Asia-pacific Journal of Convergent Research Interchange
Vol.9, No.9, September 30 (2023), pp.313-322
http://dx.doi.org/10.47116/apjcri.2023.09.24

A Study on the Precedence of Financial Markets Using YouTube Data

유튜브 데이터를 이용한 금융 시장의 선행성 분석

Jae Pil Yu[1], Jong Min Park[2]

유재필[1], 박종민[2]

[1] Professor, Management Engineering, Sangmyung University, Korea, jaepilyu@smu.ac.kr
[2] Researcher, RM Team, KOREA INVESTORS SERVICE, Korea, qkrwhdals1gh@naver.com

Corresponding author: Jae Pil Yu

Abstract: Emotional analysis is a way to predict the financial market, which is generally analyzed using things such as Google search trends and Twitter data. Recently, the frequency of use of YouTube has increased compared to these media, and since video data is the main environment, research on emotional analysis using YouTube is not easy to find. Financial market participants can search and watch related videos through YouTube according to market changes and leave comments. In other words, the frequency of YouTube use may vary depending on changes in investor sentiment. Therefore, in this study, data such as the number of views and comments of major YouTube channels were collected through a crawling algorithm, and statistical unit root verification and parallax analysis were performed. In addition, through Grandeur causal analysis, the priorities with financial assets were quantitatively analyzed, and as a result, data such as YouTube's number of views and comments were statistically proven to precede the financial product market. In particular, YouTube data was found to have a stronger precedent for trading volume than the price of financial assets. The results of this study are expected to serve as a reference for future studies on emotional analysis through various media.

Keywords: Grandeur Causal Analysis, Antecedent Analysis, Emotional Analysis, YouTube, Finance

요약: 금융 시장을 예측하는 방법으로 감성분석이 있는데 이는 일반적으로 구글 검색 트랜드나 트위터 데이터와 같은 것을 활용하여 분석한다. 최근에 이러한 매체보다 유튜브의 사용 빈도가 높아지고 있는데 영상 자료가 주된 환경이다 보니 유튜브를 활용한 감성분석에 관한 연구는 쉽게 찾아볼 수 없다. 금융 시장 참여자들은 시장 변동에 따라서 유튜브를 통해서 관련 영상을 검색하여 시청할 수 있으며 댓글 등을 남기는 활동을 할 수 있다. 즉 투자자의 심리 변화에 따라서 유튜브 이용 빈도가 달라질 수 있다. 따라서 본 연구에서는 크롤링 알고리즘을 통해서 주요 유튜브 채널들의 조회 수, 댓글 수 등의 자료를 수집하고 이를 통계적인 단위근 검증과 시차 분석을 수행했다. 또한 그랜저 인과분석을 통해서 금융 자산과의 선행성을 정량적으로 분석했으며, 그 결과 유튜브의 조회 수, 댓글 수 등과 같은 데이터는 금융 상품 시장을 선행한다는 것을 통계적으로 입증했다. 특히 유튜브 데이터는 금융 자산의 가격보다는 거래량에 더 강한 선행성을 보이는 것으로 나타났다. 본 연구의 결과는 향후 후행될 다양한 매체를 통한 감성분석에 관한 연구에 참고 문헌이 되기를 기대한다.

Received: May 14, 2023; 1st Review Result: June 18, 2023; Accepted: August 25, 2023

핵심어: 그랜저 인과분석, 선행성 분석, 감성분석, 유튜브, 금융

1. 서론

금융 시장은 높은 수익을 기대할 수 있는 동시에 불확실성과 위험성이 높다. 금융 자산에 대한 투자는 투자자들에게 매력적인 수익 기회를 제공하지만 동시에 자본 손실의 위험도 함께 따른다. 이에 따라 금융 시장의 위험 요소를 이해하고 관리하는 것은 투자자 보호와 금융 안정성 강화를 위해 필수적인 과제이다. 특히 주식 시장은 일상적으로 변동성이 큰 시장으로써 가격 추이를 예측하기 어렵기 때문에 주식 시장은 투자자들에게 많은 위험을 안겨줄 수 있다[1]. 이러한 변동성은 급격한 가격 하락으로 인해 투자자들이 자본 손실을 입을 수 있는 위험을 내포하며 투자자들은 이를 적절히 관리하는 전략이 필요하다. 또한 금융 시장은 다양한 시장 리스크에 노출되는데 정치적, 경제적 그리고 사회적 변화 등의 요인들은 금융 시장에 막대한 영향을 미칠 수 있으며, 이러한 잠재적 위험은 금융 자산의 가격에 대한 불확실성을 증가시킬 수 있다. 예컨대 금리 변동, 세제 정책 변경, 자연재해 등의 이벤트는 금융 시장에 영향을 미쳐 투자자들에게 위험을 초래할 수 있으며, 이에 따라 금융 시장의 다양한 위험을 인식하고 대비하는 방안이 필요하다. 더불어 투자자들은 자본을 투자하고 의사결정을 하는 과정에서 감정적인 요소들에 의해 영향을 받을 수 있고, 이는 합리적이지 않은 행동과 판단을 초래할 수 있다. 즉 과도한 낙관성이나 비관적인 태도는 투자자의 판단을 흐리게 하고 잘못된 투자 결정을 유발할 수 있으며, 무엇보다 투자자의 심리적 요인을 최소화할 수 있는 과학적이고 정량적인 예측 기술이 무엇보다 중요하다.

주가 예측은 금융 시장에서 매우 중요한 역할을 하는데 첨단 정보통신 기술이 발전하면서 정량적인 방법으로 주가의 움직임을 예측하여 효과적인 투자 전략을 수립하고 손실을 최소화하는 연구와 개발이 활성화되고 있다[2]. 예컨대 SNS(Social Network Service)의 데이터를 이용해서 주가의 방향성을 예측하는 연구가 있는데 이는 감성분석의 하나로 투자자 감정 파악, 대중 의견 파악, 정보 비대칭 문제 해결 등에서 활용되고 있다. 이러한 감성분석은 주식 시장의 흐름을 예측하고 투자자들의 행동을 이해하는 데 도움을 줄 수 있다[3]. SNS를 이용하여 주가 움직임을 예측하는 방법론은 감성분석, 텍스트 마이닝, 소셜 네트워크 분석, 기계 학습 등이 있으며, 이를 통해 투자자들의 감정과 관심사를 파악하고 주가 움직임과의 상관관계를 분석하여 예측 모델을 구축함으로써 더욱 정확한 주가 예측을 실현할 수 있다.

최근에는 유튜브가 인터넷상에서 가장 인기 있는 영상 공유 플랫폼 중 하나로, 수많은 사용자가 영상을 업로드하면서 동시에 시청도 가능한 공간이다. 이러한 영상 데이터는 다양한 주제와 콘텐츠를 다루며 많은 수의 시청자와 상호작용을 한다. 이러한 유튜브 데이터는 사용자들의 관심과 반응을 반영하고 대중의 의견과 감정을 파악하는 데에 유용한 정보를 제공할 수 있다[4]. 특히 [그림 1]은 유튜브라는 단어를 제외하고 나머지 글 제목을 모자이크 처리한 삼성전자의 주식 게시판인데 유튜브와 관련된 글들이 자주 올라오고 있으며, 이는 주식과 유튜브와의 정성적 관계는 유의미하다는 것을 보여주는 사례다.

유튜브는 국민의 약 80% 이상이 매일 사용하고 있으며, 월평균 사용 시간도 약 30시간 정도에 달하는데 이는 인스타그램과 페이스북 등 다른 SNS에 비해서도 월등하게

높다[5]. 그만큼 유튜브의 조회 수, 댓글 수 등의 데이터는 대중들의 유행과 관심사 등을 확인하는데 중요한 자료가 될 것이며, 이를 정량적으로 분석하는 방안에 관해서 연구하는 것은 매우 의미가 있을 것으로 판단된다.

2023.07.26 18:02	유튜브 무단 [illegible]	jaun****	73	10	1
2023.07.25 04:29	유튜브로 고도화 [illegible]	yobg****	191	0	0
2023.07.16 14:46	유튜브 [illegible]	cyh****	279	3	0
2023.07.14 08:42	유튜브 [illegible]	jait****	228	1	2
2023.07.13 10:47	유튜브 [illegible]	choz****	211	14	0
2023.07.07 18:30	유튜브 [illegible]	jaun****	182	6	2
2023.07.04 00:13	유튜브 vs [illegible]	cons****	497	1	2
2023.07.02 22:50	유튜브에서 BBC뉴스 코리아 [2]	tm******	372	5	4
2023.06.26 16:45	└유튜브 가짜뉴스에 [illegible]	dia1****	319	2	3
2023.06.25 16:39	└유튜브의 가짜뉴스에 [illegible]	dia1****	385	0	3
2023.06.24 06:35	유튜브에서 드디어 [illegible]	jait****	682	2	0
2023.06.24 06:36	└유튜 [illegible]	panj****	513	2	1
2023.06.21 16:54	└유튜브랑 똑같이 [illegible]	dia1****	259	0	2
2023.06.10 01:26	유튜브 이서진 중국에서 [illegible] [1]	rian****	965	13	1
2023.06.09 19:57	유튜브 어디선가 [illegible] [2]	jaun****	398	13	6
2023.06.07 18:39	유튜브 [illegible]	jaun****	293	3	1
2023.06.02 13:50	유튜브에 [illegible] [1]	sood****	319	5	0
2023.05.31 14:54	유튜브에 [illegible]	js51****	240	0	1
2023.05.22 14:56	유튜브에 [illegible] [1]	yili****	340	1	1
2023.04.25 16:46	유튜브 [illegible]	ashg****	172	0	0

[그림 1] 주식 게시판에서의 유튜브 관련 글

[Fig. 1] YouTube Posts on Stock-Related Boards

따라서 본 연구에서는 앞서 설명한 대표적인 콘텐츠 채널인 유튜브의 구독자 및 조회 수 등의 데이터를 크롤링(Crawling) 알고리즘을 통해서 수집하고, 이를 금융 시장의 대표적인 지표인 KOSPI 및 KOSDAQ 지수와의 인과 관계를 그랜저 인과분석을 통해서 분석하고자 한다. 본 연구를 통해 유튜브 데이터를 활용한 주가 예측 모델의 유효성과 가능성을 검증함으로써 새로운 시각과 기여를 제공하고, 향후 관련 분야에 대한 더 나은 이해와 기술 발전을 도모하기를 기대한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 선행 연구를 고찰하고 3장에서는 연구 방법을 소개한다. 그리고 4장에서는 실험계획 및 결과 분석을 기술하고, 마지막으로 5장에서는 결론 및 시사점을 제시한다.

2. 선행연구 고찰

감성분석은 다양한 분야에서 의사결정 문제를 과학적으로 해결하기 위해 매우 중요한 분석 기법이다[6]. 최근에는 인간과 컴퓨터의 상호작용 확대로 인해서 SNS에서 사용자의 감정을 추출하는 방식이 매우 중요해지고 있다[7].

시장 참여자들은 자신이 투자한 금융 자산에 대한 정보를 취득하기 위해서 인터넷 포털 사이트 등을 활용하는데 이는 호황과 불황에 따라서 검색량에 차이가 있다[8]. Huang et al.(2020)의 연구에서도 S&P500 지수의 움직임에 따라서 특정 단어들의 구글 검색량에 차이가 있음을 입증했다[9].

Bank et al.(2011)는 독일 주식 시장과 주식 관련 단어에 대한 구글 검색량과의 인과 관계를 분석한 결과, 검색량의 증가는 후행적으로 주식 시장의 유동성이 증가했다[10]. Liu et al.(2015)는 중국 주식의 예측을 위해 인터넷 데이터의 전처리 모형인 CLSI(Composite Leading Search Index)를 적용했으며, 이는 구글의 검색량 지수인 SVI(Search Volume Index)에 비해 우수한 성능을 보였다[11]. Smith(2012)는 구글에서 경제위기와 같은

단어의 검색량과 금융 시장 가치 변동의 함수 모형이 GARCH(1, 1) 수준으로 검색량 데이터의 예측력을 입증했다[12]. GARCH 모형의 특성상 금융 시계열 값 자체를 예측할 수 없지만 검색량 추이 데이터를 통해서 금융 시장의 변동성을 예측할 수 있다는 것에서 의미가 있다.

Chechley et al.(2017)은 Granger 인과 관계 모형을 이용해서 트위터의 활동성 강도와 주식의 가격, 변동성 그리고 거래량과의 선행적 관계를 분석했는데, 특히 가격에 비해서 변동성과 거래량은 트위터의 활동성 강도가 강한 선행을 보인다고 나타났다[13]. Zhang et al.(2011)의 연구에서는 트위터 사용자에 대한 활동 내용을 바탕으로 주가와 상관관계를 분석했는데, 트위터의 주식 관련 데이터의 증가는 주식 시장과 음의 관계를 보인다는 결과를 보였다[14].

주식 시장이 아닌 부동산 시장에 적용한 관련 문헌도 있는데, Venkataraman et al.(2018)은 인도의 4개 지역의 부동산 가격을 구글의 SVI를 바탕으로 예측했으며, 그 결과 부동산 관련 단어의 검색량 증가는 대도시 부동산의 매수세가 강했으며, 동시에 작은 도시의 부동산은 매도세가 강했던 것으로 나타났다[15]. 또한 Beracha et al.(2020)의 연구에서도 주택 구매의 매수세가 강해지기 이전에 특정 지역에 관한 단어의 검색량이 증가한다는 것을 알 수 있었다[16]. 그 밖에 구글의 SVI 데이터를 바탕으로 실업률과 같은 경제적 지표를 예측하는데 실효성이 있다는 연구 결과도 있다[17]. 특정 시계열을 예측하는데 주로 구글의 검색량과 트위터 데이터를 적용한 연구에 비해서 유튜브를 이용한 연구는 극히 부족한데 유튜브에 기록되는 댓글의 NLP(Natural Language Processing) 분석은 감성분석에 실효성이 있다는 연구 결과가 있다[18]. 즉 유튜브의 조회 수, 좋아요 수, 댓글 수 등은 사용자의 감정이 반영된 데이터이며, 이를 시계열 예측에 다양한 측면으로 적용하는 시도는 매우 의미가 있다. Ehliz(2022)는 유튜브의 주식 관련 인플루언서(Influencer) 채널에 대한 데이터가 주식 시장에 미치는 영향을 분석했는데, 장기적인 관점보다는 단기적으로 매우 강한 인과 관계가 있음을 입증했다[19]. 이처럼 유튜브를 비롯한 다양한 인터넷 매체의 데이터는 감성분석의 성능을 높이는데 중요한 자료가 된다. 다만 구글 포털과 트위터와 같은 인터넷 매체보다 대중성이 다소 늦었던 유튜브를 활용한 감성분석 연구는 매우 드물다. 따라서 본 연구에서는 유튜브 데이터를 이용해서 주식 시장에 적용하는 방안과 각각의 인과관계에 대해서 실험하고자 한다.

3. 연구방법

본 장에서는 실험에 필요한 유튜브 데이터 수집을 위한 방법과 그랜저 인과분석에 대해서 설명한다.

3.1 자료 수집

본 연구에서는 유튜브의 데이터를 구글 API(Application Programming Interface)를 이용하여 수집하는데 이처럼 과학적인 방법으로 웹 사이트에서 원하는 정보를 편리하게 추출하는 기술을 스크래핑(Scraping)이라고 한다[20]. 즉 스크래핑은 웹 페이지로부터 데이터를 추출하는 기술을 의미하는데 일반적으로 웹 페이지는 HTML, CSS, JavaScript 등으로 구성되어 있고, 스크래핑은 이러한 웹 페이지를 자동으로 탐색하여 필요한 정보를 추출하는 프로세스를 말한다. 스크래핑은 다양한 프로그래밍 언어와 라이브러리를 사용하여 수행할 수 있으며, 주로 Python의 BeautifulSoup, Scrapy, Selenium

등이 많이 사용된다. 본 연구에서는 Python에서 Google Client API Python 라이브러리를 이용하였고, YoutTube Data API V3을 사용한다.

실험을 위해서 주식 콘텐츠로 등록된 채널 중에서 2020년 1월 1일을 기준으로 구독자 수가 가장 많은 5개의 채널을 선정하고 해당 채널의 조회 수, 좋아요 수, 댓글 수의 데이터를 일별로 수집하는데 값들의 수준이 서로 다르므로 식(1)과 같이 정규화(Normalization) 작업을 수행한다.

$$A = \left(9 * \frac{a - Min(a)}{Max(a) - Min(a)} + 1\right) \tag{1}$$

정규화를 통해 산출된 조회 수, 좋아요 수, 댓글 수는 각각 V, L, C로 표기하며, 이를 식(2)에 적용해서 하나의 값으로 산출한다. 다만 조회 수 데이터가 가장 중요하다는 것을 고려해서 가중치 0.8을 그리고 나머지는 각각 0.1을 부여한다. 이렇게 산출된 값을 본 연구에서 감성지수라고 표현하며, 편의상 YouTube Sentiment Index의 약자를 참조하여 YSI로 표기하는 것을 제안한다.

$$YSI = V * 0.8 + L * 0.1 + C * 0.1 \tag{2}$$

```python
def get_video_stats(video_id):
    youtube = build('youtube', 'v3', developerKey=API_KEY)

    response = youtube.videos().list(
        part='statistics',
        id=video_id
    ).execute()

    if 'items' in response:
        video_data = response['items'][0]['statistics']
        return {
            'views': video_data['viewCount'],
            'likes': video_data['likeCount'],
            'dislikes': video_data['dislikeCount'],
            'comments': video_data['commentCount']
        }
    else:
        print("Failed to get video statistics.")
        return None

if __name__ == "__main__":
    video_id = input("Enter the YouTube video ID: ")
    stats = get_video_stats(video_id)

    if stats:
        print("Video Statistics:")
        print(f"Views: {stats['views']}")
        print(f"Likes: {stats['likes']}")
        print(f"Dislikes: {stats['dislikes']}")
        print(f"Comments: {stats['comments']}")
    else:
        print("Failed to get video statistics.")
```

[그림 2] 유튜브 데이터 수집 예시

[Fig. 2] YouTube Data Collection Example

3.2 그랜저 인과분석

그랜저 인과분석은 변수들 사이에서 서로 인과 관계 여부를 분석할 수 있으며 특히 하나의 변수가 다른 변수 움직임의 원인에 해당하는지 확인할 수 있다[21]. 그랜저 인과분석은 두 변수의 예측력에 해당하는 정보가 시계열 데이터에만 포함된다고 가정하기 때문에 이를 해결하기 위해서 대칭적인 회귀방정식을 정의하는데 이는 식(4)과

같이 F-검정을 사용한다[22]. 식 (6)에서 k는 제약조건이 없는 회귀계수이며, n은 관측치의 수 그리고 q는 제약조건이 있는 회귀계수이다. SSE_R은 a_i=0과 b_i=0의 조건이 성립했을 때를 의미하고 SSE_{UR}은 그 반대이다. 식 (4)와 (5)에서 YSI_t와 X_t는 각각 유튜브 데이터와 주식 관련 데이터인데 X_t에 대해서는 실험계획에서 추가로 설명하고자 한다. 마지막으로 $\boldsymbol{\varepsilon_t}$ 는 상호 독립적인 오차항으로써 등분산을 내포한다. 본 연구에서 실험 변수 간의 인과성 분석을 위해서 다음과 같이 가설을 설계한다.

$$\boldsymbol{YSI_t = a_0 \sum_{i=1}^{m} a_i X_{t-1} + \sum_{i=1}^{n} a_i X_{t-1} + \varepsilon_i} \quad (4)$$

$$\boldsymbol{X_t = b_0 \sum_{i=1}^{m} b_i YSI_{t-1} + \sum_{i=1}^{n} b_i YSI_{t-1} + \varepsilon_i} \quad (5)$$

$$\boldsymbol{F = \frac{(SSE_R - SSE_{UR})/q}{SSE_{UR}/(n-k)}} \quad (6)$$

① $\boldsymbol{\sum a_i \neq 0, \sum b_i = 0}$, 이면, $\boldsymbol{X \rightarrow YIS}$ 인과성 성립

② $\boldsymbol{\sum a_i = 0, \sum b_i \neq 0}$, 이면, $\boldsymbol{YSI \rightarrow X}$ 인과성 성립

③ $\boldsymbol{\sum a_i \neq 0, \sum b_i \neq 0}$, 이면, $\boldsymbol{X \leftrightarrow YSI}$ 인과성 성립

④ $\boldsymbol{\sum a_i = 0, \sum b_i = 0}$, 이면, 서로의 독립성 성립

4. 실험계획 및 분석

본 장에서는 연관분석을 위한 실험계획과 결과 분석 등에 대해서 설명한다.

4.1 실험계획

본 절에서는 앞에서 설명한 내용을 바탕으로 실험계획을 수립한다. 실험에 필요한 유튜브 채널은 금융 및 경제와 관련된 채널 중에서 구독자가 높은 5개의 채널을 선택한다. 각각의 채널에서 2018년 1월부터 2022년 12월까지의 업로드된 영상들 기준으로 조회 수, 좋아요 수, 댓글 수를 크롤링 알고리즘을 통해서 수집한다. 수집된 데이터는 식(2)를 통해서 하나의 일별 시계열 데이터로 만드는데 이는 앞서 설명했듯이 YSI라고 표기한다. 더불어 YSI와 인과 관계 여부를 분석하기 위한 주식 관련 데이터로 KOSPI 지수, KOSDAQ 지수 그리고 각각의 거래량을 수집하며, 편의상 이를 각각 KP_S, KD_S, KP_V, KD_V로 표기한다. 수집된 데이터는 인과분석 과정에서 데이터의 안정성을 검증하기 위해서 ADF 및 PP 단위근 검증을 수행한다. 앞서 설명한 실험계획을 간단하게 정리하면 [표 1]과 같다.

[표 1] 실험계획

[Table 1] Experimental Planning

Category	Explanation
Subject of Experiment	KP_S, KD_S, KP_V, KD_V
Data acquisition Cycle	Daily Data
Experimental Period	2018.01~2022.12
Research Model	Granger Analysis
Analysis Tools	Google API and Python

4.2 단위근 검증 및 시차 선정

인과분석은 서로 간의 인과 관계를 파악하기 위한 분석 방법으로 원인과 결과 사이의 인과적인 연관성을 확인할 수 있다. 인과분석을 하기 위해서는 실험 데이터에 대한 단위근 검증(Unit Root Test)을 분석해야 하며, 이는 시계열 데이터의 안정성 여부를 파악하기 위한 중요한 과정이다. 비정상적인 실험 데이터는 시간 경과에 따라 평균이나 분산이 변하는 비정상적인 패턴을 가지고 있으며, 이러한 비정상적인 데이터를 인과분석에 사용하면 잘못된 결과를 도출할 수 있다. 단위근 검증은 주로 ADF(Augmented Dicky-Fuller) 검정과 PP(Phillips-Perron)와 같은 통계적 방법을 사용한다[23]. 이러한 검증은 시계열 데이터가 정상성을 가지는지 여부를 판단하며, 정상성을 가지지 않는 시계열 데이터는 차분(Differencing)을 통해 정상성을 만족하는 데이터로 변환할 수 있다. 본 연구에서 사용되는 데이터를 분석한 결과 [표 2]와 같이 1차 로그 차분을 통해서 1% 유의수준으로 데이터의 단위근 안정성을 확인할 수 있었다.

[표 2] 단위근 검증 결과

[Table 2] Unit Root Test Results

Time Series		ADP		PP	
		T	P	t	P
Actual Time Series	*YSI*	0.65	0.45	0.69	0.55
	KP_S	0.56	0.24	0.67	0.41
	KD_S	0.45	0.31	0.74	0.38
	KP_V	0.81	0.58	0.78	0.53
	KD_V	0.73	0.42	0.84	0.46
Differential Time Series	*YSI*	-4.75	0.00	-3.49	0.00
	KP_S	-5.16	0.00	-7.74	0.00
	KD_S	-5.07	0.00	-4.97	0.00
	KP_V	-6.76	0.00	-5.48	0.00
	KD_V	-5.46	0.00	-6.45	0.00

더불어 인과분석을 수행하기 위해서 적정 시차(Time Lag)를 산출해야 하는데 본 연구에서는 AIC(Akaike Information Criteria)와 SC(Schwarz Criterion) 기준을 통해 산출한다[24]. 시차 선정은 인과분석을 통해서 적정 시차를 계산할 수 있지만 많은 경우의 수를 고려한다는 단점과 연구의 간결성을 위해서 인과분석 전에 적정 시차를 정의한다.

[표 3] 시차 분석 결과

[Table 3] Time Lag Analysis Results

Time Series	Model	1 Time Lag	2 Time Lag	3 Time Lag	4 Time Lag
YSI-KP_S	AIC	-8.72	-4.48	-17.14	-11.45
	SC	-6.47	-17.87	-12.51	-4.74
YSI-KD_S	AIC	-9.84	-11.42	-17.74	-11.94
	SC	-7.52	-5.71	-14.40	-7.47
YSI-KP_V	AIC	-10.04	-7.52	-9.85	-8.97
	SC	-8.41	-10.20	-10.87	-10.14
YSI-KD_V	AIC	-4.97	-4.75	-17.54	-8.48
	SC	-5.74	-10.46	-10.81	-9.24

[표 3]은 시차 분석의 결과를 정리한 표인데 안정성이 높은 AIC 기준을 적용하며, 우수한 성능을 보이는 3 시차로 결정한다.

4.3 그랜저 인과분석 결과

본 실험은 벡터자기회귀모형(Vector Autoregression)을 통해서 인과분석을 수행하기 위해서 VAR 모형을 설계하고 최적의 시차를 적용한다. Granger 인과분석은 적정 시차 정보인 계수에 대해서만 카이제곱 검정(chi-Squared test)을 수행하는 왈드검정(Modified Wald test)을 통해 계산한다[25]. [표 4]는 각 시계열 간의 Granger 인과분석에 관한 결과를 정리한 표이며, 이는 [표 3]을 통해서 선정된 적정 시차에 대해서만 보여준다. Granger 인과 관계 검증은 각 방향으로의 관련이 없다는 것을 귀무가설로 하고 카이제곱 통계치 옆의 **, *는 각각 1%와 5% 이내의 유의수준으로 귀무가설을 기각한다는 것을 나타내고 있다. 실험 결과를 보면 *YSI*는 KOSPI와 KOSDAQ 시장의 거래량에 강한 선행 관계를 보인다. 특히 *YSI*가 KOSPI 보다는 KOSDQ 지수에 더욱 강한 선행 관계를 보이는데 이는 일반적인 투자자들이 각 시장의 유동성에 차지하는 비중이 KOSPI에 비해서 KOSDAQ 시장에 더 크게 작용하기 때문이라고 판단된다. 예컨대 KOSPI 시장의 경우에는 외국의 대형 사모펀드 기업과 다양한 기관들이 투자하는 비중이 KOSDAQ 보다는 훨씬 크다. 이러한 관점은 *KDV→YSI*가 높은 인과성을 보이는 것에도 비슷한 이유가 될 것으로 판단된다.

[표 4] 실험결과

[Table 4] Experiment Result

Direction	Chi-Square	PV	Time Lag
$YSI \rightarrow KP_S$	214.17**	0.00	3
$YSI \rightarrow KD_S$	187.55**	0.00	3
$YSI \rightarrow KP_V$	387.14**	0.00	3
$YSI \rightarrow KD_V$	401.61**	0.00	3
$KP_S \rightarrow YSI$	67.11	0.00	3
$KD_S \rightarrow YSI$	17.46	0.01	3
$KP_V \rightarrow YSI$	47.94	0.01	3
$KD_V \rightarrow YSI$	145.57*	0.00	3

본 실험을 통해서 대중적으로 인기가 높은 유튜브는 주식 시장에 미치는 영향이 있다는 것을 입증했다. 또한 이를 통해서 금융 시장을 분석하는 전통적인 방법론과 함께 시장 참여자들의 대중적 심리를 객관적으로 분석하는 것이 무엇보다 중요하다는 것을 알 수 있었다.

5. 결론

본 연구는 가장 대중적으로 사용도가 높은 유튜브 데이터가 금융 시장의 참여자들의 심리적 영향과 밀접한 관계가 있을 것이라는 문제를 객관적으로 분석하기 위해서 관련 데이터를 바탕으로 인과분석을 수행했다. 이를 위해 가장 인기가 높은 유튜브 채널의 데이터를 크롤링 알고리즘을 통해서 수집하고, 이를 KOSPI와 KOSDAQ 시장 데이터와 인과성을 분석했다. 그 결과 유튜브의 조회 수와 같은 데이터는 금융 시장을 선행한다는 것을 통계적으로 입증했으며, 이는 대중의 감성분석에 있어서 유튜브 데이터가 실효성이

있다는 것을 의미한다. 특히 유튜브 데이터는 주가에 비해서 거래량에 더욱 유의미한 결과를 보였다.

이미 전통적인 경제 심리학 이론을 근거로 포털 사이트의 검색 트랜드 정보가 금융 자산에 선행적 인과 관계를 보인다는 것은 많은 선행연구로 입증이 되었지만, 최근 포털 사이트의 사용량을 넘어선 유튜브의 수치적 정보를 활용한 연구는 다소 미약한 편이다. 따라서 본 연구의 실험 결과는 향후 후행 될 감성분석에 관한 연구에 도움이 될 것으로 사료된다. 본 연구에서는 유튜브 데이터를 활용한 주가 선행성 분석의 유용성을 입증하였지만, 데이터의 한계와 불확실성을 인지해야 한다. 향후 연구에서는 더 다양한 SNS 플랫폼 데이터와의 통합, 더 정교한 감성 분석 모델, 그리고 인공지능과의 결합을 통해 보다 정확한 주가 예측 모델을 연구할 예정이다.

References

[1] B. M. Henrique, V. A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, Expert Systems with Applications, (2019), Vol.124, No.15, pp.226-251.
DOI: https://doi.org/10.1016/j.eswa.2019.01.012

[2] L Salim, Entropy-Based Technical Analysis Indicators Selection for International Stock Markets Fluctuations Prediction Using Support Vector Machines, Fluctuation and Noise Letters, (2014), Vol.13, No.2, 1450013.
DOI: https://doi.org/10.1142/S0219477514500138

[3] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of Computational Science, (2011), Vol.2, No.1, pp.1-8.
DOI: https://doi.org/10.1016/j.jocs.2010.12.007

[4] P. Sanna, YouTube:Audience emotional reactions and convergent alignment, Internet Pragmatics, (2022), Vol.6, No.1, pp.42-66.
DOI: https://doi.org/10.1075/ip.00085.pel

[5] R. G. Hwang, Effect of The Thumbnail Service Type on Continuous Usage Intention of The OTT Platform : Focusing on YouTube Users, Journal of Korea Entertainment Industry Association, (2022), Vol.16, No.3, pp.13-27.
DOI: https://10.21184/jkeia.2022.4.16.3.13

[6] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowledge and Information Systems, (2019), Vol.60, pp.617-663.
DOI: https://doi.org/10.1007/s10115-018-1236-4

[7] E. Cambria, D. Das, S. Bandyopadhyay, A. Feraco, Affective Computing and Sentiment Analysis, Socio-Affective Computing, (2017), Vol.5.
DOI: https://doi.org/10.1007/978-3-319-55394-8_1

[8] D.Michael, Google search-based metrics, policy-related uncertainty and macroeconomic conditions, Applied Economics Letters, (2015), Vol.22, No.10, pp.801-807.
DOI: https://doi.org/10.1080/13504851.2014.978070

[9] M. Y. Huang, R. R. Rojas, P. D. Convery, Forecasting stock market movements using Google Trend searches, Empirical Economics, (2020), Vol.59, pp.2821-2839.
DOI: https://doi.org/10.1007/s00181-019-01725-1

[10] M. Bank, M. Larch, G. Peter, Google search volume and its influence on liquidity and returns of German stocks, Financial Markets and Portfolio Management, (2011), Vol.25, pp.239-264.
DOI: https://doi.org/10.1007/s11408-011-0165-y

[11] Y. Liu, Y. Chen, S. Wu, G. Peng, B. Lv, Composite leading search index: a preprocessing method of internet search data for stock trends prediction, Annals of Operations Research, (2015), Vol.234, pp.77-94.

DOI: https://doi.org/10.1007/s10479-014-1779-z

[12] G. P. Smith, Google Internet search activity and volatility prediction in the market for foreign currency, Finance Research Letters, (2012), Vol.9, No.2, pp.103-110.
DOI: https://doi.org/10.1016/j.frl.2012.03.003

[13] M. S. Checkley, D. A. Higon, H. Alles, The hasty wisdom of the mob: How market sentiment predicts stock market behavior, Expert Systems with Applications, (2017), Vol.77, pp.256-263.
DOI: https://doi.org/10.1016/j.eswa.2017.01.029

[14] X. Zhang, H. Fuehres, P. A. Gloor, Predicting Stock Market Indicators Through Twitter “I hope it is not as bad as I fear”, Procedia - Social and Behavioral Sciences, (2011), Vol.26, pp.55-62.
DOI: https://doi.org/10.1016/j.sbspro.2011.10.562

[15] M. Venkataraman,V. Panchapagesan, E. Jalan, Does internet search intensity predict house prices in emerging markets? A case of India, Property Management, (2018), Vol.36, No.1, pp.103-118.
DOI: https://doi.org/10.1108/PM-01-2017-0003

[16] E. Beracha, M. B. Wintoki, Forecasting Residential Real Estate Price Changes from Online Search Activity, Journal of Real Estate Research, (2013), Vol.35, No.3, pp.283-312.
DOI: https://doi.org/10.1080/10835547.2013.12091364

[17] S. Mihaela, Improving unemployment rate forecasts at regional level in Romania using Google Trends, Technological Forecasting and Social Change, (2020), Vol.155.
DOI: https://doi.org/10.1016/j.techfore.2020.120026

[18] H. Jelodar, Y. Wang, M. Rabbani, S. B. B. Ahmadi, L. Boukela, R. Zhao, R. S. A Larik, A NLP framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on youtube comments, Multimedia Tools and Applications, (2020), Vol.80, pp.4155-4181.
DOI: https://doi.org/10.1007/s11042-020-09755-z

[19] M. Ehliz, Prediction Accuracy of YouTube Influencers Measured Against Subscriber Counts, Journal of Student Research, (2022), Vol.11, No.3.
DOI: https://doi.org/10.47611/jsrhs.v11i3.3332

[20] M. Airoldi, D. Beraldo, A. Gandini, Follow the algorithm: An exploratory investigation of music on YouTube, Poetics, (2016), Vol.57, pp.1-13.
DOI: https://doi.org/10.1016/j.poetic.2016.05.001

[21] A. Shojaie, E. B. Fox, Granger Causality: A Review and Recent Advances, Annual Review of Statistics and Its Application, (2022), Vol.9, pp.289-319.
DOI: https://doi.org/10.1146/annurev-statistics-040120-010930

[22] W. Enders, Improved critical values for the Enders-Granger unit-root test, Applied Economics Letters, (2001), Vol.8, No.4, pp.257-261.
DOI: https://doi.org/10.1080/135048501750104033

[23] Z. Adali, A. Toygae, U. Yildirim, Assessing the stochastic behavior of fishing grounds footprint of top ten fishing countries, Regional Studies in Marine Science, (2023), Vol.63, 103015.
DOI: https://doi.org/10.1016/j.rsma.2023.103015

[24] S. Tang, Measurement and Management of Interest Rate Risk of Commercial Banks: Based on VaR-GARCH Model of a Case Study of SHIBOR, Scientific and Social Research, (2022), Vol.4, No.1.
DOI: https://doi.org/10.36922/ssr.v4i1.1318

[25] J. H. Kim, C. H. AnA Study on Estimation and Prediction of Vector Time Series Model Using Financial Big Data, Turkish Journal of Computer and Mathematics Education, (2021), Vol.12, No.5.
DOI: https://doi.org/10.17762/turcomat.v12i5.951