金融工學硏究, 제23권 제1호 (2024. 3.) 1~21 http://dx.doi.org/10.35527/kfedoi.2024.23.1.001

자산가격변화와 비정보기반투자자의 심리에 대한 선도-지연 관계 분석

황상욱* · 김은총**

〈요 약〉

본 연구의 목적은 비정보기반투자자의 심리를 BERT 모형을 통해 계량화함으로써 투자자의 "심리"가 자산가격변화에 대해 예측력을 갖는지 선도-지연 관계에 대한 실증분석을 통해 살펴보는 것이다. 2018년 1월 1일부터 2020년 12월 31일까지의 시가총액 기준 상위 20개 개별종목의 일간 수익률, 국고채 3년물 금리, 그리고 각 개별종목에 대한 게시판 댓글들로 산출한 감성지수를 사용하여 감성지수와 위험조정수익률 간의 상관성을 확인하고자 우선 단순선형회귀분석을 진행하였다. 단순선형회귀분석 결과 감성지수와 위험조정수익률 사이에 통계적으로 유의한 관계가 있음을 확인하였다. 이 결과를 토대로 위험조정수익률과 감성지수의 선도-지연 관계를 분석하기 위해 VAR 분석, 그랜저 인과 분석, 충격반응 분석 그리고 예측오차분산분해를 진행하였다. 분석 결과, 감성지수는 위험조정수익률에 선도하는 관계를 갖지 않고, 오히려 위험조정수익률의 감성지수에 대한 예측력이 통계적으로 유의미함을 확인하였다. 해당 분석 결과의 강건성을 확인하고자 Fama-French 3 Factor 모형과 자본자산가격결정 모형에서 도출한 회귀계수를 바탕으로 젠센의 알파를 유도하여 젠센의 알파와 감성지수 간의 선도-지연 관계를 살펴보기 위해 감성지수에 각각 -1, 0, +1 시차를 두고 회귀분석을 진행함으로써 감성지수가 초과수익률에 지연된 관계가 존재한다는 것을 확인하였다. 결론적으로, 비정보기반투자자투자자 심리의 과거 정보가 현재의 자산가격변화를 설명하지 못하고, 오히려 비정보기반투자자들의 현재 심리가 자산가격변화의 과거 정보에 의해 설명되고 있음을 발견하였다.

주제어 : 비정보기반투자자, 감성지수, 선도-지연, BERT, VAR, Jensen's Alpha

논문접수일 : 2023. 06. 09. 1차 수정일 : 2023. 06. 17. 게재확정일 : 2024. 01. 04.

* 제1저자, 한양대학교 경영대학 경영학과 재무금융 석사과정, 02-2220-2439,
E-mail: sanguk.hwang@kakao.com

** 교신저자, 한화자산운용 ETF매니저, 한양대학교 컴퓨테이셔널 파이낸스공학화 겸임교수, 02-6950-0085,
E-mail: eunchong01@gmail.com

Ⅰ. 서 론

투자자들의 감정이 금융시장 움직임의 중요한 원동력 중 하나라는 것은 선행연구들을 통해 입증된 바가 있다. 고전적인 금융이론은 투자자가 합리적이라고 가정하지만 광범위한 연구들에서 이미 투자자들의 비이성적이고 비합리적인 행동이 빈번하게 발생한다는 것을 밝혀 내었다(Lee et al., 1991; Baker and Wurgler, 2006). 이와 같은 이유로 머신러닝 및 딥러닝 분석기법이 발달하기 전에는 투자자 설문조사를 통해 감성을 측정하고자 하였다. 투자자에게 설문지를 보내어 시장환경과 시장동향에 대한 여론을 수집하는 것은 전통적인 방법이었다. 하지만 이러한 방법들은 설문조사에 들어가는 시간과 비용 때문에 데이터의 시의적절성이 낮아 과거의 감성에 대한 후속적인 분석만 가능하고, 미래에 대한 예측은 어렵다는 한계가 있다.

위와 같은 이유로 트위터(Bollen et al., 2011), 온라인뉴스(Chen et al., 2011)와 같이 인터넷에 존재하는 방대한 양의 감성 텍스트를 분석하여 설문조사를 통한 감성분석에 대한 대용변수로 사용하였다. 더불어 자연어처리(Natural Language Process)의 비약적인 발전과 함께 전자상거래, 소셜미디어, 인터넷 리뷰 댓글, 인터넷 기사 댓글 등을 통해 생성된 데이터들 중의 80%을 차지하는 비정형화된(unstructured) 데이터인 텍스트를 분석하여 사회현상을 이해하고자 하는 연구가 활발하게 이루어지고 있다. 컴퓨터 그래픽카드 성능의 발전과 함께 엄청난 크기의 단순계산이 신속해짐에 따라 딥러닝 및 머신러닝이 세상의 빛을 보게 되었다. 이 과정에서 여러 관점에서의 텍스트 분석방법론이 개발되었고, 정보처리의 속도가 향상되었다.

고전 금융분야의 연구들에서는 자산가격의 결정요인을 설명하기 위해 다양한 위험요인들을 고려했지만, 여전히 이 위험요인들로 자산의 가격변동이 설명되는가에 대한 의문은 남아있다. 자산가격은 투자자들의 심리적인 요인에서 기인한다는 행태재무 분야에서는 시장의 이상현상에 대해 심리적 편향을 가진 투자자들로 하여금 비합리적인 의사결정이 발생하기 때문에 거래자의 심리가 자산가격의 결정에 반영된다고 주장하고 있다(Baker, Wang, and Wurgler, 2008). 투자자 심리란 투자자의 공통된 판단오류로 인해 가격을 잘못 측정하는 현상을 의미한다(Shlefier, 2000). Stambaugh, Yu, and Yuan(2012)은 개인투자자가 기관 투자자에 비해 투자 경험이 적고 전문지식이 부족하여 자산의 가격을 잘못 평가할 가능성이 크다고 주장했다. Black(1986)은 가짜 정보에 반응하거나 특정한 정보에 대해 잘못된 방향으로 반응하는 투자자들을 잡음투자자로 지칭하였다. 고전적인 금융분야의 연구들에서 비합리적인 투자자의 행위는 일시적이고 차익거래자에 의해 흡수되므로 자산가격의 결정에 영향을 미치지 않는다고 한다. 하지만, Brown and Cliff(2004)는 투자자의 심리가 지속적으로 가격결정오류를 야기하기 때문에 자산가격의

결정에 영향을 미칠 수 있다고 주장하였다. 뿐만 아니라 Stambaugh et al.(2012)는 투자자의 심리가 주식시장에서 발생하는 이상 현상에 영향을 준다는 사실을 발견하였다. 국내 주식시장에 한해서는 김현아, 정상창(2010)의 연구에서 핫마켓 IPO에서의 이상현상이 낙관적 투자자들의 기대로 인한 것임을 밝혀내었다. 이러한 기존 연구의 결과는 국내외 주식시장에 비합리적이고 비정보기반의 투자자(거래자)가 존재하며, 투자자의 심리가 시장에서 자산의 가격결정에 영향을 미친다는 것을 알 수 있다. 하지만, 기존의 연구들에서는 투자자(거래자)의 심리라는 측정 불가능한 정성지표를 다양한 대용변수를 활용하여 측정하였기 때문에 “심리(감성)” 이 선정된 대용변수로 실제로 대체가 되는지 입증하기 어렵다는 한계가 존재한다. Costola et al. (2023)은 미국 주식 시장에 한하여 3개 언론사의 기사들에 대해 BERT를 사용한 감성과 S&P 500의 수익률 사이에 유의미한 양의 관계가 있다는 사실을 밝혀내었다. 따라서 본 연구에서는 BERT 모형을 사용해 직접적으로는 정량화할 수 없었던 실제 투자자 (거래자)의 감성(심리)을 측정한다. 측정결과인 감성지수를 위험조정수익률과의 선도-지연 관계 분석을 통해 비정보기반투자자(거래자)들의 감성이 실질적으로 자산가격결정에 영향을 미치는지를 살펴보고자 한다. 즉, 본 연구에서 위험조정수익률과 감성지수 사이의 선도-지연 관계 분석을 통해 “비정보기반 투자자들의 심리는 현재의 자산가격변화에 대해 예측력을 갖고 있다.”라는 가설을 검증하고자 한다.

본 연구는 기존 연구들에서 해결하지 못했던 시장참여자 감성 측정의 문제를 BERT 모형을 활용하여 직접 감성을 분석하고 추가적인 선도-지연관계를 분석해 가설을 검증한다. 본 연구의 구성은 제Ⅱ장 연구 방법에서 자료에 대한 설명과 함께 분석에 사용된 모형을 소개하고, 제Ⅲ장에서는 실증분석 결과를 살펴본다. 마지막으로 제Ⅳ장에서는 결론 및 한계를 제시한다.

Ⅱ. 연구 방법

1. 데이터

본 연구에서는 2018년 1월 1일부터 2020년 12월 31일까지의 시가총액 기준 상위 20개 개별종목의 일간 수익률에서 국고채 3년물 수익률을 제한 위험조정수익률에 영향을 미치는 감성지수에 대한 분석에 초점을 맞추고 KOSPI에 상장된 회사들에서 시가총액 기준 상위 20개 기업에 대해 2018년 1월부터 2020년 12월 31일까지의 네이버 종목토론실 댓글을 텍스트 마이닝하여 감성지수를 산출하였다. 감성분석 대상이 된 표본의 수는 다음의 <표 1>과 같다.

<표 1> 종목 코드와 댓글 개수

총 테스트 데이터는 20개 종목 1,559,246개의 종목 토론실에서 발췌한 댓글이며, 아래의 표는 각 종목별로 발췌된 댓글의 수를 의미함.

종목	댓글 수
012330	21,580
207940	15,908
105560	29,559
096770	103,840
068270	214,814
066570	147,966
051910	79,284
051900	6,976
036570	96,560
035720	41,294
035420	24,889
028260	39,669
017670	22,113
000270	118,620
005930	228,737
006400	64,540
005490	40,520
000660	40,333
015760	159,200
005380	62,844
Total Comments	1,559,246

표본의 텍스트 마이닝은 네이버에서 컴퓨터의 접근을 방어하는 부분을 user agent 설정을 통해 해제하였고, 각 페이지의 '날짜', '댓글', '조회', '공감', '비공감'을 마이닝하였다. 마이닝한 데이터에서 '날짜'와 '댓글'만 남긴 후, 날짜를 datetime index로 변환하였다. 숫자와 특수문자 및 오타와 같은 것들은 정규표현식 알고리즘을 사용하여 전처리를 완료하였다. 네이버 종목토론실 댓글은 전문용어의 사용이 잦지 않으므로 금융관련 전문용어로 fine-tuning을 하지 않고 감성적인 용어의 사용이 빈번한 네이버 영화 리뷰 데이터나 쇼핑몰 댓글로 fine-tuning을 진행하여 테스트를 진행하였다. 산출물은 로짓 회귀의 결과이므로 argmax를 취해 긍정은 1, 부정은 0으로 라벨링 하였다(Ke, 2019).

Date	comment	sentiment
2017-08-01	[속보 뉴욕증시 실적 호조 기대에 다우 사상 최고 출발]	1
2017-08-01	[연말 맞아요]	0
2017-08-01	[속보 삼전 자사주 매입 진짜 이유]	0
2017-08-01	[갑자기 왜 떨어짐]	0
2017-08-01	[아 도배하지마라 시바]	0
...	...	...
2021-03-31	[개인들이 너무 사랑해서 못가는 종목]	1
2021-03-31	[올해는 주식하지마라]	0
2021-03-31	[배당금]	1
2021-03-31	[실적발표일]	1
2021-03-31	[경북 일베 토착왜구 이것 모르면 간첩]	0

[그림 1] koBERT 감성분석 결과

1과 0으로 라벨링 된 감성을 아래의 식에 따라 개별종목의 일별 감성지수를 산출하고, 이 결과에 동일 가중평균을 하여 표본 전체의 감성지수를 도출하였다. 아래의 알고리즘에서 i 주식의 t 일에 관측된 긍정 감성의 개수를 $n(Positive_{i,t})$ 라 정의하였다. w 는 가중치 행렬이다.

$$sentiment_{i,t} = \frac{n(Positive_{i,t})}{n(Positive_{i,t}) + n(Negative_{i,t})}$$
$$sentiment_{t} = w \cdot sentiment_{i,t}$$

표본 전체의 감성지수의 분포는 표준편차가 작아 첨도가 높은 형태였기 때문에 표준편차를 늘려 정규분포에 근사시키기 위해 감성지수를 극대-극소 스케일링을 진행하였고, 그 결과는 다음의 <표 2>와 같다. 따라서, 다음의 감성지수는 0에 가까울수록 부정적인 감성을 나타내고, 1에 가까워질수록 긍정적인 감성을 나타내게 된다.

<표 2> 감성지수 산출 결과

날짜	감성지수
2020-12-23	0.585045
2020-12-24	0.635248
2020-12-28	0.627858
2020-12-29	0.457108
2020-12-30	0.694080

<표 3>은 위험조정수익률과 감정지수의 기초통계량을 나타낸 표이다. 평균은 각각 위험조정수익률이 -1.43%, 감성지수는 0.49이다. 즉, 2018년 1월 1일부터 2020년 12월 31일까지의 개별 자산의 수익률이 무위험 수익률보다 크지 않은 것이고, 감성지수는 평균적으로 중립을 유지함을 확인할 수 있다. 다만, 위험조정수익률의 경우 75분위수와 최댓값의 차이가 크기 때문에 long tail이 존재함을 알 수 있고, 감성지수 역시 최솟값은 0, 최댓값은 1이므로 감성이 완전 긍정이거나 완전 부정으로 측정되는 극단적인 사례가 나타났다. 또한, 위험조정수익률의 왜도가 양수이기 때문에 위험조정수익률의 분포는 정규분포보다 오른쪽 꼬리가 긴 형태를 갖는 반면, 감성지수의 왜도는 음수이기 때문에 감성지수의 분포가 정규분포보다는 왼쪽 꼬리가 긴 형태임을 확인할 수 있다.

첨도와 표준편차는 스케일링을 진행한 감성지수와 위험조정수익률에서 차이가 확연히 드러난다. 위험조정수익률의 분포는 첨도가 정규분포 보다 약 7정도 높은 상태이지만 감성지수의 분포는 스케일링을 통해 첨도가 정규분포가 유사해졌으므로 2.93이다. 따라서, 하르케-베라 검정 결과 감성지수의 분포는 5% 유의수준 내에서 귀무가설을 기각하지 않으므로 정규분포에 근사함을 알 수 있다. 다만, 위험조정수익률의 경우 1% 유의수준에서 귀무가설을 기각하므로 정규분포에 근사할 수 없음을 확인할 수 있다.

<표 3> 위험조정수익률과 감성지수의 기초통계량

	위험조정수익률	감성지수
관측치 수	737.000000	737.000000
평균	-0.014304	0.493992
표준편차	0.014267	0.165042
극솟값	-0.098584	0.000000
25%	-0.021716	0.380312
50%	-0.014551	0.510776
75%	-0.007367	0.604310
극댓값	0.078799	1.000000
왜도	0.563012	-0.201048
첨도	10.296801	2.928825
하르케 베라 통계량	1673.949363	5.120505

<표 4> 위험조정수익률과 감성지수의 상관계수

	위험조정수익률	감성지수
위험조정수익률	1.000000	
감성지수	0.087127	1.000000

<표 4>는 위험조정수익률과 감정지수의 피어슨 상관계수를 나타낸 표이다. <표 4>의 결과에 따르면 위험조정수익률과 감성지수는 0.088 정도의 상관관계를 가지고 있다.

시계열 데이터는 데이터의 정상성을 확보하는 것이 중요하다. 그 이유는 첫째, 정상 시계열 데이터는 예측, 추론 및 가설 검정과 같은 통계적 작업을 가능하게 한다. 둘째, 시계열 데이터는 계열 상관이 존재하는 경우가 빈번하므로 정상성이 확보되지 않으면 시간에 따른 평균과 분산이 일정하지 않아 회귀계수가 편향될 우려가 있다. 셋째, 많은 시계열 모델은 정상성 가정에 기반하여 개발되었다. 또한 시계열 데이터의 정상성에 따라서 본 연구에서 사용할 적절한 VAR 모형 혹은 VECM 모형을 선정할 기준을 마련할 수 있다. 마지막으로, 시계열 데이터의 정상성이 확보되지 않은 비정상 시계열 데이터는 추세, 계절성, 그리고 자기 상관과 같은 구성요소를 포함하고 있을 수 있고, 이는 다시 회귀계수 추정의 불편성을 하락시킬 여지가 있다. 이러한 이유로 위험조정수익률과 감성지수의 정상성을 확인하기 위해 Augmented Dickey-Fuller 단위근 검정을 진행하였다.

<표 5> 위험조정수익률과 감성지수 Augmented Dickey-Fuller 단위근 검정

각각 ***, **, *는 1%, 5%, 10% 유의수준을 의미함.

	위험조정수익률	감성지수
유의수준	0.05	0.05
검정 통계량	-4.1687	-3.5244
1% 임계값	-3.439	-3.439
5% 임계값	-2.866	-2.866
10% 임계값	-2.569	-2.569
p-value	0.0007***	0.0074***
	Reject H_0	Reject H_0

<표 5>는 위험조정수익률과 감성지수의 Augmented Dickey-Fuller 단위근 검정의 결과이다. 위험조정수익률과 감성지수 모두 1% 유의수준 하에서 귀무가설인 '시계열 데이터에 단위근이 존재하므로 비정상 시계열 데이터이다.'를 기각하기 때문에 정상성을 갖는 것으로 해석할 수 있다. 따라서 정상 시계열인 위험조정수익률과 감성지수 사이의 선도-지연 관계를 분석하기 위해서는 VAR 모형을 사용해야 한다.

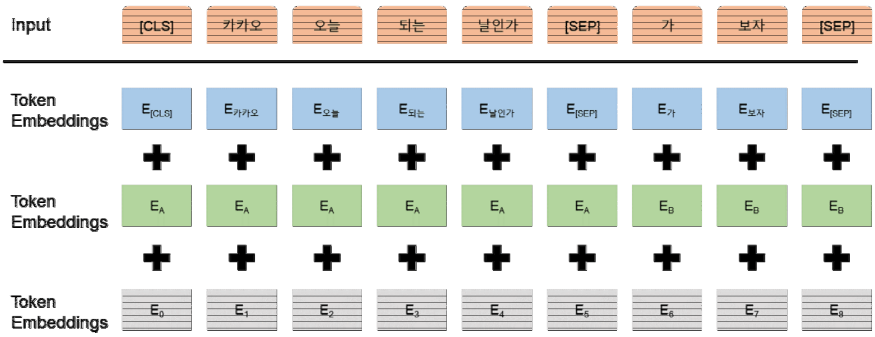
2. 연구 방법

2.1 BERT(Bidirectional Encoder Representations from Transformers)

자연어처리 기법의 시초는 Bag-of-words이다. Bag-of-words 방법은 수집한 텍스트를

개별적인 단어들의 집합으로 학습하고 이를 단어의 표현 빈도로 나타내는 방법이다. 그러나 Bag-of-words 방법은 특이한 단어들에 대해서는 원활하게 작동하지 않는다는 한계가 있다(Im et al., 2018). Bag of Words의 한계를 해결하고자 Word Embedding 방법이 고안되었다. Word Embedding은 한국어 기준으로 어절이나 구(phrase)를 실수 공간에 맵핑(mapping)함으로써 특이치에 대비하는 방법이다. 가장 대표적으로 사용되는 알고리즘은 Mikolov et al.(2013)가 개발한 word2Vec이다. word2Vec에서는 신경망을 통해 하나의 구(phrase)에 대해서 하나의 어절이 다른 어절들과 함께 존재할 확률을 계산하여 각 어절의 의미를 실수 공간으로 구체화하는 방법을 사용한다. 하지만 word2Vec 역시 Word Embedding을 할 때 각 어절의 위치를 고려할 수 없다는 한계가 있다.

word2Vec이나 Glove 알고리즘이 가진 static representation의 한계를 극복하고자 ELMo, GPT, BERT와 같은 contextualized representation이 가능한 모델이 발표되었다. 그 중에서도 정확도가 가장 뛰어난 BERT를 통해 네이버 종목토론실의 댓글을 분석함으로써 감성지수가 초과수익률에 유의미한 영향이 있는지를 살펴보고자 하였다. 한국어로 pre-training 한 koBERT는 token embedding, segment embedding, position embedding을 통해 input을 하게 된다. 임베딩을 하여 인코딩한 데이터를 MLM(Masked Language Model)과 NSP(Next Sentence Prediction) 2가지 방식을 통해 pre-training을 진행한다. 학습된 언어모델을 transfer learning을 시켜 모델을 구성한다.



[그림 2] BERT의 구조[1)]

2.2 VAR(Vector Autoregression)

본 연구에서는 비정보기반투자자의 심리가 자산가격변화에 대해 예측력을 가지는지

1) Devlin, J., Chang, M. W., Lee, K., and Toutanova, K.(2018), Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

살펴보기 위해 VAR 모형을 사용하여 위험조정수익률과 감성지수 간의 선도-지연 관계를 분석한다. 만약 심리가 자산가격변화에 대해 예측력을 갖고 있지 않다면 감성지수가 위험조정수익률에 대해 적어도 0 시차 이상의 지연 관계를 갖게 될 것이다. 반면, 위험조정수익률이 심리에 예측력을 갖는다면 위의 경우와 위험조정수익률은 감성지수에 대해 적어도 1 시차 이상의 선도 관계를 갖게 될 것이다.

$$\begin{bmatrix} \mathrm{Return_t} \\ \mathrm{Sentiment_t} \end{bmatrix} = \begin{bmatrix} \alpha_{\mathrm{r}} \\ \alpha_{\mathrm{s}} \end{bmatrix} + \begin{bmatrix} \delta_{\mathrm{rr},1} & \delta_{\mathrm{rs},1} \\ \delta_{\mathrm{sr},1} & \delta_{\mathrm{ss},1} \end{bmatrix} \begin{bmatrix} Return_{t-1} \\ Sentiment_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \mathrm{Return_{t-p}} \\ \mathrm{Sentiment_{t-p}} \end{bmatrix} + \begin{bmatrix} \varepsilon_{\mathrm{r,t}} \\ \varepsilon_{\mathrm{s,t}} \end{bmatrix}$$

상기한 식을 이용하면 위험조정수익률의 현재값과 감성지수의 현재값은 각각 자기자신의 과거값과 상대방의 과거값으로부터 현재의 값이 예측되는지를 파악할 수 있다.

2.3 그랜저 인과 검정

하지만 VAR 분석만으로는 위험조정수익률과 감성지수 간의 선도-지연 관계가 있다는 주장을 강화하는 것에 한계가 존재하므로 그랜저 인과 분석을 진행하였다. 그랜저 인과 분석은 위험조정수익률을 설명하는 항 중에서 감성지수와 관련이 있는 항들의 계수 δ_{rs}들이 0인지에 대한 가설을 검정하는 방향으로 이루어졌다.

$$\delta_{\mathrm{rs},1} = \delta_{\mathrm{rs},2} = \cdots = \delta_{\mathrm{rs,p}} = 0$$

마찬가지로, 감성지수를 설명하는 항 중에서 위험조정수익률과 관련이 있는 δ_{sr} 들이 0이 되는지에 대한 가설 검정을 수행한다.

$$\delta_{\mathrm{sr},1} = \delta_{\mathrm{sr},2} = \cdots = \delta_{\mathrm{sr,p}} = 0$$

만약, $\delta_{\mathrm{rs,p}}$와 $\delta_{\mathrm{sr,p}}$들 중 적어도 하나 이상의 δ에서 귀무가설을 기각하게 되면 그랜저 인과(Grange cause) 한다고 볼 수 있다. 즉, $\delta_{\mathrm{rs,p}}$에 대해 귀무가설이 기각되면 감성지수가 위험조정수익률을 그랜저 인과하고, $\delta_{\mathrm{sr,p}}$에 대한 귀무가설이 기각되면 위험조정수익률이 감성지수를 그랜저 인과함을 확인할 수 있다.

2.4 OLS(Ordinary Least Squares) Regression

앞서 진행한 분석의 주장을 강화하기 위해 추가적인 회귀분석을 진행하여 위험조정수익률과 감성지수 간의 선도-지연 관계를 분석하였다(Meligkotsidou, 2009). 분석은 CAPM에서 도출한 β에서 계산한 α에 대해 감성지수의 시차를 -1, 0, +1로 수행한 회귀분석과

Fama-French 3 Factor Model에서의 β에서 계산한 α에 대해 감성지수의 시차를 -1, 0, 1로 수행한 회귀분석의 두 가지 방식으로 나누어 진행하였다. CAPM에서의 젠센 α와 Fama-French 3 Factor Model의 젠센 α는 다음과 같이 도출되었다.

$$\alpha_t^{CAPM} = R_t - \beta_{MKT,t} MKT_t - \varepsilon_t$$
$$\alpha_t^{FF} = R_t - \beta_{MKT,t} MKT_t - \beta_{HML,t} HML_t - \beta_{SMB,t} SMB_t - \varepsilon_t$$

위의 식에서 산출한 α_t와 $SENT_{t-1,}$ $SENT_{t,}$ $SENT_{t+1}$ 사이의 OLS를 수행하여 각각 β의 통계적 유의성을 확인한다. $SENT_{t-1}$의 관계에서 통계적으로 유의미하다면 감성지수가 초과수익률을 설명한다고 볼 수 있고, 마찬가지로 $SENT_{t,}$ $SENT_{t+1}$의 β가 통계적으로 유의미한 경우 오히려 감성지수가 초과수익률로 설명됨을 알 수 있게 된다. 설정한 회귀모형은 다음과 같다.

$$y_{t(Jensen's\ \ Alpha)} = \alpha_t + \beta_{-1\ lag}\ SENT_{t-1} + \varepsilon_t$$
$$y_{t(Jensen's\ \ Alpha)} = \alpha_t + \beta_{0\ lag}\ SENT_{t-1} + \varepsilon_t$$
$$y_{t(Jensen's\ \ Alpha)} = \alpha_t + \beta_{+1\ lag}\ SENT_{t-1} + \varepsilon_t$$

본 연구에서는 감성지수와 위험조정수익률의 선도-지연 관계를 살펴보고 있으므로, 젠센 α의 회귀분석을 통해 위험조정수익률에 더해 초과수익률과 감성지수의 선도-지연 관계를 분석함으로써 비정보기반투자자들의 감성이 자산 가격에 움직임을 일으키는지 혹은 자산 가격의 변화에 기여를 거의 하지 못하고 단순히 반응만 하는 것인지를 확인할 수 있다.

Ⅲ. 실증분석

1. 횡단면 단순선형회귀분석

VAR 모형을 활용한 선도-지연 관계 분석에 앞서 위험조정수익률과 감성지수에 유의미한 관계가 있는지 확인해보기 위해 횡단면 회귀분석을 진행하였다. 개별자산 위험조정수익률의 전체 관측기간에 대한 평균을 종속변수로 개별자산 감성지수의 전체 관측기간에 대한 평균을 독립변수로 두어 분석을 진행하였고, 설정한 회귀모형은 아래와 같다.

$$y_i = \alpha + \beta x_i + \varepsilon$$

회귀계수가 통계적으로 유의미한 값을 갖는다면 위험조정수익률과 감성지수가 상호

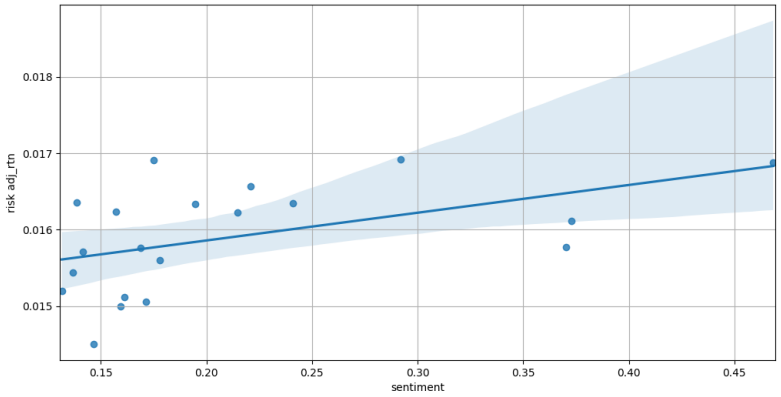
연관이 있음을 확인할 수 있다. 다만, 시간에 대한 변수는 고려되지 않았기 때문에 선도-지연 관계를 이 회귀분석에서는 살펴볼 수 없으므로 이는 후술할 VAR 모형을 통한 분석에서 설명한다.

<표 6> 위험조정수익률과 감성지수의 횡단면 회귀분석 결과

각각 ***, **, *는 1%, 5%, 10% 유의수준을 의미함.

R-squared:		0.235		Adj. R-squared:		0.193
Date:		Wed, 31 May 2023		Prob (F-statistic):		0.0301
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0151	0.000	42.536	0.000***	0.014	0.016
x1	0.0036	0.002	2.355	0.030**	0.000	0.007

분석 결과, 회귀계수는 0.0036, p-value는 0.03으로, 5% 유의수준 내에서 $\beta = 0$이라는 귀무가설을 기각하였다. 즉, 위험조정수익률의 평균은 감성지수의 평균에 0.0036만큼 영향을 받고 있음을 알 수 있다. 아래의 [그림 3]은 회귀분석 결과를 도식화한 그림이다.



[그림 3] 위험조정수익률과 감성지수의 횡단면 회귀분석 그래프

2. 선도-지연 관계 분석

앞서 진행한 단위근 검정 결과에 따라 VECM 모형이 아닌 VAR 모형을 사용하였다. VAR 모형을 활용한 추정에서 우선적으로 진행해야 하는 과정은 적정시차 설정이다. VAR 모형에서는 정보기준에 따라 적정시차를 설정하게 된다. 정보기준을 통한 적절한 시차 설정으로 모델의 과적합을 방지함과 동시에 예측능력을 증가시킬 수 있으므로 필수적인

과정이다. 이때 각 정보기준은 모델의 복잡성을 표현하는 페널티 항과 데이터 적합도를 고려하여 모델을 평가하게 되고 그 결과는 아래 <표 7>과 같다.

<표 7> VAR Order Selection

Lag	AIC	BIC	FPE	HQIC
0	-12.12	-12.10	5.462e-06	-12.11
1	-12.23	-12.20	4.866e-06	-12.22
2	-12.34	**-12.27***	4.383e-06	**-12.31***
3	**-12.35***	-12.26	**4.344e-06***	-12.31

정보기준을 통한 시차 결정 결과 AIC와 FPE는 시차 3이, BIC와 HQIC에서는 시차 2가 적정 시차임을 나타내었다. 본 연구에서는 위험조정수익률과 감성지수의 동적 관계를 파악함으로써 비정보기반투자자들의 심리와 자산가격변화 간의 선도-지연 관계에서 정보의 반영은 얼마 동안의 시간을 두고 일어나는지를 분석하기 위해 가능한 범위 내에서 시차를 최대한 늘리고자 적정 시차를 3 시차로 설정하였다.

<표 8> VAR(3) 추정 결과

각각 ***, **, *는 1%, 5%, 10% 유의수준을 의미함.

	위험조정 수익률	p-value	감성지수	p-value
상수항	-0.009155	0.000***	0.341653	0.000***
위험조정수익률(-1)	0.091154	0.014***	0.884879	0.043**
위험조정수익률(-2)	0.294681	0.000***	0.846768	0.044**
위험조정수익률(-3)	0.118967	0.001***	0.061415	0.107
감성지수(-1)	0.003168	0.314	0.231041	0.000***
감성지수(-2)	-0.001177	0.714	0.037982	0.062*
감성지수(-3)	0.002261	0.469	0.059618	0.889

<표 9> 위험조정수익률과 감성지수의 그랜저 인과 검정

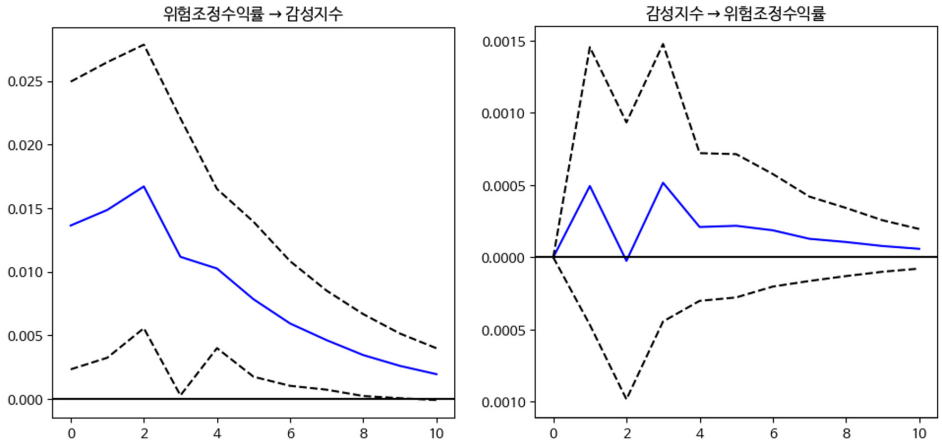
각각 ***, **, *는 1%, 5%, 10% 유의수준을 의미함.

	H_0: 위험조정수익률 $\not\rightarrow$ 감성지수	H_0: 감성지수 $\not\rightarrow$ 위험조정수익률
검정 통계량	3.541	0.561
임계값	2.611	2.611
p-value	0.014**	0.641
df	(3, 1454)	(3, 1454)
	Reject H_0	Do not reject H_0

적정 시차를 설정하였으므로, VAR(3) 모형을 통해 위험조정수익률과 감성지수 간의 선도-지연 관계에 대한 분석을 실시하였다. <표 8>은 시차 1과 시차 2에서 위험조정수익률이 감성지수에 5% 유의수준 내에서 예측력을 갖고 있음을 나타내고 있다. 그러나 감성지수의 위험조정수익률에 대한 예측력은 모든 시차에서 10% 유의수준 내에서도 예측력을 갖지 못하고 있다. 즉, 감성지수가 위험조정수익률에 선도하는 관계는 어떠한 시차에서도 존재하지 않으며 오히려 위험조정수익률이 감성지수에 선도하고 감성지수는 위험조정수익률에 지연되는 관계가 있음을 VAR(3) 모형 추정을 통해 확인하였다.

<표 9>의 그랜저 인과 분석에서는 위험조정수익률이 감성지수에 그랜저 인과(Granger cause) 하지 않는다는 귀무가설이 5% 유의수준 내에서 기각되어, 실질적으로 위험조정수익률의 과거값이 감성지수의 현재값에 대해 예측력을 갖고 있다는 VAR(3) 모형의 분석 결과를 강화하고 있다.

충격반응분석은 시계열 변수 간의 움직임에 대한 동적인 효과를 분석하는 데에 유용한 분석 기법이다. 충격반응분석을 통해 각 시계열 변수에서 한 변수에 대한 충격(shock)이 다른 변수에 어떤 영향을 미치는지 살펴봄으로써 변수 간의 인과관계와 동적인 상호작용을 분석할 수 있다. 아래의 [그림 4]는 위험조정수익률과 감성지수를 이용한 충격반응분석의 결과이다.



[그림 4] VAR 충격반응분석

[그림 4]의 충격반응분석에 따르면 감성지수 1단위 충격이 위험조정수익률에 미치는 충격은 신뢰구간이 0을 포함하고 있으므로 충격이 발생한 즉시 소멸한다. 반면, 위험조정수익률 1단위 충격은 2일 후 0.016의 영향을 감성지수에 미쳤다가 시간이 지날수록 충격이

소멸하게 된다. 즉, <표 8>의 VAR(3) 추정, <표 9>의 그랜저 인과 분석 결과로 미루어 보아 감성지수의 위험조정수익률에 대한 예측력보다 위험조정수익률의 감성지수에 대한 예측력이 더 크다는 사실을 확인할 수 있다.

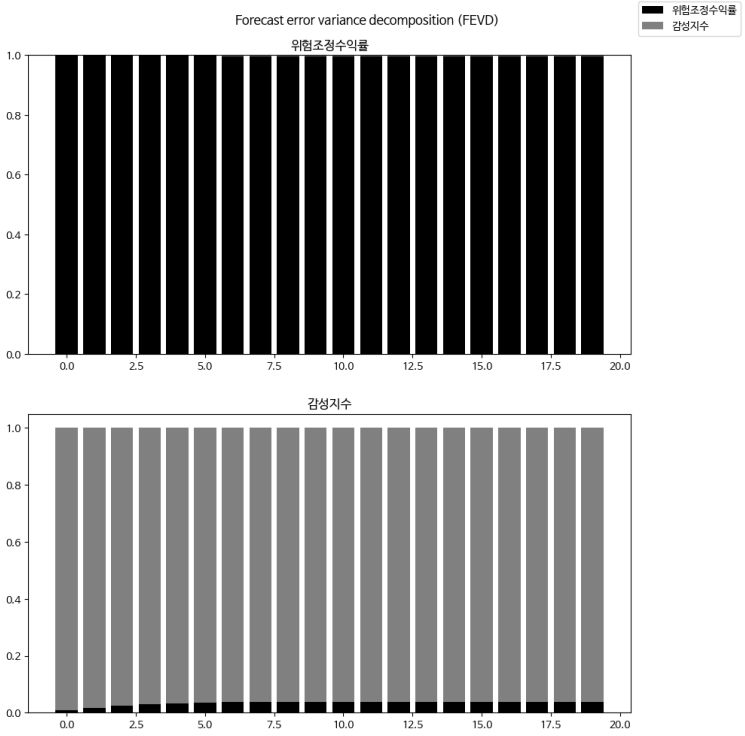
추가적으로 예측오차분산분해를 통해 위험조정수익률과 감성지수가 서로 얼마나 영향을 미치는지 파악하고자 하였다. 예측오차분산분해서에서는 변수의 오차 분산을 다른 변수에 대한 외생적 충격으로부터 기인한 설명력과 자기 자신의 충격으로부터 기인한 설명력으로 분리한다. 따라서, 전체 충격 중 다른 변수에서 기인한 충격이 어느 정도인지 파악할 수 있다는 이점이 있다. <표 10>과 <표 11>에 따르면 위험조정수익률의 오차 분산은 자기 자신의 충격에 의해서 대부분 설명되는 것에 비해서, 감성지수의 오차 분산은 위험조정수익률에 의해 설명되는 부분이 0기에는 0.76%, 4기에는 3.46%까지 증가한다. 하지만, 위험조정수익률은 4기에도 감성지수에 의해 설명되는 오차 분산의 비중이 0.27%에 지나지 않는다. 결과적으로, 위험조정수익률이 비정보기반투자자의 심리(감성지수)에 영향을 주는 것과 달리, 비정보기반투자자의 심리(감성지수)는 위험조정수익률에 거의 영향을 주지 못한다.

<표 10> VAR 예측오차분산분해

위험조정수익률에 대한 예측오차분산분해		
	위험조정수익률	감성지수
0	1.000000	0.000000
1	0.998620	0.001380
2	0.998734	0.001266
3	0.997433	0.002567
4	0.997247	0.002753
감성지수에 대한 예측오차분산분해		
	위험조정수익률	감성지수
0	0.007591	0.992409
1	0.015633	0.984367
2	0.025699	0.974301
3	0.029920	0.970080
4	0.033594	0.966406

마지막으로 위험조정수익률이 아닌 초과수익률과 감성지수 간의 관계를 살펴봄으로써 감성지수의 각 시차에 따른 자산의 초과수익률에 대한 설명력을 OLS를 통해 분석하였다. OLS 분석 결과, 감성지수의 +1 시차에서 초과수익률을 설명하는 회귀계수가 통계적으로 유의미하다면 앞서 진행했던 위험조정수익률은 감성지수에 예측력이 존재한다는 분석 결과와 함께 비정보기반투자자들의 감성이라는 정보는 자산가격결정에 영향을 미치지 않고

반대로 자산의 가격결정이 비정보기반투자자들의 감성을 설명함으로써 비정보기반투자자들의 심리는 자산가격변화에 영향을 주지 않는다는 주장을 강화할 수 있다. <표 12>에서는 각 시점에서 계산한 CAPM 모형에서 계산한 젠센의 알파와 Fama-French 3 Factor 모형에서 계산한 젠센의 알파를 종속변수로 두고 감성지수에 시차를 두어 개별적으로 OLS를 진행한 결과이다.



[그림 5] VAR 예측오차분산분해

<표 11> 젠센의 알파와 감성지수 간의 선도-지연 관계 회귀분석 결과

각각 ***, **, *는 1%, 5%, 10% 유의수준을 의미함.

	lag for sentiment	coefficient	p-value
FF3F Alpha	sentiment (-1)	0.0020	0.164
	sentiment (0)	0.0028	0.047**
	sentiment (+1)	0.0026	0.068*
CAPM Alpha	sentiment (-1)	0.0018	0.214
	sentiment (0)	0.0026	0.067*
	sentiment (+1)	0.0025	0.077*

<표 12>에 따르면 CAPM 모형에서 계산한 젠센의 알파와 Fama-French 3 Factor 모형에서 계산한 젠센의 알파에 대한 감성지수의 설명력을 분석하는 회귀분석 결과들에서 일관적으로 -1 시차에서의 회귀계수는 통계적으로 유의미하지 않고 0 시차와 +1 시차에서의 회귀계수가 10% 유의수준 내에서 통계적으로 유의미한 것으로 나타났다. 즉, 0 시차에서 감성지수는 각각 CAPM 모형의 초과수익률에는 0.0026, Fama-French 3 Factor 모형의 초과수익률에는 0.0028 만큼의 영향이 있다. 또한, +1 시차에서 감성지수는 각각 CAPM 모형의 초과수익률에는 0.0025, Fama-French 3 Factor 모형의 초과수익률에는 0.0026 만큼의 영향이 있다. 따라서 비정보기반투자자들의 감성은 자산가격결정에 0에서 +1 시차 만큼의 간격을 두고 영향이 있다는 사실로 말미암아 앞선 VAR(3) 모형의 추정 결과와 함께 비정보기반투자자들의 심리는 자산가격변화에 대해 예측력과 설명력을 갖지 못함을 시사하고 있다.

Ⅳ. 결 론

본 연구에서는 2018년 1월 1일부터 2020년 12월 31일까지의 시가총액 기준 상위 20개 기업의 온라인 주식게시판의 댓글들에 BERT 모형을 이용해 산출한 감성지수와 일별 수익률에서 무위험수익률을 제한 위험조정수익률 그리고 Market 팩터, Value 팩터, Size 팩터를 이용해 시장 참여자 중 비정보기반거래의 대다수를 차지하는 개인의 감성과 자산가격의 결정 사이의 선도-지연 관계를 분석하였다. 선도-지연 관계 분석을 통해 한국의 코스피 시장에서 비정보기반투자자들의 심리가 자산가격의 변화를 일으키는지 혹은 단순히 자산가격의 변화에 반응만 하는 것인지에 대해 알아보고자 하였다.

이를 위해, 위험조정수익률과 감성지수의 단위근 검정을 통해 데이터의 정상성을 확인하였고, 위험조정수익률과 감성지수 모두 정상 시계열로 나타났다. 이에 따라, VAR(3) 모형을 통해 선도-지연 관계를 살펴보았다. 동시에 그랜저 인과 분석, 충격반응분석, 그리고 예측오차분산분해를 수행하였다. 또한, CAPM 모형과 Fama-French 3 Factor 모형으로 계산한 젠센의 알파를 사용하여 시차를 준 감성지수들과 OLS 회귀분석을 진행하였다.

분석 결과, VAR(3) 추정 및 젠센의 알파와의 회귀분석 결과 감성지수는 위험조정수익률 및 초과수익률을 선도하는 관계가 존재하지 않았고 이는 곧 시장에서 비정보기반투자자들의 심리라는 정보는 자산가격에 반영되지 못했다고 해석할 수 있다. 따라서 한국의 코스피 시장에서 비정보기반거래자들의 심리는 자산가격변화에 예측력이 있다고 가정하였던 본 연구의 가설이 기각되었다.

단, 본 연구는 감성지수를 산출할 때, 온라인 주식 게시판의 댓글에서 감성은 구분할 수 있지만 시장의 가격결정에는 어떠한 영향도 주지 않는 댓글들도 포함이 되었기 때문에 감성지수 자체에 잡음이 포함되는 한계가 있다. 그럼에도 불구하고 기존에는 컴퓨터 공학

및 통계학의 한계로 인해 시장참여자들의 감성이라는 비정형적인 지표를 측정할 수 없었기 때문에 감성을 대변하는 모멘텀 (Fama, 2014), 개인투자자의 매수-매도 불균형(김가람, 2018) 등의 대용변수(proxy variable)를 사용하여 독립변수로 사용했다. 하지만 BERT와 같은 양방향 자연어 학습 모형으로 인해 자연어처리 성능이 비약적으로 발전함에 따라 시장 참여자들의 감성을 대용변수를 활용하지 않고 직접적인 측정이 가능하게 되었다. 따라서 본 연구는 감성지수를 변수로 한 선도-지연 관계 분석을 통해 비정보기반투자자들의 심리가 자산가격변화에 예측력과 설명력이 있는지 확인하였다는 점에 의의가 있다. 향후 연구과제로는 한국 언론사들이 생산해내는 음성데이터와 문자데이터, 그리고 증권사 보고서들과 같이 잡음이 적은 비정형데이터를 입력 데이터로 사용하여 감성지수를 계산해보는 것이다. 이러한 정제된 데이터를 통해 계산된 감성지수를 활용하여 시장에서의 심리와 자산수익률 간의 유의미한 상관관계가 있는지를 확인한다면 투자자의 심리라는 변수가 믿을만하고 유용하게 쓰일 수 있는 계기가 될 것이다.

참 고 문 헌

김가람・류두진・양희진 (2018), “투자심리지수의 대용변수와 유용성: 개별기업 주식수익률에 미치는 영향을 바탕으로,” 경영학연구, 제47권 제5호, 1231-1260.

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

김은총・이낙영・강형구 (2022), “비트코인 현물과 선물시장 사이의 가격발견기능과 전이효과에 관한 연구,” 금융공학연구, 제21권 제2호, 79-102.

김현모・류경석・윤호영・소리・박재홍 (2014), “온라인 주식게시판 정보와 주식거래량에 관한 인과관계 연구: 온라인 주식게시판의 역할을 중심으로,” 한국IT서비스학회 학술대회 논문집, 1-7.

김현아・정성창 (2010), “낙관적 투자자의 기대가 핫마켓상황 IPO 시장의 이상현상에 미치는 영향력 검증,” 재무관리연구, 제27권 제2호, 1-33.

연소영・정정현 (2023), “시장지표를 이용한 산업별 주식수익률의 예측성과,” 금융공학연구, 제22권 제1호, 81-1196.

장은아・최회련・이홍철 (2020), “BERT를 활용한 뉴스 감성분석과 거시경제지표 조합을 이용한 주가지수 예측,” 한국컴퓨터정보학회논문지, 제25권 제5호, 47-56.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

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

Black, F. (1986), “Noise,” *The Journal of Finance*, 41(3), 528-543.

Bollen, J., H. Mao, and X. Zeng (2011), “Twitter mood predicts the stock market,” *Journal of Computational Science*, 2(1), 1-8.

Brogaard, J., T. H. Nguyen, T. J. Putnins, and E. Wu (2022), “What moves stock prices? The roles of news, noise, and information,” *The Review of Financial Studies*, 35(9), 4341-4386.

Brown, G. and M. Cliff (2005), “Investor sentiment and asset valuation,” *Journal of Business*, 78(2).

Brown, G. W. and M. T. Cliff (2004), “Investor sentiment and the near-term stock market,” *Journal of Empirical Finance*, 11(1), 1-27.

Costola, M., O. Hinz, M. Nofer, and L. Pelizzon (2023), “Machine learning sentiment analysis, COVID-19 news and stock market reactions,” *Research in International*

Business and Finance, 64, 101881.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Easley, D., M. O'Hara, and P. S. Srinivas (1998), "Option volume and stock prices: Evidence on where informed traders trade," *Journal of Finance*, 53(2).

Fama, E. F. and K. R. French (1993), "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33, 3-56.

Fama, E. F. and K. R. French (2015), "A five-factor asset pricing model," *Journal of Financial Economics*, 116(1), 1-22.

Fraiberger, S. P., D. Lee, D. Puy, and R. Ranciere (2018), "Media sentiment and international asset prices (No. w25353)," *National Bureau of Economic Research*.

Fromentin, V., M. S. H. Lorraine, C. P. E. R. Ariane, and T. Alshammari (2022), "Time-varying causality between stock prices and macroeconomic fundamentals: Connection or disconnection?," *Finance Research Letters*, 49, 103073.

Han, B. (2008), "Investor sentiment and option prices," *The Review of Financial Studies*, 21(1).

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., & Xu, Y. (2019). BERT-based financial sentiment index and LSTM-based stock return predictability. arXiv preprint arXiv:1906.09024.

Hochreiter, S. and J. Schmidhuber (1997), "Long short-term memory," *Neural Computation*, 9(8), 1735-1780.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Jung, S. H., G. J. Gu, D. Kim, and J. W. Kim (2020), "Predicting Stock Prices Based on Online News Content and Technical Indicators by Combinatorial Analysis Using CNN and LSTM with Self-attention," *Asia Pacific Journal of Information Systems*, 30(4), 719-740.

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

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

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert:

A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

Lee, C. M. C., A. Shleifer, and R. H. Thaler (1991), "Investor sentiment and the closed-end fund puzzle," *Journal of Finance*, 46(1).

Lee, Y. and J. H. Rhee (2022), "A VECM analysis of Bitcoin price using time-varying cointegration approach," *Journal of Derivatives and Quantitative Studies*, (ahead-of-print).

Levy, O. and Y. Goldberg (2014), "Neural word embedding as implicit matrix factorization," *Advances in Neural Information Processing Systems*, 2177-2185.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., & Du, X. (2018). Analogical reasoning on chinese morphological and semantic relations. arXiv preprint arXiv:1805.06504.

Meligkotsidou, L., I. D. Vrontos, and S. D. Vrontos (2009), "Quantile regression analysis of hedge fund strategies," *Journal of Empirical Finance*, 16(2), 264-279.

Opitz, D. and R. Maclin (1999), "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, 11, 169-198.

Shleifer, A. (2000), *Inefficient markets: An introduction to behavioural finance*, Oup Oxford.

Stambaugh, R. F., J. Yu, and Y. Yuan (2012), "The short of it: Investor sentiment and anomalies," *Journal of Financial Economics*, 104(2), 288-302.

Xu, H. C. and W. X. Zhou (2018), "A weekly sentiment index and the cross-section of stock returns," *Finance Research Letters*, 27, 135-139.

Abstract

Analysis of Lead-Lag Relationship between Risk-Adjusted Return and Sentiment of Uninformed Investors

Sang Uk Hwang and Eun Chong Kim***

The purpose of this study is to examine whether the investor's sentiment has an predictive and explanatory power for asset price changes through an empirical analysis of the lead-lag relationship between uninformed investor's sentiment and risk-adjusted returns. VAR model, Granger causality analysis, Impulse response analysis, and forecast error variance decomposition were conducted to analyze the lead-lag relationship between risk-adjusted returns and sentiment index. As a result of the analysis, it was confirmed that the sentiment index did not have a leading relationship with the risk-adjusted return rate, but that there was a statistically significant relationship in the delayed lag. In conclusion, it was found that the past information on investor sentiment of uninformed investors does not predict current change of asset price, but rather, the current sentiment of uninformed investors is predicted by lagged information on the change of asset price.

Key words: Uninformed Investor, BERT, Sentiment, Lead-Lag, VAR

* First Author, Master Course Department of Finance, Hanyang University, +82-2-2220-2439, E-mail: sanguk.hwang@kakao.com

** Corresponding Author, ETF Manager, Hanhwa Asset Management, Adjunct Professor, Department of Computer Science and Financial Engineering, Hanyang University, +82-2-2220-2439, E-mail: eunchong01@gmail.com