SNS 여론과 주가지수의 상관관계 분석

김현지*[1], 오성주*[2]
[1]서강대학교 경영학과 학부생
[2]서강대학교 경제학과 석사과정

mescsful@gmail.com, dhtjdwn98@naver.com
*본 저자는 모두 공동 제1저자로서, 저자들 간 연구에 대한 기여도는 동일함.

Correlation Analysis Between Online Public Opinion and Stock Price

Hyun-Ji Kim*[1], Sung-Ju Oh*[2]
[1]Dept. of Business, Sogang University
[2]Dept. of Economics, Sogang University
*These authors contributed equally to this work.

요 약

"이성적이며 이상적인 합리적 인간"을 가정하는 기존 경제학의 이론이 항상 실제 상황과 일치하지는 않는 것으로 알려져 있다. 이의 대안으로 나온 행동경제학은, 인간의 경제적 의사결정에 심리, 인지, 감정, 사회문화적 배경 등이 영향을 미친다고 본다. 본 연구에서는 행동경제학에 의거하여, 개인의 감정과 경험이 경제적 의사결정에 영향을 미치는지 여부를 빅데이터 모델을 활용하여 분석하였다. SNS 여론으로는 Reddit, 주가지수로는 S&P 500 을 선정하였다. 수집한 텍스트 데이터를 전처리와 감정분석을 통해 독립변수 값으로 사용했고, 주가지수 등락의 방향성을 종속변수로 사용하여 로지스틱 모형을 구성했다. 모델을 활용하여 분석한 결과 Public sentiment 와 Market sentiment 간 양의 상관관계를 확인할 수 있었다. 또한, lag 를 설정하는 모델이 정확도가 더욱 높음을 확인해, 기존 경제학의 EMH 와 대립되는 바를 확인할 수 있었다. 하지만 최적의 lag 산정을 위해, 더 광범위한 데이터를 바탕으로 한 후속연구가 필요하다.

1. 서론

1.1 주제 선정 이유

행동경제학은 기존 경제학의 대안으로 나온 경제학으로, "경제심리학"으로도 불리우며 실제적인 인간의 행동이 의사결정에 미치는 영향을 규명하기 위한 학문이다. 행동경제학에서는 인간이 기존 EMH 의 "이상적이며 합리적인 경제주체"가 아닌, 감성적인 측면 즉 심리적인 요인에 의해서도 많은 영향을 받는 존재라고 본다. 따라서 행동경제학의 측면에서, 개인의 감정과 경험(SNS 여론)이 의사결정에 영향을 미치는지(주가등락)을 확인해보기 위해 본 연구를 기획하였다.

1.2 분석대상

게임스톱 집중매수 사태에서 개인투자자들의 영향력을 확인할 수 있던 Reddit 을 SNS 여론의 분석대상으로 선택하고, 시장전체에 대한 대표성을 확보하기 위해 S&P 500 을 주가지수 분석 대상으로 선택하였다.

2. 본문

2.1 데이터 수집 및 처리 과정

2.1.1 데이터셋 구축

praw 라는 API 를 사용하여 "Stock"이름의 subreddit 의 1 년치 글을 "Top"기준(Score 와 Comments 가 많은 순)으로 총 999 개의 자료를 스크래핑하여 데이터로 변환하였다.

2.1.2 전처리

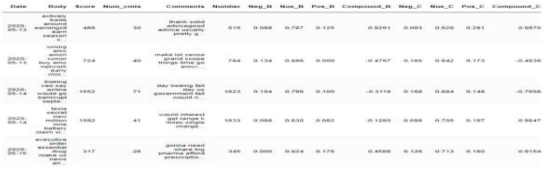
자연어 감정분석을 위해 불필요한 단어 삭제 및 변형된 단어를 명확한 형태로 변환하는 전처리과정을 거쳤다. 첫째 enter, tap, URLs, HTMLs, emoticons 등과 불용어를 제거하였다. 둘째, slang, Acronyms, abbreviations 등을 적절한 의미의 단어로 반환하고 contractions 과 apostrophe 로 축약된 단어를 원형태로 복구하였다. 셋째, 모든 글자를 소문자로 치환하고 알파벳을 제외한 모든 단어를 삭제하였다.

전처리 이전	전처리 이후
'Tesla₩'s secret new ₩'million mile₩' battery 2020/05/14/tesla-secret-electric-car-battery-m ans to introduce a new low-cost, long-life bat xt that it expects will bring the cost of elec ries to have second and third lives in the ele n Musk has been teasing investors, and rivals, ology during a "Battery Day" in late May.'	'tesla s secret new million mile battery cla troduce new low cost long life battery model ehicles line gasoline model allow ev batteri utive elon musk tease investors rival promis may '

<그림 1> 전처리 과정 예시

2.1.3 감정분석

Vader 라이브러리를 사용하여 negative, Neutral, Positive 감정의 각 값을 구해 -1 에서 1 사이 compound 값을 구했다. 단, Reddit 의 게시글마다 스코어나 댓글 수가 차이나기 때문에 그 영향력이 다르다. 따라서 각 게시글의 댓글+스코어의 개수를 'Number'라는 변수로 만들었고, 이를 가중치로 활용해 compound 를 평균하여 Sentiment_score 라는 변수를 만들어 활용했다.



<그림 2>감정 분석 결과 예시

2.2 가설 및 모델 설정

2.2.1 가설설정

첫째, Public Sentiment 가 Market Sentiment 에 영향을 주며, 양의 상관관계를 가진다. 둘째, 정보는 시장에 즉각적으로 반영되지 않는다.

2.2.2 모델설정

독립변수에 따른 종속변수의 값을 0 과 1 이진법으로 반환하는 로지스틱 모형을 사용하였다. 주가 자체의 예측보다는 여론에 따른 주가 등락의 연관성을 확인하는 것이 타당하다고 판단했기 때문이다. 로지스틱 모형을 활용한 모델 1,2,3 과 OLS 모형을 비교하여 로지스틱 모형의 타당성을 검증하였다. 모델 1 과 2 는 각각 독립변수가 body 와 comments 의 감정분석 평균값과 각각의 감정분석 값이고, lag 가 0 이고 종속변수가 up=1, down=0 인 모델이다. 모델 3 은 모델 1 과 같되 lag=1 인 모델이며, OLS 는 각각의 감정분석 값을 독립변수로 하며 change 값을 종속변수로 한다.

2.3 모델검증

sklearn 의 model selection, preprocessing, linear model 라이브러리를 이용하여 분석하였으며, 전체 데이터를 75% 크기의 train set 과 25% 크기의 test set 으로 나눠주었다. 그 결과에 대해 정규화 스케일링 이후 train 과 test set 각각을 로지스틱 회귀분석하여 score 와 coefficient 를 구했다. 마지막으로 최적의 규제정도와 규제 방법을 찾은 후 test set 의 score 와 coefficient 값을 구하며 어떠한 모델이 가장 적합한지 알아보기 위해 모델들을 비교하였다.

	Independent variable	Train score	Test score	Coefficient	규제방법/정도 변경	
					Score	Coefficient
Model1	Avg_0	0.616	0.638	0.303	X	
Model2	Body_0, cmts_0	0.605	0.638	0.104 & 0.309	0.621	0.09 & 0.254
Model3	Avg_1	0.579	0.603	0.291	0.585	0.240
Linear regression	Body_0, cmts_0	0.087	-0.102	0.003 & 0.006	X	

<그림 3> 로지스틱 및 OLS 분석결과

sklearn 의 model selection, preprocessing, linear model OLS 의 train 과 test 모두 스코어가 눈에 띄게 작고, 심지어 test score 가 음의 값이 나왔는데, 이것은 모델 자체가 아예 잘못 설정되었다는 것을 의미한다. 따라서 OLS 보다 로지스틱 모형이 더욱 적절함을 알 수 있다.

2.4 Lag 설정 및 분석

단, 앞선 결과의 모델 1~3 에서 lag 를 설정할수록 test score 가 작아졌는데, 이는 통상적인 의견과 대치된다. 금융데이터에서 Lag 는 장마감을 구분해주는 역할을 하기 때문이다. 따라서 시간적 결측치를 줄이기 위해 Year 데이터를 Month 데이터로 바꾸어 각각 lag=0, lag=1 인 모델을 만들어 다시 검토하였다.

	Independent variable	Train score	Test score	Coefficient	규제방법/정도 변경	
					Score	Coefficient
Model0(1)	Original_0	0.667	0.5	-0.349	0.5	-0.02
Model0(2)	Origianl_1	0.8	0.0	-0.744	0.2	-0.03
Model1	Avg_0	0.867	0.667	1.338	X	
Model2	Avg_1	0.867	1.0	1.234	1.0	1.986

<그림 4>추가모델 분석결과

분석결과 model 2, month 데이터와 Lag=1 의 모델이 가장 우수한 값을 보였다.

3. 결론

Public Sentiment 와 Market Sentiment 간 양의 상관관계를 찾을 수 있으며, 로지스틱 모형이 분석에 적합하다는 것을 알 수 있다. 또한 lag=1 의 모델이 가장 정확도가 높으므로 즉각적인 정보반영(lag=0)을 주장하는 EMH 와는 상반되는 내용을 보였다. 하지만 절대적인 데이터셋의 양적 부족 및 비관련 데이터로 인한 모델의 한계가 있다.

이 논문은 서강대학교 학부 금융시장행태주의(Behavioral)분석과정 캡스톤디자인 프로젝트의 일환으로 작성되었음

참고문헌

[1] A Mitta, A Goel: Stock Prediction Using Twitter Sentiment Analysis (Standford University, CS229)

[2] VS Pagolu, KN Reddy, G Panda: Sentiment Analysis of Twitter Data for Predicting Stock Market Movements (IEEE 2016)

[3] Hutto,C., & Gilbert,E.VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text,AAAI,2014