 기업의 게시글들을 매도의견으로 가정한다. 상·하위 수익률을 구분하기 위한 기준점은 일별로 정렬된 수익률의 양 극단 2%에 해당하는 수익률로 한다.[5] 일별 수익률을 이용하여 부여된 투자의견을 이용한 학습데이터를 통해 게시글에 투자의견을 부여하여 포트폴리오 분석을 실시한 결과 일반 감정사전으로 게시글의 투자의견을 분류하여 포트폴리오를 구축한 것보다 더 높은 성과를 내는 것을 확인하였다.[6] 최종적으로 전체 게시글은 매수 게시글이 7,566,441개, 매도 게시글이 8,585,682개로 분류되었다.

본 연구에서 사용할 머신러닝의 분류모델을 선택하기 위해 위에서 구축된 학습데이터를 사용하여 다양한 분류모델의 정확도를 측정한 결과 모든 분류모델에서 비슷한 정확도가 확인되었다.[7] 이에 본 논문은 전통적인 분류모델 중 하나로 텍스트 분류를 비롯하여 많은 분류 문제를 잘 해결하는 것으로 알려져 있는 서포트 벡터 머신 분류모델을 통해 분류한 게시글의 투자 의견을 이후의 연구 결과를 도출하는데 사용한다. <표 1>은 서포트 벡터 머신 분류모델을 통해 분류된 게시글 예시를 보여준다.

---

5) 일별로 정렬된 수익률의 분포의 양극단 1%부터 10%까지 1%씩 순차적으로 늘려가며 투자 의견을 부여하여 학습데이터를 구축 한 후 나머지 글에 투자의견을 분류하여 포트폴리오의 성과를 샤프비율(Sharpe Ratio)로 비교한 결과 양 극단 2%에 해당하는 수익률에 해당하는 글을 이용하여 투자의견을 부여한 경우의 포트폴리오가 가장 높은 샤프비율을 갖는다는 것을 확인하였다. 투자의견을 구축한 각 수익률 비율에 대한 비교 결과는 부록 4에 제시한다.

6) 비교분석대상 일반 감정사전으로는 한국에서 일반 감정사전으로 알려진 서울대 감정사전(http://word.snu.ac.kr/kosac)을 사용하였다. 비교분석 결과는 부록 5에서 제시한다.

7) 비교분석을 위해 사용된 분류모델은 나이브 베이즈(Naïve Bayes), 로지스틱 회귀(Logistic Regressionn), 서포트 벡터 머신(Support Vector Machine)으로 비교분석 결과는 부록 3에 제시한다.

조선대학교
CHOSUN UNIVERSITY

여기에 <표 1> 삽입

2. 투자의견(Bullish) 변수 및 의견 불일치(Disagree) 변수

본 연구는 Antweiler and Frank(2004)의 논문에서 제시한 투자심리지표인 매수의견점수(Bullish)와 의견 불일치정도(Disagree)를 게시글의 투자심리지표로 사용하였다. 각 기업별 게시글의 투자의견을 일별로 종합하여 각 기업별 일별 매수의견점수와 의견 불일치정도를 측정한다. 먼저 매수의견점수를 나타내는 $Bullish_{i,t}$는 각 기업의 일별 게시글의 의견을 종합하여 만든 기업에 대한 매수·매도 의견 측정치로 다음과 같이 정의된다.

$$Bullish_{i,t} = \ln\frac{1+\ M_{i,t}^{BUY}}{1+\ M_{i,t}^{SELL}} \quad (7)$$

매수의견점수의 측정치인 $Bullish_{i,t}$는 기업 i의 일별(t) 매수의견 게시글 수($M_{i,t}^{BUY}$)와 매도의견 게시글 수($M_{i,t}^{SELL}$)를 통해 계산된다. 이 측정치는 매수의견의 게시글 수가 증가할수록 큰 값을 가지게 된다. 매수의견점수를 계산할 때 사용된 일별 게시글 수는 [그림 2]에 나타난 것처럼 일반적으로 게시판에 글을 쓰는 글쓴이들은 주식시장이 개장되어있을 때 글을 쓰는 경향이 있기 때문에 거래일의 장중 게시글 수만을 포함하여 계산한다.

의견 불일치정도를 나타내는 $Disagree_{i,t}$는 Antweiler and Frank(2004)가 제시한 것과 동일하게 게시글에 나타난 투자의견의 표준편차를 이용하여 정의한다. 이때 사용되는 투자의견 $Opinion_{i,t}$는 게시판에서 투자의견의 불일치 정도를 측정하기 위해 투자의견에 대한 표준 편차로 계산되는 $Disagree_{i,t}$의 범위를 [0,1]로 제한하기 위해 범위가 (-1,1)인 $(M_{i,t}^{BUY}+M_{i,t}^{SELL})/(M_{i,t}^{BUY}-M_{i,t}^{SELL})$로 계산한다. 의견 불일치정

조선대학교
CHOSUN UNIVERSITY

도는 식 (7)과 동일한 일별 게시글 수를 사용하여 다음과 같이 계산된다.

$$Disagree_{i,t} = 1 - \sqrt{1 - Opinion_{i,t}^2} \qquad (8)$$

의견 불일치정도를 나타내는 $Disagree_{i,t}$는 투자의견이 불일치할수록 1에 가까워지며, 투자의견이 매수의견이나 매도의견 어느 한쪽으로 동질적일수록 0에 가까운 값을 갖게 된다.

조선대학교
CHOSUN UNIVERSITY

제5장 실증분석결과

이 장에서는 게시판에서 형성된 센티멘트를 이용하여 측정된 투자심리지표인 매수의견점수와 의견 불일치정도가 주식수익률과 변동성에 대해 정보력과 예측력을 갖는지 살펴본다. 첫 번째로 게시판을 통해 측정된 투자심리가 주식수익률과 변동성에 설명력을 갖는지를 확인하기 위해 기업 고유 특성치를 통제한 후 주식수익률과 변동성을 투자의견점수와 의견 불일치정도에 대하여 회귀분석한다. 두 번째로 매수의견점수로 나타난 투자심리가 주식수익률에 대하여 예측력을 갖는지 검증하기 위해 포트폴리오 분석을 실시한다. 이러한 분석 결과에 개인투자자의 거래량이 영향을 미치는지 확인하기 위해, 개인투자자의 거래량 크기에 따라 분석대상집단을 구분한 후 회귀분석과 포트폴리오 분석을 실시한다. 마지막으로 게시판의 정보력이 지속되는 기간을 확인하기 위해 주식수익률에 매수의견점수 측정기간과 포트폴리오 보유기간을 변화시키며 시장포트폴리오 대비 샤프비율을 확인한다.

제1절 기술통계

본 연구는 실증분석을 위하여 2018년 1월 2일부터 2019년 12월 30일까지 계속 상장된 기업 중 금융업과 자본 잠식된 기업을 제외하고, 적어도 거래일 평균 게시글이 10개 이상 존재하는 기업을 대상으로 한 결과 350개의 기업이 표본으로 선정되었다. 본 연구에서는 2018년도부터 2019년도까지 2년 동안의 게시판의 게시글을 파이썬을 이용해 웹 크롤링 방법으로 수집하였다. 표본기업에 대한 재무제표 자료 및 주가자료는 FnGuide에서 제공하는 DataGuide에서 수집하였고, 무위험 이자율의 대용치로는 한국은행 경제통계시스템에서 제공하는 통화안정증권 364일물의 금리를 사용하였다.

조선대학교
CHOSUN UNIVERSITY

<표 2>은 회귀분석에서 사용된 변수들의 기초통계량으로 기업별로 일별 데이터를 평균한 값을 가지고 계산된 전체 기업에 대한 기초통계량을 나타낸다. <표 2>의 패널 A를 통해 게시판의 투자심리를 나타내는 매수의견점수인 Bullish변수는 주식시장이 불황기였던 2018년도부터 2019년도까지 2년동안의 평균 또한 음(-)의 값이 나옴을 확인할 수 있다. 이는 게시판에서 형성되는 투자심리가 주식시장의 움직임과 관련이 있음을 유추해볼 수 있다. 게시판의 데이터로부터 형성된 변수인 Bullish와 Disagree, Messages 모두 3보다 큰 첨도 값이 확인되었고, 이는 정규분포에 비해 꼬리가 두껍다는 것을 의미하며 기업에 따라 게시판이 갖는 영향력이 모두 동일하지 않다는 것을 의미함으로 연구에서 게시판의 데이터를 사용하는 것이 유의미함을 알 수 있다. <표 2>의 패널 B는 회귀분석에 사용된 변수 간 상관관계를 나타낸다. 게시판의 텍스트 데이터를 통해 측정된 투자심리지표인 매수의견점수를 나타내는 Bullish와 의견 불일치정도를 나타내는 Disagree간에 상관관계가 매우 낮음을 알 수 있다. 이를 통해 두 변수를 동시에 회귀분석에 사용하는 데 문제가 없음을 확인할 수 있다. 포트폴리오 분석에서 사용된 위험요인의 기초통계량은 <표 3>에 제시하였다.

여기에 <표 2> 삽입

여기에 <표 3> 삽입

제2절 투자심리지표의 정보력 검증

1. 전체기업에서 투자심리지표와 주식수익률, 변동성 간의 회귀분석

게시판에서 측정되는 투자심리가 주식수익률과 변동성에 정보력을 갖는지를 살펴보기 위해 회귀분석을 실시하였다. <표 2>의 패널 B를 통해서 매수의견점수는

조선대학교
CHOSUN UNIVERSITY

주식수익률과 양(+)의 관계가 있고, 의견 불일치정도는 변동성과 양(+)의 관계가 있음을 확인할 수 있었다. 이러한 관계는 회귀분석을 통해서도 동일하게 나타났고 <표 4>에서는 통계적으로도 그 관계가 유의함을 확인할 수 있었으며 이는 게시판에서 긍정적인 센티멘트, 즉 매수의견이 많을수록 기업의 주식수익률이 증가하고 의견이 불일치될수록 주식가격의 변동성이 높아짐을 보여준다. 결과적으로 게시판에서 형성되는 투자심리는 주식수익률에 대해 정보력이 있음을 알 수 있다.

여기에 <표 4> 삽입

투자의 목적으로 게시판에서 나타나는 투자심리가 주식수익률에 대해 예측력이 있는지를 확인하는 것 또한 중요하다. 이를 위해 본 연구는 t시점의 투자심리지표를 가지고 그 다음날인 t+1시점의 주식수익률 및 변동성과 어떤 관계가 나타나는지를 회귀분석을 통해 확인하였다. <표 5>는 투자심리지표와 주식수익률 및 변동성을 1일의 시간차를 두고 회귀분석한 결과를 제시하고 있다. 분석 결과 <표 4>와 마찬가지로 매수의견점수는 주식수익률과 통계적으로 유의한 양(+)의 관계가 나타나고, 의견 불일치정도는 변동성과 통계적으로 유의한 양(+)의 관계가 확인되었다. 그러나 주식수익률에 대한 회귀분석 모형의 설명력은 50.5%에서 5.5%로 크게 감소되어 매수의견점수의 주식수익률에 대한 예측력은 존재하지만 그 설명력은 크지 않은 것으로 나타났다.

여기에 <표 5> 삽입

게시판에서 형성된 투자심리가 주식수익률과 변동성에 대해 통계적으로 유의한 관계가 확인됨으로써 게시판의 텍스트 데이터가 주식시장에 대해 정보력을 가지고 있음을 확인하였다. 이는 게시판이 글쓴이들의 투자심리에 대한 정보를 반영하고 있다고 추정될 수 있고, 한국 주식시장과 같이 개인투자자의 거래가 활발한 시장에서는 기관투자자에 비해 정보가 부족한 개인투자자가 정보를 얻기 위해 게시판

조선대학교
CHOSUN UNIVERSITY

을 이용하는 것으로 생각해 볼 수 있다. 이에 개인투자자의 거래량이 많은 기업이 실제로 게시판에서 형성된 투자심리에 영향을 더 크게 받는지 검증할 필요가 있다.

2. 기업특성에 따른 투자심리지표와 주식수익률, 변동성 간의 회귀분석

한국 주식시장은 개인투자자의 거래비중이 크고, 개인투자자와 기관투자자는 정보비대칭이 존재하기 때문에 상대적으로 정보에 있어 열약한 위치에 존재하는 개인투자자는 정보를 얻기 위해 온라인 상의 정보에 의존할 수 있다. 그러므로 게시판은 주로 개인투자자가 이용할 가능성이 높고, 이에 본 연구에서는 게시판의 투자심리가 주식시장에 미치는 영향력이 개인투자자의 거래량으로 나타낼 수 있는 기업특성에 따라 달라지는 지를 분석하기 위해 개인투자자 거래량의 크기에 따라 분석대상기업을 3분위로 나눈 후, 각 집단 별로 투자심리지표와 주식수익률 및 변동성 간의 관계를 회귀분석을 통해 살펴보았다.

<표 6>은 개인투자자의 거래량 크기에 따라 구분한 집단 별로 투자심리지표가 주식수익률과 변동성에 미치는 영향력을 회귀분석을 통해 확인한 결과이다. 주식수익률에 대한 매수의견점수의 영향력은 모든 집단에서 통계적으로 유의한 양(+)의 관계가 나타났는데, 개인투자자 거래량이 큰 집단에서 설명력이 거래량이 가장 낮은 집단의 설명력보다 약 1.5배 더 높은 것으로 보아 개인투자자 거래량이 높은 기업일수록 게시판에서 형성되는 투자심리에 영향을 더 받는 다는 것을 알 수 있다. 이를 통해 게시판은 주로 개인투자자가 정보를 얻기 위해 사용할 가능성이 높고, 비합리적 투자자로 알려진 개인투자자의 의견이 주식게시판에서 상호작용함으로써 정보력을 갖게 될 수 있음을 시사한다.

여기에 <표 6> 삽입

조선대학교
CHOSUN UNIVERSITY

<표 7>은 개인투자자의 거래량 크기에 따라 집단을 구분한 후 집단에 따라 예측력이 달라지는지를 확인하기 위해 집단 별로 투자심리지표와 주식수익률 및 변동성을 1일의 시차를 두고 표본 외 회귀분석을 실시한 결과이다. 회귀분석 결과 <표 6>의 결과와 마찬가지로 모든 집단에서 매수의견점수가 주식수익률과 통계적으로 유의한 양(+)의 관계를 갖고, 개인투자자의 거래량이 큰 기업집단 일수록 그 설명력은 더 높아지는 결과가 나타났다. 그러나 주식수익률에 대한 두 집단의 설명력 모두 <표 6>의 결과에 비해 큰 폭으로 감소하여 매수의견점수가 주식수익률에 대해 예측력은 존재하지만 그 영향력은 크지 않은 것으로 판단된다.

여기에 <표 7> 삽입

개인투자자의 거래량 크기에 따라 온라인에서 형성된 투자심리가 주식시장에 갖는 정보력과 예측력을 비교분석한 결과 개인투자자의 거래량이 많은 기업집단에 속할수록 게시판의 게시글이 갖는 정보력과 예측력이 높아짐을 확인할 수 있었다. 이는 게시판이 기관투자자나 외국인투자자에 비해 정보적 열위에 놓여있는 개인투자자들의 정보채널이 되어 이곳에서 개인투자자들의 상호작용을 통해 형성된 투자심리가 주식가격에 영향을 줄 수 있음을 시사한다. 더불어 개인투자자의 거래가 많은 기업 일수록 게시판을 통해 개인투자자들이 시장 참여자로써 영향을 미칠 수 있음을 암시한다.

제3절 매수의견점수의 3분위 포트폴리오 분석

1. 매수의견점수에 따른 단일정렬 포트폴리오

게시판을 통해 측정된 투자심리가 정보력과 예측력이 있다면, 이를 통해 투자이익을 얻을 수 있는지 분석하기 위해 다음과 같이 포트폴리오 분석을 실시한다. 먼

조선대학교
CHOSUN UNIVERSITY

저 게시판의 텍스트 데이터를 통해 추정한 Antweiler and Frank(2004)의 투자심리 지표인 매수의견점수 변수를 기준으로 일별로 정렬한 후 정렬 순서에 따라 분석대상기업을 3분위로 구분하여 포트폴리오를 구성한다. 포트폴리오 전략은 각 포트폴리오를 이후 1일 보유한다고 가정하고, t+1일의 가치가중 포트폴리오 및 동일가중 포트폴리오의 수익률을 측정한다. <표 8>은 매수의견점수를 기준으로 단일정렬된 포트폴리오의 샤프비율과 위험요인을 통제 한 후 초과수익률을 제시한다. ExtraSR은 시장포트폴리오 대비 매수의견점수 포트폴리오의 투자성과를 측정하기 위해 매수의견점수 포트폴리오의 샤프비율에서 시장포트폴리오의 샤프비율을 차감한 값이다. 포트폴리오의 수익률이 위험요인을 통제 한 이후에도 통계적으로 유의한 초과수익률을 얻는지 확인하기 위하여 CAPM, FF3 모형, FF5 모형으로 부터 추정된 알파를 측정하였다.

여기에 <표 8> 삽입

<표 8>의 패널 A는 가치가중한 매수의견점수 포트폴리오의 투자성과를, 패널 B는 동일가중한 매수의견점수 포트폴리오 투자성과를 나타낸다. 두 가중치로 분석한 포트폴리오 모두 매수의견점수가 낮은 포트폴리오부터 높은 포트폴리오까지 모두 시장포트폴리오 대비 추가 성과를 얻는 것이 확인되었고, 매수의견점수가 높은 주식들로 구성된 포트폴리오가 가장 높은 샤프비율을 갖는 것으로 나타남으로 게시판에서 센티멘트를 이용하여 추정한 매수의견점수가 주식수익률에 대해 예측력이 존재함을 알 수 있다. 위험요인을 통제한 후의 초과수익률인 알파를 보면 모든 포트폴리오에서 위험요인을 통제한 후에도 통계적으로 유의한 양(+)의 값이 확인되었고, 위험요인의 설명력은 매수의견점수가 높은 기업으로 구성된 포트폴리오일수록 설명력이 낮아짐을 확인할 수 있다. 이는 게시판을 통해 측정된 매수의견점수를 이용하여 구축한 포트폴리오 중에서 매수의견점수가 높은 포트폴리오 일수록

조선대학교
CHOSUN UNIVERSITY

위험요인에 의해 설명되지 않으며 이를 통해 더 큰 투자 수익을 얻을 수 있는 것으로 해석할 수 있다.

이러한 매수의견점수에 따라서 포트폴리오를 구성할 경우 투자수익을 얻을 수 있는지 알아보기 위하여, 매수의견점수에 따라 정렬한 3개의 포트폴리오의 표본기간 동안의 누적수익률을 살펴보았다. 즉, 2018년 1월 2일에 각 포트폴리오에 1원을 투자하고 매일 매수의견점수를 새로 추정하여 정렬한 후 포트폴리오를 재조정하여 2019년 12월 30일까지 계속 보유하는 투자전략을 가정하고 누적수익률을 계산하였다. [그림 3]은 표본기간동안 매수의견점수에 따라 정렬된 포트폴리오의 누적수익률의 시계열을 나타낸다. 패널 A는 가치가중한 매수의견점수 포트폴리오의 누적수익률, 패널 B는 동일가중한 매수의견점수 포트폴리오의 누적수익률을 나타내는 것으로 두 가중치에서 모두 매수의견점수가 큰 주식으로 구성한 High 포트폴리오에서 가장 큰 누적수익률을 얻었다. 투자성과의 비교를 위해 시장 포트폴리오에 대해서도 동일한 투자전략으로 누적수익률을 계산한 결과 매수의견점수 포트폴리오는 모두 시장 포트폴리오에 비해 높은 누적수익을 얻으며, 특히 High 포트폴리오는 시장 포트폴리오보다 더 빠른 속도로 누적수익이 증가하는 것을 확인할 수 있다. 결과적으로 매수의견점수를 이용한 투자전략은 상당한 투자성과를 얻을 수 있을 것으로 예상된다는 점에서 경제적으로도 게시판에서 형성된 투자심리의 예측력이 유의함을 시사한다.

여기에 [그림 3] 삽입

매수의견점수에 따른 단일정렬 포트폴리오의 수익률을 분석한 결과를 요약하자면, 게시판으로부터 투자심리를 추정하였을 때, Antweiler and Frank(2004)에서 제시한 매수의견점수를 투자심리지표로 사용한 결과 시장 포트폴리오 대비 샤프비율은 높게 나타나고, 위험요인을 통제한 후 초과수익률은 통계적으로 유의한 양(+)

조선대학교
CHOSUN UNIVERSITY

의 값을 갖는다는 것이 확인되었다. 그러나 매수의견점수가 높은 기업들로 구축된 포트폴리오일수록 위험요인의 설명력이 낮아지는 현상이 발견되었고, 이를 통해 위험요인에 의해서는 매수의견점수 포트폴리오의 수익률이 유의적으로 설명되지 못하고 거시경제의 위험 요인이 아닌 개별기업의 위험이 존재할 것을 유추해볼 수 있다. 이후의 논의에서는 개인투자자의 비중을 개인투자자의 거래량의 크기로 보고 개인투자자의 거래량 크기에 따라 기업을 두 집단으로 먼저 구분한 후, 매수의견점수로 정렬하여 이중정렬된 포트폴리오를 구성하여 개인투자자 비중으로 나타나는 기업의 특성에 따라 게시판의 투자심리가 미치는 영향력을 살펴보고자 한다.

2. 매수의견점수와 개인투자자 거래량에 따른 이중정렬 포트폴리오

앞서 게시판에서 측정된 매수의견점수를 정렬하여 포트폴리오를 구축하여 분석한 결과 각 포트폴리오는 모두 시장포트폴리오보다 높은 성과를 얻는 것을 확인할 수 있었다. 반면 CAPM, FF3 모형, FF5 모형으로 위험요인을 통제한 후 초과수익률은 통계적으로 유의하지만, 매수의견점수가 높은 포트폴리오일수록 위험요인에 의한 설명력이 낮아졌다. 이는 기업 특성에 따라 게시판이 주식수익률에 미치는 영향력이 달라질 수 있기 때문에 발생한다고 해석할 수 있다. 이 가설을 간접적으로 확인해보기 위하여 게시판의 주된 이용자로 고려되는 개인 투자자의 거래량을 개인투자자의 비중으로 보고 개인투자자의 거래량과 매수의견점수에 의해 이중정렬된 포트폴리오를 구성하고 그 수익률들 간의 관계를 분석한다. 만약, 개인투자자의 거래량이 많은 기업이 개인투자자의 거래량이 적은 기업보다 게시판에서 측정된 투자심리지표인 매수의견점수에 영향을 더 많이 받는다면 예측력이 더 클 것이고 이로 인해 개인투자자의 거래량이 많은 기업으로 구분된 기업집단에서 더 큰

조선대학교
CHOSUN UNIVERSITY

투자성과를 얻을 수 있을 것이다.

일별로 분석대상 표본주식을 개인투자자의 거래량 크기에 따라 전체표본기업을 먼저 3분위로 나눈 후 이어서 각 분위별 포트폴리오에 포함된 기업들을 다시 매수의견점수를 기준으로 정렬하여 3분위로 구분한다. 이렇게 기업규모와 매수의견점수에 의해 이중정렬된 총 9개의 포트폴리오에 대하여 수익률을 측정하고 샤프비율과 위험요인을 통제한 후의 초과수익률을 확인하고자 한다. 가치가중 및 동일가중한 이중정렬 포트폴리의 샤프비율 및 위험요인을 통제한 후 초과수익률은 <표 9>에 보고한다.

여기에 <표 9> 삽입

<표 9>의 패널 A는 시장가치로 가중평균한 포트폴리오의 수익률을, 패널 B는 이중정렬 포트폴리오의 동일가중 수익률을 나타낸다. 결과적으로, 먼저 모든 포트폴리오에서 시장 포트폴리오 대비 높은 샤프비율을 얻음을 확인할 수 있고, 특히 개인투자자 거래량이 적은 기업보다 큰 기업의 포트폴리오 수익률이 더 크게 나타났다. 즉, 게시판에서 형성되는 매수의견점수는 주식수익률을 예측하는데 도움이 되며 그 영향력은 개인투자자의 거래량이 많은 기업일수록 커진다는 것을 알 수 있다. 위험요인을 조정한 후 초과수익률을 나타내는 알파를 살펴보면 모든 기업에서 통계적으로 유의적인 양(+)의 알파를 가졌고, 개인투자자의 거래량이 많은 기업일수록 그리고 그 중에서도 매수의견점수가 높은 기업들로 포트폴리오가 구성될수록 위험요인에 의한 설명력은 낮아짐을 확인 할 수 있다. 이를 통해 개인투자자의 거래량이 많은 집단에서도 상위의 매수의견점수를 갖는 기업으로 구성된 포트폴리오는 비이성적 투자자인 개인투자자가 게시판을 통해 시장 참여자로써 영향을 미칠 수 있고, 이는 위험요인에 의해 여전히 설명되지 않음을 알 수 있다. 개인투자자의 거래량이 낮은 집단에서는 위험요인의 설명력이 가장 높은 것을 알 수 있는데

조선대학교
CHOSUN UNIVERSITY

이는 개인투자자 비중이 낮은 기업이 일반적으로 기업규모가 큰 기업임을 생각해 볼 때 김석진, 김지영(2000)의 연구에서 제시하는 것처럼 기업규모가 큰 기업일수록 총변동성이 작아져 체계적 위험인 거시 경제 요인에 의해 잘 설명될 수 있기 때문으로 해석할 수 있다.

개인투자자의 거래량과 매수의견점수를 이용하여 이중정렬 포트폴리오 전략을 구사한 경우 경제적으로 의미 있는 투자수익을 얻을 수 있는지 포트폴리오간 비교분석을 통해 확인하고자 한다. 포트폴리오 투자전략은 매수의견점수만을 이용해 정렬하여 포트폴리오를 구성한 단일포트폴리오의 투자전략과 동일하게 개인투자자 거래량 크기에 따라 3등분된 각 집단별로 표본기간동안 매일 일별로 매수의견점수를 정렬하여 포트폴리오를 재조정하여 1일 보유한 경우 누적수익을 계산한다. [그림 4]는 위의 전략에 따라 가치가중과 동일가중을 사용하여 수익률을 계산한 이중정렬 포트폴리오의 누적수익률 시계열 분포이다. 결과적으로 두 가중치에서 모두 전반적으로 시장 포트폴리오 보다 높은 투자 수익을 얻었으며, 개인투자자의 거래가 많고, 매수의견점수가 높은 High-top 포트폴리오가 가장 높은 누적 수익을 얻는 것으로 확인된다. 즉, 게시판에서 형성되는 투자의견은 개인투자자 거래량이 큰 기업집단과 작은 기업집단 각각에서 다른 영향력을 가지며 그 영향력은 개인투자자가 많은 기업집단에서 증가함을 알 수 있다.

여기에 [그림 4] 삽입

3. 보유기간에 따른 정보력 검증

마지막으로 게시판에서 측정된 투자심리인 매수의견점수의 정보력과 예측력이 얼마나 지속되는지 검증하기 위해 매수의견점수를 1일, 2일, 3일, 1주일, 2주일, 1개월의 기간을 두고 순차적으로 늘려가며 각 기간 동안 평균 매수의견 점수를 측

조선대학교
CHOSUN UNIVERSITY

정하였다. 각각의 기간에 대한 매수의견점수의 정보력과 예측력이 지속되는 기간을 확인하기 위해 포트폴리오 수익률 또한 1일, 2일, 3일, 1주일, 2주일, 1개월의 기간을 놓고 매수의견점수와 교차하여 보유기간 수익률을 계산하여 각 포트폴리오별로 샤프비율을 분석에 사용하였다.

여기에 <표 10> 삽입

포트폴리오 보유기간에 따른 매수의견점수의 정보력을 검증한 결과는 <표 10>에서 샤프비율을 통해 제시한다. <표 10>에 제시된 샤프비율은 투자성과를 확인하기 위해 각각의 포트폴리오의 샤프비율에서 시장 포트폴리오의 샤프비율을 차감한 값이다. 분석 결과 매수의견점수는 가치가중 및 동일가중 포트폴리오 모두에서 측정기간이 길어질수록 샤프비율이 증가하는 것으로 나타났다. 이를 통해 게시판에서 형성되는 투자의견은 모멘텀 효과가 존재하여 예측력이 증가하는 것을 추정할 수 있다. 각 기간별 매수의견점수를 이용하여 보유기간을 조정해가며 예측력이 지속되는 기간을 확인한 결과 1일만 보유하는 포트폴리오에서 모두 가장 높은 시장 대비 샤프비율을 확인할 수 있었다. 요약하자면, 게시판의 글을 사용하여 투자심리를 측정할 경우 긴 기간의 글을 사용할수록 그 정보력은 높아지고, 이를 통한 투자수익은 1일만 보유하는 단기투자전략에서 가장 높은 예측력이 존재함을 의미한다.

조선대학교
CHOSUN UNIVERSITY

제6장 결론

인터넷과 모바일 기기가 대중화됨에 따라 텍스트 데이터의 양은 급증하고 있고, 컴퓨터의 성능과 데이터 분석 기술이 향상됨에 따라 비정형 데이터인 텍스트 데이터를 다루는 데 있어 계산비용이 감소하며 텍스트 데이터를 이용하여 주식수익률을 분석하고자 하는 시도들이 증가하고 있다. 특히 글쓴이들의 심리를 직접적으로 나타낼 수 있는 텍스트 데이터는 투자심리지표를 구축하기 위한 데이터로써 각광을 받고 있으며, 실제로도 시장데이터나 설문 등을 이용하여 측정된 간접적인 투자심리지표보다 텍스트 데이터를 이용하여 측정된 투자심리지표가 주식수익률에 대해 더 높은 설명력을 갖는다고 실증 분석 결과가 제시되고 있다.

이에 본 연구는 게시판의 게시글 데이터를 이용하여 Antweiler and Frank(2004)가 제시한 투자심리지표인 매수의견점수와 의견 불일치정도 변수를 구축하고 한국 주식시장에서 주식수익률과 변동성에 대해 온라인에서 형성된 투자자들의 투자심리가 영향력을 갖는지 검증하였다. 즉, 게시판에서 형성된 투자심리가 주식수익률에 대해 정보력이 존재하는지 검증하기 위해 매수의견점수와 의견 불일치정도가 주식수익률과 변동성과 갖는 관계를 회귀분석을 통해 확인하였다. 더불어 게시판의 정보를 이용하여 주식수익률을 예측하고 이를 통해 투자성과를 얻을 수 있는지 확인하기 위해 일별로 매수의견점수 크기에 따라 개별기업을 정렬한 후 3분위로 포트폴리오를 구성하여 샤프비율을 계산하고, 위험 요인을 조정 한 후의 초과수익률을 살펴보았다. 그리고 게시판의 주된 이용자로 유추할 수 있는 개인 투자자의 거래량 크기에 따라 기업집단을 구분한 후 각 집단별로 회귀분석과 포트폴리오 분석을 실시하여 게시판에서 측정된 투자심리가 기업 특성에 따라 미치는 영향력에 차이가 있는지를 조사하였다. 마지막으로 게시판에서 형성되는 투자심리의 정보력

조선대학교
CHOSUN UNIVERSITY

과 예측력이 지속되는 기간을 확인하기 위해 보유기간을 1일부터 1달까지 증가시켜 가며 보유수익률을 계산한 후 샤프비율을 측정하였다.

본 연구의 주요 결과를 요약하면 다음과 같다. 첫째, 게시판의 센티멘트를 통해 측정된 투자심리는 주식수익률과 변동성에 대해 통계적으로 유의한 양(+)의 관계를 가진다. 매수의견점수는 주식수익률과, 의견 불일치정도는 변동성과 통계적으로 유의한 양(+)의 관계를 가짐으로 게시판에서 형성되는 투자 심리는 주식시장에 대해 정보력이 있음을 확인하였다. 둘째, 개인투자자 거래량이 큰 기업집단에서 매수의견점수와 의견 불일치정도가 주식수익률과 변동성에 대해 갖는 설명력이 개인투자자의 거래량이 작은 기업보다 더 높아짐을 확인하였다. 이는 개인투자자가 많은 기업일수록 게시판에서 형성된 투자심리에 더 크게 영향을 받으며, 게시판의 주된 이용자가 개인투자자임을 암시한다. 셋째, 매수의견점수를 이용하여 다음날의 수익률을 예측하는 포트폴리오는 시장 포트폴리오 대비 높은 샤프비율을 가짐으로 게시판의 투자심리는 주식수익률에 대해 예측력이 있음을 확인하였다. 넷째, 개인투자자의 거래량과 매수의견점수에 따른 이중정렬 포트폴리오를 구성하여 수익률을 분석한 결과 개인투자자의 거래량이 크고 매수의견점수가 가장 높은 포트폴리오의 시장 대비 샤프비율이 가장 큼을 확인했다. 즉, 개인투자자들은 게시판을 통해 의견을 공유함으로써 비합리적 투자자인 개인투자자들의 의견은 집단지성의 한 종류가 되어 주식수익률에 대한 예측력이 증가하고, 이러한 영향력은 개인투자자가 많은 기업에서 더 크게 나타남을 의미한다. 마지막으로 매수의견점수를 구축하는 기간이 길어질수록 매수의견점수의 정보력은 증가하였으며, 포트폴리오를 보유하는 기간은 1일일 때 가장 높은 투자 성과가 나타났다. 이를 통해 매수의견점수는 모멘텀 효과가 있어 계산하는 기간이 길어질수록 정보력이 높아짐을 확인하였다.

조선대학교
CHOSUN UNIVERSITY

따라서, 본 연구의 결과는 게시판의 글을 이용하여 구축한 투자심리는 한국 주식시장에서 주식수익률과 변동성에 대해 정보력과 예측력을 가지며, 그 영향력은 기관투자자에 비해 상대적으로 정보가 부족하여 게시판의 주된 이용자로 예상되는 개인투자자가 많은 기업에 대해 더 증가한다고 볼 수 있다.

본 연구의 시사점은 첫째, 주식시장에 특화된 감정사전이 존재하지 않아 텍스트 데이터를 이용하여 주식시장을 분석하는데 어려움이 있는 환경에서 머신러닝의 분류모델을 사용하기 위한 학습 데이터를 구축하는 방법을 제시하였다는 점에서 기여하는 바가 있다. 둘째, 비합리적 투자자로 알려진 개인투자자들의 의견이 게시판을 통해 상호작용 됨으로써 집단지성으로써 정보력을 갖게 됨을 확인함으로써 게시판이 개인투자자들의 정보채널이 될 수 있음을 제시하였다. 셋째, 다양한 투자자가 존재하는 한국 주식시장에서 게시판에서 발생하는 정보에 영향을 받는 기업의 특성을 규명함으로써 한국 주식시장에서 투자자간 정보환경에 대한 이해를 높였다는 점에 의의가 있다.

본 연구의 한계점은 다음과 같이 존재한다. 첫째, 본 연구에서 수집한 게시판의 글에서 학습데이터를 구축 시, 주식수익률의 부호에 따라 텍스트를 긍정 및 부정으로만 분류하여 구축하였다. 그러나 학습데이터가 필요하지 않는 머신러닝 분류모델을 사용함으로써 긍정 및 부정의 맥락을 갖는 텍스트와 구분되는 텍스트를 파악하여 분석하면 더욱 효과적일 것이라 생각된다. 둘째, 포트폴리오의 성과를 시장 포트폴리오 하나와 비교분석하였다. 포트폴리오의 투자성과가 시장 포트폴리오와 비교해서도 높은 성과를 얻었지만, 향후 연구에서는 보다 다양한 비교분석 포트폴리오를 사용하면 게시판에서 형성된 투자심리의 투자성과를 이해하는데 도움이 될 것으로 생각된다. 셋째, 현실에서는 포트폴리오를 재조정할 때 거래비용이 존재하기 때문에 일별로 재조정하는 투자전략을 사용하기 어렵다는 한계점이 있기

조선대학교
CHOSUN UNIVERSITY

때문에 거래비용을 고려하여 포트폴리오의 투자성과를 측정할 필요가 있다. 향후 이러한 연구의 한계점을 보완한 연구를 통하여 게시판이 갖는 정보력과 예측력을 보다 정확하게 측정하여 투자의사결정에 효율적으로 활용할 수 있기를 기대한다.

조선대학교
CHOSUN UNIVERSITY

[참고문헌]

구맹회, 이윤선, 2001, 투자자 유형과 주가의 관계에 관한 연구, *재무관리연구*, 18(1), 43-66.

김석진, 김지영, 2000, 기업규모와 장부가/시가 비율과 주식수익률간의 관계, *재무연구*, 13(2), 21-47

김영민, 정석재, 이석준, 2014, 소셜 미디어 감성분석을 통한 주가 등락 예측에 관한 연구. *Entrue Journal of Information Technology*, *13*(3), 59-69.

김용석, 조성욱, 2019, 한국어 텍스트 분석과 적용: 머신러닝을 통한 증권발행신고서의 비정형화된 텍스트 분석. *한국증권학회지*, *48*(2), 215-235.

남달우, 박진우, 김민경, 조현, 김성희, 2012, 인터넷 주식게시판을 통한 집단지성과 주식시장과의 상관관계 연구. *인터넷전자상거래연구*, *12*(2), 149-164.

이효정, 2014, 개인투자자의 투자심리와 주식수익률간의 관계에 관한 횡단면적 분석: 거주지모델 (Habitat Model) 을 중심으로. *금융지식연구*, *12*(1), 177-204.

정지선, 김동성, 김종우, 2015, 온라인 언급이 기업 성과에 미치는 영향 분석: 뉴스 감성분석을 통한 기업별 주가 예측. *지능정보연구*, *21*(4), 37-51.

Antweiler, W. and Frank, M. Z., 2004, Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, *59*(3), 1259-1294.

Baker, M. and Stein, J. C., 2004, Market liquidity as a sentiment indicator. *Journal of Financial Markets*, *7*(3), 271-299.

Baker, M. and Wurgler, J., 2006, Investor sentiment and the cross-section of stock returns. *Journal of Finance*, *61*(4), 1645-1680.

Baker, M. and Wurgler, J, 2007, Investor sentiment in the stock market. *Journal of Economic Perspectives*, *21*(2), 129-152.

Calomiris, C. W. and Mamaysky, H., 2019, How news and its context drive risk and returns around the world. *Journal of Financial Economics*, *133*(2), 299-336.

Chen, H., De, P., Hu, Y. J., and Hwang, B. H., 2014, Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, *27*(5), 1367-1403.

Cookson, J. A. and Niessner, M., 2020, Why Don't We Agree? Evidence from a Social Network of Investors. *Journal of Finance*.

Cortes, C. and Vapnik, V., 1995, Support-vector networks. *Machine learning*, *20*(3), 273-297.

조선대학교
CHOSUN UNIVERSITY

Da, Z., Engelberg, J., and Gao, P., 2015, The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, *28*(1), 1-32.

Das, S. R. and Chen, M. Y., 2007, Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, 53(9), 1375-1388.

De Long, J. B., Shleifer, A., Summers, L. H., and Waldmann, R. J., 1990, Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703-738.

Fama, E. F. and French, K. R., 1993, Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33, 23-49.

Frank, M. Z. and Sanati, A., 2018, How does the stock market absorb shocks?. *Journal of Financial Economics*, *129*(1), 136-153.

Garcia, D., 2013, Sentiment during recessions. *Journal of Finance*, *68*(3), 1267-1300.

Gentzkow, M., Kelly, B., and Taddy, M., 2019, Text as data. *Journal of Economic Literature*, *57*(3), 535-74.

Giannini, R., Irvine, P., and Shu, T., 2018, Nonlocal disadvantage: An examination of social media sentiment. *Review of Asset Pricing Studies*, *8*(2), 293-336.

Giannini, R., Irvine, P., and Shu, T., 2019, The convergence and divergence of investors' opinions around earnings news: Evidence from a social network. *Journal of Financial Markets*, *42*, 94-120.

Henry, E. and Leone, A. J., 2016, Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *Accounting Review*, *91*(1), 153-178.

Hillert, A., Jacobs, H., and Müller, S., 2014, Media makes momentum. *Review of Financial Studies*, *27*(12), 3467-3501.

Hillert, A., Jacobs, H., and Müller, S., 2018, Journalist disagreement. *Journal of Financial Markets*, *41*, 57-76.

Hirshleifer, D., Lim, S. S., and Teoh, S. H., 2011, Limited investor attention and stock market misreactions to accounting information. R*eview of Asset Pricing Studies*, *1*(1), 35-73.

Hofstede, G., 2001, Culture's recent consequences: Using dimension scores in theory and research. *International Journal of cross cultural management*, *1*(1), 11-17.

조선대학교
CHOSUN UNIVERSITY

Huang, D., Jiang, F., Tu, J., and Zhou, G., 2015, Investor sentiment aligned: A powerful predictor of stock returns. *Review of Financial Studies*, *28*(3), 791-837.

Jiang, F., Lee, J., Martin, X., and Zhou, G., 2019, Manager sentiment and stock returns. *Journal of Financial Economics*, *132*(1), 126-149.

Ke, Z. T., Kelly, B. T., and Xiu, D., 2019, *Predicting returns with text data* (No. w26186). National Bureau of Economic Research.

Kearney, C. and Liu, S., 2014, Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171-185.

Kothari, S. P., Li, X., and Short, J. E., 2009, The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *Accounting Review*, *84*(5), 1639-1670.

Kraus, M. and Feuerriegel, S., 2017, Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38-48.

Kumar, A. and Lee, C. M., 2006, Retail investor sentiment and return comovements. *Journal of Finance*, *61*(5), 2451-2486.

Lemmon, M. and Portniaguina, E., 2006, Consumer confidence and asset prices: Some empirical evidence. *The Review of Financial Studies*, *19*(4), 1499-1529.

Li, F., 2010, Survey of the Literature. *Journal of Accounting Literature*, *29*, 143-165.

Loughran, T. and McDonald, B., 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35-65.

Malkiel, B. G. and Fama, E. F., 1970, Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383-417.

Malkiel, B. G., 1999, *A random walk down Wall Street: including a life-cycle guide to personal investing*. WW Norton & Company.

Preis, T., Moat, H. S., and Stanley, H. E., 2013, Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, *3*, 1684.

Schmeling, M., 2009, Investor sentiment and stock returns: Some international evidence. *Journal of Empirical Finance*, *16*(3), 394-408.

Shiller, R. J., Fischer, S., and Friedman, B. M., 1984, Stock prices and social dynamics. *Brookings papers on economic activity*, *1984*(2), 457-510.

조선대학교
CHOSUN UNIVERSITY

Shleifer, A. and Summers, L. H., 1990, The noise trader approach to finance. *Journal of Economic Perspectives*, *4*(2), 19-33.

Shleifer, A. and Vishny, R. W., 1997, The limits of arbitrage. *Journal of Finance*, *52*(1), 35-55.

Tetlock, P. C., 2007, Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S., 2008, More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, *63*(3), 1437-1467.

조선대학교
CHOSUN UNIVERSITY

[그림 1] 네이버 금융 종목토론실의 게시글

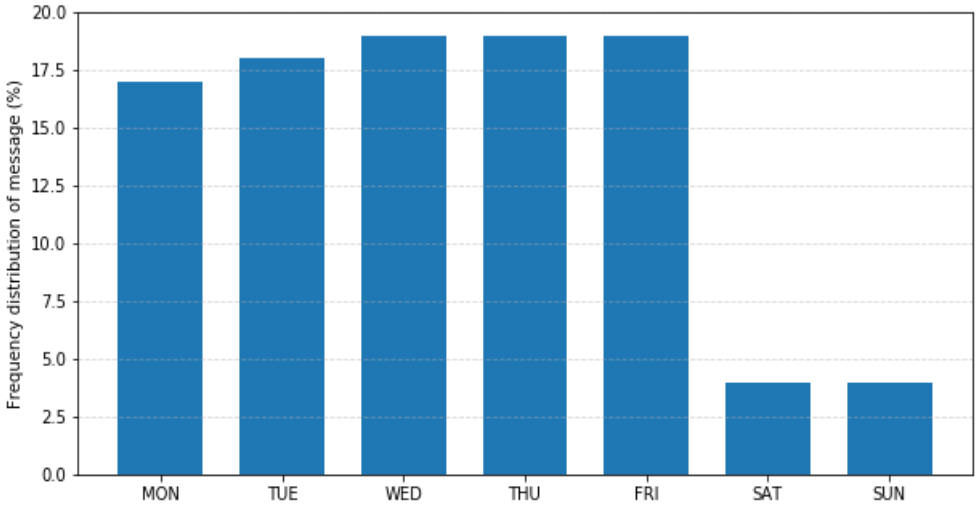
네이버 금융(http://finance.naver.com)에서 제공하는 서비스인 종목토론실은 각 기업별로 주식 투자자들이 정보와 의견을 교환할 수 있는 게시판으로, 이 그림들은 종목토론실의 이용자들이 쓴 글을 스크린 캡쳐한 이미지이다. 본 연구에서는 종목토론실의 게시글에서 게시글과 글쓴이의 아이피 및 아이디 일부 정보, 조회수, 공감 그리고 비공감 수를 분석하는데 사용하였다.

조선대학교
CHOSUN UNIVERSITY

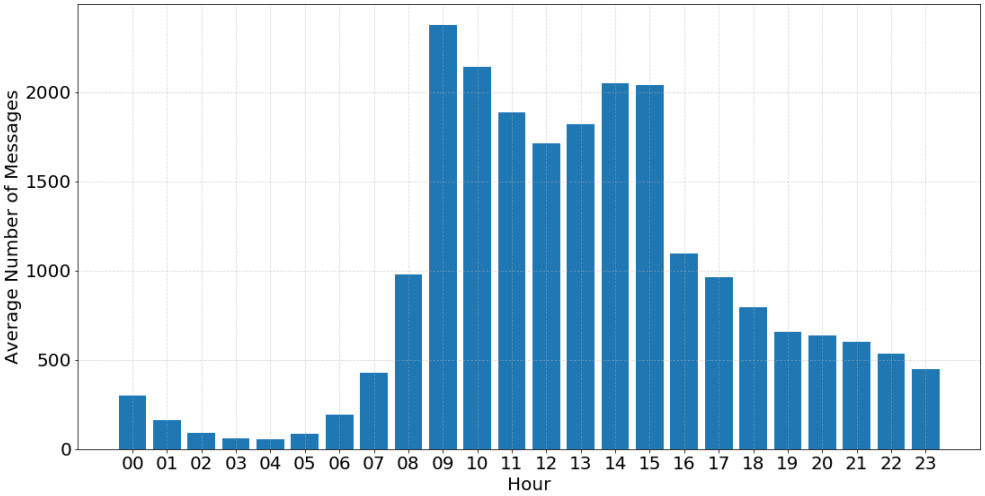
[그림 2] 요일별 게시글 분포, 표본기간동안 시간대별 게시글 분포

이 그림은 네이버 금융의 종목토론실에 게시된 요일별 게시글 분포(패널 A)와 시간대별 게시글 분포(패널 B)를 보여준다.

패널 A. 요일별 게시글 분포



패널 B. 시간대별 게시글 분포

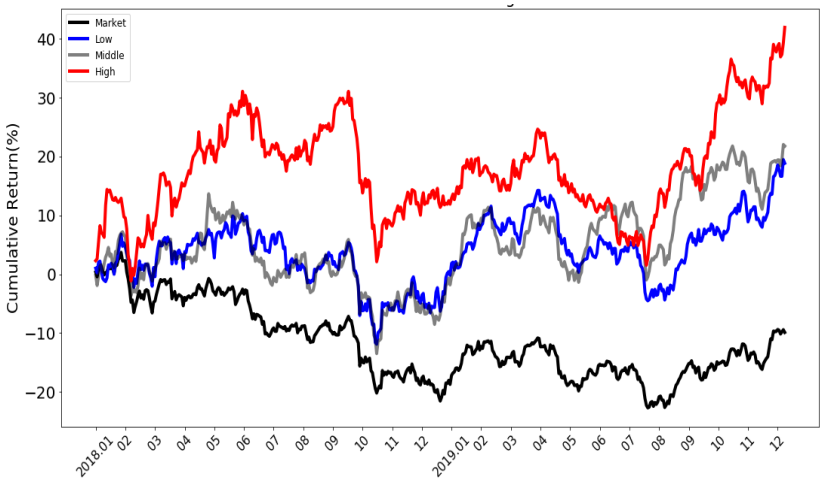


조선대학교
CHOSUN UNIVERSITY

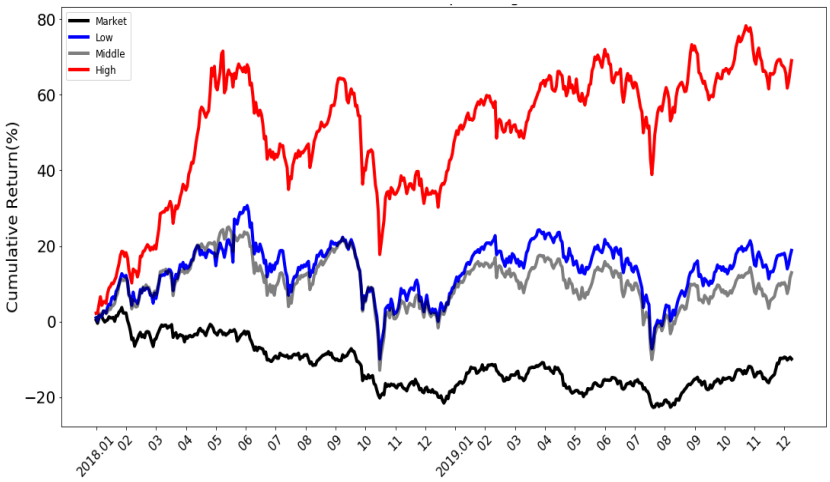
[그림 3] 단일정렬 포트폴리오의 누적수익률

이 그림은 매수의견점수에 따라 3분위로 분류한 포트폴리오의 표본기간 동안 누적수익률을 나타낸다. 빨간색 그래프는 가장 상위의 매수의견점수를 가진 기업들로 구성된 포트폴리오이고, 회색 그래프는 중간의 매수의견점수를 가진 기업들로 구성된 포트폴리오, 파란색 그래프는 가장 하위의 매수의견점수를 가진 기업들로 구성된 포트폴리오, 그리고 검정색은 시장포트폴리오이다. 패널 A는 기업의 규모로 가중치를 사용한 포트폴리오들의 누적수익률분포를, 패널 B는 포트폴리오 구성 기업들에 모두 동일한 가중치를 사용한 포트폴리오의 누적수익률 분포를 나타낸다.

패널 A. 가치가중 포트폴리오



패널 B. 동일가중 포트폴리오

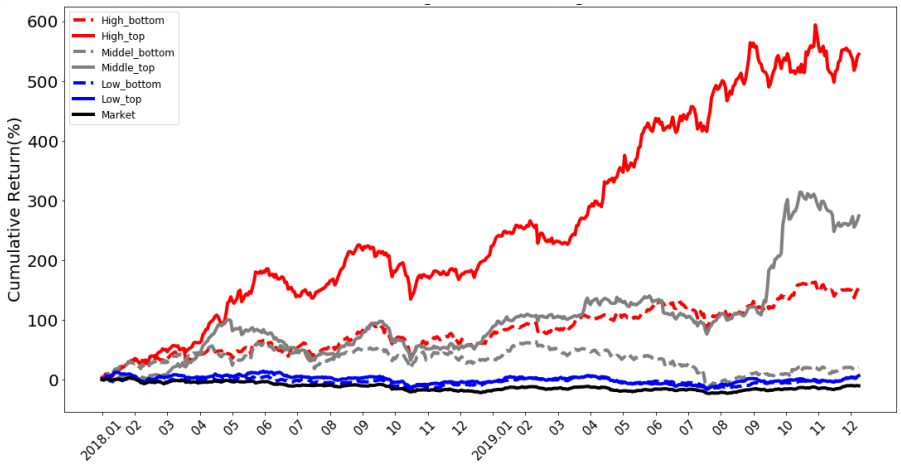


조선대학교
CHOSUN UNIVERSITY

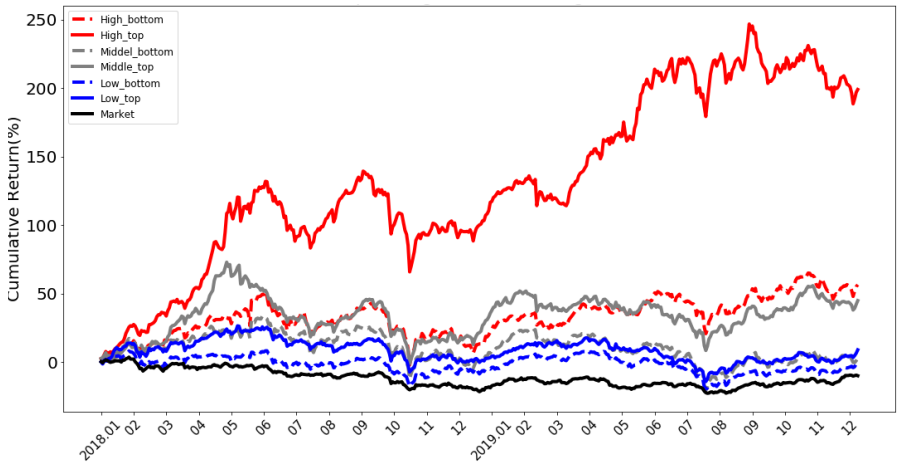
[그림 4] 개인투자자의 거래량과 매수의견점수에 따른 이중정렬 포트폴리오의 누적수익률

이 그림은 개인투자자의 거래량에 따라 3분위로 먼저 분류한 후, 매수의견점수에 따라 3분위로 분류한 이중정렬 포트폴리오의 표본기간 동안 누적수익률을 나타낸다. 개인투자자의 거래량은 각 기업의 전체 거래량 대비 개인투자자의 거래량 비율을 사용하였다. 굵은 선 그래프는 개인투자자의 거래량이 많은(High) 기업들로 구성된 포트폴리오이고, 점선 그래프는 개인투자자의 거래량이 적은(Low) 기업들로 구성된 포트폴리오이다. 빨간색 그래프는 가장 상위(top)의 매수의견점수를 가진 기업들로 구성된 포트폴리오, 회색 그래프는 중간(middle)의 매수의견점수를 가진 기업들로 구성된 포트폴리오, 파란색 그래프는 가장 하위(bottom)의 매수의견점수를 가진 기업들로 구성된 포트폴리오, 그리고 검정색은 시장포트폴리오이다. 패널 A는 기업의 규모로 가중치를 사용한 포트폴리오들의 누적수익률분포를, 패널 B는 포트폴리오 구성 기업들에 모두 동일한 가중치를 사용한 포트폴리오의 누적수익률 분포를 나타낸다.

패널 A. 가치가중 포트폴리오



패널 B. 동일가중 포트폴리오



조선대학교
CHOSUN UNIVERSITY

<표 1> 매수 및 매도로 분류된 게시글 예시

이 표는 본 연구에서 구축한 학습데이터를 이용해 학습시킨 분류모델을 통해 매수 및 매도로 분류된 게시글의 예시를 보여준다.

Bullish Messages
“코스피2위먹으러가시죠”
“가즈아”
“지금은 거래량터지고 수급좋고 각자그릇만큼드시면됩니다 25만원까지분할매수”
Bearish Messages
“일단 팔고 보자 오늘은 하락 마감 예상”
“오를 이유가 없네여”
“하루종일 프로그램매도로 공매를 치니 얼마까지 내릴지 감이 안온다”

조선대학교
CHOSUN UNIVERSITY

<표 2> 기초통계량

이 표는 회귀분석에 사용된 변수들의 기초통계량을 보고한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수 의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 패널 A는 회귀분석에 사용된 변수들의 요약정보를 나타내고, 패널 B는 회귀분석에 사용된 변수들의 상관관계를 나타낸 값이다.

패널 A. 회귀분석 변수의 기초통계량

Variable	Mean	Std	Median	Min	Max	Skew	Kurt
lnReturn	-0.00	0.39	-0.00	-0.19	0.22	0.72	9.48
Volatility	0.03	0.02	0.03	0.01	0.09	1.34	2.22
Messages	32.38	56.99	14.76	1.01	574.38	5.30	43.55
Bullish	-0.17	1.19	-0.16	-5.47	5.48	0.25	7.56
Disagree	0.12	0.19	0.05	0.00	0.91	2.56	7.44
Market	-0.00	0.01	0.00	-0.05	0.04	-0.53	1.91
lnSize	25.46	0.27	26.45	25.88	27.08	0.09	-0.18
lnM/B	7.58	0.31	7.56	6.97	8.25	0.10	-0.41
lnVolume	13.04	1.10	12.91	9.66	16.76	0.36	3.61

패널 B. 회귀분석 변수 간 상관관계

Variable	lnReturn	Volatility	Messages	Bullish	Disagree	Market	lnSize	lnM/B	lnVolume
lnReturn	1								
Volatility	0.03	1							
Messages	0.04	0.42	1						
Bullish	0.67	0.06	-0.01	1					
Disagree	0.07	0.24	0.24	-0.06	1				
Market	0.23	-0.00	-0.01	0.05	-0.02	1			
lnSize	0.01	-0.20	0.37	-0.09	-0.05	0.00	1		
lnM/B	0.02	0.25	0.27	0.02	0.05	-0.00	0.11	1	
lnVolume	0.11	0.42	0.62	0.12	0.22	-0.02	-0.00	0.04	1

조선대학교
CHOSUN UNIVERSITY

<표 3> 위험요인의 기초통계량

이 표는 Fama and French(2015) 5요인(FF5)의 평균수익률(%)과 상관계수를 보고한다. MKT, SMB, HML, RMW, CMA는 각각 5요인 모형의 시장요인, 규모요인, 가치요인, 수익성요인, 투자도요인이다.

Factors	Average Return(%)	Std(%)	Correlation coefficients				
			MKT	SMB	HML	RMW	CMA
MKT	-1.72	0.94	1				
SMB	0.00	0.42	-0.25	1			
HML	-0.07	0.68	-0.25	0.20	1		
RMW	0.01	0.43	0.08	-0.08	0.54	1	
CMA	0.00	0.31	-0.25	0.33	0.59	-0.02	1

조선대학교
CHOSUN UNIVERSITY

<표 4> 회귀분석 결과

이 표는 게시글로부터 얻어진 투자심리지표인 Bullish와 Disagree의 주식수익률과 변동성에 대한 회귀분석을 실시한 결과를 보고한다. 게시글로부터 얻은 변수인 Bullish, Disagree, Messages는 2018년 1월 2일부터 2019년 12월 31일까지 표본기간동안 거래일 중 장 중 시간인 09:00부터 15:30까지의 관측치를 이용한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수 의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t 값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다. 모든 회귀분석에서는 산업고정효과와 월별고정효과를 고려하였다.

	Return	Volatility
Intercept	-0.0395*** (-29.690)	0.0722*** (51.841)
Bullish	0.0218*** (181.824)	0.0001*** (2.621)
Disagree	0.0245*** (37.355)	0.0087*** (28.807)
lnNo. Messages	-0.0002** (-2.523)	0.0059*** (75.387)
Market	0.8898*** (108.317)	
lnSize	0.0018*** (38.427)	-0.0046*** (-112.895)
lnM/B	-0.0006*** (-5.842)	0.0042*** (56.468)
lnVolume		0.0022*** (35.414)
adjR^2	0.505	0.404
No. Observations	171,500	171,500

조선대학교
CHOSUN UNIVERSITY

<표 5> 표본 외 회귀분석 결과

이 표는 게시글로부터 얻어진 투자심리지표인 t시점 Bullish와 Disagree의 t+1시점 주식수익률과 변동성에 대한 표본 외 회귀분석을 실시한 결과를 보고한다. 게시글로부터 얻은 변수인 Bullish, Disagree, Messages는 2018년 1월 2일부터 2019년 12월 31일까지 표본기간동안 거래일 중 장 중 시간인 09:00부터 15:30까지의 관측치를 이용한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다. 모든 회귀분석에서는 산업고정효과와 월별고정효과를 고려하였다.

	Return	Volatility
Intercept	-0.0277*** (-15.487)	0.0649*** (52.963)
Bullish	0.0007*** (4.708)	0.0003*** (5.652)
Disagree	0.0043*** (5.175)	0.0103*** (34.134)
lnNo. of Messages	-0.0014*** (-13.886)	0.0057*** (87.040)
Market	1.0054*** (89.938)	
lnSize	0.0007*** (11.426)	-0.0045*** (-116.731)
lnM/B	0.0019**** (14.097)	0.0041*** (55.738)
lnVolume		0.0028*** (55.072)
adjR^2	0.056	0.423
No. Observations	171,150	171,150

조선대학교
CHOSUN UNIVERSITY

<표 6> 개인투자자 그룹별 회귀분석 결과

이 표는 개인투자자 거래량 크기에 따라 구분된 기업 집단별로 회귀분석을 실시한 결과를 보고한다. 개인투자자의 거래량은 각 기업의 전체 거래량 대비 개인투자자의 거래량 비율을 사용하였고, 개인투자자의 거래량 크기에 따라 3분위로 구분된 기업 집단 중 가장 낮은 개인투자자 거래량을 가진 기업 집단을 Low, 중간의 개인투자자 거래량을 가진 집단을 Middle, 가장 높은 개인투자자 거래량을 가진 집단을 High로 구분하였다. 게시글로부터 얻은 변수인 Bullish, Disagree, Messages는 2018년 1월 2일부터 2019년 12월 31일까지 표본기간동안 거래일 중 장 중 시간인 09:00부터 15:30까지의 관측치를 이용한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다. 모든 회귀분석에서는 산업고정효과와 월별고정효과를 고려하였다.

	Low		Middle		High	
	Return	Volatility	Return	Volatility	Return	Volatility
Intercept	-0.0254*** (-9.740)	0.0721*** (31.485)	-0.0417*** (-6.971)	0.1076*** (22.960)	-0.0353*** (-4.938)	0.1278*** (21.834)
Bullish	0.0179*** (86.779)	0.0008*** (9.600)	0.0221*** (107.176)	-0.0003*** (-3.162)	0.0242*** (126.345)	-0.0004*** (-4.618)
Disagree	0.0153*** (13.635)	0.0057*** (12.695)	0.0214*** (18.901)	0.0056*** (10.806)	0.0278*** (26.065)	0.0083*** (15.798)
lnNo. of Messages	-0.0005 (-0.685)	0.0032*** (33.503)	-0.0006*** (-4.069)	0.0056*** (40.001)	-0.0006*** (-4.452)	0.0084*** (54.337)
Market	0.0193* (1.670)		-0.0074 (-0.505)		0.0122 (0.784)	
lnSize	0.0013*** (13.457)	-0.0038*** (-59.098)	0.0018*** (6.354)	-0.0072*** (-31.518)	0.0014*** (4.241)	-0.0076*** (-26.889)
lnM/B	-0.0004*** (-2.883)	0.0025*** (25.699)	0.0001 (0.468)	0.0062*** (36.097)	0.0002 (0.811)	0.0064*** (29.420)
lnVolume		0.0019*** (23.315)		0.0032*** (28.400)		0.0018*** (14.426)
adjR^2	0.350	0.362	0.460	0.357	0.529	0.391
No. Observations	57,330	57,330	56,840	56,840	57,330	57,330

조선대학교
CHOSUN UNIVERSITY

<표 7> 개인투자자그룹별 표본 외 회귀분석 결과

이 표는 개인투자자 거래량 크기에 따라 구분된 기업 집단별로 표본 외 회귀분석을 실시한 결과를 보고한다. 개인투자자의 거래량은 각 기업의 전체 거래량 대비 개인투자자의 거래량 비율을 사용하였고, 개인투자자의 거래량 크기에 따라 3분위로 구분된 기업 집단 중 가장 낮은 개인투자자 거래량을 가진 기업 집단을 Low, 중간의 개인투자자 거래량을 가진 집단을 Middle, 가장 높은 개인투자자 거래량을 가진 집단을 High로 구분하였다. 게시글로부터 얻은 변수인 Bullish, Disagree, Messages는 2018년 1월 2일부터 2019년 12월 31일까지 표본기간동안 거래일 중 장 중 시간인 09:00부터 15:30까지의 관측치를 이용한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다. 모든 회귀분석에서는 산업고정효과와 월별고정효과를 고려하였다.

	Low		Middle		High	
	Return	Volatility	Return	Volatility	Return	Volatility
Intercept	-0.0207*** (-6.226)	0.0693*** (36.964)	-0.0932*** (-11.228)	0.1133*** (24.919)	-0.1386*** (-12.766)	0.1354*** (23.719)
Bullish	-0.0000 (0.422)	0.0008*** (10.555)	0.0007*** (2.717)	-0.0000 (-0.781)	0.0011*** (4.658)	-0.0002*** (-2.262)
Disagree	0.0003 (0.198)	0.0066*** (14.631)	0.0049*** (3.565)	0.0071*** (13.911)	0.0081*** (5.578)	0.0100*** (19.041)
lnNo. of Messages	-0.0006*** (-3.573)	0.0032*** (37.384)	-0.0023*** (-11.627)	0.0058*** (48.947)	-0.0031*** (-15.346)	0.0079*** (61.862)
Market	0.0221 (1.527)		0.0084 (0.429)		-0.0128*** (-0.585)	
lnSize	0.0006*** (4.762)	-0.0039*** (-60.194)	0.0039*** (9.818)	-0.0077*** (-34.327)	0.0056*** (11.079)	-0.0085*** (-30.315)
lnM/B	0.0012*** (6.171)	0.0025*** (25.760)	0.0011*** (3.542)	0.0062*** (37.188)	0.0025*** (6.342)	0.0064*** (30.364)
lnVolume		0.0023*** (30.402)		0.0038*** (39.911)		0.0028*** (28.550)
adjR^2	0.003	0.373	0.007	0.383	0.010	0.416
No. Observations	57.213	57.213	56,724	56,724	57.213	57.213

조선대학교
CHOSUN UNIVERSITY

<표 8> 매수의견점수에 따른 3분위 포트폴리오

이 표는 고안된 투자심리지표 중 매수의견점수를 이용하여 구성된 3분위 포트폴리오의 성과를 보고한다. 매수의견점수는 t-1일에서 게시글의 센티멘트를 이용하여 추정되고, 매수의견점수가 가장 낮은 기업으로 구성된 포트폴리오를 Low, 중간 매수의견점수를 가진 기업으로 구성된 포트폴리오를 Middle, 가장 높은 매수의견점수를 가진 기업으로 구성된 포트폴리오를 High로 표시한다. 검증결과는 가치가중법(패널 A)와 동일가중법(패널 B)로 구분하고, 각각의 포트폴리오 평균수익률을 이용하여 계산된 Sharpe Ratio, 포트폴리오의 샤프비율에서 시장 포트폴리오의 샤프비율을 차감한 ExtraSR, 연간수익률 및 CAPM과 Fama-French 3요인 모형(FF3), Fama-French 5요인 모형(FF5)의 위험조정 수익률인 α를 제시한다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다.

패널 A. 가치가중 포트폴리오

Sentiment Portfolio	Sharpe Ratio	Extra SR	Annual Return(%)	CAPM		FF3		FF5	
				α	R^2	α	R^2	α	R^2
Low	0.03	0.06	18.86	0.02***	0.65	0.02***	0.69	0.02***	0.69
Middle	0.03	0.06	21.77	0.02***	0.65	0.02***	0.67	0.02***	0.67
High	0.06	0.09	41.93	0.02***	0.44	0.01***	0.49	0.01***	0.49
Market	-0.03	-	-9.87						

패널 B. 동일가중 포트폴리오

Sentiment Portfolio	Sharpe Ratio	Extra SR	Annual Return(%)	CAPM		FF3		FF5	
				α	R^2	α	R^2	α	R^2
Low	0.03	0.06	18.79	0.02***	0.35	0.02***	0.58	0.01***	0.58
Middle	0.02	0.05	12.89	0.02***	0.34	0.01***	0.61	0.01***	0.62
High	0.07	0.1	68.96	0.02***	0.24	0.01***	0.49	0.01***	0.50
Market	-0.03	-	-9.87						

조선대학교
CHOSUN UNIVERSITY

<표 9> 개인투자자 그룹별 매수의견점수 포트폴리오

이 표는 매수의견점수와 개인투자자의 거래량에 따라 이중정렬된 포트폴리오의 성과를 보고한다. 매수의견점수는 t-1일에서 게시글의 센티멘트를 이용하여 추정되고, 매 t-1일 매수의견점수를 기준으로 정렬하여 전체 주식을 Top, Middle, Bottom으로 3등분 한 뒤, 이어서 각 3분위로 나뉜 포트폴리오 내의 주식들을 다시 개인투자자의 거래량에 따라 High, Middle, Low로 정렬하여 3등분한다. 개인투자자의 거래량은 각 기업의 전체 거래량 대비 개인투자자의 거래량 비율을 사용하였다. 검증결과는 가치가중법(패널 A)와 동일가중법(패널 B)로 구분하고, 각각의 포트폴리오 평균수익률을 이용하여 계산된 Sharpe Ratio, 포트폴리오의 샤프비율에서 시장 포트폴리오의 샤프비율을 차감한 ExtraSR, 연간수익률 및 CAPM과 Fama-French 3요인 모형(FF3), Fama-French 5요인 모형(FF5)의 위험조정수익률인 α를 제시한다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다.

패널 A. 가치가중 포트폴리오

Sentiment Portfolio	Sharpe Ratio	Extra SR	Annual Return(%)	CAPM		FF3		FF5	
				α	R^2	α	R^2	α	R^2
Low bottom	0.01	0.04	5.61	0.02***	0.59	0.02***	0.64	0.02***	0.64
Low middle	0.04	0.07	26.01	0.02***	0.68	0.02***	0.68	0.02***	0.68
Low top	0.01	0.04	7.12	0.02***	0.52	0.01***	0.54	0.01***	0.54
Middel bottom	0.03	0.06	21.19	0.02***	0.18	0.01***	0.43	0.01***	0.44
Middle middle	-0.00	0.03	-12.34	0.02***	0.21	0.01***	0.50	0.01***	0.50
Middle top	0.12	0.15	274.50	0.02***	0.16	0.02***	0.32	0.02***	0.33
High bottom	0.09	0.12	153.28	0.02***	0.19	0.02***	0.41	0.02***	0.42
High middle	0.11	0.14	167.03	0.02***	0.18	0.02***	0.44	0.02***	0.46
High top	0.19	0.22	545.22	0.02***	0.09	0.01***	0.26	0.01***	0.29
Market	-0.03	-	-9.87						

패널 B. 동일가중 포트폴리오

Sentiment Portfolio	Sharpe Ratio	Extra SR	Annual Return(%)	CAPM		FF3		FF5	
				α	R^2	α	R^2	α	R^2
Low bottom	0.00	0.03	0.31	0.02***	0.49	0.01***	0.62	0.01***	0.62
Low middle	0.02	0.05	12.25	0.02***	0.50	0.01***	0.63	0.01***	0.63
Low top	0.01	0.04	9.03	0.02***	0.39	0.01***	0.54	0.01***	0.54
Middle bottom	0.01	0.04	3.05	0.02***	0.26	0.02***	0.54	0.02***	0.55
Middle middle	-0.02	0.01	-19.63	0.02***	0.26	0.01***	0.57	0.01***	0.57
Middle top	0.05	0.08	44.95	0.02***	0.21	0.01***	0.46	0.01***	0.47
High bottom	0.05	0.08	56.59	0.02***	0.21	0.02***	0.48	0.02***	0.50
High middle	0.06	0.09	55.41	0.01***	0.19	0.01***	0.51	0.01***	0.52
High top	0.13	0.16	199.09	0.01***	0.13	0.01***	0.38	0.01***	0.41
Market	-0.03	-	-9.87						

조선대학교
CHOSUN UNIVERSITY

<표 10> 보유기간에 따른 정보력 검증

이 표는 고안된 투자심리지표 중 매수의견점수를 이용하여 구성된 3분위 포트폴리오 중 가장 높은 매수의견점수를 갖는 기업들로 구성된 포트폴리오의 보유기간에 따른 정보력을 보고한다. 정보력을 측정하기 위해서 1일부터 1달까지 평균 매수의견점수인 Average Bullish를 측정하고, 각 기간에 대해 포트폴리오 보유 기간 또한 1일부터 1달까지 교차하여 보유기간 별 포트폴리오 평균수익률을 이용하여 계산된 포트폴리오의 샤프비율에서 시장 포트폴리오의 샤프비율을 차감한 ExtraSR을 제시한다. 검증결과는 가치가중법(패널 A)와 동일가중법(패널 B)로 구분하여 표시하였다.

$$ExtraSR_{b,h} = BullishPf_SR_{b,h} - MarketPf_SR_{h}$$

b = Period for calculating average bullish, b∈{1day, 2day, 3day, 1week, 2week, 1month}
h = Portfolio holding period, h∈{1day, 2day, 3day, 1week, 2week, 1month}

패널 A. 가치가중 포트폴리오

		Market SR	Average Bullish					
			1day	2day	3day	1week	2week	1month
Holding Period	1day	-0.03	0.09	0.63	0.87	1.12	1.42	1.60
	2day	-0.04	0.06	0.52	0.75	1.04	1.37	1.60
	3day	-0.05	0.05	0.46	0.69	0.98	1.32	1.57
	1week	-0.06	0.07	0.41	0.61	0.93	1.28	1.52
	2week	-0.09	0.14	0.41	0.58	0.84	1.14	1.46
	1month	-0.17	0.19	0.39	0.54	0.77	1.06	1.39

패널 B. 동일가중 포트폴리오

		Market SR	Average Bullish					
			1day	2day	3day	1week	2week	1month
Holding Period	1day	-0.03	0.10	0.65	0.86	1.06	1.27	1.28
	2day	-0.04	0.10	0.55	0.76	1.00	1.24	1.27
	3day	-0.05	0.12	0.50	0.71	0.96	1.20	1.26
	1week	-0.06	0.12	0.45	0.65	0.90	1.13	1.23
	2week	-0.09	0.17	0.43	0.58	0.79	1.02	1.17
	1month	-0.17	0.24	0.42	0.53	0.70	0.93	1.14

조선대학교
CHOSUN UNIVERSITY

[부 록]

부록 1. 텍스트 데이터 전처리

비정형 데이터 형태의 게시글 데이터는 분석을 위해 각 게시글을 단어 벡터로 변환하는 정제 과정이 필요하다. 텍스트 데이터를 정제하는 것은 일반적으로 다음과 같은 과정을 통해 진행된다. 먼저 한글과 숫자가 아닌 모든 언어 및 특수문자를 제거한다. 두 번째로 토큰화 과정을 통해 문장 성분을 분석하여 형태소를 추출한 후 투자 의견에 영향을 주지 않는 조사를 제외하고 명사, 동사, 형용사, 일반부사, 감탄사만을 텍스트 분석에 사용한다.[8] 세 번째로 각 게시글을 단어들의 벡터 형식으로 표현하는 문서 단어 행렬로 변환한다. 이 때 단순히 게시글에서 나온 빈도수를 사용하여 단어를 벡터화 하지 않고 문서 내에서 특정 단어가 특정 문서 내에서 어느 정도 많이 사용되었는지를 상대적으로 나타내어 특정단어가 해당 게시글에서 얼마나 중요한지를 판단할 수 있는 TF-IDF(Term Frequency-Inverse Document Frequency) 값을 이용하여 단어 문서 행렬을 구축한다(Salton and McGill, 1983).

TF-IDF는 TF(Term Frequency)에 IDF(Inverse Term Frequency)의 가중치를 고려한 값으로 문서 내 단어를 벡터화 시키는 방법을 말한다. TF는 특정 단어가 특정 문서 내에서 나타난 빈도수로 이 값이 클수록 특정문서에서 많이 사용되는 단어라는 것을 알 수 있다. 그러나 특정 단어가 전체문서에서도 많이 나타난다면 그 단어가 특별한 의미를 갖지 않고 흔하게 사용된다고 볼 수 있다. 따라서 단어 빈도수만 고려하는 것이 아니라 특정 단어가 전체 문서에서 얼마나 등장 했는지를 나타내는 DF(Document Frequency)를 함께 고려해야 하는데 이 값에 역수를 취한 값인 IDF이다. 결과적으로 TF-IDF의 값이 높을수록 특정 단어가 특정 문서에서만 빈

8) 형태소 분석을 위해 MeCab 형태소 분석기를 사용하였다.

조선대학교
CHOSUN UNIVERSITY

번하게 나타나 상대적으로 해당 문서의 다른 단어들에 비해 중요하다는 것을 알 수 있다. 본 논문에서는 이 과정을 거칠 때 연구에 사용된 전체 게시글에서 10번 이상 나온 단어들만 선별하여 TF-IDF값을 계산하는데 이용하였다. 모든 전처리 과정은 파이썬(python)을 이용하여 진행되었다.

부록 2. 수익률을 이용한 훈련데이터 구축

머신러닝의 지도학습 기반 분류모델을 사용하기 위해서는 학습 데이터가 잘 구성되어 있어야 한다. 본 논문에서는 게시판의 게시글이 의미가 있다면 주식수익률에 대해 설명력과 예측력을 가질 것이라고 보기 때문에 수익률이 높은 날에는 긍정적인 글이 나타나고 수익률이 낮은 날에는 상대적으로 부정적인 글이 나타날 것이라고 생각해 볼 수 있다. 따라서 우리는 학습데이터를 구축하기 위해 먼저 개별 기업의 게시글에 개별 기업의 일별 수익률을 매칭시킨 후 일별 수익률을 수익률 크기에 따라 일별로 정렬한 후 가장 상위의 수익률을 가지고 있는 기업에 쓰인 게시글들을 긍정적인 센티멘트를 갖는 매수의견을 나타내는 글로 가정하고 가장 하위의 수익률과 함께 쓰인 기업의 게시글들을 부정적인 센티멘트를 갖는 매도의견으로 가정하여 양 극단의 수익률과 함께 나타난 글들을 추출하였다. 극단치에 해당하는 수익률과 함께 나타난 게시글일수록 투자의견을 잘 나타내고, 평균적인 수익률이 나타난 날 쓰인 게시글들은 상대적으로 투자의견이 희석될 수 있다. 그러므로 우리는 일별로 정렬된 수익률 분포의 양극단 1%부터 10%까지 1%씩 순차적으로 늘려가며 해당 수익률과 함께 쓰인 글에 투자의견을 부여하여 학습데이터를 구축 한 후, 매수의견과 매도의견이 가장 잘 구분되어 있는 학습 데이터 집합을 찾기 위해 투자의견이 잘 구분된 학습 데이터 일수록 나머지 게시글을 잘 분류하여 주식수익률에 대해 정보력이 높아질 것이라 보고 각 학습 데이터 집합별로 나머지

조선대학교
CHOSUN UNIVERSITY

게시글의 투자의견을 분류하여 주식수익률에 대한 예측력을 비교분석 하였다. 예측력을 살펴보는 자세한 과정은 부록 4에서 보고하고 있다. 각 학습 데이터 집합에서 분류된 매수의견 게시글과 매도의견 게시글 개수는 <표 A.1>에서 보고한다.

여기에 <표 A.1> 삽입

부록 3. 분류모델

게시글을 수익률에 따라 각 학습 데이터 집합으로 나눈 후, 학습 데이터를 사용하여 분류모델을 선택하기 위해 각 학습 데이터 집합에 대해 분류모델별로 성능을 측정한다. 본 논문에서는 게시글을 투자의견에 따라 분류하기 위한 분류모델로 서포트 벡터 머신(Support Vector Machine), 나이브 베이즈(Naïve Bayes), 그리고 로지스틱 회귀(Logistic Regression)를 이용하여 각 학습 데이터 집합에 대해 분류 성능을 비교분석하였다. 분류 성능을 확인하기 위해 각 학습 데이터 집합에서 학습 데이터의 20%를 검증 데이터로 다시 분류하여 80%의 학습 데이터로 모델을 학습한 뒤 20%의 검증 데이터에 대한 분류 예측의 정확도를 측정하였다. <표 A.2>에서 확인되는 것처럼 1~2% 극단치에서 모든 모델에서 전반적으로 높은 정확도를 얻을 수 있었다. 이에 본 연구에서는 텍스트 분류를 비롯하여 많은 분류문제를 잘 해결하는 것으로 알려져 있는 서포트 벡터 머신을 이용하여 학습 데이터를 훈련시키고 나머지 모든 글에 투자의견을 부여한다.

여기에 <표 A.2> 삽입

부록 4. Sharpe ratio

본 논문에서는 각 학습 데이터 집단을 이용해 다양한 머신러닝 분류모델을 학습시키고 나머지 게시글의 투자의견을 분류하였다. 게시판이 의미를 가지고 있다면

조선대학교
CHOSUN UNIVERSITY

게시판에서 나타난 투자의견을 이용하여 포트폴리오를 구성했을 때 높은 수익률을 얻을 수 있을 것이다. 따라서 우리는 수익률을 이용하역 구축된 각 학습 데이터 집합으로 서포트 벡터 머신(Support Vector Machine), 나이브 베이즈(Naive Bayes), 로지스틱 회귀(Logistic Reression) 분류모델을 학습시킨 후 나머지 게시글을 매수의견과 매도의견으로 분류한 전체 게시글로 개별 기업의 일별 매수의견 점수를 계산하고 이를 통해 포트폴리오를 구성하여 샤프비율(Sharpe ratio)를 살펴보았다. <표 A.3>은 분류모델별로 각 수익률 분포에 따른 학습 데이터 집단에서 매수의견점수에 따라 3분위 포트폴리오를 구축한 후 가장 매수의견점수가 높은 기업들로 구성된 포트폴리오의 샤프비율을 보여주고 있다.

여기에 <표 A.3> 삽입

포트폴리오의 성과를 나타내는 샤프비율은 수익률 분포의 양 극단치를 이용해 학습데이터를 구축 할수록 높게 나타났으며, 서포트 벡터 머신을 적용하여 동일가중 포트폴리오를 구축한 경우 가장 높게 나타났다. 이는 양 극단의 수익률을 통해 구축된 학습 데이터가 매수의견과 매도의견으로 잘 구분되었으며, 게시판이 주식수익률을 예측하는 데 있어 정보를 가지고 있음을 나타낸다. 이에 본 연구에서는 부록 3에서 선택된 서포트 벡터 머신을 분류모델로 이용한 경우 가치가중 포트폴리오와 동일가중 포트폴리오에서 가장 높은 샤프비율을 갖는 양 극단 2%의 수익률과 함께 나타난 글로 구축된 학습 데이터를 사용하여 투자의견이 부여된 게시글을 실증분석에 사용한다.

부록 5. 학습 데이터의 강건성 검증

수익률을 이용하여 구축된 학습 데이터의 정확도를 검증하기 위해 분석에 사용된 수익률 양 극단 2%의 글로 구축된 학습 데이터의 센티멘트를 한국에서 일반적

조선대학교
CHOSUN UNIVERSITY

인 공통감정사전으로 알려져 있는 서울대 감정사전(http://word.snu.ac.kr/kosac)을 이용하여 분류한 후 수익률의 상승 및 하락 방향과 일치하는지 비교하였다. 이를 위해 일별로 양 극단의 수익률을 갖는 기업의 모든 게시글을 종합하여 그 날 게시글에서 형성되는 기업에 대한 전반적인 센티멘트를 서울대 감정사전을 통해 측정하였다. 만약 수익률을 이용한 학습 데이터 구축이 정확하다면 수익률이 크게 상승하여 긍정 센티멘트를 갖을 것으로 예상되는 기업의 서울대 감정사전을 이용하여 측정된 게시글의 종합적인 센티멘트 또한 긍정적으로 나타날 것이고, 수익률이 크게 하락하여 부정적인 센티멘트를 갖을 것으로 예상되는 기업의 게시글의 종합적인 센티멘트는 부정적으로 나타날 것이다. 서울대 감정사전을 통해 일별 수익률 상위 2%에 해당하는 기업의 글과 하위 2%에 해당하는 기업의 글의 센티멘트를 분류한 결과 수익률이 상승한 기업에 쓰여진 게시글에 대해서는 74%를 긍정 센티멘트로 분류한 반면, 수익률이 하락한 날 기업에 쓰여진 게시글에 대해서는 32%만을 부정 센티멘트로 분류하는 것으로 확인되었다.

이러한 결과가 나온 이유는 두 가지 측면에서 생각해볼 수 있는데, 먼저 수익률이 하락한 기업의 게시글에서 서울대 감정사전에서 부정적인 센티멘트를 갖는 단어가 나타나지 않기 때문으로 이러한 경우 수익률이 하락한 날의 게시글은 부정적인 감정을 갖는다고 보기 어려울 수 있다. 다음으로 게시글 전반에 걸쳐 지배적으로 나타나는 긍정 센티멘트를 가진 공통단어가 존재하여 단어 수를 기반으로 센티멘트를 분류하는 감정사전을 이용한 분류방법이 지배적인 단어에 영향을 받아 잘못 분류 할 경우가 존재한다. 이러한 경우 수익률이 하락하는 날 기업에 쓰인 게시글이 부정 센티멘트를 가지고 있다면 지배적으로 나타나는 공통단어를 감정사전에서 제외한 감정사전을 이용하여 센티멘트를 분류하는 경우 수익률이 하락하는 날 기업의 게시글이 부정 센티멘트를 갖는 것으로 나올 것이다.

조선대학교
CHOSUN UNIVERSITY

수익률이 하락한 기업에 쓰인 게시글이 부정 센티멘트로 분류되지 않는 이유가 지배적으로 나타나는 공통단어 때문인지, 실제로 부정 센티멘트가 존재하지 않기 때문인지 확인하기 위해 본 연구에서는 일별 수익률 상위 2%에 해당하는 기업에 쓰인 게시글의 단어와 일별 수익률 하위 2%에 해당하는 기업에 쓰인 게시글의 단어의 서울대 감정사전에서 제시된 긍정점수에서 부정점수를 차감하여 긍정 센티멘트의 정도를 파악하고, 상위 2%와 하위 2% 각 집단에서 단어별로 단어가 나타난 횟수를 가중치로 곱하여 각 집단에서 지배적인 단어들을 확인하였다. 두 집단의 상위 200개의 지배적인 단어를 비교한 결과 181개가 동일한 것으로 나타나 게시판에는 공통적으로 사용되는 긍정적인 센티멘트를 갖는 단어가 존재함을 확인하였고, 이러한 단어는 모든 글에서 공통적으로 나타나기 때문에 센티멘트를 분류하는데 노이즈를 발생시킬 수 있으므로 감정사전에서 제외할 필요가 있다. 이에 본 연구에서는 공통적으로 사용되는 지배적인 단어를 1개부터 20개까지 1개씩 늘려가며 감정사전에서 제외시킨 후 지배적인 단어가 제외된 감정사전으로 수익률이 하락한 기업에 쓰인 게시글의 센티멘트를 측정하여 살펴보았고 각 감정사전에 의해 부정 센티멘트로 분류된 비율은 [그림 A.1]에서 보고한다.

여기에 [그림 A.1] 삽입

분석 결과 게시글에서는 공통적으로 나타나는 긍정 센티멘트를 갖는 지배단어가 존재하며 이 단어들을 서울대 감정사전에서 제외함으로써 게시글의 센티멘트를 분류할 때 고려되지 않도록 하면 수익률이 하락한 날 기업에 쓰인 게시글의 센티멘트 또한 부정 센티멘트를 갖는 것으로 나타났다. 결과적으로 수익률을 이용하여 학습데이터를 구축하는 경우 게시글에 전반적으로 나타나는 공통단어가 포함되지만 이를 제외하면 수익률이 상승한 날은 긍정 센티멘트 게시글로, 수익률이 하락한 날은 부정 센티멘트 게시글로 잘 분류됨을 검증하였다. 감정사전 기반 센티멘

조선대학교
CHOSUN UNIVERSITY

트 분류 방법은 감정사전에 존재하는 단어에 의존하여 텍스트의 센티멘트를 분류하기 때문에 이를 사용하기 위해서는 노이즈를 발생시킬 수 있는 단어가 감정사전에 존재하는 경우 이를 제외시키고 감정사전을 사용하여야 하는데, 이 때 공통단어를 몇 개까지 제외시킬 것인가에 연구자의 주관이 개입되야 하는 한계가 존재한다. 그러나 학습 데이터를 이용하여 분류모델을 학습시켜 텍스트의 센티멘트를 분류하는 머신러닝 방법은 단어 하나의 의미보다 텍스트의 맥락을 보고 센티멘트를 분류하기 때문에 개별 단어의 영향력이 감정사전에 의한 방법보다 작아 공통단어를 제외시키지 않고도 사용할 수 있다. 이에 본 연구에서는 수익률을 이용하여 구축한 학습 데이터를 사용하는 머신러닝 분류방법으로 게시글의 센티멘트를 분류하고 투자심리지표를 측정하여 주식수익률에 대한 정보력과 예측력을 분석하였고, 머신러닝 분류방법의 성능을 비교하기 위해 서울대 감정사전으로 분류된 센티멘트를 이용하여 측정된 투자심리지표 주식수익률에 대한 정보력과 예측력을 부록에서 보고한다.

서울대 감정사전으로 게시글의 센티멘트를 분류하는 것은 게시글에 나타난 단어가 감정사전에 존재하는 경우 해당하는 감정점수를 부여하는 방식으로 이를 이용한 투자심리지표 측정 과정은 다음과 같다. 먼저 서울대 감정사전에 존재하는 분석 대상 게시글의 단어들에 서울대 감정사전에서 제시하는 긍정점수와 부정점수를 부여한 후 이를 각각 종합하여 게시글의 긍정점수와 부정점수를 측정한다. 이 때 개별 게시글의 긍정점수와 부정점수는 해당 게시글에서 서울대 감정사전에 존재하는 단어 수로 나눠줌으로써 0~1의 값을 갖도록 하였다. 이 후 긍정점수와 부정점수를 비교하여 큰 값을 갖는 센티멘트를 게시글에 부여하고, 이를 이용하여 본 연구에서 사용하였던 각 기업의 일별 투자심리지표를 동일한 방법으로 측정한다.

만약 본 연구에서 사용한 학습 데이터 구축방법이 게시판의 텍스트 데이터를 분

조선대학교
CHOSUN UNIVERSITY

석하는데 적합하다면 이를 통해 측정한 투자심리지표가 서울대 감정사전으로 센티멘트를 분류하여 측정한 센티멘트보다 주식수익률에 대해 더 높은 설명력과 예측력을 가질 것이다. 이를 확인하기 위해 서울대 감정사전으로 측정된 투자심리지표를 사용하여 주요 분석방법과 동일한 방법으로 4.1절의 회귀분석 모형을 사용하여 설명력을 확인하고, 투자심리지표 중 매수의견점수를 이용하여 3분위 포트폴리오를 구축한 후 샤프비율을 살펴본다. <표 A.4>는 서울대 감정사전으로 측정된 투자심리지표를 사용하여 주식수익률과 변동성에 대하여 회귀분석과 표본 외 회귀분석을 실시한 결과를 나타낸다.

여기에 <표 A.4> 삽입

서울대 감정사전을 이용하여 측정된 투자심리지표는 회귀분석에서 수익률에 대하여 5.5%의 설명력을 갖는 것으로 나타났다. 이는 <표 3>에 제시된 수익률을 이용하여 구축한 학습 데이터로 분류 모델을 학습시킨 후 센티멘트를 분류하여 측정한 투자심리지표가 수익률에 대하여 갖는 설명력인 50.5%보다 크게 작은 값으로 서울대 감정사전으로 게시판의 센티멘트를 분류하는 경우 게시글의 센티멘트를 정확하게 분류하지 못하여 이를 통해 계산된 투자심리지표의 설명력이 크게 떨어진다는 것을 확인할 수 있었다. 서울대 감정사전을 이용하여 구축된 투자심리지표가 예측력을 갖는지를 확인하기 위해 매수의견점수를 이용하여 3분위 포트폴리오를 구축한 뒤 본 연구에서 사용한 포트폴리오 투자전략과 동일하게 보유기간이 1일인 포트폴리오의 수익률을 통해 샤프비율을 계산한 결과 또한 상위의 매수의견점수를 갖는 기업으로 구성된 가치가중 및 동일가중 포트폴리오의 샤프비율이 0과 0.06으로 <표 7>에 제시된 0.06과 0.07보다 작음으로 예측력 또한 본 연구에서 사용한 학습 데이터로 인해 측정된 투자심리지표보다 낮음이 확인되었다. 이러한 결과는 주식수익률을 이용하여 학습 데이터를 구축하는 방법이 게시판의 텍스트 데이터를

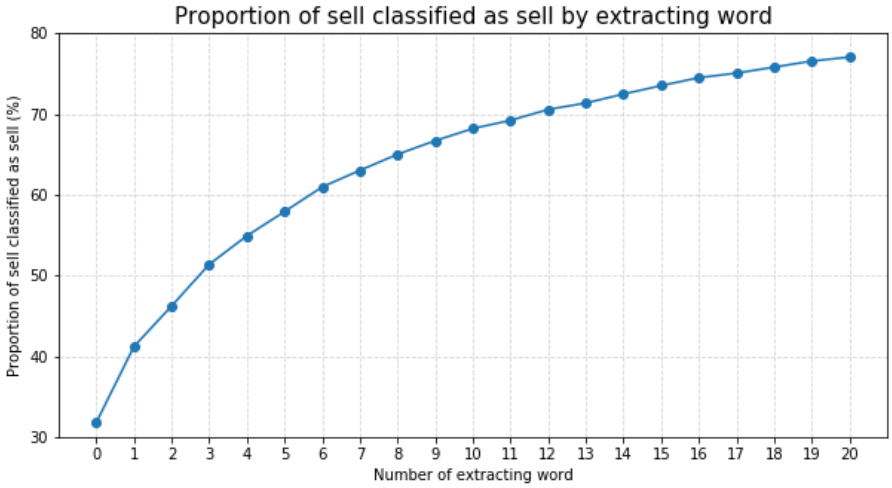
조선대학교
CHOSUN UNIVERSITY

분석하는데 적합하며 서울대 감정사전을 이용하는 감정사전 기반 센티멘트 분류방법은 주식시장에 특화된 글과 온라인 상에서 생성되는 텍스트 데이터의 감정을 측정하는데 한계가 있음을 확인시켜 준다.

조선대학교
CHOSUN UNIVERSITY

[그림 A.1] 수익률이 하락한 기업의 게시글을 부정 센티멘트로 분류한 비율 : 서울대 감정사정으로 분류

이 그림은 게시글에서 전반적으로 나타나는 긍정 센티멘트를 갖는 공통단어가 제외된 서울대 감정사전으로 수익률이 하락한 기업의 게시글의 센티멘트를 측정하여 부정 센티멘트로 나타난 비율을 나타낸다. x축은 서울대 감정사전에서 제외되는 공통단어의 개수를 나타내고, y축은 수익률이 하락한 날 게시글이 부정 센티멘트로 분류된 비율을 나타낸다.



조선대학교
CHOSUN UNIVERSITY

<표 A.1> 학습 데이터 집합에서의 각 투자의견에 해당하는 게시글 수

이 표는 수익률을 이용하여 구축된 각 학습 데이터 집합에서 분류된 매수의견 게시글과 매도의견 게시글의 개수 및 전체 게시글 대비 학습데이터의 비율을 보고한다. 학습데이터는 개별 기업의 게시글에 개별 기업의 일별 수익률을 매칭시킨 후 수익률 크기에 따라 일별로 수익률을 정렬한 후 양 극단 1%부터 10%까지 1%씩 순차적으로 늘려가며 상위 수익률 기업에 쓰인 게시글을 긍정 센티멘트를 갖는 매수의견 게시글, 하위 수익률 기업에 쓰인 게시글을 부정 센티멘트를 갖는 매도의견 게시글로 분류하여 학습데이터를 구축하였다.

Sort: Return	Number of buy (Highest Return)	Number of sell (Lowest Return)	Number of label
±1%	1,380,283	1,341,903	2,722,186 (16.9%)
±2%	1,878,713	1,896,636	3,775,349 (23.4%)
±3%	2,238,762	2,270,159	4,508,921 (27.9%)
±4%	2,524,378	2,562,282	5,086,660 (31.5%)
±5%	2,758,828	2,798,622	5,557,450 (34.4%)
±6%	2,947,813	3,000,766	5,948,579 (36.8%)
±7%	3,134,049	3,200,417	6,334,466 (39.2%)
±8%	3,307,737	3,374,000	6,681,737 (41.4%)
±9%	3,461,889	3,538,800	7,000,689 (43.3%)
±10%	3,615,296	3,691,297	7,306,593 (45.2%)

조선대학교
CHOSUN UNIVERSITY

<표 A.2> 분류모델 성능 비교

이 표는 수익률을 이용하여 구축된 각 학습 데이터 집합으로 서포트 벡터 머신(Support Vector Machine), 나이브 베이즈(Naive Bayes), 로지스틱 회귀(Logistic Reression) 분류모델의 분류 성능을 정확도로 측정한 결과를 보고한다. 각 분류모델의 분류 성능을 측정하기 위해 각 학습 데이터 집합에서 학습 데이터의 20%를 검증 데이터로 다시 분류하여 80%의 학습 데이터로 모델을 학습한 뒤 20%의 검증 데이터에 대한 분류 예측의 정확도를 측정하였다. 학습 데이터는 개별 기업의 게시글에 개별 기업의 일별 수익률을 매칭시킨 후 수익률 크기에 따라 일별로 수익률을 정렬한 후 양 극단 1%부터 10%까지 1%씩 순차적으로 늘려가며 상위 수익률 기업에 쓰인 게시글을 긍정 센티멘트를 갖는 매수의견 게시글, 하위 수익률 기업에 쓰인 게시글을 부정 센티멘트를 갖는 매도의견 게시글로 분류하여 학습 데이터를 구축하였다.

Sort : Return	Support Vector Machine	Naïve Bayes	Logistic Regression
±1%	64.12	64.05	64.42
±2%	63.17	62.64	63.46
±3%	56.32	61.95	62.69
±4%	54.40	61.78	62.56
±5%	52.06	61.55	62.31
±6%	52.06	61.31	62.09
±7%	50.58	61.13	61.85
±8%	50.77	60.96	61.59
±9%	50.78	60.69	61.42
±10%	50.62	60.69	61.35

조선대학교
CHOSUN UNIVERSITY

<표 A.3> 샤프비율을 이용한 포트폴리오 성과 분석

이 표는 수익률을 이용하여 구축된 각 학습 데이터 집합으로 서포트 벡터 머신(Support Vector Machine), 나이브 베이즈(Naive Bayes), 로지스틱 회귀(Logistic Reression) 분류모델을 학습시킨 후 나머지 게시글의 센티멘트를 분류하여 개별 기업의 일별 매수의견점수를 계산하고 이를 통해 구축한 3분위 포트폴리오 중 가장 상위의 매수의견점수를 갖는 기업으로 구성된 포트폴리오의 투자 성과를 샤프비율로 보고한다. 학습 데이터는 개별 기업의 게시글에 개별 기업의 일별 수익률을 매칭시킨 후 수익률 크기에 따라 일별로 수익률을 정렬한 후 양 극단 1%부터 10%까지 1%씩 순차적으로 늘려가며 상위 수익률 기업에 쓰인 게시글을 긍정 센티멘트를 갖는 매수의견 게시글, 하위 수익률 기업에 쓰인 게시글을 부정 센티멘트를 갖는 매도의견 게시글로 분류하여 학습 데이터를 구축하였다.

패널 A. 가치가중 포트폴리오

Sort : Return	Support Vector Machine	Naïve Bayes	Logistic Regression
±1%	0.01	0.02	0.01
±2%	0.06	0.01	0.05
±3%	0.04	0.00	0.05
±4%	0.04	0.00	0.04
±5%	0.04	0.03	0.04
±6%	0.05	0.02	0.04
±7%	0.02	0.02	0.03
±8%	0.04	0.02	0.03
±9%	0.04	0.02	0.05
±10%	0.04	0.02	0.04

패널 B. 동일가중 포트폴리오

Sort : Return	Support Vector Machine	Naïve Bayes	Logistic Regression
±1%	0.07	0.06	0.07
±2%	0.07	0.07	0.07
±3%	0.07	0.07	0.07
±4%	0.07	0.07	0.07
±5%	0.07	0.07	0.07
±6%	0.07	0.07	0.07
±7%	0.06	0.07	0.06
±8%	0.06	0.07	0.06
±9%	0.06	0.07	0.06
±10%	0.06	0.07	0.06

조선대학교
CHOSUN UNIVERSITY

<표 A.4> 주식수익률과 변동성에 대한 회귀분석 결과: 서울대 감정사전을 이용하여 측정된 투자심리지표

이 표는 서울대 감정사전을 이용하여 게시글로부터 측정된 투자심리지표인 Bullish와 Disagree의 주식수익률과 변동성에 대해 회귀분석을 실시한 결과를 보고한다. 종속변수인 Return과 Volatility에 대하여 회귀분석의 경우 첨자t를, 표본 외 분석의 경우 첨자t+1를 표시하였다. 게시글로부터 얻은 변수인 Bullish, Disagree, Messages는 2018년 1월 2일부터 2019년 12월 31일까지 표본기간동안 거래일 중 장 중 시간인 09:00부터 15:30까지의 관측치를 이용한다. lnReturn은 수정주가로 계산된 일별 수익률에 로그를 취한 값을 나타내고, Volatility는 17일간 일별 수익률의 표준편차이다. Messages는 게시글 수, Bullish와 Disagree는 투자심리지표로 매수의견점수와 의견 불일치정도를 나타낸다. lnSize는 시가총액에 로그를 취한 값이고, lnM/B는 시가총액을 장부가로 나눈 값에 로그를 취한 값, lnVolume은 거래량에 로그를 취한 값이다. 괄호 안의 값은 Newey and West(1987)의 표준오차를 이용하여 조정한 t값을 나타내고, 유의수준 1%, 5%, 10%에서 유의한 통계량에는 각각 ***, **, *를 붙여서 표시하였다. 모든 회귀분석에서는 산업고정효과를 고려하였다.

	$Return_t$	$Volatility_t$	$Return_{t+1}$	$Volatility_{t+1}$
Intercept	-0.0069*** (-3.071)	0.0765*** (50.238)	-0.0277*** (-13.804)	0.0725*** (52.849)
Bullish	0.0022*** (18.418)	0.0001** (2.319)	0.0002 (1.142)	0.0002*** (2.785)
Disagree	-0.0014** (-1.972)	-0.0042*** (-10.362)	0.0002 (0.233)	-0.0046*** (-11.462)
lnNo. Messages	0.0011*** (7.319)	0.0060*** (75.090)	-0.0012*** (-11.125)	0.0061*** (87.177)
Market	1.0209*** (91.661)		1.0139*** (91.040)	
lnSize	-0.0001 (-1.536)	-0.0047*** (-113.629)	0.0006*** (9.150)	-0.0047*** (-118.577)
lnM/B	-0.0011*** (7.888)	0.0045*** (59.072)	0.0020*** (14.546)	0.0044*** (57.845)
lnVolume		0.0026*** (42.184)		0.0030*** (57.792)
adjR^2	0.055	0.373	0.054	0.389
No. Observations	171,500	171,500	171,150	171,150