트위터 감정 분석과 한미 비트코인 시장에 미치는 영향 비교분석

Comparing Twitter Sentiments Impact On the Korean and US Bitcoin Markets

저자 (Authors)
이석원, 김래현, 강재우
Seok-Won Yi, Raehyun Kim, Jaewoo Kang

출처 (Source)
한국정보과학회 학술발표논문집 , 2019.6, 1920-1922(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08763708

APA Style
이석원, 김래현, 강재우 (2019). 트위터 감정 분석과 한미 비트코인 시장에 미치는 영향 비교분석. 한국정보과학회 학술발표논문집 , 1920-1922

트위터 감정 분석과 한·미 비트코인 시장에 미치는 영향 비교·분석*

이석원[O1], 김래현[2], 강재우[2,3]
고려대학교 국제학부[1]; 고려대학교 대학원 컴퓨터·전파통신공학과[2];
고려대학교 대학원 바이오협동과정[3]
{seankala, raehyun, kangj}@korea.ac.kr

Comparing Twitter Sentiment's Impact On the Korean and US Bitcoin Markets

Seok-Won Yi[O1], Raehyun Kim[2], Jaewoo Kang[2,3]
Division of International Studies, Korea University[1];
Department of Computer Science and Engineering, Korea University[2];
Interdisciplinary Graduate Program in Bioinformatics, Korea University[3]

요 약

시장의 움직임을 예측하려는 연구는 꾸준한 관심을 받아왔고 비트코인을 비롯한 암호화폐도 시장의 신참자로서 최근에 뜨거운 관심을 받고 있다. 2017년 암호화폐 열풍을 비롯해 대한민국이 전 세계의 주목을 받는 가운데 "김치 프리미엄"과 같은 현상을 보고 한국 시장은 감정에 의존한다는 이야기가 만연했다. 본 논문은 트위터 감정분석을 통해 과연 대한민국 암호화폐 시장이 미국 시장보다 감정의 영향을 더 받는지 탐색해보려 시도했으나 트위터 감정과 시장의 상관관계가 한·미 양쪽 시장에서 거의 같게 나와 대한민국 암호화폐 시장은 다른 시장보다 감정의 영향을 더 받지 않는다고 결론을 짓는다.

1. 서 론

한국의 암호화폐 시장은 2017년 전 세계 시장보다 높게는 50% 더 높은 가격으로 암호화폐가 거래되기도 했는데 사람들은 이를 "김치 프리미엄(Kimchi Premium)"이라고 일컬었다. 이런 현상이 발생하는 이유를 대한민국 시장의 특성에서 찾는 이들도 있었는데 이들이 이야기하기를 "한국의 시장은 다른 시장보다 감정에 의존하며 그로인해 이런 움직임을 보이는 것"이라고 했다.

본 연구에서는 그 주장이 일리가 있는지 없는지를 통계학적으로 검증해보려 한다. 그 방법으로는 트위터 데이터에 감정분석(sentiment analysis)을 적용 시키고 한국과 미국을 각자 대표하는 암호화폐 거래소인 코인베이스(Coinbase)와 코빗(Korbit)의 비트코인 가격 데이터를 이용하여 상관계수를 구하고 그 상관계수가 통계적으로 유의성(significance)이 있는지 *t*-test를 통해 검증한다.

2. 관련 연구

관련 연구는 크게 정통증시와 암호화폐 시장으로 나눌 수 있다. 일반적으로 감정분석이 정통증시에 미치는 영향에 관한 연구가 가장 많았으며 암호화폐 시장에 관한 연구도 있지만 상대적으로 그 양이 적은 편이고 놀랍게도 한국 시장에 초점이 맞춰진 연구는 찾을 수 없었다.

감정분석을 이용해 시장을 예측하려는 연구는 보통 효율적 시장가설(efficient market hypothesis – EMH)을 반박하는 차원에서 쓰인다.

정의: 효율적 시장가설

효율적 시장 가설(이하 EMH)은 1960년대에 미(美) 경제학자 유진 파마(Eugene Fama)에 의해 정의된 가설로, 시장은 현 시점에 있는 모든 정보를 참고해 가장 효율적으로 움직이며 시장을 예측하는 것은 불가능하며 시장은 랜덤 워크(random walk) 형태로 움직인다고 정의한다.

[1]의 경우 트위터 데이터와 다우존스산업평균지수 (Dow Jones Industrial Average) 데이터를 가지고 트위터 감정이 주식시장에 끼치는 영향을 그랜저 인과관계(Granger causality analysis)와 self-organizing fuzzy neural network(SOFNN)을 이용해 분석했다. 그 결과 단순히 기존 주식 데이터를 이용했을 때보다 감정 데이터를 대입하면 예측 오차를 6%까지 줄일 수 있다고 주장한다. 그렇지만 트위터 데이터는 대부분 영어로 구성됐다는 점, 그리고 "감정 데이터"에 관한 실젯값(ground truth)을 알 수 없다는 점 등과 같은 보완점이 아직 존재한다고 한다.

그 뒤에 이어지는 [2]와 [3]도 기술적인 방식이나 접근법에 약간의 변경을 주었을 뿐 [1]과 크게 다르지 않다. [2]의 경우 *k*-fold 교차 검증법을 이용했다는 점, [3]의 경우는 단순히 트위

* 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2017R1A2A1A17069645, NRF-2017M3C4A7065887)

터 감정분석뿐만 아니라 트위터 데이터의 볼륨(한 시점에 게시되는 트윗의 수)까지 감안 했다는 점에서 다르다.

[4]는 앞선 연구와는 다소 다른데 그가 주장하는 바를 한 문장으로 요약하면 "상관관계가 있다고 인과관계가 성립되는 것은 아니다" 이다. 이들이 내리는 결론은 감정이 시장에 영향을 끼치는지 혹은 그 역인지 알 수 없다는 것이다.

[5]와 [6]은 감정분석을 이용해 실제로 암호화폐에 적용한 연구 사례이다. 안타깝게도 이들이 내린 결론 또한 암호화폐 시장은 시장의 특성상 일반증시보다 높은 주가변동성(volatility)을 보이는데 단순 감정분석 가지고는 효율적으로 시장을 예측하지 못한다고 결론짓는다.

3. 감정분석 도구 및 데이터

본 연구에서 쓰인 감정분석 도구는 TextBlob과 NLTK이다. 데이터세트는 크게 두 트위터 데이터와 비트코인 데이터로 나눌 수 있으며 총 12개의 데이터세트가 쓰였다.

3.1 감정분석 도구

TextBlob은 MIT 라이센스 하에 운영되고 있는 오픈소스 Python 라이브러리이며 NLTK는 자연어처리 기술을 사용하는 여러 라이브러리를 종합한 대표적인 라이브러리이다.

3.1.1 TextBlob

TextBlob은 문자열 (여기서는 트윗)을 입력값으로 해당 문자열에 대해

(polarity, subjectivity)

형태의 tuple을 출력한다. 여기서 polarity는 [-1, 1] 사이의 값인데 1에 가까울수록 긍정적인 감정이 강한 것이고 subjectivity는 말 그대로 해당 문자열이 얼마나 주관적인지를 알려준다. subjectivity는 [0, 1] 사이의 값을 가지며 1에 가까울수록 주관성이 강한 트윗이다.

3.1.2 NLTK Vader

NLTK Vader는 TextBlob처럼 감정 수치를 [-1, 1] 사이의 값으로 출력하는 점에서는 비슷하나 차이점은 단순히 (polarity, subjectivity)가 아니라

(negative, neutral, positive, compound)

처럼 총 네 가지의 감정 분류를 출력한다. 여기서 compound는 앞선 세 가지 수치들의 평균값이다. [-1, 1] 사이의 수치를 해석하는 원리는 TextBlob과 동일하다.

3.2 데이터

앞서 언급된 것처럼 본 연구에서 쓰인 데이터세트는 모두 코인베이스와 코빗 각 거래소별로 여섯 개, 감정 데이터와 합치면 총 12개의 데이터세트가 쓰였다. 트위터와 비트코인 가격 데이터는 2017년 5월 10일부터 2019년 1월 7일 총 608일간 데이터를 사용했다. 데이터를 수집하고 처리하는 과정을 다음과 같이 요약할 수 있다:

데이터 처리 과정

1. 트위터 데이터와 거래소별 비트코인 데이터 수집하고 기본적인 전처리 과정을 시행한다.
2. 각 트윗 감정분석 데이터를 거래소 데이터와 합친다.
3. 2.에서의 데이터에 시간지연을 적용시킨다. 여기서 각 거래소별로 세 개의 데이터세트가 만들어지는데 하나는 시간지연을 적용 안 시킨 세트, 하나는 트위터 감정을 하루 앞당긴 세트 (Lag -1), 하나는 뒤로 하루 지연시킨 세트이다 (Lag +1).
4. 거래소별 여섯 개의 데이터세트 (2개의 감정분석 × 3개의 시간지연 방식)가 만들어지고 거래소가 두 개인 점을 감안하면 총 12개의 데이터세트가 존재한다.

4. 상관관계 계산·분석 및 통계검정

본 연구에서는 코인베이스와 코빗 데이터 간 피어슨, 켄달, 그리고 스피어만 상관관계(각 Pearson, Kendall, Spearman)를 이용해 상관관계를 구하며 최초에 세웠던 가설대로 한국 시장이 미국보다 더 높은 상관관계를 보이는지 탐색한 후 통계검정(statistical testing)을 통해 그 유의성(significance)을 검증한다.

4.1 상관관계 계산 및 분석

그림 1과 그림 2는 코인베이스와 코빗 데이터에 TextBlob 감정분석을 접목시킨 데이터세트의 피어슨 상관계수 행렬을 계산하고 heat map으로 표현한 것이다:

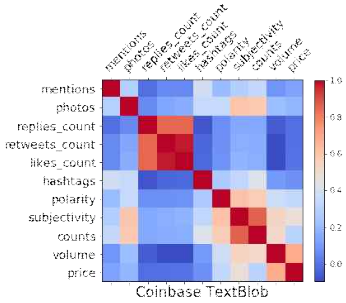

그림 1 코인베이스 · TextBlob 피어슨 상관계수 행렬

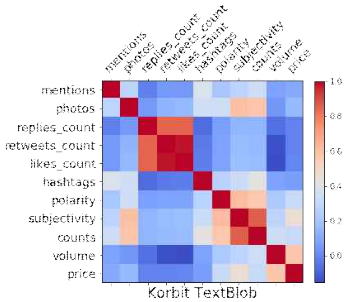

그림 2 코빗 · TextBlob 피어슨 상관계수 행렬

두 그림에서 본 논문은 마지막 열, 즉, price에 초점을 맞춘다. 코빗에서 price와 감정을 나타내는 polarity 간의 상관관계가 높게 나오기를 기대했으나 보이는 결과는 이와 다를 뿐만 아니라 양쪽 거래소에서 모두 상관계수 값이 낮게 나와 처음에

세운 가설이랑 오히려 반대의 결과가 나온다.

상관계수 계산법 다양화, 감정분석 도구 중 Vader 추가, 트위터 데이터와 비트코인 데이터 간 시간지연 적용 등과 같은 접근법을 추가해 price와 polarity (TextBlob), price와 compound (Vader) 간 상관계수를 계산한 결과를 표 1 에 정리한다.

표 1 상관계수 값 및 차잇값

Tool	Lag	Corr	Coinbase	Korbit	Difference
TB	No lag	Pearson	0.334407	0.325896	0.008511
		Kendall	0.200962	0.200988	0.000026
		Spearman	0.296076	0.296015	0.000061
	Lag +1	Pearson	0.339047	0.331259	0.007787
		Kendall	0.202382	0.201974	0.000407
		Spearman	0.298124	0.296848	0.001276
	Lag -1	Pearson	0.338796	0.329257	0.009540
		Kendall	0.204701	0.204500	0.000201
		Spearman	0.300716	0.301836	0.001119
NLTK	No lag	Pearson	0.298266	0.305032	0.006766
		Kendall	0.176093	0.174038	0.002056
		Spearman	0.261459	0.258722	0.002737
	Lag +1	Pearson	0.302851	0.313087	0.010236
		Kendall	0.179963	0.176368	0.003595
		Spearman	0.267099	0.263005	0.004095
	Lag -1	Pearson	0.294363	0.302763	0.008400
		Kendall	0.172333	0.171296	0.001037
		Spearman	0.255832	0.254950	0.000881

표를 보면 차이(difference)의 평균이 약 0.0038로서 양쪽 거래소 간 차이가 크지 않다는 것을 확인할 수 있으나 이를 통계검정을 통해 그 유의성을 입증한다.

4.2 통계검정을 통한 significance 검증

본 논문에서 설정한 귀무가설(null hypothesis)은 "코인베이스와 코빗은 트위터 감정과의 상관관계를 측정하면 같게 혹은 유사하게 나올 것"이다. 해당 귀무가설을 *t*-test로부터 얻어지는 *p*-value를 통해 기각하는데 실패 혹은 성공하는지 보인다.

정의: *p*-value

일반적으로 두 변수 간 *t*-검증을 실행하고 *p*-value를 분석할 때 다음과 같은 세 분류로 나뉜다:

1. $p > 0.5$: 두 변수 간 유의성이 거의 없음. (귀무가설 기각 실패)
2. $p \leq 0.5$: 두 변수 간 유의성이 존재함. (귀무가설 기각 성공)
3. $p \approx 0.5$: 유의성의 존재가 불명확함.

t-test를 코인베이스와 코빗 양 데이터에 적용시켜서 계산하면 $p = 0.9637$의 값을 구할 수 있다. 0.5보다 상당히 높은 값이 나왔는데 위에 *p*-value의 정의를 고려하면 이는 코인베이스와 코빗 데이터 두 변수가 상호 유의성을 거의 안 띄며 앞서 세운 귀무가설, 즉, 두 변수 간 상관계수 차이가 크지 않을 것이라는 가설이 성립한다는 결론을 내릴 수 있다.

5. 결론 및 향후 연구

본 연구가 최초에 세운 가설은 "한국의 비트코인 시장은 미국보다 SNS 감정의 영향을 더 받을 것"이였다. 그 가설을 입증하기 위해 감정분석, 상관계수 계산 및 통계검정을 시도하였지만 아쉽게도 결과는 가설을 입증하지 못하고 오히려 반대로 나왔다.

본 논문에서 세운 가설이 논문의 범위 안에서 입증이 안 되었지만 여기서 고려해야할 부분이 몇 가지 있다:

1. 연구에 사용된 데이터의 양이 상대적으로 적다. 앞서 언급했던 관련 연구에서는 평균적으로 몇 십 혹은 몇 백만 개의 트윗 ([1]의 경우는 천만 트윗 가까이 사용)을 사용했다는 점을 생각하면 본 연구에서 사용한 약 20만 개의 트윗은 상대적으로 너무 적다.
2. 트위터 데이터는 기본적으로 영어이고 모국어가 영어가 아닌 한국에서 그 영향을 측정하기에는 부족함이 있다. 트위터가 한국에서도 널리 사용된다고 해도 미국 시장만큼의 영향을 받을지는 의문이 든다.

흥미로운 부분은 2번을 감안했을 때 코인베이스에서의 상관계수 값이 코빗보다 눈에 띄게 높게 나와야하는 것이 상식에 맞는데 그렇지 않은 것을 보면 암호화폐 시장은 단순히 SNS 감정뿐만 아니라 다른 작용요인들이 있다는 것이라고 생각해볼 수 있다. 이를 밝히는 것이 차후 연구 과제이다.

참고문헌

[1] Bollen et al., "Twitter Mood Predicts the Stock Market," in *Journal of Computational Science*, vol. 2, issue 1, 1-8, 2011

[2] Anshul Mittal and Arpit Goel, "Stock Prediction Using Twitter Sentiment Analysis," 2011

[3] Ranco et al., "The Effects of Twitter Sentiment on Stock Price Returns," in *PLOS ONE*, vol. 10, issue 9, 2015

[4] Andrius Mudinas, Dell Zhang, and Mark Levene, "Market Prediction Using Sentiment Analysis: Lessons Learned and Paths Forward," 2019

[5] Galeshchuk et al., "Bitcoin Response to Twitter Sentiments," in *ICTERI Workshops 2018*, 2018

[6] Connor Lamon, Eric Neilson, Eric Redondo, "Cryptocurrency Price Prediction Using News and Social Media Sentiment," 2017