인터넷 검색량을 활용한 주식/암호화폐 시장 예측 비교 연구*

이 세 윤** (프라이드랩)
박 준 기*** (프라이드랩)

Abstract

본 연구는 주식 시장과 암호화폐 시장에서 시장 참여자들의 투자 심리와 행동이 검색 행위와 관련성이 있을 것으로 보고, 시장 가격에 선행하는 검색어 키워드를 추출하고 시계열 분석과 가상의 투자 모형에 적용하여 효과성을 검증하였다. 이를 위해서 주식 시장과 암호화폐 관련 리포트로부터 상위 빈도의 키워드를 추출하여, 키워드별 검색량과 자산시장 지수 간의 시계열 분석을 통해 상관성이 높은 키워드를 선정하였다. 이어서 선정된 키워드들의 검색량을 기반으로 한 가상의 투자 모형에 적용하여 수익률을 검증하였다. 분석 결과, 자산시장별로 직접적인 관련성이 있는 텍스트에서 키워드를 추출하는 것이 그렇지 않은 경우에 비해 수익성에 있어서 유의미한 차이가 있었다. 주식 시장의 경우 매입 후 보유 전략과 비교해 검색량 기반의 투자 모형이 높은 투자 수익률을 기대할 수 있어서 유의미한 투자전략으로 볼 수 있었다. 반면에 암호화폐 시장에서는 검색량 기반의 투자 모형이 매입 후 보유 전략보다 우수한 투자전략으로 보는 데는 한계가 있었다.

[1] 서론

자산시장의 빅데이터를 분석하는 인공지능 알고리즘을 활용한 포트폴리오 구성을 통한 투자가 투자자문기관과 자산 운용회사들에서 다양하게 시도되고 있다. 퀀트 투자 방식에서부터 주식 가격과 거래량과 같은 투자데이터와 PBR, PER과 같은 지표 그리고 경영데이터를 분석하여 최적의 투자 모델을 찾기 위한

* 본 연구는 2021년 하반기 펀드평가 3사(한국펀드평가, FnGuide, KG제로인)의 성균관대학교 자산운용센터(CAPM) 연구비 지원으로 수행되었습니다.
주제어 : 검색트렌드, 빅데이터, 주식투자전략, 코스피 지수, 코스닥 지수, 암호화폐, 투자 모델
JEL 분류기호 : C32, G17
** 프라이드랩 전문위원, E-mail : suyfj77@gmail.com
*** 프라이드랩 연구소장, E-mail : warrenpak@warrenpak.com

노력은 오랜 기간 이루어졌다. 최근에는 온라인 커뮤니티와 소셜 네트워크 서비스에 존재하는 텍스트를 분석해서 투자 모델로 구체화하는 서비스도 출시가 이루어지고 있다. 최근 밴아크(VanEck)사에서 출시한 ETF인 BUZZ가 주목받고 있다.[1] 이들은 SNS에서 1년 동안 언급되는 종목 250~350개를 선정한 후 상위 75개에 해당하는 기업을 모아서 ETF 기초지수를 설정하고 투자하는 방식의 상품을 선보였기 때문이다. 서비스 시작이 2021년 3월이기 때문에 아직은 충분한 투자실적이 축적되지 않았으나, 투자사에서 제시한 지난 5년간의 테스트 결과에 따르면, S&P500이 113% 상승하는 동안 BUZZ는 215% 상승했다고 한다. 투자자에게 빅데이터를 활용하는 것은 다양한 투자 아이디어가 실제 투자 모델로 활용되고 수익을 창출할 수 있는 다양한 시도로써 의미를 갖는다. 빅데이터의 범위가 단순히 주식 시장과 연관된 것에만 머무는 것이 아니라, 투자자 행동 데이터를 연결하여 다양한 투자 모델과 상품을 만들어서 제공한다면 투자자 선택을 높일 수 있다는 점에서 시장에서 매력적인 상품이 될 수 있다.

구글, 네이버와 같은 인터넷 검색은 일상의 필수가 되었기 때문에 개인들의 행동 패턴이 집단적으로 반영된 지표로써 검색량을 활용한 연구가 진행되고 있다. 기존에는 주식 투자자들이 투자 종목에 대한 관심도를 알아내는 방법으로 뉴스와 같은 극단적 사건이 발생했을 때 주식거래량이나 수익률의 변화를 살펴보았으나 직접적 연관성이 떨어진다는 한계가 있었다. Da *et al.*(2011)은 이러한 한계에 대한 대안으로 인터넷 검색량에 주목하였다. 인터넷 검색량은 투자자들이 투자 정보를 획득하기 위한 능동적 행위의 결과가 집단적으로 나타난 것이므로 실질적인 투자자의 관심도(Attention)에 해당한다고 볼 수 있기 때문이다. 실제로 주식 투자자가 경제 상황에 대해 갖는 관심에 대한 민감도는 경제 관련 인터넷 검색량과 관계가 있으며, 특히 시장이 크게 흔들리는 불안한 국면에서는 인터넷 검색이 더 많이 늘어난다는 실증 결과도 존재한다(Preis *et al.* 2013).

투자자 관심을 대표할 수 있는 텍스트를 분석하기 위해서 선행되어야 하는 것은 어떠한 키워드를 추출하는 것이 적합한가 하는 것이다. 주식이나 암호화폐와 같은 개별 투자 자산을 대상으로 하는 투자전략에 적용하기 위해서 일반적으로 투자 대상의 종목명과 그 의미를 유추할 수 있는 키워드 혹은 종목 코드를 기반으로 투자 모형을 구성하는 방법이 고려될 수 있다. 반면에 투자 시장 전체를 대표하는 코스피, 코스닥과 같은 지수 혹은 암호화폐를 대표하는 비트코인 같은 경우에 투자자 관심에 따른 행동 패턴을 찾아내기 위한 키워드는 특정 기업이나 종목 코드가 아니라 시장 전체의 동태와 상관성이 높은 키워드이어야 할 것이다. 즉, 자산시장 전체를 일정 수준으로 설명할 수 있는 검색어 키워드는 개별 종복에 대한 키워드와는 차이가 있을 것이다. 이러한 키워드를 도출하고 실증적 분석을 통해 자산시장의 동태적 트렌드와의 상관성을 발견할 수 있을 것이다. 이러한 과정을 거쳐 투자자 관심에 따른 모델을 검증함으로써 투자전략을 수립할 수 있다(Zhu *et al.* 2021). 따라서 본 연구에서는 주식과 암호화폐 시장을 대상으로 시장을 설명하는 검색어 키워드를 찾아내고 이를 활용한 투자 모델이 얼마나 효과성이 있는지를 검증하고자 한다. 이를 위해서 주식 시장과 암호화폐 관련 리포트로부터 상위 빈도의 키워드를 추출하여, 키워드별 검색어 트렌드와 자산시장 지수 간의 시계열 분석을 통해 상관성이 높은 키워드를 선정하였다. 이어서 선정된 키워드들의 검색량을 기반으로 한 가상의 투자 모형에 적용하여 수익률을 검증하였다. 주식 시장과 암호화폐 시장의 비교를 통해서 시장 간의 투자자 관심에 따른 차이점을 살펴보고 자산시장에 활용 가능한 포트폴리오 투자 모형을 구성하는데 인터넷 검색량을 응용하는 방법을 제공하고자 하였다.

1) 'Buzz' ETF tracking social media sentiment launches Thursday amid Reddit manias in stocks(CNBC, 2021.3.4.) https://www.cnbc.com/2021/03/04/buzz-etf-tracking-social-media-sentiment-launches-thursday-amid-reddit-manias-in-stocks.html

[2] 선행연구

주식 시장은 대규모 참여자들의 의사결정과 행위들의 총합이 반영되어 주가와 거래량과 같은 데이터로 나타나는 전형적인 영역이라고 할 수 있다. 이러한 주식 시장의 변동을 빅데이터를 활용하여 예측하거나 계량화하기 위한 시도들이 활발히 이루어지고 있다. 우민철·김지현(2017)은 인터넷상의 증권 게시판의 게시물과 주가와의 관계 분석을 통해 통계적 유의성을 확인하였다. 특히, 긍정 혹은 부정적 키워드가 포함된 게시물의 건수가 주가에 더욱 영향을 준다고 하였다. 이처럼, 시장 참여자들이 인터넷을 통해 교환하는 정보를 대표할 수 있는 키워드에 주목하는 연구들이 시도되고 있다. 그중 인터넷 검색 엔진을 통해 대규모로 이루어지는 검색 행위에 관한 데이터는 검색 포털 서비스에서 지수화하여 제공하기 때문에 이를 다양한 연구에 활용할 수 있다.

Preis *et al.*(2013)은 경제와 관련한 단어들과 주가지수 수익률을 정량적으로 모델링하는 방법론을 제시하였다. 주식 투자자들은 의사결정에 필요한 정보를 수집하기 위해 특정 단어에 대한 검색 행동을 할 것이기 때문에 인터넷 검색량에 대한 데이터는 주식 시장에 대해 선행하는 패턴을 가질 것이라는 가정을 할 수 있다. Preis et al.(2013)은 주식 시장과 관련해 선정된 단어들에 대해 일정 기간 동안 검색량 증감에 따라, 매수와 매도를 결정하는 방식을 반복하여 수익률을 도출하였고, 검색어 데이터를 활용한 의사결정이 무작위나 단순히 보유하는 방식의 투자 전략(매수 후 유지 전략)보다 효과가 있음을 보였다. 김민수·구평회(2013)는 Preis et al.(2013)의 연구를 한국 시장에서 적용하였는데, KOSPI200 지수를 대상으로 하여, 국내 검색 포털인 네이버에서 제공하는 검색 트렌드 데이터를 활용하였다. Preis et al.(2013)의 연구에 사용된 총 84개의 검색어 데이터를 기반으로 한 투자 전략이 시장 평균보다 높은 경우는 25% 수준에 불과하였다. 또한, 2007년부터 2011년까지를 학습 데이터로 하여 도출한 수익률과 2012년부터 2013년 7월 말까지의 예측 수익률의 상관관계를 분석한 결과 상관성이 높지 않아 한국에서 적용하는 데에는 한계가 있다는 결과를 도출하였다. 다만, 한국 시장에 맞는 검색어 도출이 필요하다는 점을 제시하였다. 구평회·김민수(2015)는 KOSPI 지수와 관련한 일반용어가 아닌, 개별 기업명에 대한 검색량 변화와 해당 기업 주가 및 거래량과의 관계를 분석했다. 이 연구에서는 2007~2013년간 KOSPI와 KOSDAQ 시장에서 검색어를 활용한 투자전략이 평균 수익 이상을 실현할 수 있었으며, 대기업보다 중소기업에서 투자 효과가 있다는 점을 확인하였다. 이처럼 개별 기업명에 대한 검색량과 주가와의 관계에 대해서, 일찍이 투자자 관심 혹은 투자자 주의(investor attention)라는 개념으로 접근이 이루어져 왔다.

Da *et al.*(2011)은 투자자 관심에 대한 측정 지표로서 기존에 활용되었던, 극단적인 수익률과 거래량, 뉴스량이 간접적인 지표에 불과하며, 인터넷 검색량이 직접적이고 실질적인 투자자 관심에 대한 지표가 될 수 있음을 제시하였다. 이후, 한국 내에서도 투자자 관심 변수로써 기업에 대한 인터넷 검색량을 활용한 연구들이 이루어졌다(장영봉 등 2015; 반주일 등 2016; 김류미 2018). 김류미(2018)는 검색 엔진을 통해 기업을 검색한, 인터넷 검색량의 합계를 투자자 관심 척도로 하여, KOSPI 시장에 상장된 모든 기업을 대상으로 투자자별 거래량과 수익률에 미치는 영향을 연구하였다. 인터넷 검색량의 증가는 거래량의 증가를 가져오며, 모든 투자자의 매수와 매도 증가와 유의한 양의 관계가 있음을 확인하였다. 특히 개인투자자의 매수 증가와의 양의 관계가 가장 강함을 확인하고, 인터넷 검색량이 많아지면 주식수익률이 일시적으로 증가할 수 있음을 밝혔다. 이처럼 검색량 빅데이터가 투자자 관심(attention)의 척도임이 확인되고, 인터넷 검색이 투자자 행동에 대해 선행할 수 있다는 점이 연구되면서 추가 연구가 진행되고 있다. 전새미 등(2016)은 비정상적인 인터넷 검색량 증가는 주가 변동

성의 유의한 증가를 가져온다는 것을 확인하고, 산업군별로 구분하였을 때, IT, 소프트웨어, 건설, 유통 산업군에서 특히 강하게 나타난다는 것을 확인하였다. 반면, 투자자 관심 지표로써 검색량 데이터와 주식수익률 관계에 집중한 기존 연구와 달리, 김민수·권혁준(2017)은 주식수익률 자체가 아니라 수익률의 분포 형태, 특히 수익률의 극단적 하락 위험에 집중했다. 네이버 검색량 지수를 이용하여 투자자 관심과 정보 전파 속도를 측정하고, 측정 변수들이 주가 급락 위험을 감소시키는 것을 관찰하였다. 결론적으로 투자자들의 관심이 높아질수록, 기업의 부정적 정보들이 주식시장에 신속하게 전파되어 주가의 극단적인 하락 위험이 감소하는 것으로 해석하였다. 하준성 등(2019)은 인터넷 검색량 지수를 투자자 관심의 척도로 선정했지만, 추가로 애널리스트 투자의견 변경을 조절 변수로 하였을 때 투자자 과소반응과 지연반응이 어떤 영향을 주는지 확인했다. 최근에는 포털 검색량과 검색의 변동성이 주가 동조성 및 총 위험, 체계적 위험에 미치는 영향에 관한 연구(김민수 등, 2020), 검색량 지수를 활용한 투자자 관심 수준이 주가 표류 현상에 미치는 영향에 대한 연구(전경민·남기만, 2020) 등 기업에 대한 인터넷 검색량 지수를 투자자 관심 지표로써 활용한 연구가 활발히 이루어지고 있다.

투자자 관심 척도로서 기업에 대한 인터넷 검색량 기반 연구를 확장해서 연관 검색어 검색량을 기반으로 한 연구가 또 다른 흐름을 형성하고 있다. 김민수와 구평회(2013)는 Preis *et al.*(2013)의 검색어가 한국에서는 한계가 있음을 밝힌 바 있다. 즉 특정 시장 또는, 특정 시기에 의미가 큰 검색어를 적용할 필요성을 제시하였다. 김범수 등(2015)은 KOSPI 지수와 관련된 검색어 트렌드가 거래량과 지수의 변화와 상관관계가 있으며, KOSPI 지수와 의미가 가까운 키워드일수록 거래량과 지수를 더 잘 설명한다고 하였다. 또한, 지수 변화와는 음의 상관관계가 있음을 제시하여, 투자자의 검색 행위가 위험 회피적인 심리가 작용하고 있음을 시사하였다.

한편, 암호화폐 시장은 홍기훈(2021)이 기존의 논문들과 최근의 사회경제적 현상들을 통해 암호화폐(코인)가 화폐가 아닌 투자 자산으로 인식되고 있다고 기고한 바와 같이 변동성이 높은 투자 자산 시장이라고 할 수 있다. 이에 따라, 암호화폐 시장에서도 투자자 관심과 연관성 연구가 최근에 활발하게 이루어지고 있다. Subramaniam and Chakraborty(2021)는 암호화폐 시장에서 발생하는 비효율성이 가격변동에 있어 행동 심리학적 측면의 영향을 살펴볼 수 있다는 관점에서 투자자 관심이 암호화폐 변동에 미치는 영향을 조사했다. 연구 결과는 투자자들은 빈번하게 보이는 뉴스와 순위에 오른 암호화폐(비트코인과 이더리움)에 주목한다는 것을 보여주었다. Choi(2021)는 비트코인 투자자 관심을 확인하기 위해서 SNS인 트위터의 트윗 수를 활용했다. 그 결과 트윗의 1% 증가는 향후 5~10분 동안 약 7%의 유동성 개선을 나타냈다. 즉 트윗과 같은 텍스트 증가가 투자자의 관심을 증가시키고 유동성에 강한 영향을 준다는 점을 나타내고 있다. 양철원(2019)은 비트코인의 국내외 가격 차이를 유발하는 변수로 이성적 요인들이 아닌, 구글 트렌드 지수와 네이버 트렌드 지수의 차이와 같은 심리적 변수가 통계적으로 유의하였다는 실증 결과를 보였다. 더 나아가, 비트코인 시장과 검색어 트렌드와의 관련성에 대해 Zhu *et al.*(2021)은 구글 검색어로 '비트코인'과 비트코인 시장 자체의 변화를 살펴봤다. 그랜저 인과관계와 VAR 분석을 통해서 투자자 관심과 시장 수익 및 변동성 간에 그래져 인과관계가 존재하는 것을 확인했다. 또한, 비선형 분석을 통해서는 비트코인 검색어와 투자 변동성은 비선형적 관계가 있다는 점도 분석해 냈다. 마지막으로 투자자 관심이 비트코인 시장 변동성 예측의 정확도를 높인다는 것도 확인했다. 즉, 비트코인과 같은 암호화폐 시장이 투자자 관심에 따라 영향을 받는다고 할 수 있다.

연관 검색어의 검색량과 투자 자산 간의 관계에 관한 연구는 인간 감정이 의사결정이나 행동에 영향을 미친다는 행동경제학(behavioral economics) 이론에 기반을 두고 있다. 따라서 시장 심리를 반영한 연관 검색어를 어떻게 선택하는가가 더 중요해지고 있다. 유재필 등(2016)은 시장의 심리를 반영한 연관 검색어를 추출하기 위해 증권사 주요 리포트로부터

텍스트마이닝 기법을 통해 산업별(섹터) 키워드를 도출하였다. 이렇게 추출한 키워드들에 대한 검색량을 기반으로 섹터별 상장지수펀드(ETF: Exchange Traded Funds)에 대한 투자전략을 취한 결과, 2011년부터 2014년까지 대체로 텍스트마이닝을 통해서 선정된 키워드를 바탕으로 매매 했을 때, 보다 더 높은 수익률을 도출하였다. Perlin *et al.*(2017)은 미국, 영국, 호주, 캐나다 4개국을 분석 대상 표본으로 하여 VAR 모형을 적용한 시계열 분석을 통해, 인터넷 검색량과 주가지수 간의 관계를 분석하였다. 분석에 사용한 키워드의 선정은 Preis *et al.*(2013)이 구글이 제공하는 연관 검색어를 활용한 것과 달리, 경제 관련 서적들로부터 키워드를 추출하는 방식을 적용했다. 이는 해당 서비스(Google sets)가 더 이상 제공되지 않고, 관련성이 의심되는 의외의 키워드들이 포함되는 문제가 있었기 때문이다. 이들 키워드를 사용한 검색어 모델이 주가 변동성과 주가 하락의 예측에 유의한 관계가 있음을 확인하였고, 시계열 분석으로 4개국에서 공통적으로 그랜저 인과관계가 검증된 키워드들이 투자 모델에서 유의미한 결과를 보였다. 암호화폐 시장에도 검색어 선정은 중요하다. 정기호·하성호(2020)는 비트코인에 초점을 맞추어 구글 트렌드가 암호화폐의 가격과 변동성에 대해 인과관계를 갖는지를 분석했다. 분석 결과 전 세계 구글 트랜드 검색지수는 비트코인 수익률과 인과관계가 있는 반면에, 미국에 한정된 구글 트렌드 검색지수는 비트코인 수익률과 인과관계가 보이지 않았다. 즉 동일한 검색어를 사용한다고 해도 지역별 카테고리별로 그 영향 관계가 다르게 나타날 수 있다는 점이다. 앞서 연구들을 종합할 때, 연관 검색어를 활용하여 더 좋은 투자 성과를 도출하기 위해서는 검색어의 추출, 그리고 사전 연관성 검증, 효과적인 투자전략의 선택이 필요하다고 하겠다. 본 연구에서는 이러한 검색어 추출, 시계열 분석을 통한 연관성 검증과, 투자 모형 적용을 통해 연관 검색어를 선택하여 효과적인 투자전략을 제안하고자 하였다.

3 데이터 및 연구모형

3.1 데이터

3.1.1 검색어 선정

주식 시장과 암호화폐 시장에 관련된 주요 키워드를 선정하기 위해, 해당 분야의 대표적인 매체의 일간 리포트를 크롤링하여 출현 빈도가 높은 키워드를 추출하였다. 크롤링에는, 주식 시장과 관련해서 한국 내 경제지 중 1위인 매일경제 사이트에서 일간으로 제공하는 '코스피 마감시황'과 '코스닥 마감시황'을 대상으로 하였다. 암호화폐에 대해서는 암호화폐 전문 매체인 코인데스크코리아에서 일간으로 제공하는 '아침브리핑'을 대상으로 하였다. 각 일간 리포트로부터 2020년 7월 1일부터 2021년 6월 30일까지 1년간의 텍스트를 수집하였으며, 매체별 수집된 텍스트의 개요는 〈표 1〉과 같다.

수집된 텍스트는 코스피, 코스닥, 암호화폐 영역별로 말뭉치를 구성한 뒤 R의 KoNLP 라이브러리를 활용하여 명사 단어의 출현 빈도를 추출하였다. 빈도 분석에 사용된 텍스트는 전처리 과정을 거쳐, 숫자와 공백, 특수문자를 제거하였으며, '거래됐', '투자하'와 같이 조사나 동사의 어미가 함께 추출된 단어는 어근인 명사만(거래, 투자 등) 추출하였다. 이와 같이 도출된 명사들 중 코스피, 코스닥, 암호화폐 영역별로 상위 빈도순으로 10개의 단어를 분석에 사용할 키워드로 선정하였다. 선정 과정에서 단순히 시장 상황을 기술하기 위해 사용된 단어(상승, 하락, 이번, 대비, 마감, 기록, 이날, 반면, 장중 등)와 특정 기업의 명칭(삼성,

〈표 1〉 분석 대상 텍스트

구분	코스피	코스닥	암호화폐
매체	매일경제	매일경제	코인데스크코리아
리포트 명	코스피 마감시황	코스닥 마감시황	아침브리핑
주기	일간	일간	일간
수집 기간	2020.07.01.~2021.06.30.	2020.07.01.~2021.06.30.	2020.07.01.~2021.06.30.
건수	246건	245건	239건

셀트리온 등)은 제외하였으며, 일간 검색량 데이터상에 검색지수가 0인 날이 99% 이상인 '주변국', '방송서비스'는 검색어로써 분석하는 것이 의미가 낮다고 판단되어 분석에서 제외하였다. 이와 같은 과정에 따라 각 영역으로 도출된 상위 10개 단어를 〈표 2〉에 정리하였으며, 영역별 중복을 제외하면 분석에 사용된 키워드는 총 24개이다.

3.1.2 키워드 검색량 데이터

키워드 검색량 데이터는 네이버에서 제공하는 데이터랩 서비스를 통해 수집하였다. 네이버는 한국에서 점유율 1위의 검색 포털로서 한국에서의 검색량 분석에 적합하다고 할 수 있다. 데이터랩에서 제공하는 키워드 검색량 데이터는 설정한 기간 중 해당 키워드의 검색량을 0부터 가장 많을 때의 값을 100으로 정규화하여 소수점 다섯째 자리까지의 데이터로 제공한다. 키워드 검색량 데이터를 수집한 기간은 2017년 6월 1일부터, 2021년 7월 10일까지로 하였으며, 19세 이상 연령대에서 전체 성별, 일간으로 설정하여 검색량 데이터를 수집하였다. 수집된 검색량 데이터는 주식 시장과 암호화폐 시장의 시계열 데이터와의 분석을 위하여 주말, 공휴일 등 주식 시장 개장일이 아닌 날은 제외하였다.

〈표 2〉 주요 키워드 추출 결과

순번	코스피		코스닥		암호화폐	
	키워드	빈도	키워드	빈도	키워드	빈도
1	계약	2,135	†기관	704	비트코인	1,692
2	†매수	1,489	†매수	687	암호화폐	1,537
3	†순매도	1,408	업종	687	†거래	1,202
4	†외국인	1,279	†순매도	579	달러	981
5	†거래	1,111	†외국인	470	가격	644
6	†기관	969	지수	407	미국	623
7	선물	783	†상승폭	405	토큰	524
8	금융	697	코스닥지수	389	블록체인	516
9	증시	661	†개인	377	투자	447
10	†개인	642	종목	356	자산	383

† 표시된 단어는 영역 간 중복되는 단어

3.1.3 주식 시장 및 암호화폐 시장 시계열 데이터

주식 시장 시계열 데이터는 한국거래소(KRX)의 정보데이터시스템에서 제공하는 코스피 지수 및 코스닥 지수 통계 데이터를 사용하였다. 데이터 추출 기간은 키워드 검색량 데이터 수집 기간과 동일하게 2017년 6월 1일부터, 2021년 7월10일까지 일별 시초가 데이터를 수집하였다. 시초가 데이터를 분석에 사용한 이유는 거래일 이전의 검색량이 자산 시장에 선행할 것이라는 본 연구의 전제에 따라, 거래 당일의 거래 정보가 반영되는 종가 보다, 거래 당일 전날까지의 정보에 의해 판단되는 시초가를 분석에 사용하는 것이 적합하다고 보았기 때문이다.

암호화폐는 주식 시장과 같은 종합 지수가 존재하지 않으므로, 주요한 암호화폐인 비트코인과 이더리움의 가격을 데이터로 사용하였다. 코스피, 코스닥 지수 데이터와 마찬가지로 비트코인과 이더리움에 대해 같은 기간의 원화 환율로 일별 시초가 데이터를 수집하였다. 데이터는 인베스팅닷컴 한국어 사이트(https://kr.investing.com/)에서 제공하는 과거 데이터에서 추출하였다. 암호화폐 시장은 주식 시장과 달리 휴장하는 날이 없으므로, 같은 기간 모든 날의 데이터를 분석에 사용하였다.

3.2 연구모형

본 연구는 크게 두 가지 단계로 수행되었다. 먼저, 선정된 24개 검색어별 검색 트렌드와 주식 시장(코스닥, 코스피 지수) 및 암호화폐(비트코인, 이더리움) 간 시계열 선·후행 관계를 VAR 모형을 기반으로 분석하여, 유의한 인과관계를 보이는 키워드를 도출하였다. 이어서, 키워드 검색량 기반의 가상의 투자전략을 적용하여 누적 수익률을 비교함으로써, 실증적인 투자 모형에 적용하였을 때 시계열 분석의 유효성을 검증하였다. 본 연구에서 시계열 분석 및 실증 모형에 적용한 기간은 종속 변수에 해당하는 주식 시장과 암호화폐 데이터를 기준으로 2017년 7월 1일부터, 2021년 6월 30일까지 4년을 적용하였다.

시계열 분석을 위해 적용한 모델은 Perlin *et al.*(2017)이 사용한 모델을 준용하여 다음의 식을 적용하였다. 분석 대상인 주식 시장과 암호화폐 시장에 대해 각각 적용하였으며, 분석에서 계절 효과는 고려하지 않았다.

$$P_t = \alpha_1 + \sum_{p=1}^{OptLag} \beta_p P_{t-p} + \sum_{p=1}^{OptLag} \lambda_p ST_{t-p} + \epsilon_{1,t} \quad \text{식 (1)}$$

$$ST_t = \alpha_2 + \sum_{p=1}^{OptLag} \gamma_p ST_{t-p} + \sum_{p=1}^{OptLag} \phi_p P_{t-p} + \epsilon_{2,t} \quad \text{식 (2)}$$

(P_t : t일의 시초가, ST_t : t의 검색량)

두 번째 단계인 가상의 투자 모형은 키워드 검색량 증가율을 활용하여 일정한 방식의 투자 의사결정 방식을 분석 기간 동안 적용하였을 때 기대할 수 있는 투자 수익률을 도출함으로써, VAR 모형 기반의 그랜저 인과관계 분석 결과와 투자 모형의 수익성을 실증적으로 비교하였다. 본 연구에서 사용한 투자 모형은 특정 검색어에 대해 Δt 기간 동안 검색량의 평균 증가율에 따라 매수 또는 매도 포지션을 결정하는 방식을 적용하였다. 이때, 의사결정의 기준이 되는 검색량의 평균 증가율은 시간 단위를 일(day) 단위로 하였을 때, Δt 일 동안 검색량의 평균 증가율을 적용하였다. 일간 평균 증가율은 Δt 일 동안 일일 증가율을 기하 평균하여 산출하였다. t일의 검색량을 ST_t라고 할 때, Δt 기간 동안 검색량의 평균 증가율을 로그를 취한 $\log \Delta ST_{t,\Delta t}$은 식 (3)과 같다. 여기서 $\log \Delta ST_{t,\Delta t}$의 값이 0보다 크면, Δt 일 평균 검색량이 상승하였음을 의미하며, 0보다 작은 경우 Δt 일 동안 평균적으로 검색량이 하락했음을 의미한다.

$$\log \Delta ST_{t,\Delta t} = \frac{\sum_{i=1}^{\Delta t} (\log ST_{t+1-i} - \log ST_{t-i})}{\Delta t} \quad \text{식 (3)}$$

검색량 일일 평균 증가율에 따라 본 연구에서는 매도와 매수 포지션을 결정하고, 다음 거래일에 이를 청산하는 독립적인 방식의 투자 모형을 적용하였다.

구체적으로, (t-1)일부터 Δt 일 만큼 과거의 기간 동안 검색량의 평균 증가율의 로그값이 음수이면, t일의 시초가 P_t에 매수하고 P_{t+1}에 매도하여 청산하며, 같은 기간 검색량의 평균 증가율의 로그값이 양수이면, P_t에 매도하고 P_{t+1}에 매수하여 청산하는 방식을 적용하였다. 이처럼 검색량이 증가할 때 매도하고, 검색량이 감소할 때 매수하는 투자전략은 시장에 우려가 높을 때, 사람들이 더 많은 정보를 모으려는 경향을 (Preis *et al.*, 2013) 반영한 것이다. 마찬가지로 유재필 등(2016)은 시장에 대한 불안감이 높아질수록 시장에 대한 관심도가 증가하며, 관심도 증가가 주식 시장 하락을 선행한다는 가정하에 검색어를 활용한 투자 모델을 검증한 바 있다.

매 거래일마다 독립적으로 투자 의사결정과 청산이 이루어지는 투자 모형으로, t일의 수익률 ΔRet_t은 다음과 같이 표현될 수 있다.

t일에 매수 포지션인 경우 수익률, 식 (4)

$$\Delta Ret_t = \frac{P_{t+1}}{P_t}$$

t일에 매도 포지션인 경우 수익률, 식 (5)

$$\Delta Ret_t = \frac{P_t}{P_{t+1}}$$

위의 수익률을 로그(log)를 취하면 아래와 같이 표현되며, 로그를 취함에 따라 수익이 증가한 경우 양의 값을, 수익이 감소한 경우 음의 값을 갖는다.

$\log \Delta ST_{t-1,\Delta t} < 0$ 이면, 식 (6)

$$\log \Delta Ret_t = \log P_{t+1} - \log P_t$$

$\log \Delta ST_{t-1,\Delta t} > 0$ 이면, 식 (7)

$$\log \Delta Ret_t = \log P_t - \log P_{t+1}$$

$$(\Delta t = 1, 2, \ldots 10)$$

이 경우, 1일부터 t일까지 누적 수익률(Ret_t)은 로그를 취하여 다음과 같이 나타낼 수 있다.

$$\log Ret_t = \sum_{i=1}^{t} \log \Delta Ret_i \quad \text{식 (8)}$$

[4] 실증분석 결과

4.1 VAR 모형 검정

4.1.1 단위근 검정

시계열 데이터가 불안정한 경우에는 자기회귀 모형으로 표현했을 때 단위근을 갖는다. 따라서 시계열 데이터를 활용하여 모형을 추정할 때는 데이터의 안정성을 확인해야 한다. 본 연구에서 수행한 단위근 검정은 ADF(Augment Dickery-Fuller) 검정과 Phillips-Perron(PP) 검정이다. 〈표 3〉은 단위근 검정 결과를 나타낸다. 코스피, 코스닥 그리고 비트코인과 이더리움 변수들은 ADF 검정 결과와 PP 검정 결과에서 단위근을 가진다는 귀무가설을 기각하지 못하고 있다. 종속 변수들은 전부 단위근을 가지고 있는 것으로 나타났다. 반면에 키워드 검색량의 경우는 ADF 검정 결과에서는 차분하지 않은 원 데이터의 경우 가격, 투자, 자산, 금융, 계약, 매수, 개인, 업종 등 8개 단어의 경우 귀무가설을 기각하지 못했으나, PP 검정 결과는 전부 기각되었다. 또한, 1차 차분에서는 모든 변수들이 5% 유의수준에서 귀무가설을 기각하여 안정적 시계열로 확인되었다. 따라서 본 연구에서는 코스피지수, 코스닥지수, 비트코인 가격, 이더리움 가격은 1차 차분된 지수 변화량과 가격 변화량을 사용하였다. 다만 변수명은 분석의 편의성을 고려하여 코스피, 코스닥,

비트코인, 이더리움으로 표현하였다. 한편 키워드는 개별 변수 검색량으로서 PP검정 결과를 기반으로 원 데이터를 그대로 사용하였다.

〈표 3〉 단위근 검정 결과

변수	ADF 검정		PP 검정	
	수준	1차차분	수준	1차차분
코스피	1.13	-20.73**	-1.15	-1075**
코스닥	0.86	-21.60**	-4.73	-1091**
비트코인	0.10	-27.88**	-5.46	-1547**
이더리움	0.08	-27.00**	-6.23	-1680**
계약	-1.53	-33.78**	-793**	-1104**
매수	-2.28	-28.29**	-67.7**	-1033**
순매도	-6.41**	-21.77**	-408**	-1015**
외국인	-1.62	-32.44**	-376**	-1098**
거래	-3.15*	-30.83	-940**	-1161**
기관	-2.82**	-29.11**	-386**	-1026**
선물	-5.47**	-25.52**	-776**	-1055**
금융	-1.67	-28.74**	-25.6**	-1100**
증시	-5.01**	-23.95**	-274**	-1043**
개인	-1.80	-32.64**	-739**	-1122**
업종	-1.70	-33.29**	-545**	-1071**
지수	-16.43**	-22.55**	-216**	-961**
상승폭	-15.46**	-19.21**	-882**	-1138**
코스닥지수	-3.78**	-25.96**	-188**	-1078**
종목	-2.14*	-31.64**	-324**	-1190**
비트코인	-14.25**	-22.18	-809**	-970**
암호화폐	-5.06**	-23.12	-240**	-1159**
달러	-4.21**	-24.07	-505**	-1027**
가격	-1.31	-33.26	-707**	-1167**
미국	-16.87**	-22.42**	-790**	-999**
토큰	-3.38**	-28.98**	-224**	-1127**
블록체인	-5.44**	-26.61**	-200**	-1136**
투자	-1.45	-32.00**	-241**	-1085**
자산	-1.80	-32.02**	-388**	-1005**

Notes: * p<0.1, ** p<0.05, *** p<0.01

4.1.2 공적분 검정

시계열 데이터가 단위근을 가지고 있는 것으로 판명된 경우 차분을 통하면 시계열 데이터는 안정화되고 VAR 모형 분석이 가능하다. 따라서 VAR 분석을 시행함에 있어 코스피, 코스닥 그리고 비트코인과 이더리움 데이터는 1차 차분 데이터를 활용하고, 나머지 키워드 검색량 데이터는 원 데이터로 분석을 진행하고자 한다. 1차 차분 데이터는 수익 변동 폭을 나타내는 데이터이기 때문에 차분을 하게 되면 발생하는 장기 변화 내용 정보의 일부가 사라져 버린다는 문제가 존재한다. 따라서 개별 변수 간의 관계에서 공적분 관계의 존재 여부를 살펴봐야 한다. 공적분 관계를 검정하는 방법으로 Johansen 공적분 검정 방법을 실시했다. 공적분 관계 여부에 대한 가설 검정은 트레이스(trace) 통계량 검정을 사용하였다. 또한 최적

〈표 4〉 요한슨 공적분 검정 결과(Trace 통계량)

변수	코스피		코스닥		비트코인		이더리움	
	r=0	r≤1	r=0	r≤1	r=0	r≤1	r=0	r≤1
계약	197.48**	4.32	198.54**	1.99	692.0**	2.58	693.18**	2.74
매수	44.97**	4.40	39.62**	1.97	391.19**	2.59	358.8**	2.75
순매도	102.1**	4.08	98.11**	2.08	255.64**	2.59	254.66**	2.73
외국인	88.33**	4.38	85.75**	1.98	358.07**	2.58	359.8**	2.74
거래	281.65**	4.33	258.19**	1.99	557.66**	2.64	660.35**	2.77
기관	133.11**	4.40	133.57**	1.99	474.43**	2.58	468.79**	2.73
선물	147.37**	4.33	146.68**	1.99	143.36**	2.58	144.23**	2.73
금융	23.23*	4.63	16.06	1.92	346.29**	2.57	345.58**	2.73
증시	114.09**	3.89	103.81**	2.11	363.05**	2.59	366.02**	2.74
개인	172.74**	4.32	169.83**	1.99	511.26**	2.58	512.03**	2.72
업종	152.15**	4.32	137.20**	2.00	680.19**	2.59	688.96**	2.73
지수	266.93**	4.30	257.99**	1.99	336.81**	2.6	323.42**	2.75
상승폭	302.68**	4.32	300.61**	2.00	513.19**	2.6	511.34**	2.73
코스닥지수	94.43**	3.81	89.80**	2.40	464.36**	2.59	467.77**	2.74
종목	80.95**	4.31	85.02**	2.00	309.61**	2.58	308.32**	2.74
비트코인	294.7**	4.32	274.00**	1.99	428.94**	2.6	406.95**	2.72
암호화폐	74.6**	4.33	81.23**	1.97	408.04**	2.58	177.02**	2.94
달러	146.48**	4.27	146.07**	2.04	183.23**	3.43	414.74**	2.74
가격	146.9**	4.31	149.24**	1.99	551.12**	2.59	552.48**	2.74
미국	329.28**	4.27	321.45**	2.00	320.35*	2.48	473.67**	2.74
토큰	102.1**	4.42	98.48**	2.00	444.13**	2.63	375.38**	2.73
블록체인	55.29**	4.27	56.30**	2.02	267.61**	2.6	275.25**	2.71
투자	69.23**	4.39	67.25**	1.98	211.69**	2.59	210.03**	2.76
자산	159.5**	4.29	150.91**	2.02	518.77**	2.59	510.90**	2.74

Notes: * p<0.05, ** p<0.01

시차를 확인하기 위해서 AIC(Akaike Information Criteria)와 SC(Schwarz Information Criteria)를 검토하고 SC를 먼저 고려했으며 포트맨토(port-manteau) 검정을 활용하였다.

4.1.3 그랜저 인과관계 분석

공적분 검정 결과에서 코스피, 코스닥 그리고 비트코인, 이더리움과 키워드 검색량 변수에 대해서 각 변수별로 공적분이 존재하지 않는 것으로 확인되었다. 따라서 VAR 모형을 기반으로 그랜저 인과관계 분석을

〈표 5〉 VAR 모형 기반 그랜저 인과관계 분석 결과

검색어	코스피			코스닥			비트코인			이더리움		
	Lag	Sum of λ	Sum of ∅	Lag	Sum of λ	Sum of ∅	Lag	Sum of λ	Sum of ∅	Lag	Sum of λ	Sum of ∅
계약	5	1.054	0.704	5	0.965	0.695	9	1.09	0.496	10	1.096	0.536
매수	3	0.999**	0.912**	3	1.039**	0.919**	8	1.11**	0.853**	10	1.09	0.878**
순매도	4	1.035**	0.827	4	0.950**	0.828	7	1.04	0.846**	10	1.08*	0.867**
외국인	6	1.861**	0.791**	5	0.960**	0.768	8	1.04	0.849	10	1.09	0.817
거래	5	0.979	0.715**	5	0.967	0.822**	8	1.03*	0.813**	10	1.08**	0.604**
기관	3	0.999**	0.692*	4	0.889*	0.740*	9	1.10*	0.725**	10	1.09	0.71
선물	5	0.977	0.617	5	0.968	0.827	7	1.04	0.864	7	1.01	0.864
금융	4	1.045**	0.911*	4	0.935**	0.922	8	1.04	0.922	10	1.09	0.981
증시	4	1.021**	0.823	4	0.995**	0.831	7	1.04	0.86	10	1.09	0.891
개인	5	0.979	0.731	5	0.969	0.735	9	1.09	0.694	10	1.09	0.643
업종	5	1.056**	0.79	5	1.069	0.801	8	1.03	0.806	10	1.09	0.797
지수	5	0.979	0.670**	5	0.967	0.678	4	1.01**	0.784**	2	0.997	0.702**
상승폭	1	0.999	0.185*	1	0.996	0.180*	1	0.997	0.170**	1	0.996	0.167**
코스닥지수	4	0.972**	0.828	4	0.995**	0.838	9	1.01	0.764	10	1.09	0.782
종목	2	0.976	0.786**	2	0.997	0.769**	9	1.02	0.81	10	1.09	0.835
비트코인	1	0.998	0.247**	1	0.995	0.273**	1	0.998*	0.267**	1	0.997	0.288**
암호화폐	3	0.979	0.843*	3	0.971	0.822**	4	0.987**	0.852**	10	0.997**	0.836**
달러	2	0.969	0.641	2	0.957**	0.639	7	1.04	0.776	10	1.09	0.737
가격	5	1.053	0.718	5	0.971	0.713	9	1.09	0.759	10	1.09	0.771
미국	1	0.999*	0.215	1	0.996	0.214	1	0.998	0.228	1	0.997	0.227
토큰	1	1.000	0.717**	1	0.995	0.719**	4	1.00**	0.400**	2	0.987*	0.367**
블록체인	2	0.978	0.857	2	0.97	0.849	9	1.09	0.837	10	1.14**	0.844
투자	5	1.755	0.392**	5	0.967	0.782*	9	1.09	0.954**	10	1.10*	0.899*
자산	5	0.977	0.847	5	0.792	0.711	8	1.03	0.779**	10	1.09**	0.780**

Notes: * p<0.05, ** p<0.01 (Lag : Optimal Lag)

시행할 수 있다.

그랜저 인과관계 분석은 한 변수가 다른 변수의 예측에 도움이 되는지를 검증하는 방법이다. 종속 변수인 코스피, 코스닥, 비트코인, 이더리움과 키워드 검색량 간 양방향 분석을 진행한 결과를 〈표 5〉에 정리하였다. '3.2. 연구모형'의 식(1)에서 $\sum\lambda$가 유의하게 나타난 검색어는 해당 키워드 검색량이 종속변수인 선행 요인으로 작용한다고 볼 수 있다. 예를 들면 코스피 지수에 대해서는 총 24개의 키워드 중에 9개의 키워드(미국, 금융, 매수, 순매도, 외국인, 기관, 증시, 업종, 코스닥지수)의 $\sum\lambda$가 유의하게 나타났다. 반면에 비트코인에 대해서는 이보다 많은 7개의 키워드(매수, 거래, 기관, 지수, 비트코인, 암호화폐, 토큰)가 $\sum\lambda$가 유의하게 나타난 키워드에 해당하였다. 코스피는 9개, 코스닥의 경우는 8개 키워드가 나타났으며, 그중 7개는 동일한 키워드였다. 반면에 비트코인과 이더리움은 각각 7개의 유의한 키워드가 도출되었으나, 3개의 키워드만 동일한 것으로 나타났다.

4.2 실증 모형의 자산시장 적용

VAR 모형 분석을 통해 주식 시장(코스피 지수, 코스닥 지수)과 암호화폐 시장(비트코인, 이더리움)에서 그랜저 인과관계가 유의하게 나타난 키워드를 도출하였다. 이어서 가상의 투자 모형을 적용하여 연구기간 동안 기대할 수 있는 투자 수익률을 산출하여 투자 유효성을 검증하였다. 본 연구에서 적용한 투자 모형은 앞서 '3.2. 연구모형'의 식 (3) ~ 식 (8)을 통해서 설명하였다. 본 연구에서는 Δt를 1일부터 10일까지로 설정하여, 각 키워드별 주식 시장과 암호화폐 시장별 수익률을 시뮬레이션하였다. 투자 모형을 적용한 기간은 2016년 7월 1일부터, 2021년 6월 30일까지이며, 무작위 방식의 투자 모형과 매입 후 보유 전략(Buy & Hold)을 대조군으로 하여 본 연구의 투자 모형에 따른 누적 수익률(Ret_t)과 비교하였다.

무작위 투자 모형은 거래 기간 동안, 매 거래일마다 무작위로 매수와 매도 결정을 내린 후, 해당 의사결정에 따라 시초가에 매수 또는 매도하고, 다음 거래일에 모두 청산하는 방식을 반복 적용하였다. 매 거래일마다 독립적으로 의사결정과 청산이 이루어지므로, 누적 수익률은 본 연구의 실증 모형에 적용한 방식과 같다. 무작위 모형은 각 자산 시장별로 총 10,000회를 반복하여, 누적 수익률의 평균과 표준편차를 산출하였다.

매입 후 보유 전략은 해당 기간의 첫 번째 거래일에 매수하여 마지막 거래일에 모두 청산하는 방식이며, 이 경우 누적 수익률은 다음 식과 같다.

$$\log Ret_{te-ts} = \log P_{te} - \log P_{ts} \quad \text{식 (9)}$$

(te : 최종 거래일, ts : 최초 거래일)

4.2.1 무작위 모형과 검색어 모형 간 t-검정

키워드별 투자 모형의 수익성을 검증하기 위해, 첫 번째로 Δt를 1일부터 10일까지 적용하여 도출한 10건의 누적 수익률 분포와 무작위 투자 모형의 누적 수익률 분포가 상이한지에 대한 t-검정을 수행하였다. 무작위 모형과 각 키워드별 투자 모델의 누적 수익률이 같다는 귀무가설에 대해 t-검정 결과를 주식 시장과 암호화폐 시장별로 〈표 6〉과 〈표 7〉에 정리하였다. 코스피 시장의 경우 그랜저 인과관계에서 유의한 관계를 보인 9개의 키워드 중 외국인, 업종, 미국의 3개를 제외하고 6개 키워드가 95% 신뢰 구간에서 유의한 차이를 보였다. 코스닥 시장의 경우는 그랜저 인과관계에서 유의한 관계를 보인 8개의 키워드 중 외국인, 기관, 달러의 3개 키워드를 제외하고 5개 키워드가 무작위 모형보다 95% 신뢰 구간에서 유의한 차이를 보였다. 이에 따라, 주식 시장에서 그랜저 인과관계가 유의한 키워드들이 무작위 모형보다 수익률에 차이를 보이는 경우가 더 많다고 할 수 있다.

한편, 암호화폐 시장은, 비트코인의 경우 그랜저 인과관계에서 유의한 7개 키워드 중 2개, 이더리움은 7개 키워드 중 3개의 키워드가 실증 모형에서 무작위 모형과 유의한 차이를 보인 것으로 나타났다. 암호화폐 시장은 주식 시장과 달리 그랜저 인과관계에서 유의한 키워드가 실증 모형에서 무작위 모형과 유의한 차이를 보인 비율이 높지 않았다.

〈표 6〉 주식 시장 키워드별 누적 수익률 t-검정

코스피 (무작위 모형 평균 0.002, 표준편차 0.165)					코스닥 (무작위 모형 평균 0.001, 표준편차 0.223)				
검색어	평균	표준편차	t값	유의확률	검색어	평균	표준편차	t값	유의확률
계약	0.096	0.172	1.804	0.105	매수	0.201	0.087	7.238	0.000***
매수	0.115	0.064	5.768	0.000***	순매도	0.141	0.149	2.961	0.016*
순매도	0.154	0.116	4.230	0.002**	외국인	-0.027	0.279	-0.314	0.760
외국인	0.102	0.199	1.650	0.133	기관	0.104	0.204	1.605	0.143
거래	0.120	0.122	3.145	0.012*	개인	-0.056	0.130	-1.378	0.201
기관	0.153	0.110	4.430	0.002**	업종	-0.173	0.195	-2.824	0.020*
선물	-0.040	0.085	-1.411	0.192	지수	0.151	0.146	3.231	0.010*
금융	0.081	0.088	2.956	0.016*	상승폭	-0.033	0.105	-1.034	0.328
증시	0.261	0.112	7.426	0.000***	코스닥지수	0.415	0.144	9.091	0.000***
개인	0.017	0.071	0.831	0.428	종목	0.168	0.138	3.824	0.004**
업종	-0.048	0.158	-0.930	0.377	계약	0.103	0.190	1.701	0.123
지수	0.022	0.102	0.745	0.475	거래	0.050	0.164	0.950	0.367
상승폭	-0.034	0.092	-1.092	0.303	선물	-0.087	0.170	-1.641	0.135
코스닥지수	0.285	0.096	9.440	0.000***	금융	0.185	0.155	3.742	0.005**
종목	0.106	0.116	2.947	0.016*	증시	0.428	0.194	6.964	0.000***
비트코인	0.055	0.118	1.533	0.159	비트코인	0.051	0.215	0.733	0.482
암호화폐	0.005	0.146	0.146	0.887	암호화폐	-0.165	0.179	-2.934	0.017*
달러	0.148	0.145	3.264	0.010*	달러	0.102	0.163	1.960	0.081
가격	-0.011	0.122	-0.237	0.818	가격	-0.060	0.161	-1.202	0.260
미국	-0.009	0.107	-0.198	0.847	미국	-0.149	0.146	-3.244	0.010*
토큰	0.069	0.058	3.825	0.004**	토큰	0.167	0.139	3.766	0.004**
블록체인	-0.061	0.119	-1.559	0.153	블록체인	-0.239	0.152	-4.962	0.001**
투자	0.131	0.103	4.072	0.003**	투자	0.228	0.179	3.997	0.003**
자산	-0.072	0.124	-1.792	0.107	자산	-0.224	0.180	-3.940	0.003**

1. Notes: * p<0.05, ** p<0.01, *** p<0.001
2. 각 영역별 상위 10개로 추출된 키워드는 빗금 음영 처리,
3. 그랜저 인과관계 분석(〈표 4〉)의 $\sum\lambda$가 유의하게 나타난 키워드는 밑줄로 표시

〈표 7〉 암호화폐 시장 키워드별 누적 수익률 t-검정

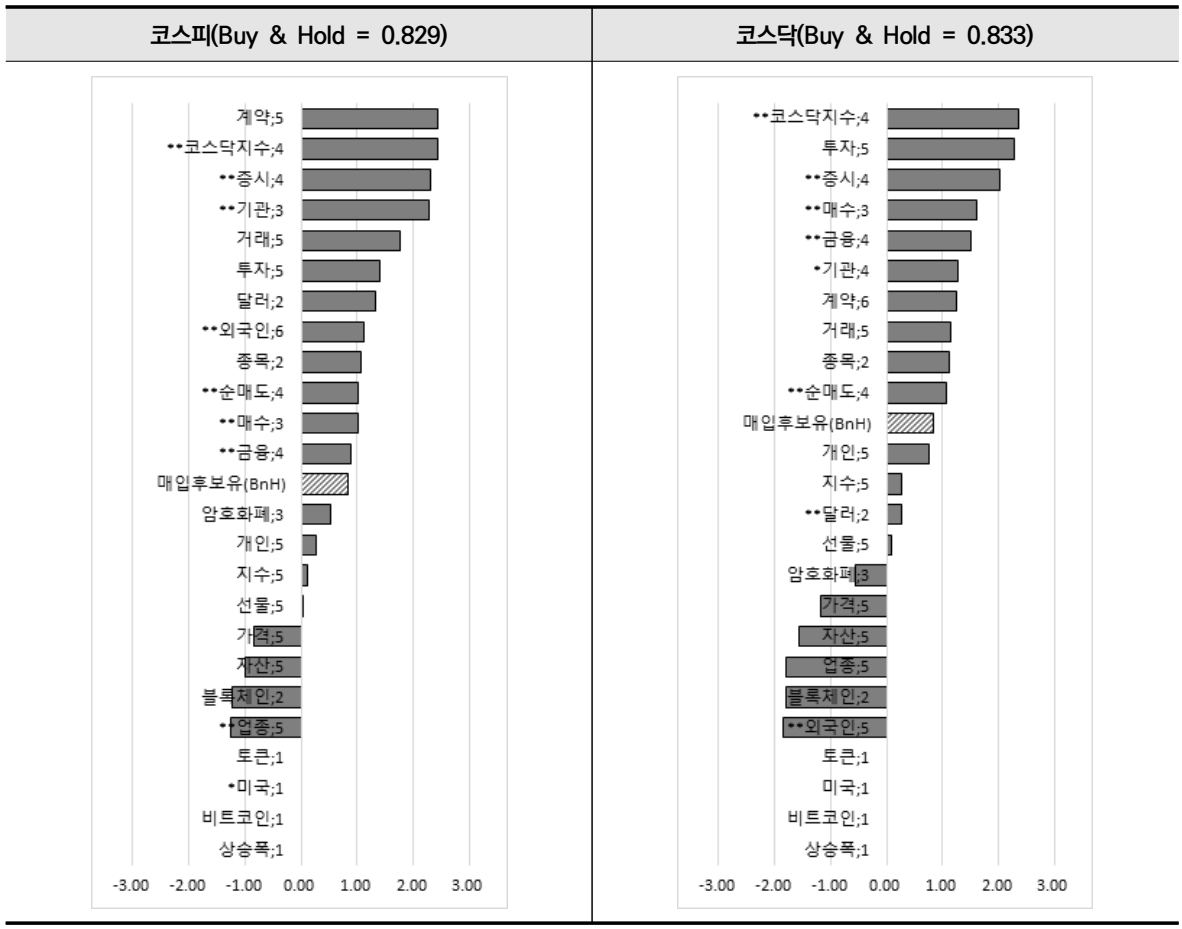
비트코인 (무작위 모형 평균 -0.004, 표준편차 0.708)					이더리움 (무작위 모형 평균 0.009, 표준편차 0.870)				
검색어	평균	표준편차	t값	유의확률	검색어	평균	표준편차	t값	유의확률
비트코인	-0.023	0.485	-0.127	0.901	비트코인	0.223	0.754	0.961	0.362
암호화폐	-0.486	0.280	-5.435	0.000***	암호화폐	-0.349	0.652	-1.660	0.131
달러	0.312	0.357	2.787	0.021*	달러	0.805	0.683	3.752	0.005**
가격	0.374	0.401	2.969	0.016*	가격	0.382	0.824	1.492	0.170
미국	0.665	0.724	2.920	0.017*	미국	0.748	0.845	2.820	0.020*
토큰	-0.315	0.627	-1.569	0.151	토큰	-0.574	0.597	-3.006	0.015*
블록체인	0.154	0.408	1.216	0.255	블록체인	0.344	0.476	2.326	0.045*
투자	-0.117	0.558	-0.646	0.534	투자	-0.132	0.802	-0.495	0.632
자산	0.221	0.335	2.113	0.063	자산	0.291	0.518	1.815	0.103
거래	0.162	0.592	0.884	0.400	거래	0.412	0.857	1.544	0.157
계약	0.313	0.521	1.919	0.087	계약	0.817	0.511	5.084	0.001***
매수	-0.030	0.294	-0.287	0.781	매수	0.084	0.579	0.496	0.632
순매도	0.678	0.615	3.501	0.007**	순매도	0.468	0.548	2.734	0.023*
외국인	0.299	0.643	1.484	0.172	외국인	0.391	0.833	1.506	0.166
기관	0.665	0.542	3.894	0.004**	기관	0.835	0.675	3.938	0.003**
선물	0.532	0.421	4.012	0.003**	선물	0.539	0.677	2.545	0.031
금융	0.338	0.347	3.100	0.013*	금융	0.829	0.545	4.842	0.001***
증시	0.559	0.408	4.349	0.002**	증시	0.953	0.646	4.692	0.001***
개인	0.534	0.383	4.425	0.002**	개인	0.678	0.596	3.628	0.005**
업종	0.154	0.562	0.884	0.400	업종	0.406	0.965	1.350	0.210
지수	0.092	0.601	0.503	0.627	지수	-0.134	0.596	-0.674	0.517
상승폭	0.351	0.279	3.994	0.003**	상승폭	-0.056	0.310	-0.499	0.629
코스닥지수	0.672	0.631	3.385	0.008**	코스닥지수	0.915	0.639	4.557	0.001***
종목	-0.120	0.273	-1.353	0.209	종목	0.309	0.732	1.361	0.207

1. Notes: * p〈0.05, ** p〈0.01, *** p〈0.001
2. 각 영역별 상위 10개로 추출된 키워드는 빗금 음영 처리,
3. 그랜저 인과관계 분석(〈표 4〉)의 $\sum\lambda$가 유의하게 나타난 키워드는 밑줄로 표시

4.2.2 적정 시차를 적용한 누적 수익률 검증

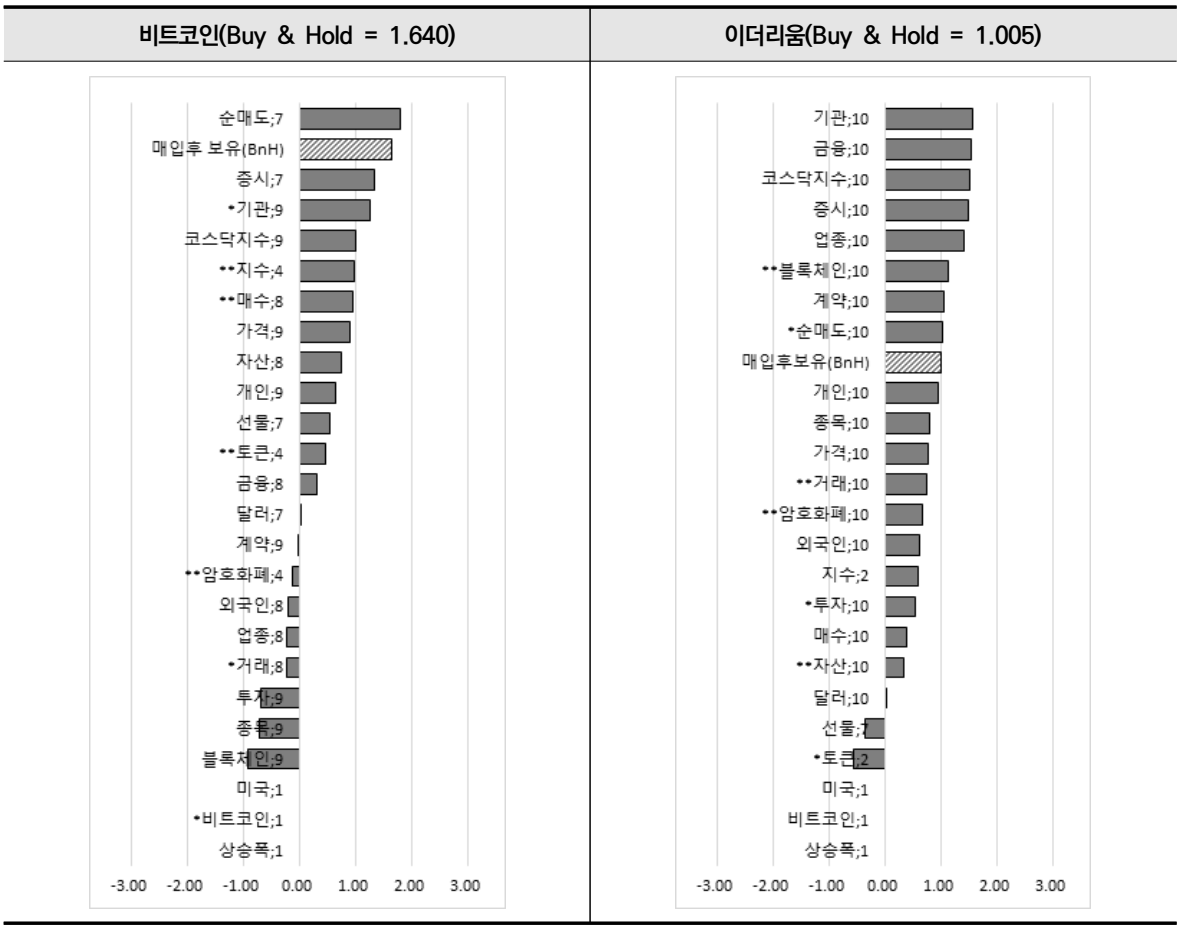
다음으로 VAR 모형에서 도출한 적정 시차에 따라 적절한 Δt 값을 적용하였을 때 실증 모형의 수익률을 매입 후 보유 전략과 비교하였다. 본 연구모형은 t일보다 Δt 일의 기간만큼 과거의 기간 동안 검색어 트렌드의 평균 증가율에 따라 t일에 투자 의사결정을 내리는 구조이다. t일의 투자 의사결정에 대한 수익률은 t+1일에 결정되므로, t+1일에 대한 적정 시차(Lag)에 해당하는 기간은 $(\Delta t+1)$이다. 즉, $\Delta t = Opt\,Lag - 1$로 적용할 수 있다. 예를 들어, 코스피 지수에 대해 '계약'의 적정 시차가 5이므로, Δt를 4로 설정하면 수익률이 결정되는 날을 기준으로 적정 시차 5일만큼 과거의 기간에 해당하는 검색어 평균 증가율로 투자 의사 결정을 내린 결과를 얻을 수 있다. 이와 같은 방법으로, 적정 시차와 키워드 검색량의 평균 증가율에 따라 투자 의사 결정을 적용하여 누적 수익률을 도출한 결과를 〈그림 1〉과 〈그림 2〉에 정리하였다. 다만, 적정 시차가 1인 경우는 Δt가 0이 되기 때문에 해당 시차에서는 시뮬레이션에 의한 누적 수익률 값이

〈그림 1〉 검색어 트랜드별 누적 수익률(코스피, 코스닥)



1. 막대그래프 : Δt (=적정 시차-1)에서 검색량 기반 투자 모델의 누적 log 수익률(무작위 모형 기준 표준화)
   ※ 적정 시차는 키워드별 세미콜론(;) 뒤 숫자로 표기
2. 빗금 막대 : 매입 후 보유 전략(Buy&Hold)의 누적 log 수익률
3. *, ** : 그랜저 인과관계 분석 $\sum\lambda$의 유의확률(* p〈0.05, ** p〈0.01)

〈그림 2〉 검색어 트랜드별 누적 수익률(비트코인, 이더리움)



1. 막대그래프 : Δt (=적정 시차-1)에서 검색량 기반 투자 모델의 누적 log 수익률(무작위 모형 기준 표준화)
   ※ 적정 시차는 키워드별 세미콜론(;) 뒤 숫자로 표기
2. 빗금 막대 : 매입 후 보유 전략(Buy&Hold)의 누적 log 수익률
3. *, ** : 그랜저 인과관계 분석 $\sum\lambda$의 유의확률(* p〈0.05, ** p〈0.01)

존재하지 않는다. 〈그림 1〉과 〈그림 2〉의 누적 수익률은 로그를 취한 뒤, 무작위 모형의 평균과 표준편차에 따라 표준화한 값이다. 대조군으로 매입 후 보유 전략(BnH)과 수익률을 비교하였다.

코스피 시장은 그랜저 인과관계에서 유의한 관계를 보인 9개 키워드 중 적정 시차가 1로 누적 수익률 값이 존재하지 않는 '미국'을 제외한 8개 키워드가 모두 대조군(표준화한 수익률 : 0.829)보다 높은 누적 수익률 값을 보였다. 다만, '업종'은 음(-)의 수익률이므로, 매도와 매수를 반대로 하였을 때, 대조군보다 높은 수익률을 기대할 수 있다. 코스닥 시장의 경우 8개 키워드 중 '달러'는 표준화한 수익률이 0.258로 대조군(0.833)보다 수익률이 낮았으며, 나머지 7개 키워드는 대조군보다 높은 수익률을 보였다. 단, '외국인'은 매수와 매도를 반대의 전략을 취할 때 대조군보다 높은 수익률을 기대할 수 있다.

암호화폐 시장은 주식 시장과 달리, 대조군보다 높은 수익률을 보인 키워드가 적었다. 비트코인 시장의 경우 그랜저 인과관계가 유의하게 나타난 7개의 키워드 중 대조군인 매입 후 보유 전략의 수익률(1.640)보다 높은 수익률을 보인 키워드가 없거나 적정 시차가 1이어서 누적 수익률 값이 도출되지 않았

다. 24개 키워드 전체에 대해서도 대조군보다 높은 수익률을 보인 키워드는 '순매도(1.794)'로 1개뿐이었다. 이더리움 시장은 그랜저 인과관계가 유의하게 나타난 7개 키워드 중 '블록체인(1.142)'과 '순매도(1.039)' 2개의 경우가 대조군(1.005)보다 수익률이 높은 것으로 나타났다. 전체 키워드 중에는 '블록체인', '순매도' 외에 '기관(1.565)', '금융(1.540)' 등 8개 키워드에서 대조군보다 높은 수익률을 보였다. 이들 8개의 키워드 중 '블록체인'을 제외한 7개의 키워드는 코스피나 코스닥과 관련한 키워드였다.

[5] 논의와 결론

인터넷 검색량을 자산 시장 참여자의 행동 패턴을 반영한 지표로 활용되는 사례가 늘어나고 있다. 인터넷 검색량의 변동은 자산 시장 투자 참여자의 심리와 행동을 반영하는 지표로서, 특정 자산 시장 또는 시장 전반을 선행하는 지표로써 활용될 가능성이 있다. 이러한 측면에서 본 연구는 주식 시장과 암호화폐 시장에서 시장 참여자들의 투자 심리와 행동이 검색 행위와 관련성이 있을 것으로 보고, 시장 가격에 선행하는 검색어 키워드를 추출하고 시계열 분석과 가상의 투자 모형에 적용하여 효과성을 검증하였다.

분석 결과, 그랜저 인과관계에서 유의한 선행 요인으로 나타난 키워드는 코스피 지수에 대해서는 총 24개 키워드 중 9개였으며, 그중 코스피 지수와 관련한 텍스트로부터 추출한 상위 10개의 키워드 중에서는 6개가 유의한 그랜저 인과관계를 보였다. 코스닥 지수에 대해서는 8개가 유의한 키워드로 도출되었으며, 이중 관련 텍스트로부터 추출된 키워드는 5개였다. 암호화폐의 경우, 비트코인 가격에 대해서는 전체 24개 키워드에서 7개의 키워드가 그랜저 인과관계가 있었으며, 암호화폐 관련 텍스트로부터 추출된 키워드는 4개가 해당하였다. 이더리움 가격에 대해서는 총 7개의 키워드가 유의한 그랜저 인과관계를 보였으며, 암호화폐 관련 키워드는 6개가 해당하였다. 따라서, 주식 시장과 암호화폐 시장 모두에서 그랜저 인과관계가 유의하게 나타난 키워드는 관련 텍스트에서 추출한 키워드 상위 10개 중에 50.0% 이상의 비율로 나타난 반면 그 외의 텍스트에 해당하는 키워드는 42.9% 이하로 나타났다. 이를 통해 각 자산 시장에 관련된 텍스트로부터 추출한 키워드에 대한 검색량이 다른 출처로부터 도출된 키워드들보다 시계열 상 인과관계가 유의하게 나타날 수 있는 가능성이 높다고 유추할 수 있다. 특히, 코스피, 코스닥, 암호화폐 시장이 서로 무관한 시장으로 보기에는 어려움에도 불구하고, 해당 자산 시장과 직접적으로 관련된 텍스트가 아닌 곳에서 추출된 키워드의 경우는 그랜저 인과관계를 유의하게 나타난 비율이 42.9% 이하로 낮았으며, 이더리움 시장의 경우 14.3%에 불과하였다. 이는 특정 자산 시장과 관련성이 높은 텍스트로부터 키워드를 추출하는 것이 시계열 상 선·후행 관계를 갖는 키워드 검색량을 찾는데 효과적이라는 점을 알 수 있다.

시계열 분석에 이어서 실시한 가상의 투자 모형 시뮬레이션에서는 무작위 투자 모형을 10,000회 실시한 결과를 대조군으로 하여, 각 키워드별 검색량 기반 투자 모형과 t-검정을 시행하였다. 코스피 시장이 11개 키워드, 코스닥 시장이 14개 키워드, 비트코인 시장과 이더리움 시장이 각각 12개와 11개의 키워드가 대조군과 유의한 차이가 있는 것으로 나타났다. 이들 키워드는 키워드 검색량의 변화에 따른 투자 의사결정 모형이 무작위 방식의 투자 보다는 유의미한 수익률을 기대할 수 있는 키워드들로 볼 수 있다. 특히, 코스피 시장의 경우 코스피 시장과 관련된 상위 10개의 키워드로서 그랜저 인과관계가 유의한 것으로 나타난 키워드에 한정하면, 6개 키워드 중 5개가 t-검정에서 유의한 키워드로 나타났다. 코스피 시장은 직

접 관련된 텍스트에서 키워드를 추출하고 시계열 상 유의미한 인과관계가 도출되는 키워드를 활용할 때, 본 연구에서 비교한 다른 자산 시장에 비해서 유의미한 수익률을 달성할 가능성이 높다고 할 수 있다.

좀 더 실용적인 측면에서 VAR 모형 분석을 통해 도출한 자산 시장별, 키워드별 적정 시차를 활용하여 투자 의사 결정에 적용한 결과를 살펴보면, 주식 시장인 코스피, 코스닥 지수에 대해서는 매입 후 보유 전략(buy & hold)에 비해 키워드 검색량 기반의 투자 모형이 높은 수익률을 거둘 가능성이 높았다. 반면에, 암호화폐 시장은 매입 후 보유 전략 보다 투자 결과가 높은 경우가 적었다. 구체적으로 보면, 코스피 지수에 대한 투자 결과에서는, 그랜저 인과관계가 유의한 키워드 9개 중 투자 수익률 값이 없는 '미국'을 제외한 8개 키워드 모두, 매입 후 보유 전략보다 높은 투자 수익률을 나타내었다. 코스닥 시장도 8개 키워드 중 7개 키워드를 활용한 투자 모형에서 매입 후 보유 전략보다 높은 수익률을 나타내었다. 영역별 직접 관련된 상위 10개의 키워드 중 그랜저 인과관계가 유의한 것으로 나타난 키워드로 한정했을 때, 코스피 시장은 6개, 코스닥 시장 5개 키워드 모두가 매입 후 보유 전략보다 높은 수익률을 보였다. 따라서 주식 시장에 대해서는 직접 관련된 텍스트로부터 추출한 키워드 중, 시계열 분석에서 인과관계가 존재하는 것으로 분석된 키워드들에 대해 적정 시차 기간 동안의 평균 검색량 증가율에 따라 투자 의사결정 내릴 때. 매입 후 보유 전략보다 높은 투자 수익률을 기대할 수 있다.

이와 달리, 암호화폐의 경우는 매입 후 보유 전략의 수익률보다 적정 시차를 적용한 투자 모형의 수익률이 높은 결과가 드물었다. 비트코인의 경우 그랜저 인과관계가 유의한 키워드 중에는 매입 후 보유 전략보다 수익률이 높은 경우가 없었으며, 이더리움은 7개 중 2개가 매입 후 보유 전략보다 수익률이 높게 나타났다. 암호화폐와 직접 관련된 텍스트로부터 추출된 키워드로 한정하면, 매입 후 보유 전략보다 높은 경우는 비트코인 시장에서는 해당하는 키워드가 없었으며, 이더리움 시장에서는 '블록체인'만 해당하였다.

이처럼 주식 시장과 암호화폐 시장에서 검색량 기반의 투자 결과가 달리 나타난 이유는, 첫째, 암호화폐 시장이 대체로 연구 대상 기간 동안 상승세가 압도적이어서, 잦은 투자 의사결정 보다 일관된 보유를 하는 경우 누적 수익률이 더 높게 나타났기 때문이다. 무작위 투자 모형을 기준으로 표준화한 값에 따르면, 비트코인 시장의 매입 후 보유 전략의 누적 수익률은 표준화한 값으로 1.640, 이더리움 시장은 1.005로 표준편차(1σ)보다 높게 나타났다. 주식 시장의 경우는 코스닥 시장이 0.829, 코스닥 시장이 0.833으로 1σ보다 낮았다. 둘째는 키워드 검색량 기반의 투자 모형의 수익률 자체가 암호화폐 시장이 주식 시장에 비해 더 낮기 때문이다. $|1\sigma|$보다 큰 수익률을 보인 키워드는 코스피 시장이 13개, 코스닥 시장이 15개인 반면, 비트코인의 경우 3개, 이더리움은 8개로 더 적었다. 적정 시차가 1이어서 수익률이 도출되지 않은 키워드들을 제외하고, 표준화한 누적 수익률의 절댓값의 평균을 비교하더라도, 코스피 시장은 1.217, 코스닥 시장은 1.287이었으나, 암호화폐 시장인 비트코인은 0.672, 이더리움은 0.869로 키워드 기반 투자 수익률이 주식 시장보다 암호화폐 시장이 낮게 나타났다. 이를 통해 볼 때, 암호화폐와 같이 시장의 전체적인 상승세가 우세한 경우에는 투자 의사결정을 자주 하는 투자전략보다는 보유하는 전략이 효과적이라고 볼 수 있다. 한편, 암호화폐 시장이 글로벌 시장과 흐름을 같이하기 때문에 국내의 키워드 검색량과의 연관성이 상대적으로 떨어지기 때문이라고 해석할 수 있으나, 시계열 분석의 결과에 의하면, 주식 시장과 암호화폐 시장이 큰 차이가 있다고 보기에는 어려움이 있다.

결론적으로, 본 연구는 연관 텍스트로부터 상위 빈도의 키워드를 추출하고, 시계열 분석을 통해 그랜저 인과관계에서 유의한 결과를 보인 키워드를 도출하였고, 이를 투자 모형에 적용하여 유효성을 검증하였다. 자산 시장별로 직접적인 관련성이 있는 텍스트에서 키워드를 추출하는 것이 유의미한 차이가 있었으며, 주식 시장의 경우 매입 후 보유 전략과 비교해 검색량 기반의 투자 모형이 높은 투자 수익률을 기대할 수 있어서 유의미한 투자전략으로 볼 수 있다. 반면에 암호화폐 시장에서는 연구 대상 기간에서는 검색량

기반의 투자 모형이 매입 후 보유 전략보다 우수한 투자전략으로 보기 어려웠다.

키워드 검색량을 활용한 투자 모형의 효과성에 대해서 기존에 연구들이 이루어져 왔으나, 본 연구는 다음과 같은 점에서 차별성을 갖는다. 먼저, 본 연구는 자산 시장별로 키워드 추출에 사용한 텍스트의 출처를 달리하여, 직접 관련된 텍스트로부터 추출된 키워드와 다른 자산 시장과 관련된 키워드를 활용한 결과를 비교하였다. 해당 자산 시장과 관련된 텍스트로부터 추출된 키워드의 검색량 데이터가 그렇지 않은 경우보다 시계열 분석상 선·후행 관계가 존재할 가능성이 높았다. 둘째, 주식 시장과 암호화폐 시장을 비교하여 키워드 검색량을 활용한 투자 모형을 검토하였다. 특히, 기존 연구들이 대부분 주식 시장에 대해서 이루어져 왔기 때문에 최근 성장하고 있는 암호화폐 시장에 대해서도 검색어를 활용한 투자 모형을 주식 시장과 비교 연구하였다는 점에서 의미가 있다. 본 연구에서 주식 시장과 암호화폐 시장에 대해 시계열 분석을 통해 검색량과 자산 시장간에 관련성이 있는 키워드들을 도출할 수 있었으나, 투자 모형 시뮬레이션에서는 암호화폐 시장의 경우 매입 후 보유 전략에 비해 키워드 검색량 기반 투자 모형이 효과적이지 못하여 주식 시장과 큰 차이를 확인하였다. 셋째, 본 연구에서는 시계열 분석에서 도출된 적정 시차를 의사결정에 활용할 수 있도록 적정 시차 기간의 평균 검색어 증가율을 투자 의사결정에 활용하였다. 선행연구들에서는 일정 기간 평균 검색량 대비 직전 검색량의 증감에 따라 의사결정 모형을 적용하여 시계열 분석과의 연계가 낮았으나, 본 연구에서는 시계열 데이터 간의 시차에 따른 영향을 반영할 수 있는 투자 의사결정 방식을 적용하였다. 주식 시장의 경우에는 적정 시차를 적용한 투자 모형의 누적 수익률이 매입 후 보유 전략에 비해 대체로 높게 나타나서 실용적인 투자 모형으로서 가능성을 보여주었다.

본 연구의 한계로는 키워드 검색량이 시장의 지수 또는 가격에 대해서 일정한 관계가 있다는 가정을 하였다는 점이다. 검색량의 증감이 시장 가격의 증감과 일정한 방향성이 있는 키워드가 있을 수 있으며, 시장 가격과는 관련성이 낮더라도 거래량과 같은 다른 지표들과 더 높은 관련성이 있는 키워드 검색량이 존재할 수 있다. 따라서 후속 연구로 거래량과 같은 다른 지표들과의 관계를 찾고 모형화함으로써, 시장 가격에 더하여 거래량 등 다른 지표들과의 관계를 복합적으로 적용하여 투자 모형을 더욱 정교히 할 수 있을 것이다. 아울러, 본 연구에서는 자산 시장 영역에 따라 적합한 텍스트를 통해 키워드를 추출할 필요성을 제기하였는데, 키워드 검색량과 시장과의 관계는 시장 상황과 관심도에 따라 변화할 수 있다. 즉, 키워드의 적합한 출처만이 아니라, 시간의 변화에 따라 투자 모형에 적합한 텍스트와 키워드가 달라질 수 있을 것이다. 키워드 검색어 기반의 투자 모형의 유효성을 지속하기 위해서는 투자 모형에 활용되는 키워드가 유효한 기간과 이들 키워드를 동적으로 적용할 수 있는 투자 모형의 개발 연구도 후속될 필요가 있다.

참 고 문 헌

구평회·김민수, 2015, 인터넷 검색추세를 활용한 빅데이터 기반의 주식투자전략에 대한 연구, 한국전자거래학회지 제20권 제2호, pp. 1-14.

김류미, 2018, 인터넷 검색량과 투자자별 거래 및 주식수익률의 관계에 대한 실증 연구, 금융공학연구 제17권 제2호, pp. 53-85.

김민수·구평회, 2013, 인터넷 검색추세를 활용한 빅데이터 기반의 주식투자전략에 대한 연구, 한국경영과학회지 제38권 제4호, pp. 53-63.

김민수·권혁준, 2017, 포털 검색 강도가 주가 급락에 미치는 영향에 관한 연구, 한국전자거래학회지 제22권 제2호, pp. 153-168.

김민수·허몽하·권혁준, 2020, 한국 포털 사이트 검색강도가 주가 동조성 및 위험에 미치는 영향, 한국전자거래학회지 제25권 제4호, pp. 125-141.

김범수·최유지·박도형, 2015, 웹검색트래픽 정보를 활용한 투자모형 개발: KOSPI 지수를 중심으로, Entrue Journal of Information Technology 제14권 제3호, pp. 63-81.

반주일·김명애·전용호, 2016, 포털사이트에서의 피검색빈도와 주식수익률, 한국산업정보학회논문지 제21권 제5호, pp. 73-83.

양철원, 2019, 비트코인의 국내외 가격차이를 이용한 차익거래에 관한 연구, 자산운용연구 제7권 제2호, pp. 1-20.

우민철·김지현, 2017, 사이버 공간의 정보가 주가에 미치는 영향 : 인공지능 알고리즘 기법을 이용하여, 자산운용연구 제5권 제2호, pp. 40-55.

유재필·한창훈·신현준, 2016, 빅테이터 트렌드를 이용한 섹터 투자 전략, 정보기술아키텍처 연구 제13권 제1호, pp. 111-121.

장영봉·권영옥·조우제, 2015, 인터넷 주의효과: 능동적 정보 검색이 투자 결정에 미치는 영향에 관한 연구, 지능정보연구 제21권 제3호, pp. 117-129.

정기호·하성호, 2020, 인터넷 검색을 통한 암호화폐 수익률 및 변동성에 대한 인과검정: 적률인과 접근, 정보시스템연구 제29권 제1호, pp. 289-301.

전경민·남기만, 2020, 투자자 관심 수준이 주가표류현상에 미치는 영향, 국제회계연구 제91권, pp. 163-186.

전새미·정여진·이동엽, 2016, 개별 기업에 대한 인터넷 검색량과 주가변동성의 관계: 국내 코스닥시장에서의 산업별 실증분석, 지능정보연구 제22권 제2호, pp. 81-96.

하준성·김도성·이영주, 2019, 투자자 관심도와 애널리스트 투자의견 변경에 대한 주가 반응, Financial Planning Review 제12권 제2호, pp. 61-77.

홍기훈, 2021, 코인 특성의 이해, 자산운용연구 제9권 제1호, pp. 66-81.

Choi, H., 2021, Investor attention and bitcoin liquidity: Evidence from bitcoin tweets, *Finance Research Letters,* Vol. 39.

Da, Z., J. Engelberg and P. Gao, 2011, In search of attention, *The Journal of Finance,* Vol. 66, No. 5, pp. 1461-1499.

Perlin, M., J. Caldeira, A. Santos and M. Pontuschka, 2017, Can we predict the financial markets based on Google's search queries?, *Journal of Forecasting,* Vol. 36, No. 4, pp. 454-467.

Preis, T., H. Moat, and H. Stanley, 2013, Quantifying trading behavior in financial markets using Google Trends, *Scientific reports,* Vol. 3, No. 1, pp. 1-6.

Subramaniam, S. and M. Chakraborty, 2020, Investor Attention and Cryptocurrency Returns: Evidence from Quantile Causality Approach, *Journal of Behavioral Finance,* Vol. 21, No. 1, pp. 103-115.

Zhu, P., X. Zhang, Y. Wu, H. Zheng and Y. Zhang, 2021, Investor attention and cryptocurrency: Evidence from the Bitcoin market, *Plos one,* Vol. 16, No, 2.

A Comparative Study of Stock & Cryptocurrency Market Prediction Using Internet Search Volume

Seyoon Lee* (PrideLab)
Jun-gi Park** (PrideLab)

Abstract

Investor sentiment and behavior of market participants in the stock market and cryptocurrency market are expected to be related to their search behavior. In this study, the effectiveness was verified by extracting keywords for search terms that precede the market price and applying them to time series analysis and hypothetical investment models. To this end, high-frequency keywords were extracted from reports related to the stock and cryptocurrency market, and keywords with high correlation were selected through time series analysis between search volume and asset market index. Subsequently, the rate of return was verified by applying it to a hypothetical investment model based on the search volume of the selected keywords. As a result of the analysis, there was a significant difference in profitability when keywords were extracted from directly related corpus to each asset market compared to others. In the case of the stock market, compared to the buy-and-hold model, the search volume-based investment model could expect a high return on investment. On the contrary, in the cryptocurrency market, it was difficult to regard the search volume-based investment model as an investment strategy superior to the buy-and-hold model.

Key words: *Search Trends, Big Data, Stock Investment Strategy, KOSPI, KOSDAQ, Cryptocurrency, Investment Model*

Article history : Received 29 September 2021, Revised 11 November 2021, Accepted 5 December 2021

JEL Classification : C32, G17

* Expert Advisor, PrideLab, E-mail: suyfj77@gmail.com
** Director of Research, PrideLab, E-mail: warrenpak@warrenpak.com