[TEAM 10] Towards Better Understanding of Cryptocurrency: A Multi-Aspect Analysis

Patara Trirat (20205642)
School of Computing, KAIST
Daejeon, South Korea
patara.t@kaist.ac.kr

Arnur Makenov (20214197)
School of Computing, KAIST
Daejeon, South Korea
arnur@kaist.ac.kr

1 INTRODUCTION

After a decade of Bitcoin's emergence [14], the cryptocurrency market has become a vital component of the international financial market. Nowadays, not the economic and social impacts of cryptocurrencies only continue to proliferate, but their values as financial assets also increase every year. Accordingly, despite the drastic market fluctuations, many people—including investors, policymakers, financial institutions, and media—have been drawn to them due to the rapid expansion of legitimate uses and fast-evolving development of blockchain technology in recent years [1, 2].

In addition to Bitcoin, Nghiem et al. [16] noted that the cryptocurrency market contains more than 5,000 kinds of coins, called altcoins, for which the volume of transactions and circulation is huge but uneven. Though the cryptocurrency market is highly complicated and risky, it can still be an alternative investment instrument with the unique characteristic of high return. As the risk of cryptocurrencies is higher than other products from the complexity of various factors that affect the change of coin prices, a comprehensive understanding of these factors is greatly important.

1.1 Related Work

Most early studies [1, 11, 13, 17–19] focused on analyzing only Bitcoin or a few altcoins data to understand the relevant factors or predict future prices. Therefore, they ignored the capabilities of the overall cryptocurrency market. Recently, there have been attempts to analyze cryptocurrency data on the market scale concerning a particular aspect. Bai et al. [2], Ho et al. [10], Jay et al. [12], Sun et al. [20], Teker et al. [21], Wołk [22] investigated coin market data regarding the price movements and trends. Similarly, some researchers analyzed the market data to forecast volatility [24] instead of price or understand certain factors that associate with coin changes such as big events [8], day of week [4], public sentiments [3], time-series compositions [23], regulation and policy uncertainty [5, 7] as well as frauds [16]. Lastly, due to the latest pandemic, various studies [6, 9, 15] have examined the impacts of COVID-19 on the coin market from a different point of view.

Notably, even though these recent papers analyzed a more extensive set of data, their analyses were still limited in particular aspects, unlike our work that studies about top-100 cryptocurrencies representing more than 95% of the total market capitalization in the different time intervals. Additionally, no prior work utilizes the coin-specific properties to examine the long-term price movements as well as their effects before, during, and after the pandemic.

KAIST CS564 Fall, December 13, 2021, Daejeon, South Korea
2021. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.2 Main Ideas

Given the above significant challenge from cryptocurrency and limitations in prior studies, our ultimate goal is to investigate and understand the factors that have a relationship with changes in cryptocurrencies from multiple aspects. Specifically, we want to answer the following four research questions in this project.

- **RQ1**. Which factors correlate with the changes in coins?
- **RQ2**. Is there a relationship in changes between coins?
- **RQ3**. Does the similarity of coin properties correlate with the changes in cryptocurrencies?

Moreover, we want to examine whether there is a difference in the association of each factor between pre-, peri-, and post-COVID-19. We consider this hypothesis as our **RQ4**.

The main contributions and findings are summarized as follows.

- We make a multi-aspect analysis with various external and internal factors for market-scale cryptocurrency data to quantify their relationships.
- We provide interesting findings of how different factors correlate with cryptocurrencies and whether there is a significant change during the COVID-19 situation as answers to the formulated research questions.
- We show that certain factors manifest correlations with the coins regarding the changes in price. However, this is not the case for the return ratio, meaning that examining the changes in return rate is far more challenging than just the movement of the prices.
- We provide the source code for reproducibility at https://github.com/itouchz/GRoGL.

2 DATA PREPROCESSING

2.1 Data Collection

This section describes how we collect the datasets and extract necessary features before the main analyses. Table 1 presents the summary of all datasets.

2.1.1 Price Datasets. Historical prices for Top-100 Cryptocurrencies[1], big technology company stocks, and world financial indices were collected. These are readily available on Yahoo Finance. We used both daily and monthly intervals of the data. Number of attributes present in this data equals five and follows **O**pen **H**igh **L**ow **C**lose **V**olume format.

2.1.2 Trend Datasets. We collect search trend data from Google Trends and counts of page visits from Wikipedia Pageviews. The attributes of these two datasets are the search keywords and wiki

[1] as of Oct 31, 2021 by coinmarketcap.com

page titles, respectively. According to the collected coin data, 107 keywords and 48 titles were selected for subsequent analysis.

2.1.3 Social Networks and News Datasets. For public opinion datasets, we looked into various sources and collected relevant daily Reddit submissions, Twitter tweets, and Google news. Due to the computational limits and provider's API restrictions, the maximum possible records for each keyword per day in the subset are 1,000 (Reddit), 100 (Twitter), and 10 (Google News), respectively. Additionally, for Twitter's tweets, we only collected the high influential ones having user interactions (i.e., retweet counts, like counts, and reply counts) greater or equal to 100.

2.1.4 Policy Uncertainty Datasets. This group of datasets represents global geopolitical and economic uncertainty indices consisting of geopolitical uncertainty, economic policy uncertainty, and Twitter-based economic uncertainty. Precisely, the geopolitical risk index measures adverse geopolitical events based on a tally of newspaper articles covering geopolitical tensions and their evolution and economic effects. Similarly, the economic policy uncertainty index reflects the relative frequency of own-country newspaper articles (or Twitter's tweets) that contain a trio of terms about the economy, policy, and uncertainty.

2.1.5 Coin-specific Features. We crawled the data coin-specific properties by the data provided by Coin properties in the form of tags and descriptions. Examples of the properties are mineability and hashing algorithms. In total, we collected 25 characteristic features.

2.2 Feature Extraction

2.2.1 OHLCV Prices. After collecting the price data, we derive the daily *change*, daily *return* ratio, and monthly *volatility* for the entire price dataset based on the ***close price*** attribute as in the previous studies [4, 13, 19]. While the volatility is the standard deviation of all daily close prices within a month, the daily change is $p_d - p_{d-1}$ and the daily return is $\frac{p_d}{p_{d-1}} - 1 \times 100$, where p_d is the close price p at day d. The daily change here will later marked as "up" or "down". If missing values exist in close prices, they are temporally forward-backward filled. Lastly, all the prices are normalized using the `scale` function.

2.2.2 Sentiment Analysis. For the text data, we first preprocess it by removing any character that matches this regular expression: `([^\'A-Za-z0-9@#_ ]+)`. We concatenate the *title*, *description*, and *main text* into one sentence or paragraph. After that, we conduct a sentence-level sentiment analysis using the `sentimentr` library. To get the sentiment polarity, each ***sentiment score*** is later changed to "positive" if the score > 0.2, "negative" if the score < 0.2, otherwise "neutral". If the concatenation still results in an empty string, it is removed from the analysis. Finally, additional time-series features such as number of topics, number of posts, number of each sentiment, and number of user interactions of tweets are computed for each day.

2.2.3 Categorical Encoding. In order to run the following clustering analysis, all features of *coin properties* (e.g., consensus mechanism, hash function, and mineability) need to be converted to sparse one-hot encoding using the `fastDummies` package due to their categorical nature.

Table 1: Summary of Data Sets

Categories	Data Sets	# Attributes (Daily / Monthly)	# Samples	Provider
OHLCV Financial Prices	Top-100 Cryptocurrencies			
	Stocks	5	~1,035 / 34	Yahoo Finance
	World Indices			
Policy Uncertainty	Economic Policy Uncertainty	NA / 26	NA / 29	policyuncertainty.com/ matteoiacoviello.com/gpr.htm
	Geopolitical Uncertainty	4 / 102	973 / 34	
	Twitter Economic Uncertainty	8 / NA	1,035 / NA	
Popularity	Google Trends	107 (Keywords)	1,035	Google
	Wikipedia Pageviews	48 (Pages)	1,035	Wikipedia
Public Opinion	Google News	3	32,732	Google
	Twitter Tweets	7	336,261	Twitter
	Reddit Submissions	3	1,030,908	Reddit
Metadata	Stock Information	5	676	Yahoo Finance
	Coin Properties	25	100	CoinMarketCap

3 METHODOLOGIES

3.1 Correlation Analysis

Given the above data sets we have collected, to answer **RQ1** and **RQ2**, we run the correlation test[2] using the `cor_mat` function in the `rstatix` library. In this analysis, we compute the correlation coefficients of a coin in price, return, and volatility aspects against world financial indices, global policy uncertainty, popularity, news and public opinion, as well as the other cryptocurrencies. As a result, we get the correlation scores and corresponding p-values for each factor.

3.2 Clustering Analysis

To answer **RQ3**, with the coin-specific properties extracted earlier, we use the `NbClust`[3] function to identify the best number of clusters and run the k-means clustering accordingly using the `kmeans` function. Here, the best number of clusters found by the `NbClust` is two. Further, we try to examine the shared characteristics of each cluster by looking deeper into its features. Lastly, we test whether there is a significant difference between these two clusters using the ANOVA test.

3.3 Ad-hoc Analysis

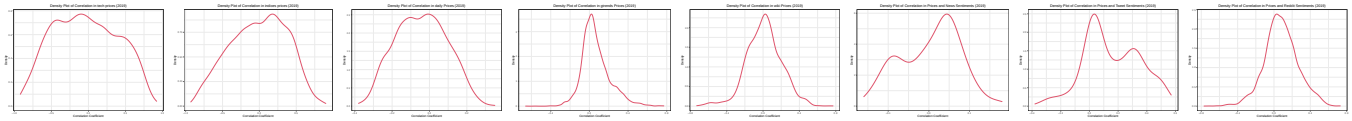
As an ad-hoc analysis, we run ANOVA tests for features extracted from close prices and do visual inspections of the results from the above analyses to answer **RQ4**. Since a particular group of coins may positively move together, we also extract the *degree*, *betweenness*, and *eigenvector* from the coin-to-coin correlation network using `igraph` library to observe potential leading coins in changes.

4 RESULTS AND DISCUSSION

4.1 RQ1: Factors that Affects the Coins

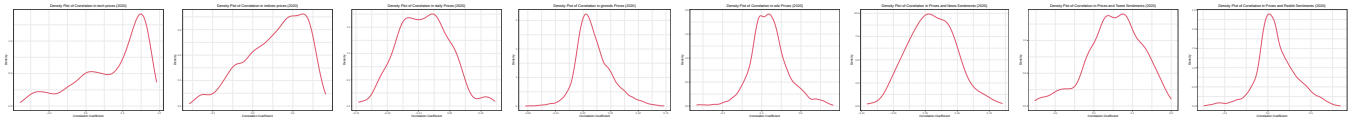
With the results from correlation matrix, we visualize the distribution of correlation coefficients for each factor in Fig. 1, 2, and 3 for price aspect as well as Fig. 4, 5, and 6 for return aspect. The findings are reported below.

[2]Even though there were hundreds of variables used here, as the goal is to observe the overall effects of many factors possible, we did not use dimensionality reduction because we believe that it could obscure direct interpretation between factors.
[3]We run NbClust with all possible indices, execpt for *ccc, scott, marriot, trcovw, tracew, friedman, rubin* due to indexing error during runtime.



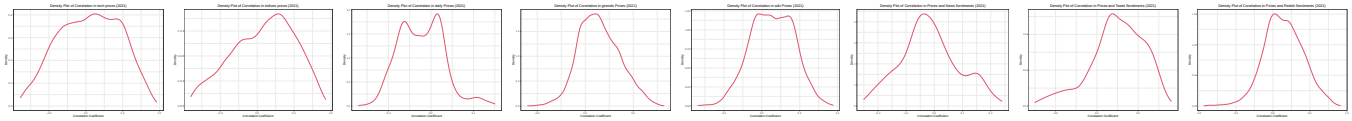
(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 1: Density Plots of Prices Correlation by Factors in 2019



(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 2: Density Plots of Prices Correlation by Factors in 2020



(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 3: Density Plots of Prices Correlation by Factors in 2021

4.1.1 Financial Indices. There are strong correlations between cryptocurrencies and financial indices (including company stocks), especially in the prices. Interestingly, during the main outbreak period, the directions of most relationships become positive, while before and after the pandemic are the frequencies of negative and positive correlations are comparable. For the return aspect, there is no strong correlation exhibits with this factor.

4.1.2 Policy Uncertainties. In general, there are only a few moderate inverse correlations with uncertainty factors strengthened during the COVID-19 year. Specifically, there are only relationships between coin prices and Twitter-based economic uncertainty. Again, for coin returns, there is no significant correlation here.

4.1.3 Popularity. In 2019, Google search trends and Wikipedia pageviews did not have any relationship with coin prices. However, their associations with cryptocurrency prices have increased over the years. The reason could be that the numbers of cryptocurrencies rapidly grow over time, and people need to know how those coins are different to make a decision on trading or mining the coins. Similar to the previous factors, popularity also has no effect to the coin returns.

4.1.4 Public Opinion and News. Contrary to common sense, News does not affect Cryptocurrencies at all. Twitter tweets substantially influence the prices, while Reddit posts, similarly to Wikipedia Pageviews, affect only specific coins with small correlation coefficients. Interestingly, the relationships between coin prices and tweets' sentiment are amplified during the COVID-19 period, like the policy uncertainty factor but in the opposite direction.

In addition to coin *prices* and *returns*, we also compute the monthly *volatility* given the daily close prices. However, due to the small samples for each year, almost the correlation scores came out to be not statistically significant regardless their values.

4.2 RQ2: Coin-to-Coin Correlations

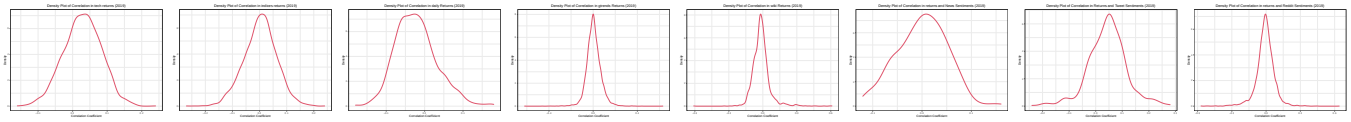
Regarding correlation analysis between cryptocurrencies, as shown in Fig. 8 and 9, we can see that there are strong relationships between coins, mostly positive both in price and returns. Also, by network analysis, we can observe that the leading coin and the direction of their relationships changed over time. Fig. 7 illustrates the very strong correlation (i.e., absolute coefficient ≥ 0.8) networks between coins.

4.3 RQ3: Similarity-based Price Movements

The visualization of clustering analysis is shown in Fig. 11. We can see that there are two possible clusters. After looking at the features, we can notice that—except for asset-backed, research, and staking features along with a few consensus mechanisms dedicated to one cluster—there is no clear distinction between these two. The Silhouette score of about 0.1 also confirms no underlying structure based on the coin-specific properties. Additionally, the ANOVA tests for the average price and return indicate no difference between these two clusters with p-values > 0.05. Therefore, the features do not helpful for indicating the changes of cryptocurrencies. Fig. 10 illustrates the movements of two clusters over year showing no significant difference.

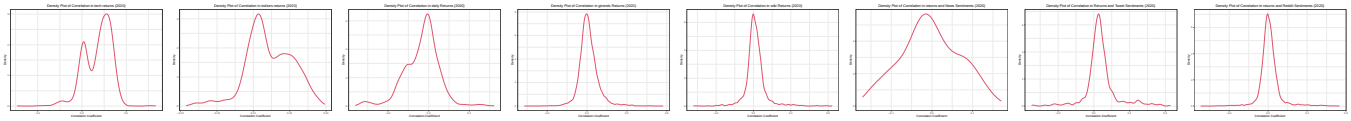
4.4 RQ4: Differences during COVID-19

According to the above results, it is evident that there are differences concerning the effect of each factor during different time



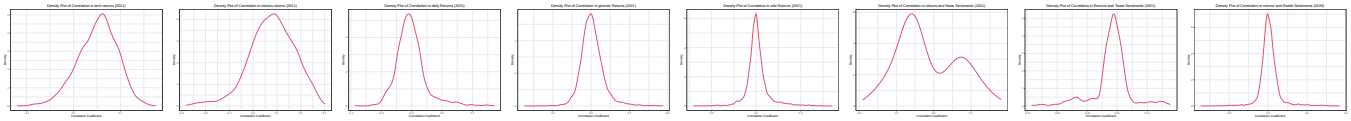
(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 4: Density Plots of Returns Correlation by Factors in 2019



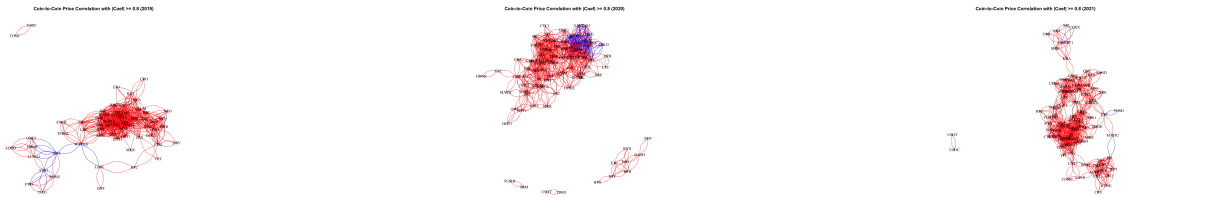
(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 5: Density Plots of Returns Correlation by Factors in 2020



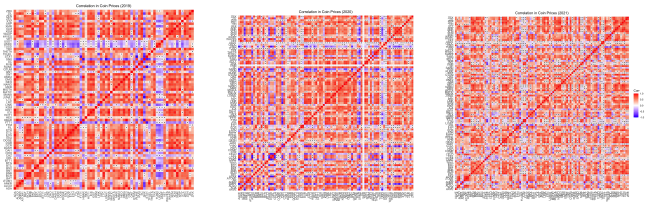
(a) Tech. Stocks (b) World Indices (c) Uncertainty (d) Search Trends (e) Wiki Views (f) News (g) Tweets (h) Reddit

Figure 6: Density Plots of Returns Correlation by Factors in 2021



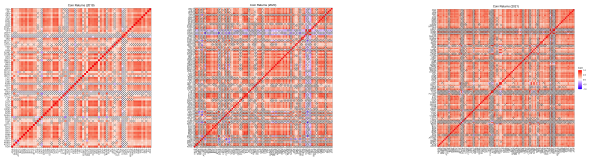
(a) 2019, with the DASH Coin Leading (b) 2020, with the ETH Coin Leading (c) 2021, with the MIOTA Coin Leading

Figure 7: Three Different Correlation Networks by Year



(a) 2019 (b) 2020 (c) 2021

Figure 8: Coin Prices across years.



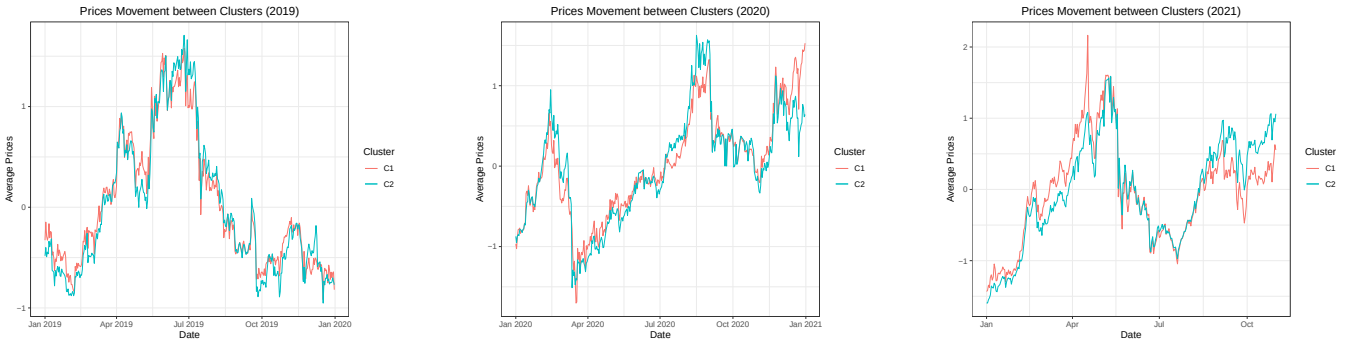
(a) 2019 (b) 2020 (c) 2021

Figure 9: Coin Returns across years.

frames. Notably, pre and post-COVID19 usually exhibit similar results, while during the COVID19 period, the impact of each factor typically has an inverse direction or becomes more intense. For the differences before and after the pandemic, ANOVA tests indicate statistically significant differences between these three years for coin prices and returns with the p-values < 0.01. Fig. 12 shows the boxplot of coin price and return of different years.

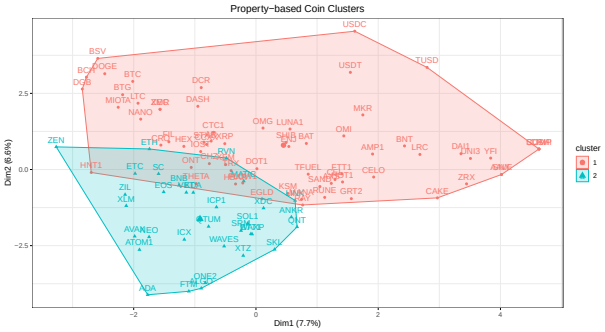
5 CONCLUSION AND FUTURE WORK

In this project, we investigate the relationship between multiple aspects of cryptocurrencies and various factors in different time frames. Our findings indicate that cryptocurrencies are highly correlated with stock prices and financial indices as well as the other

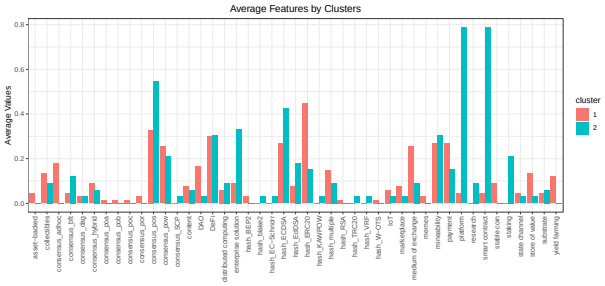


(a) 2019 (b) 2020 (c) 2021

Figure 10: Price Movement between Cluster by year.

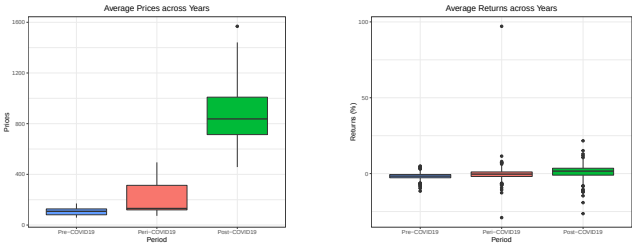


(a) Coin Clusters



(b) Cluster Avg. Features

Figure 11: Clusters and their corresponding average features.



(a) Coin Prices (b) Coin Returns

Figure 12: Boxplots of Prices and Returns over Pre-, Peri-, Post-COVID19.

coins. However, these relationships are also changed over time, especially during the pandemic. Surprisingly, the news and the underlying technical attributes of the coins do not provide us with any information on the changes in coin prices and returns. Therefore, to conjecture the changes of coins, we should utilize the changes of financial indices and other coins that are strongly connected with the target coin. While the former is generally inverse, the latter normally move together.

Limitations. This project only conducts descriptive analyses with the ultimate goal of understanding the relationships between cryptocurrencies and other factors in general. The results here are superficial, with no detailed analysis for specific events, such as sudden drop or all-time high.

Future Work. Nevertheless, as it is clear that various factors from different aspects affect the cryptocurrency prices, using the combination of these factors will be helpful for the subsequent predictive analysis. For text data, other types of analysis still can be done, such as co-occurrence analysis, to understand the public opinion in a more fine-grained manner. For example, what coins people usually tweet together and how tweets affect the cryptocurrency market for a specific event.

ACKNOWLEDGMENTS

We sincerely appreciate the insightful comments and suggestions from Prof. Meeyoung Cha and our mentor Dr.Luiz Felipe Vecchietti.

REFERENCES

[1] Gourang Aggarwal, Vimal Patel, Gaurav Varshney, and Kimberly Oostman. 2019. Understanding the social factors affecting the cryptocurrency market. *arXiv preprint arXiv:1901.06245* (2019).
[2] Chongyang Bai, Thomas White, Linda Xiao, Venkatramanan Siva Subrahmanian, and Ziheng Zhou. 2019. C2P2: a collective cryptocurrency up/down price prediction engine. In *2019 IEEE International Conference on Blockchain (Blockchain)*. IEEE, 425–430.
[3] Silvia Bartolucci, Giuseppe Destefanis, Marco Ortu, Nicola Uras, Michele Marchesi, and Roberto Tonelli. 2020. The Butterfly "Affect": impact of development practices on cryptocurrency prices. *EPJ Data Science* 9, 1 (2020), 21.
[4] Guglielmo Maria Caporale and Alex Plastun. 2019. The day of the week effect in the cryptocurrency market. *Finance Research Letters* 31 (2019).
[5] Francisco Colon, Chaehyun Kim, Hana Kim, and Wonjoon Kim. 2021. The effect of political and economic uncertainty on the cryptocurrency market. *Finance Research Letters* 39 (2021), 101621.

[6] Shaen Corbet, Yang Greg Hou, Yang Hu, Charles Larkin, Brian Lucey, and Les Oxley. 2021. Cryptocurrency liquidity and volatility interrelationships during the COVID-19 pandemic. *Finance Research Letters* (2021), 102137.

[7] Brian D Feinstein and Kevin Werbach. 2021. The impact of cryptocurrency regulation on trading markets. *Journal of Financial Regulation* 7, 1 (2021), 48–99.

[8] Paulo Ferreira and Éder Pereira. 2019. Contagion effect in cryptocurrency market. *Journal of Risk and Financial Management* 12, 3 (2019), 115.

[9] Zied Ftiti, Wael Louhichi, and Hachmi Ben Ameur. 2021. Cryptocurrency volatility forecasting: What can we learn from the first wave of the COVID-19 outbreak? *Annals of Operations Research* (2021), 1–26.

[10] Kin-Hon Ho, Wai-Han Chiu, and Chin Li. 2020. A Short-Term Cryptocurrency Price Movement Prediction Using Centrality Measures. In *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 369–376.

[11] Stefan Hubrich. 2017. 'Know When to Hodl'Em, Know When to Fodl'Em': An Investigation of Factor Based Investing in the Cryptocurrency Space. *Know When to Fodl'Em': An Investigation of Factor Based Investing in the Cryptocurrency Space (October 28, 2017)* (2017).

[12] Patel Jay, Vasu Kalariya, Pushpendra Parmar, Sudeep Tanwar, Neeraj Kumar, and Mamoun Alazab. 2020. Stochastic neural networks for cryptocurrency price prediction. *IEEE Access* 8 (2020), 82804–82818.

[13] Connor Lamon, Eric Nielsen, and Eric Redondo. 2017. Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev* 1, 3 (2017), 1–22.

[14] Jiaqi Liang, Linjing Li, Weiyun Chen, and Daniel Zeng. 2019. Towards an understanding of cryptocurrency: a comparative analysis of cryptocurrency, foreign exchange, and stock. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 137–139.

[15] Emna Mnif, Anis Jarboui, and Khaireddine Mouakhar. 2020. How the cryptocurrency market has performed during COVID 19? A multifractal analysis. *Finance Research Letters* 36 (2020), 101647.

[16] Huy Nghiem, Goran Muric, Fred Morstatter, and Emilio Ferrara. 2021. Detecting Cryptocurrency Pump-and-Dump Frauds using Market and Social Signals. *Expert Systems with Applications* (2021), 115284.

[17] Ross C Phillips and Denise Gorse. 2018. Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PloS one* 13, 4 (2018), e0195200.

[18] Nico Smuts. 2019. What drives cryptocurrency prices? An investigation of google trends and telegram sentiment. *ACM SIGMETRICS Performance Evaluation Review* 46, 3 (2019), 131–134.

[19] Darko Stosic, Dusan Stosic, Teresa B Ludermir, and Tatijana Stosic. 2018. Collective behavior of cryptocurrency price changes. *Physica A: Statistical Mechanics and its Applications* 507 (2018), 499–509.

[20] Xiaolei Sun, Mingxi Liu, and Zeqian Sima. 2020. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters* 32 (2020), 101084.

[21] Dilek Teker, Suat Teker, and Mustafa Ozyesil. 2019. Determinants of cryptocurrency price movements. In *14th Paris international conference on marketing, economics, education and interdisciplinary studies, MEEIS-19*. 12–14.

[22] Krzysztof Wołk. 2020. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems* 37, 2 (2020), e12493.

[23] Boyu Yang, Yuying Sun, and Shouyang Wang. 2020. A novel two-stage approach for cryptocurrency analysis. *International Review of Financial Analysis* 72 (2020), 101567.

[24] Kuang-Chieh Yen and Hui-Pei Cheng. 2021. Economic policy uncertainty and cryptocurrency volatility. *Finance Research Letters* 38 (2021), 101428.