Bitcoin and News Around the World in Twenty-Six Languages*

Lucia Alessi[†] Eric Ghysels[‡] Marco Petracco[§] Zhe Wang[¶]

This Draft: April 30, 2021

PRELIMINARY AND INCOMPLETE

ABSTRACT

The question to what extend cryptocurrencies are different from other asset classes is a topic of keen interests to academics, policy makers and practitioners. There is a large literature documenting the impact of U.S. macroeconomic news announcements on returns of various financial markets. In comparison, far less is known about the impact of news on cryptocurrency markets. While in general it is not clear what exactly determines the value of say Bitcoin, it is fair to say that one should not exclusively focus on macroeconomic events to study impact of news on cryptocurrencies as many other events - including regulatory announcements - matter. We use a unique and comprehensive data set to study the impact of news on Bitcoin. More specifically, we use the Europe Media Monitor (EMM) data, which gathers and aggregates more than 300,000 news articles per day from news portals world-wide in up to 70 languages. It is this unique character of our data that allows us to study the impact of news on Bitcoin beyond what has been considered so far in the literature.

*We gratefully acknowledge the financial support of the University Blockchain Research Initiative (UBRI) of Ripple.

[†]Joint Research Centre (JRC), Ispra

[‡]Department of Economics, University of North Carolina, Gardner Hall CB 3305, Chapel Hill, NC 27599-3305, USA, Department of Finance, Kenan-Flagler Business School and CEPR

[§]Joint Research Centre (JRC), Ispra

[¶]Kenan Institute, University of North Carolina, Chapel Hill

1 Introduction

The question to what extend cryptocurrencies are different from other asset classes is a topic of keen interests to academics, policy makers and practitioners. One angle to address the question is to study how cryptocurrencies might behave differently from equities, commodities, foreign exchange and fixed income securities. We know that the pricing of any asset is driven by the present value of all future payoffs. Therefore, any news pertaining to future payoffs should have an impact on the price and there is indeed a large literature documenting the impact of U.S. (macroeconomic) news announcements on returns of various financial markets.

In comparison, far less is known about the impact of news on cryptocurrency markets. Corbet, Meegan, Larkin, Lucey, and Yarovaya (2018) document that the response to macroeconomic news is weaker for cryptocurrencies in comparison to conventional assets and currencies. The same authors found evidence that digital assets can be influenced by the US Federal Fund interest rates and quantitative easing announcements (see Corbet, Meegan, Larkin, Lucey, and Yarovaya (2020)).

We focus exclusively on Bitcoin. While in general it is not clear what exactly determines the value of Bitcoin, see e.g. Ghysels and Nguyen (2019), Entrop, Frijns, and Seruset (2020) and Ghysels, Nguyen, Shin, and Wang (2021) for further discussion, it is fair to say that one should not exclusively focus on macroeconomic events to study impact of news on cryptocurrencies as many other events - including regulatory announcements - matter. We use a unique and comprehensive data set to study the impact of news on Bitcoin. Our data set is special in many regards as it covers a real-time vast set of news (any news!) sources in multiple languages, including all major news organizations around the globe as well as social media. More specifically, we use the Europe Media Monitor (EMM) data, which is a fully automated system that analyzes both traditional and social media (see Steinberger, Podavini, Balahur, Jacquet, Tanev, Linge, Atkinson, Chinosi, Zavarella, and Steiner (2015)). The automated system was developed by the Joint Research Centre (JRC) – the European Commission's science and knowledge service which employs scientists to carry out research in order to provide independent scientific advice and support to EU policy development and decision making. EMM gathers and aggregates more than 300,000 news articles per day from news portals world-wide in up to 70 languages and processes news feeds from over twenty press agencies. It monitors governmental and non-governmental news-like web pages as well as social media such as Twitter and Facebook.[1]

From the EMM data base we extract all news pertaining to Bitcoin. Hence, the analysis in our paper casts a wide net across many media platforms and across many languages. It is this unique character of our data that allows us to study the impact of news on Bitcoin beyond what has been considered so far in the literature.

[1]See for instance https://emm.newsbrief.eu/overview.html.

2 Data Description

Data was gathered from JRC's Europe Media Monitor (EMM). EMM is an automatic system that analyses traditional and social media. It gathers and analyses about 300,000 articles per day in up to 26 languages. We obtained the publishing time, language, and tonality of Bitcoin-related news between April 16, 2014, and August 31, 2020, from the EMM. It includes 403,112 articles in 26 languages (184,354 of which in English). The articles were processed and tagged by EMM-newsbrief to generate structured (meta) data which could then be used for further analysis in conjunction with high-frequency price and volume data from exchanges in different countries. In order to generate structured (meta)data, EMM-NewsBrief retrieves the full text of articles from the web based on a set of keywords organized in categories or alerts, groups related items, categorizes them into thousands of classes, extracts information such as language, topic, publication location, named entities and timestamps them and extracts sentiment polarity (tonality), and emotions (see European Commission (2013) for a public version of newsbrief). We are in particular interested in the selection criteria for the articles and in three metadata fields: the timestamp, the language, and the sentiment polarity score (tonality in EMM parlance). The selection criteria depends on the article being published in one of a pre-determined set of sources (see European Commission (2013) for a list of sources) and on it being selected by the system as belonging to the "Bitcoin" class or category. This is detected automatically based on the presence of a set of criteria and keywords determined by the user of the system. The criteria for the category were developed by subject matter experts in the Commission in order to monitor developments related to Bitcoin and Crypto currencies (see Steinberger, Pouliquen, and van der Goot (2009) for an introduction to the system and an explanation on the creation of categories). After an article is selected, the system time-stamps it with the time of selection (sources are checked at different frequencies and times of the day depending on frequency of updates, times of major updates, and importance), determines its language and records it in metadata, and starts extraction and tagging based on an analysis of the full text. EMM can apply several algorithms for the detection of sentiment, emotions and sentiment polarity/tonality, either in general terms for the whole article, or referring to a named entity contained in it. In the current application we are interested in using article going back in time for a rather long period, so we rely on the "JRC tonality" approach described in Balahur, Steinberger, Kabadjov, Zavarella, van der Goot, Halkia, Pouliquen, and Belyaeva (2010). This algorithm uses a set of language-specific dictionaries to assign a score to a set of sentiment related words: +1 and -1 for slightly positive or negative terms, and +4 and -4 for strongly positive and negative terms. The scores are then aggregated by article and normalized by the number of words, to produce a numerical "Tonality" score. This score has already been used in the past to detect early signals of distress on European Systemic Banks (see Nardo and van der Goot (2014) and Nardo, Petracco-Giudici, and Naltsidis (2016) for a survey on the use of sentiment data in market prediction). In order to minimize the impact of noise and of possible detection errors, articles are then aggregated in 30-minutes windows, and an aggregate average tonality score is calculated for every time window.

Bitcoin price and trading volume are obtained from Kaiko. We focus the trading activities at six trading exchanges of which three are located in Asian (OkCoin in China, Bitfinex in Hong Kong, and Quoine in Japan), one in Europe (Bitstamp in Luxembourg), and two in US (Coinbase and Kraken). We choose these exchanges because they have long trading history to match with our EMM sample period, and high enough trading volume to alleviate liquidity issues.

3 Empirical Models

To formulate the empirical models, we average the tonality of news in English and other languages from the EMM dataset every hour (0:00-1:00am, 1:00-2:00am,...) based on UTC time. We exclude those hourly time blocks that contain news only in English or only in other languages and keep 66.65% of all time blocks during our sample period.

Table 1 presents the average tonality of all time blocks. The tonality of English news is generally lower than that of news in other language.

Table 1: Summary Statistics of Tonality

Hourly Block	TonalityEN	TonalityOT	Hourly Block	TonalityEN	TonalityOT
0	-2.638	-0.119	12	-1.186	0.357
1	-3.408	0.364	13	-1.182	0.385
2	-3.259	0.022	14	-1.592	0.337
3	-3.226	0.218	15	-1.945	0.251
4	-2.596	0.232	16	-1.661	0.257
5	-2.023	0.511	17	-2.152	0.294
6	-2.246	0.495	18	-2.285	0.026
7	-2.274	0.432	19	-2.134	-0.104
8	-1.839	0.445	20	-2.563	0.275
9	-1.626	0.458	21	-2.451	-0.061
10	-1.797	0.476	22	-2.868	0.033
11	-1.860	0.866	23	-3.029	0.139
			All	-2.164	0.292

Notes: Entries to the table are the average tonality of all hourly time blocks.

We implemented the following logistic regressions on a daily basis for each of the hourly time blocks separately (for example, the time between 8:00 am and 9:00 am every day):

$$\text{Sign}(\text{Ret}_{t+k}) = \text{Logistic}(\alpha + \beta_1 \text{TonalityEN}_t + \beta_2 \text{TonalityOT}_t + \gamma \text{Controls}_t + \epsilon_{t+1} \tag{3.1}$$

where Ret_{t+k} is the future return on Bitcoin with k = 30 mins, 60 mins, 1 day, and 7 days, starting from the end of the time interval t. TonalityEN_t is the average tonality of all English news over time period t. TonalityOT_t is the average of tonality of news published in other languages over the same time

Table 2: Summary Statistics of Return and NetBuy

Ret

	30 mins		60 mins		1 day		7 days	
	N	Mean	N	Mean	N	Mean	N	Mean
Bitfinex	30148	2.68E-05	30146	7.84E-05	30098	1.75E-03	29930	0.016
Bitstamp	30103	2.29E-05	30100	8.15E-05	30036	1.85E-03	29802	0.017
Coinbase	28479	9.80E-06	28463	6.81E-05	28360	2.22E-03	28237	0.018
Kraken	29439	3.12E-05	29369	5.56E-05	28989	2.02E-03	28795	0.017
OkCoin	26106	7.20E-05	26091	6.93E-05	25943	1.52E-03	25621	0.014
Quoine	27257	-8.95E-05	27167	6.27E-05	26681	2.39E-03	26519	0.019

NetBuy

	30 mins		60 mins		1 day		7 days	
	N	Mean	N	Mean	N	Mean	N	Mean
Bitfinex	29934	-0.004	30061	-0.009	30150	-0.020	30140	-0.023
Bitstamp	25893	0.074	25897	0.073	25976	0.054	26090	0.048
Coinbase	28363	0.084	28388	0.078	28638	0.068	28604	0.063
Kraken	27407	0.009	28041	0.003	30143	-0.011	30153	-0.017
OkCoin	25120	-0.055	25615	-0.056	26240	-0.068	26439	-0.061
Quoine	24189	-0.028	25379	-0.034	28347	-0.019	28977	-0.019

Notes: Entries are the statistics of Ret_{t+k} and $NetBuy_{t+k}$ for k=30 mins, 60 mins, 1 day and 7 days.

interval. The controls include a macroeconomic sentiment index constructed by the Federal Reserve Bank of San Francisco.[2] Other controls include dummies measuring the difference in tonality between English and other language news. Specifically, for each time block, Tp=1 if the average tonality of English news published in the block is more than one standard error higher than the average tonality of news published in other languages, and 0 otherwise. Tn=1 if the average tonality of English news is more than one standard error lower than the average tonality of news published in other languages, and 0 otherwise.

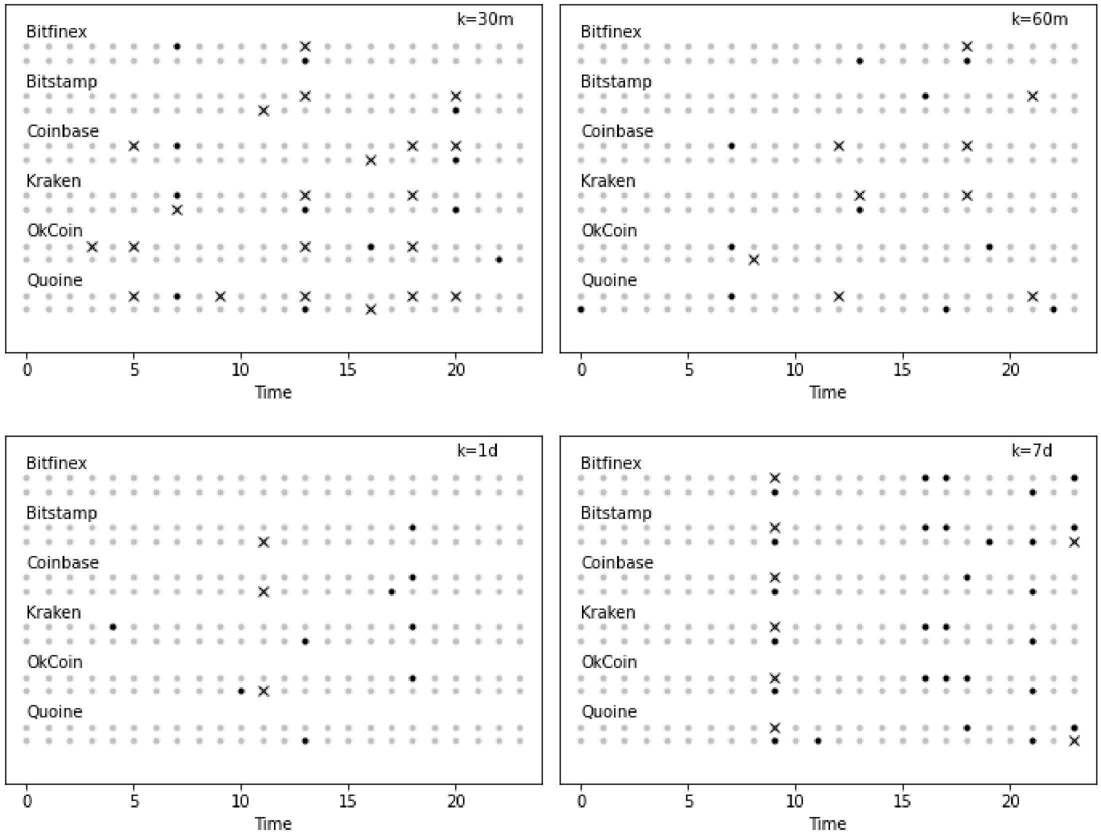
In addition to the sign of returns we also estimate net buy volume. In particular we define the variable NetBuy_t which is calculated as (Buying Volume - Selling Volume)/(Buying Volume + Selling $\text{Volume})_k$ for time k. Using net buy volume we run the following trading activity regressions for the same horizons k :

$$\text{NetBuy}_{t+k} = \alpha + \beta_1 \text{TonalityEN}_t + \beta_2 \text{TonalityOT}_t + \gamma \text{Controls}_t + \epsilon_{t+1} \tag{3.2}$$

Table 2 presents the summary statistics of Ret_{t+k} and $NetBuy_{t_k}$ at the six exchanges in our sample. We notice that return is highly correlated across the six exchanges. Comparing to return, $NetBuy$

[2] The Daily News Sentiment Index is a high frequency measure of economic sentiment based on lexical analysis of economics-related news articles. The index is described in Buckman, Shapiro, Sudhof, Wilson, et al. (2020) and based on the methodology developed in Shapiro, Sudhof, and Wilson (2020). For details, see `https://www.frbsf.org/economic-research/indicators-data/daily-news-sentiment-index/`.

Table 3: Tonality Predicts Future Returns.



The table presents results for the logistic regression appearing in equation (3.1) on a daily basis for each of the half-hourly time blocks separately. The controls include a macroeconomic sentiment index constructed by the Federal Reserve Bank of San Francisco, and dummies measuring the difference in tonality between English and other language news. The trading activities at six different exchanges are examined. Black crosses, black dots, and grey dots denote positive and significant, negative and significant, and insignificant β_1 (upper line) and β_2 (lower line) for each exchange. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

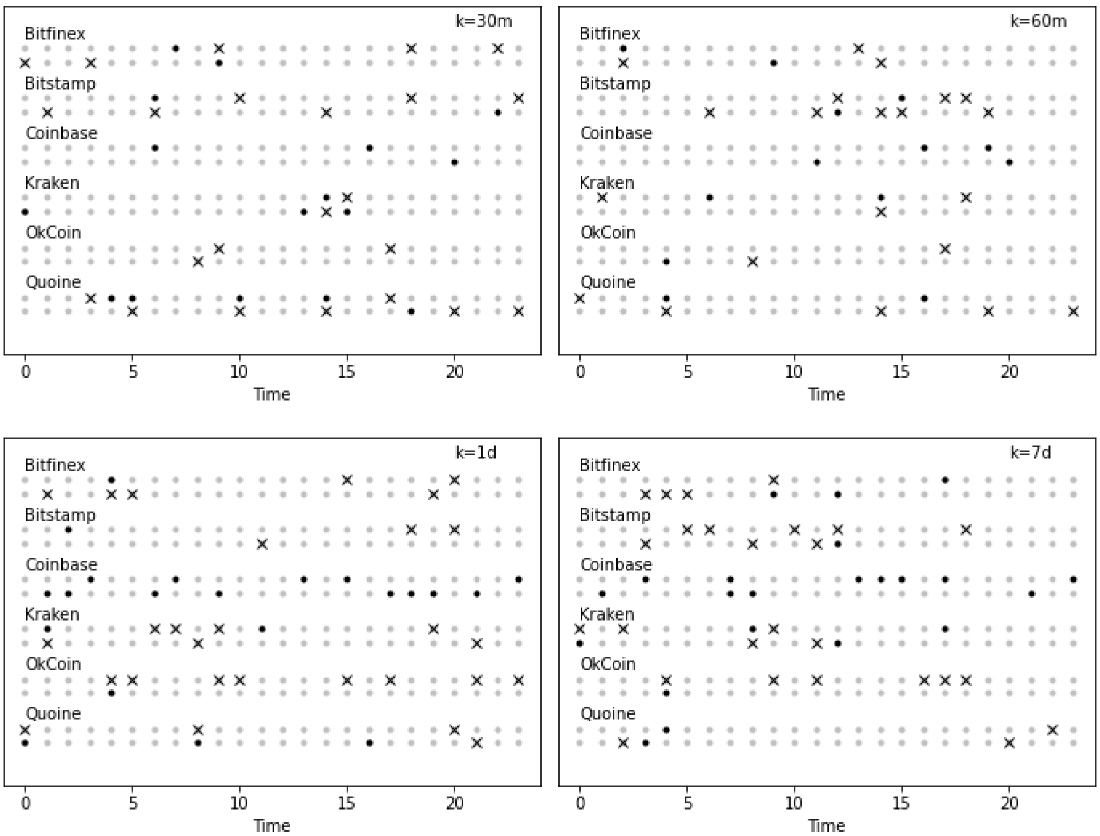
shows large variations.

4 Results

We implement regression (3.1) to explore the return-predictability of the news at six different exchanges and present the results in Table 3. Black crosses, black dots, and grey dots denote positive and significant, negative and significant, and insignificant β_1 (upper line) and β_2 (lower line) for each exchange. We observe more crosses in the panels for k=30m, 60m, and 1d, and more dots in the panel for k=7d. These results suggest that tonality is positively correlated with short-term returns but negatively with long-term return. This correlation pattern is similar across the exchanges we examined, consistent with the fact that return series is highly correlated across different exchanges (Shams (2020)).

We implement regression (3.2) to explore the implication of tonality for trading activities at six different exchanges and present the results in Table 4. In all four panels for k=30m, 60m, 1d and 7d, more crosses are observed. We conclude that tonality is positively correlated with net buy over both short and long horizon in the future. This is different from return-predictions where a long-term reversal is observed. A second difference is that this correlation pattern is different across different exchanges, suggesting some exchange-specific features in the trading activities.

Table 4: Tonality Predict Future Trading Activities.



The table presents results for the OLS regressions appearing in equation (3.2). The controls include a macroeconomic sentiment index constructed by the Federal Reserve Bank of San Francisco, and dummies measuring the difference in tonality between English and other language news. The trading activities at six different exchanges are examined. Black crosses, black dots, and grey dots denote positive and significant, negative and significant, and insignificant β_1 (upper line) and β_2 (lower line) for each exchange. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

References

BALAHUR, A., R. STEINBERGER, M. KABADJOV, V. ZAVARELLA, E. VAN DER GOOT, M. HALKIA, B. POULIQUEN, AND J. BELYAEVA (2010): "Sentiment Analysis in the News," in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, pp. 2216–2220, Valletta, Malta.

BUCKMAN, S. R., A. H. SHAPIRO, M. SUDHOF, D. J. WILSON, ET AL. (2020): "News sentiment in the time of COVID-19," *FRBSF Economic Letter*, 8, 1–05.

CORBET, S., A. MEEGAN, C. LARKIN, B. LUCEY, AND L. YAROVAYA (2018): "Exploring the dynamic relationships between cryptocurrencies and other financial assets," *Economics Letters*, 165, 28–34.

——— (2020): "Cryptocurrency reaction to fomc announcements: Evidence of heterogeneity based on blockchain stack position," *Journal of Financial Stability*, 46, 100706.

ENTROP, O., B. FRIJNS, AND M. SERUSET (2020): "The determinants of price discovery on bitcoin markets," *Journal of Futures Markets*, 40(5), 816–837.

EUROPEAN COMMISSION (2013): "Europe Media Monitor - NewsBrief," EU Science Hub - European Commission, https://ec.europa.eu/jrc/en/scientific-tool/europe-media-monitor-newsbrief.

GHYSELS, E., AND G. NGUYEN (2019): "Price discovery of a speculative asset: Evidence from a Bitcoin exchange," *Journal of Risk and Financial Management*, 12(4), 164.

GHYSELS, E., G. NGUYEN, D. SHIN, AND Z. WANG (2021): "Bloggers and Bitcoin Prices: A Textual Machine Learning Analysis," Discussion Paper UNC Chapel Hill.

NARDO, M., M. PETRACCO-GIUDICI, AND M. NALTSIDIS (2016): "Walking down Wall Street with a Tablet: A Survey of Stock Market Predictions Using the Web," *Journal of Economic Surveys*, 30(2), 356–369, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/joes.12102.

NARDO, M., AND E. VAN DER GOOT (2014): "Does web anticipate stocks?: analysis for a subset of systemically important banks.," Discussion Paper 27023, European Commission. Joint Research Centre., LU.

SHAMS, A. (2020): "The Structure of Cryptocurrency Returns," Fisher College of Business Working Paper No. 2020-03-011, Charles A. Dice Center Working Paper No. 2020-11, Available at SSRN: https://ssrn.com/abstract=3604322 or http://dx.doi.org/10.2139/ssrn.3604322.

SHAPIRO, A. H., M. SUDHOF, AND D. J. WILSON (2020): "Measuring news sentiment," *Journal of Econometrics*, forthcoming.

Steinberger, R., A. Podavini, A. Balahur, G. Jacquet, H. Tanev, J. Linge, M. Atkinson, M. Chinosi, V. Zavarella, and Y. Steiner (2015): "Observing Trends in Automated Multilingual Media Analysis," in *ACA, Zürich, Switzerland*, pp. 1–3.

Steinberger, R., B. Pouliquen, and E. van der Goot (2009): "An introduction to the Europe Media Monitor family of applications," in *Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009)*, pp. 1–8, Boston, MA, USA. arXiv: 1309.5290.