디지털자산 사회정서 지수
(Haksoon Index)

서울대학교 건설환경공학부 김학순

1. 개요

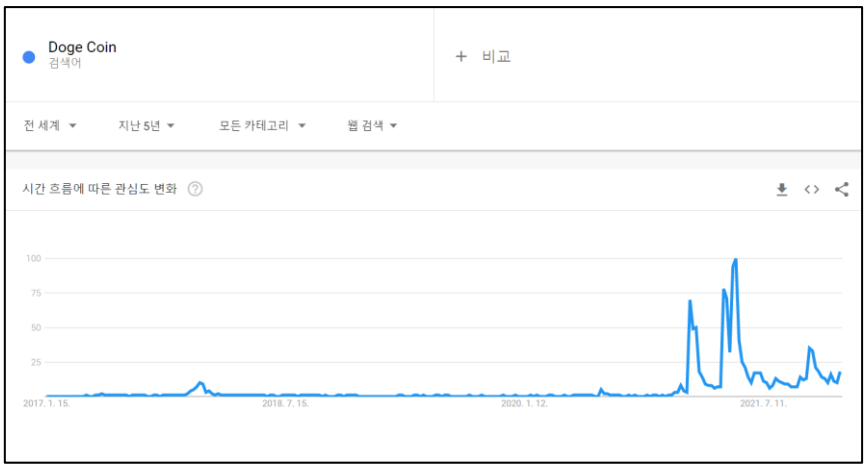
암호화폐 시장은 종종 주식시장과 비교되지만 암호화폐의 둘 다 가격의 등락이 있고 사람들의 심리가 차트에 영향을 미친다는 점은 공통적이다. 하지만 암호화폐의 내재적 가치가 있는지에 대해서는 여전히 논란이 있고 주식에 비해 참고할 지표가 현저히 적다. 그래서 이 프로젝트는 감성분석 기법을 적용해서 Social Sentiment 를 반영한 암호화폐 지수(index)를 고안하는 것이 목표이다.

2. 방법론

2.1. 데이터 수집

지수 산출에 반영될 암호화폐의 종류를 도지 코인(DOGE)과 바이낸스 코인(BNB) 두 가지로 한정했다. 비트코인과 이더리움과 같이 시가총액이 큰 암호화폐는 가격의 변동이 작고 너무 많은 요인으로부터 영향을 받기 때문에 제외했고 상대적으로 볼륨이 작은 마이너 알트코인도 제외했다. 결국 일론 머스크의 트윗 등으로 가격이 심하게 변동하고 뉴스에도 자주 등장했던 대표적인 밈 코인인 도지 코인을 선정하고 세계 최대의 암호화폐 거래소인 바이낸스에서 발행하여 시가총액 기준 현재 4 위까지 급성장한 알트코인인 바이낸스 코인도 포함하여 두 가지의 암호화폐로 한정하여 지수를 고안하는 것으로 결정했다.

데이터는 검색량, 거래량, 트위터 이렇게 세 가지를 활용하는 것으로 정했다. 검색량은 구글 트렌드에서 Binance Coin 과 Doge Coin 검색어의 지난 5 년 간 전세계의 관심도 변화를 0-100 으로 나타낸 시계열 데이터로부터 얻을 수 있었다.



[구글트렌드 검색량 데이터]

거래량은 앞서 언급한 암호화폐 거래소 바이낸스에서 제공하는 API 를 사용하여 도지 코인과 바이낸스 코인의 발행 이래 일일 거래 데이터로부터 확보할 수 있었다.

Date	Open	High	Low	Close	Volume
2017-11-06	1.5	1.799	0.5	1.571	120615.3
2017-11-07	1.571	1.8	1.5389	1.8	134129.7
2017-11-08	1.7901	1.99	1.7479	1.9889	347290.3
2017-11-09	1.9781	2.1997	1.9	1.99	287037.2
2017-11-10	1.99	2.02	1.6019	1.7133	484658.9
2017-11-11	1.79	1.925	1.58	1.655	258838.4
2017-11-12	1.6567	1.7	1.34	1.49	457159.7
2017-11-13	1.5	1.7381	1.5	1.68	166770.7
2017-11-14	1.6801	1.7381	1.5321	1.5911	177246.5
2017-11-15	1.5721	1.6794	1.49	1.5532	297638.3
2017-11-16	1.5533	1.6699	1.511	1.5599	198029.7

[거래량 데이터]

```
from binance.client import Client
import pandas as pd

api_key = ""
api_secret = ""
client = Client(api_key, api_secret)

asset = "BNBUSDT"
start = "2017.8.1"
end="2021.12.31"
timeframe="1d"
df= pd.DataFrame(client.get_historical_klines(asset, timeframe, start, end))
df=df.iloc[:, :6]
df.columns=["Date", "Open", "High", "Low", "Close", "Volume"]
df = df.set_index("Date")
df.index = pd.to_datetime(df.index, unit="ms")
df = df.astype("float")
df.to_csv('C:/Users/Soon/Desktop/bnb.csv')
```

[Binance API 를 활용한 코드]

트위터 데이터를 수집하는 과정을 설명하면 우선 Influencer, News, Cryptocurrency 이렇게 세가지 유형을 설정하고 팔로워 랭킹을 바탕으로 영향력이 가장 큰 트위터 계정들을 선별했다. 전세계적인 영향력과 자연어 처리의 용이성을 고려하여 언어는 영어로 한정했다. 도지 코인과 바이낸스 코인이 발행되었던 2017 년 이후로 모든 트위터 데이터를 스크래핑하여 사용하려 했으나, 시간이 너무 오래 걸리고 두 암호화폐의 가격 변동이 주로 2021 년 이래로 관찰이 되어서 2021 년으로 한정하여 분석 범위를 한정했다. 공식 트위터 API 로는 스크래핑에 한계가 있어서 'Scweet'라는 모듈을 사용하여 데이터를 수집했다.

Influencer:
@elonmusk, @VitalikButerin, @BillyM2k

News:
@BBCBreaking, @TIME, @cnnbrk, @washingtonpost, @nytimes, @BBCWorld, @TheEconomist, @Reuters

Cryptocurrency:
@binance, @CoinDesk, @crypto, @ForbesCrypto, @dogecoin, @dogecoin_dev, @ethereum, @Bitcoin, @BTCTN

[트위터 계정 리스트]

```python
from Scweet.scweet import scrape
from Scweet.user import get_user_information, get_users_following, get_users_followers

influencer = ['elonmusk', 'VitalikButerin', 'BillyM2k']
news = ['BBCBreaking', 'TIME', 'cnnbrk', 'WSJ', 'washingtonpost', 'nytimes', 'BBCWorld', 'TheEconomist', 'Reuters']
cryptocurrency = ['binance', 'CoinDesk', 'crypto', 'ForbesCrypto', 'dogecoin', 'dogecoin_devs', 'ethereum', 'Bitcoin', 'BTCTN']

users1 = []
for account in influencer:
    data = scrape(hashtag=None, since="2020-12-27", until="2021-12-31", from_account=f'@{account}', interval=1,
                  headless=True, display_type="Top", save_images=False,
                  resume=False, filter_replies=True, proximity=False)
    user = f'@{account}'
    users1.append(user)
    data.to_csv(f'C:/Users/Soon/Desktop/{account}.csv', encoding='utf-8-sig')
users_info1 = get_user_information(users1, headless=True)
```

[트위터 스크래핑 코드]

검색량과 거래량은 시계열 데이터로 잘 확보가 되어서 다른 전처리 작업이 필요하지는 않았다. 하지만 트위터 데이터는 사용한 API 가 걸러내지 못한 불필요한 단어 (멘션, 해시태그, 리트윗, 조회수, 하이퍼링크 등)들을 제거하고 불용어와 화이트스페이스를 제거한 텍스트만을 추출하는 전처리 작업을 수행했다.

```python
# Clean The Data
def cleantext(text):
    text = re.sub(r"@[A-Za-z0-9]+", "", text) # Remove Mentions
    text = re.sub(r"#", "", text) # Remove Hashtags Symbol
    text = re.sub(r"RT[₩s]+", "", text) # Remove Retweets
    text = re.sub(r"https?:₩/₩/₩S+", "", text) # Remove The Hyper Link
    return text

# Preprocessing Text Data
def preprocessing(text):
    sentences = []
    for sentence in text:
        x = sentence.split('₩n')
        if len(x) != 3:
            tweet = ''
            for word in x:
                if word == '':
                    continue
                elif word[0] == 'ㅇ':
                    continue
                elif word[-1] not in '천만년월일수글기과다자표0123456789':
                    tweet = tweet + ' ' + word
                else:
                    continue
            sentences.append(tweet.strip())
    return sentences
```

[전처리 코드]

index	Unnamed: 0	UserScreenName	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets	Image link
0	0	Elon Musk	@elonmusk	2020-12-27T22:18:58.000Z	Elon Musk @elonmusk · 2020년 12월 28일	Try playing Polytopia in your Tesla! Great game. Multiplayer online version coming soon. 5천 4.7천 14.1만	NaN	5천	4.7천	14.1만	[]
1	1	Elon Musk	@elonmusk	2020-12-29T20:48:04.000Z	Elon Musk @elonmusk · 2020년 12월 30일	Not everything is made of cake 2.6천 1.2만 17.7만	NaN	2.6천	1.2만	17.7만	['https://pbs.twimg.com/media/EqbtOhfWMAAqNuR?f
2	2	Elon Musk	@elonmusk	2020-12-29T20:40:49.000Z	Elon Musk @elonmusk · 2020년 12월 30일	Such a weird game when you think about it 3.8천 3.4만 34만	NaN	3.8천	3.4만	34만	['https://pbs.twimg.com/media/EqbrkaDW4AEirD0?for
3	3	Elon Musk	@elonmusk	2020-12-29T21:32:09.000Z	Elon Musk @elonmusk · 2020년 12월 30일	The Last Kingdom show is great 4.6천 6.8천 13.2만	NaN	4.6천	6.8천	13.2만	[]
4	4	Elon Musk	@elonmusk	2020-12-29T20:37:54.000Z	Elon Musk @elonmusk · 2020년 12월 30일	This is not CGI youtube.com Do You Love Me? Our whole crew got together to celebrate the start of what we hope will be a happier year: Happy New Year from all of us at Boston Dynamics. www.BostonDyna... 4.4천 1.3만 6.1만	NaN	4.4천	1.3만	6.1만	['https://pbs.twimg.com/card_img/1478423614407938 format=jpg&name=240x240']
5	5	Elon Musk	@elonmusk	2020-12-29T22:16:14.000Z	Elon Musk @elonmusk · 2020년 12월 30일	All Tesla cars delivered in the final three days of the year will get three months of the Full Self-Driving option for free. Delivery & docs must be fully complete by midnight Dec 31st. 3.4천 4.3천 7만	NaN	3.4천	4.3천	7만	[]
6	6	Elon Musk	@elonmusk	2020-12-30T09:51:07.000Z	Elon Musk @elonmusk · 2020년 12월 30일	Snake-head dog had my undivided attention until winder-head ostrich came gliding through all nonchalant 2.5천 4.6천 9.7만	NaN	2.5천	4.6천	9.7만	['https://pbs.twimg.com/media/EqegeMTVQAAzSth?f
7	7	Elon Musk	@elonmusk	2020-12-30T05:33:19.000Z	Elon Musk @elonmusk · 2020년 12월 30일	Destiny, destiny No escaping that for me 5.9천 1.1만 15만	NaN	5.9천	1.1만	15만	[]

[트위터 Raw 데이터]

index	text
0	Try playing Polytopia in your Tesla! Great game. Multiplayer online version coming soon.
1	Not everything is made of cake
2	Such a weird game when you think about it
3	The Last Kingdom show is great
4	This is not CGI youtube.com Do You Love Me? Our whole crew got together to celebrate the start of what we hope will be a happier year: Happy New Year from all of us at Boston Dynamics. www.BostonDyna...
5	All Tesla cars delivered in the final three days of the year will get three months of the Full Self-Driving option for free. Delivery & docs must be fully complete by midnight Dec 31st.
6	Snake-head dog had my undivided attention until winder-head ostrich came gliding through all nonchalant
7	Destiny, destiny No escaping that for me
8	Great interview with Mathias Döpfner businessinsider.com Elon Musk reveals Tesla's plan to be at the forefront of a self-driving-car revolution — and why he... Tesla CEO Elon Musk sat down with Axel Springer CEO Mathias Döpfner for an interview during his visit to Germany.
9	So proud of the Tesla team for achieving this major milestone! At the start of Tesla, I thought we had (optimistically) a 10% chance of surviving at all. Tesla In 2020, we produced and delivered half a million cars. Huge thanks to all those who made this possible.
10	Snow falling on Giga Berlin
11	Snow falling on Giga Berlin
12	This is called the domino effect
13	Use Signal
14	The most entertaining outcome is the most likely
15	Launch underway SpaceX Watch Falcon 9 launch the Turksat 5A mission →
16	Btw, critical feedback is always super appreciated, as well as ways to donate money that really make a difference (way harder than it seems)
17	My 14-year-old son, Saxon, said he feels like 2021 will be a good year. I agree. Let us all make it so.
18	Hey you … Yeah you Queen … You're gonna make it!
19	Mus protec yoda

[전처리 후 트위터 text 데이터]

2.2. 감성분석을 통한 트위터 INDEX 도출

주어진 문장에 대해 복합적인 Sentiment Score 를 계산해 준다는 점에서 이점이 있는 VADER Sentiment Analysis 모듈을 감성분석에 활용했다. 수집한 모든 트윗으로부터 평균 Sentiment Score 를 계산하여 지표로 활용하고자 했다. 우선 구글 트렌드로부터 암호화폐와 관련되어 검색량이 많았던 단어들을 키워드로 선정했다. 그리고 그 키워드가 존재하는 트윗들만 모아서 데이터프레임을 만들고 VADER 를 적용하여 모든 문장의 Sentiment Score 와 그 평균 점수를 구한다. 다음으로 키워드가 모든 텍스트에 얼마나 존재하는지를 카운트한다. 마지막으로 모든 키워드들에 대하여 키워드 등장 횟수와 해당 키워드의 평균 Sentiment Score 로부터 weighted average 를 구하여 최종 Sentiment Score 를 얻었다.

Cryptocurrency	Cryptocurrency, Bitcoin, Ethereum, Coinbase, Doge, NFT, Binance, DeFi
News	Fed, Price, Market, Trading, Exchange, Invest, Wallet, Stock, Money, Digital, Liquidity, Crisis, Tapering, Regulation, Crackdown, Plunge
Influencer	Reddit, Meme, SpaceX, AMC, GME

[암호화폐 관련 트위터 키워드]

neg	neu	pos	compound	sentence
0.058	0.942	0.0	-0.2732	Robinhood curbed trading in cryptocurrencies for some of its customers on Friday, the platform's latest restriction in a frenzied week of activity centered on GameStop's soaring stock. nytimes.com Robinhood curbs trading again, this time in cryptocurrencies.
0.0	0.885	0.115	0.7184	Tesla disclosed that it had purchased $1.5 billion worth of Bitcoin and expected to accept the cryptocurrency as a form of payment "in the near future." nytimes.com Tesla says it bought $1.5 billion of Bitcoin, sending the cryptocurrency to a record high. The carmaker's chief executive, Elon Musk, is known for promoting cryptocurrencies on his widely followed Twitter feed.
0.033	0.793	0.174	0.8024	"We need to make sure that those who have been most affected aren't permanently scarred by this crisis," Treasury Secretary Janet Yellen told on Monday during a discussion about pandemic relief at the DC Policy Project. nytimes.com Janet Yellen on stimulus, the lessons of GameStop and what's good (and bad) about cryptocurrency.
0.0	0.958	0.042	0.3612	The $69 million Beeple NFT was bought by an investor known only by a pseudonym and who paid for it with cryptocurrency. nytimes.com The $69 Million Beeple NFT Was Bought With Cryptocurrency "I feel like I got a steal," the buyer, who calls himself Metakovan, said in an interview about the "nonfungible token," or NFT, he bought at an online auction.
0.0	0.98	0.02	0.0964	When asked the artist known as Beeple if an NFT he sold for nearly $70 million in cryptocurrency was real, he held up his phone. An app showed he had $56,635,781.41 in cash. "You and I know this number is as real as anything else," he said. nytimes.com Are NFT Purchases Real? The Dollars Are. Dive down a rabbit hole and explore nonfungible tokens, multimillion-dollar digital art and the nature of reality.
0.037	0.851	0.112	0.4939	Coinbase, the cryptocurrency exchange, is set to begin trading on Wednesday — instantly becoming a financial giant on Wall Street, and likely to be valued at more than $65 billion. Here's why Coinbase matters so much — and why there are huge risks, too. nytimes.com Here's what you need to know about the Coinbase debut.
0.0	0.888	0.112	0.7003	Glauber Contessoto, a 33-year-old in Los Angeles, used his life savings and borrowed money to invest about $250,000 in Dogecoin, a cryptocurrency started as a joke, in February. The value now? Roughly $2 million. nytimes.com He's a Dogecoin Millionaire. And He's Not Selling. Glauber Contessoto went looking for something that could change his fortunes overnight. He found it in a joke cryptocurrency.

[VADER 적용 후 문장 별 Sentiment Score]

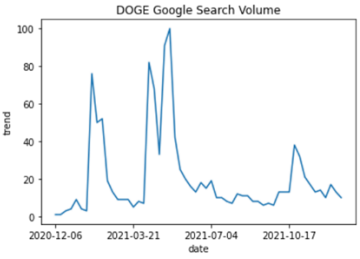
키워드	등장 횟수	점수
Cryptocurrency	5035	0.173
Bitcoin	18534	0.197
Ethereum	2228	0.211
Coinbase	754	0.215
Doge	2656	0.391
NFT	3225	0.314
Binance	5815	0.348
Price	3755	0.128
Market	5127	0.175
Regulation	983	0.119
plunge	159	-0.141
Reddit	262	0.344

→ **Sentiment Index: 0.184**

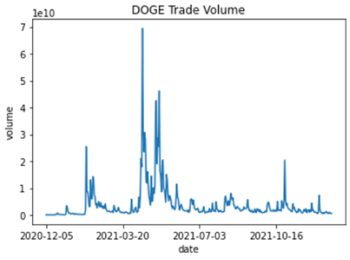
[최종 Sentiment Score]

2.3. 가격과의 상관분석을 통한 검색량, 거래량 INDEX 도출

검색량과 거래량은 모두 시계열 데이터이고 가격과 상관관계를 정량화한 수치를 INDEX 로 도출하는 방식으로 진행했다. 우선 각 시계열의 first difference 를 구하고 다음으로 Pearson 상관계수를 계산하는 식으로 검색량과 가격, 그리고 거래량과 가격의 상관 계수를 구할 수 있었다. 도지 코인과 바이낸스 코인 모두에 대해 값들을 계산하고 둘을 산술 평균하여 검색량과 거래량에 대한 INDEX 를 구했다.



검색량



거래량

가격

	검색량	거래량
BNB 가격	0.519446	-0.00251
DOGE 가격	0.466369	0.666303
평균	**0.492908**	**0.331897**

[검색량과 거래량 INDEX]

2.4. 최종 INDEX 계산

검색량, 거래량, 트위터 각각에 대한 정량적인 수치를 얻었고 세 값을 평균하여 최종 INDEX 를 계산했다.

	검색량	거래량	트위터	HI
디지털 자산 (BNB, DOGE)	0.493	0.332	0.184	**0.336**

[최종 INDEX]

3. 결론 및 느낀 점

지난 1 년 간의 데이터를 바탕으로 도지 코인과 바이낸스 코인에 대한 디지털자산 사회정서 지수를 계산했다. 검색량, 거래량, 트위터 이렇게 세 가지 유형의 데이터를 지표로 활용하는 것으로 정했는데 거래량과 검색량은 매매 기록의 종가를 기준으로 상관관계를 계산하여 Index 를 구하고, 자연어로 된 트위터에 대해서는 감성분석을 수행하여 그 평균 점수를 구했다.

정규화를 수행하고 0~1 사이의 값으로 된 정량적인 지표를 도출하는 것 자체가 생각보다 쉽지 않은 과정이었다. 거래량과 검색량은 정량적 데이터이므로 시계열의 first difference 를 구하고 가격과의 Pearson 상관계수를 구해서 Index 로 사용했지만 time lag 을 고려하거나 계열 내 상관성 등을 고려하지는 않았고 현재 사용한 방법은 어떠한 인과 관계를 반영하는 기법이 아니라 상당한 한계점들이 존재한다. 또한 Index 를 구하는 과정에서 단순하게 산술 평균을 해서 값을 계산했지만 실제로는 가중치가 다르게 반영되어야 할 것이다.

향후 더 적절한 통계적인 기법을 적용하고 지수 산출 방법을 보완한다면 실제적으로 더 나은 지표를 산출할 수 있을 것이다. 그리고 지금 계산된 최종 Index 는 지난 1 년 간의 데이터를 바탕으로 한 것인데 실시간으로 데이터를 크롤링하여 Index 를 계산하고 그것과 시장 상황을 비교하여 검증하게 된다면 고안된 지수의 적정성에 대한 평가도 가능할 것이다.

4. 참고자료

[1] S&P Twitter Sentiment Indices Methodology Book, S&P Global, 2021.11.

[2] 머신러닝 기반의 온라인 미디어 감성분석을 통한 자동차 제조사 주가 해석에 관한 연구, 박민수, 서울대 공학전문대학원 학위 연구보고서, 2020.08.

[3] 디지털자산 공포-탐욕 지수 Methodology Book, Dunamu Datavalue Team, 2022.01.
https://datavalue.dunamu.com/feargreedindex

[4] 트위터 데이터 수집 API, 2022.01.
https://github.com/Altimis/Scweet

[5] 바이낸스 API docs, 2022.01.
https://binance-docs.github.io/apidocs/spot/en/#change-log