Cross-sectional asset pricing: Theory and estimation

The sentiment behind cryptocurrency return

A cross-sectional analysis

Group Number: 31
Student 1
Flow ID Number: 93
Student 2
Flow ID Number: 115

Supervisors
Jonas Nygaard Eriksen
Associate Professor, Aarhus University
Daniel Borup
Postdoctoral Researcher, Aarhus University

Aarhus University , Department of Economics and Buisness Economics

M.Sc. Oecon

The paper may be published

Hand-in date: 4/24/2019

Authors take equal responsibility of the report and its content.

Total number of characters: 34,563

Text: 32,963

Figures: 2 x 800 = 1,600

Contents

1 Introduction **3**

2 Data **5**

2.1 Financial data 5

2.2 Social media data 6

3 Methodology **7**

4 Empirical Analysis **8**

5 Limitations and extensions **13**

6 Conclusion **14**

References **15**

1 Introduction

Over the latter part of the recent decade, the cryptocurrency market has emerged and evolved as an investment opportunity. Each day investors attempt to anticipate how this speculative market will evolve. However, how is this young and volatile market interpreted in the setting of our classic perception of asset pricing, and can our established knowledge of characteristics previously shown to aid us in explaining the cross-section of equity returns also be of substantial use in the cryptocurrency market?

This paper seeks to examine this speculative market, by analyzing the cross-section of cryptocurrency returns formed on sentiment portfolios and further study the effect of a sentiment factor, in a formal asset pricing setting. Cryptocurrencies have emerged as a completely new type of asset, they are, in essence, a currency, but unlike their traditional counterparts, they are not backed by any legal entity, such as a government. This means the market operates as a free market void of any central banking system to issue or a commercial bank to store, thereby eradicating trading barriers and decentralizing currencies [Bouoiyour & Selmi, 2015].

These characteristics have spurred on an immense growth in the cryptocurrency market capitalization, which reached 750 billion USD in January of 2018; however, it has been somewhat steadily located for the past year around the 250 billion USD mark. Related to the scale of the established financial market this, of course, is not a sizable amount, but accompanied by the fact that cryptocurrencies were introduced as a potential substitute for our current monetary system makes the market important to understand [Galeshchuk *et al.* , 2018]. The opinions on cryptocurrencies are many, including severe concerns regarding the system as being an environment for money laundering and organized crime [Kristoufek, 2015], or the fact that with no intrinsic value the evolution of the prices of coins, such as the well-known and disputed Bitcoin, constitutes solely on speculative bubbles [Bouoiyour & Selmi, 2015]. However, despite these concerns, the possibility remains that cryptocurrencies, and the technology they are a product of, will play an integral part in the future evolution of our investment markets [Liu *et al.* , 2019]. For this reason, many have already examined the cryptocurrency market using a standard empirical asset pricing approach. Liu et al. (2020) find nine cryptocurrency factors that form successful long-short strategies with statistically significant excess return and show that all of these strategies can be accounted for by their cryptocurrency three-factor model. A model containing factors for the cryptocurrency market, size and momentum, and analyzed similarly to the analysis conducted by Fama and French [Fama & French, n.d.].Their analysis used factors only constructed on price and market information of all cryptocurrencies with a market capitalization greater than 1 million USD from 2014 to 2018. Their results show that the cross-section of cryptocurrencies can be meaningfully analyzed using standard asset pricing tools.

[Gregoriou, 2019] finds that investors can obtain abnormal returns when trading cryptocurrencies on the London Stock Exchange, after accounting for systematic size value and momentum using Fama-French inspired models. Further, [Borri & Shakhnov, 2019] utilize a Fama-MacBeth regression framework to examine the price differences observed in the cryptocurrency market and find a two-factor model to explain a large fraction of the cross-sectional variation in portfolio returns. [Liu & Tsyvinski, 2018] and [Ling & Zhu, 2019] both examine the examine cryptocurrency returns using three-

factor models, and include in their work what they respectively refer to as 'investor attention' and 'network hype', by using data on the number of Google searches as an expression for this type of investor based factor.

We view investor sentiment as a natural 'next step' to this type of factor and utilize actual statements in the form of Reddit comments on a number of cryptocurrencies to extract sentiment and analyze returns in the light of the current atmosphere surrounding cryptocurrencies. In conducting this empirical asset pricing analysis, we seek to expand some of the previously referred work [Liu *et al.* , 2019]. We explore the effect of investor sentiment as a relevant factor in the cryptocurrency market. The idea of relating sentiment to asset pricing is however not unique to the cryptocurrency market, and has been studied before any cryptocurrencies even existed, in works such as [Neal & Wheatley, 1998], [Shiller, 2000], [Baker & Wurgler, 2000], and again in [Baker & Wurgler, 2006].

As this is an unprecedented market it is, by nature, subject to much speculation from investors, and multiple papers show significant results in linking returns of cryptocurrencies and investor sentiment. It has been shown that investors' interest in the cryptocurrency drives the price of Bitcoin and that in times of explosive prices this interest drives prices further up and vice versa. Prior to this, Ladislav Kristoufek has also shown that bubble and bust cycles of Bitcoin can partially be explained by the interest in the currency [Kristoufek, 2013]. [Galeshchuk *et al.* , 2018] tested the hypothesis that the exchange rate of cryptocurrencies depends on behavioral signals as opposed to any fundamental conditions, that is, that investor sentiment may cause the market shocks which can not be explained by the efficient market hypothesis. They find a significant influence of Twitter signals on Bitcoin fluctuations, and that including the sentiment score much improved the prediction accuracy for Bitcoin directional changes. Importantly, this relation is not only found concerning Bitcoin, [Li *et al.* , 2019] show that social media platforms can serve as powerful signals for predicting price movements for alternative cryptocurrencies. Other papers have established likewise relations between cryptocurrency prices and investor sentiment [Phillips & Gorse, 2017, Kim *et al.* , 2016, Garcia & Schweitzer, 2015]. From these relations, it can be concluded that cryptocurrencies form a unique asset that possesses properties of both a standard financial asset and a speculative one [Kristoufek, 2015]. We conduct an analysis by incorporating this relation into our asset pricing setting in order to address the versatility of established asset pricing theory by adhering to the intriguing new relations found in the cryptocurrency market. Our sentiment analysis is based on comments extracted from the forum website "Reddit.com", from which we have extracted all comments containing the respective coin's name or abbreviation. A relation between the amounts of daily comments and cryptocurrency price can be seen in Appendix Figure A.1.

This paper examines the cryptocurrency market with an outset in traditional empirical asset pricing, to assess potential similarities with the asset classes for which our theory is based, namely equities. More specifically, we investigate the potential role sentiment plays in this market by analyzing the significance of the excess returns on a zero-investment strategy based on daily sentiment sorted portfolios. We then examine if this cross-sectional cryptocurrency return predictor can be explained using a low number of common factors, utilizing the Fama-Macbeth regression setup and related factors [Fama & MacBeth, 1973], [Fama & French, 1992], [Fama & French, n.d.]. Finally, we construct a sentiment factor and test its relevance in risk pricing when controlling for established factors. Portfolios and common factors are constructed

using data from 01.01.14-29.30.19, consisting of twelve cryptocurrencies which historically have constituted at least 80% of the entire market capitalization. Computational restrictions have resulted in the use of a relatively low amount of coins, which entails that the constructed portfolios likely suffer from a bit of noise and idiosyncratic risk. We have opted to use quartile portfolios to balance this noise while still obtaining a proper amount of cross-sectional variability in portfolio returns.

The structure of this paper is as follows; Section 2 will go over the data used to conduct our analysis, how it has been obtained, and processed, along with descriptive properties and summary statistics. Section 3 covers the methodology used to analyze our data; this includes a thorough description of the Fama-MacBeth regression method, which will be the main driver behind our analysis. Section 4 will then contain our empirical analysis findings; this includes a discussion of results and what can be inferred about the cryptocurrency market in terms of both the incorporation of sentiment and how our results align with established asset pricing theory. The fifth section of the paper is then a short discussion about the limitations of our analysis as well as potentially relevant extensions to our work. Lastly is a conclusion to the paper. Additional tables, graphs, and auxiliary results can be found in the Appendix and will be referenced throughout when relevant.

2 Data

This section will provide an insight into the data together with their source and the prepossessing. The data-set created for this paper consist of two elements, cryptocurrency financial data, and social sentiment data. The study starts in January 2014, as limited trading and liquidity and the number of active cryptocurrencies beforehand, and ends October 2019. A cryptocurrency enters the sample at its initial coin offering. We group the data on both weekly and daily granularity to study how the changes in frequency affect the studied factor(s). Data collecting, cleaning, and estimation are all handled in Python 3.8.1. For specifics see enclosed .py scripts and reference citations for the major external libraries.[1]

2.1 Financial data

The financial data is collected using cryptocompare.com's REST API. For each day, we collect closing price and market capitalization in U.S Dollars for each cryptocurrency. The reported price is the volume-weighted average of all major exchanges. The sample comprises of twelve cryptocurrencies based on their market capitalization and relevance for asset pricing study. See Appendix Table A.1 for the full list. [2] We compute daily returns by the relative difference in the closing price of the previous day. Furthermore, a market index is built by value weighting prices for the cryptocurrencies for each day. The excess return of cryptocurrency markets, CMKT, is the difference in the returns of the market and the risk-free rate from Kenneth French's website with weekly/daily frequency which is the one-month Treasury bill rate, French2020.

[1] Fundamental packages like Matplotlib for plotting or Pandas for data handling and like can be found in scripts.
[2] Cryptocurrency similar to Tether or Binance USD with a value seeking to match the U.S. Dollar is excluded as the nature of the currency makes it unsuited for this analysis.

A summary of descriptive statistics of the financial data is shown in Appendix Table A.2.

2.2 Social media data

To create a sentiment measure, we use user comments from Reddit.com, which is a social news aggregation and forum website. Here registered users can submit content into moderated "subreddits" which is topic-specific forums; afterward, other users can reply with comments in addition to up- or down vote submission depending on subjective relevance or opinion of the matter. Reddit.com is the fifth most visited website in the U.S and 13th in the world. The subject sectioned forums make Reddit ideal for sentiment analysis as comments can be filtered conveniently to be relevant.

Raw comments are queried from Big Query, which is part of Google's Cloud platform. A comment is deemed relevant to a cryptocurrency if it contains its name or abbreviation, fx., Bitcoin or BTC. We then proceed to filter and clean the data in three steps to improve the performance of the natural language processor; Deleting irrelevant signs and formatting. Apply language filtering by use of Google's language detection to detect non-English comments in addition to deleting comments from known foreign subreddits.[Shuyo, 2010] Finally all comments from users flagged as bots is filtered out. In certain cases we filter out subreddits if it is deemed irrelevant and account for a substantial amount of comments for a search term. [3]

The sentiment is derived from cleaned comment data using the Valence Aware Dictionary and sEntiment Reasoner, VADER, which is a lexicon for rule-based sentiment analysis designed for social media introduced by C.J. Hutto and Eric Gilbert [Hutto & Gilbert, 2015]. Each comment is classified as either positive, negative, or neutral in sentiment by assigning words and phrases a polarity score and a final score. Additions have been made to the VADER lexicon to enhance its ability to detect cryptocurrency and trading by manually adding related sentiment terms and using the Loughran-McDonald financial broad sentiment corpus[4]. Polarity score is then translated into a sentiment using a decision rule of a comment being positive if the polarity is above 0.05, negative if below -0.05 and neutral if between the breakpoints. Finally are comments grouped with sum by date to match price observations.

Noise and quality in sentiment analysis are critical factors and must be addressed for the study to be valid. Investors are capable of distinguishing between comments from bots and sarcasm, but this is not features the algorithm possesses. A conscious choice weighting automated filters and quality of sentiment has been made.

We create four series based on sentiments scores, share of positive comments relative to the total comments of a specific day, share of negative comments relative to total comments, and two additional where the share is divided with the rolling mean with varying lengths. The reason for these measures is both to test if the current sentiment has an effect, and the daily/weekly change to the sentiment compared to prior periods seems relevant to study.

[3] For example, EOS is both a cryptocurrency and the abbreviation of a popular mirrorless camera from Canon.
[4] For example, authors has manually added "bearish" as a negative and "hodl" as positive.

3 Methodology

The following section will be an overview of the methods used in the empirical analysis. It will be centered around the Fama-MacBeth [Fama & MacBeth, 1973] two-pass cross-sectional regression method, which will be thoroughly explained. As mentioned in the introduction to this paper, this methodology is heavily implanted in the empirical asset pricing literature and has been used in a multiple of the studies motivating this analysis.

The method is set up as a two-step procedure, first off consisting of a time series regression for which the estimated coefficients are saved. These are then used in the second step, which is a series of cross-sectional regressions in which the saved coefficients are used as explanatory variables for regressions at each time step. The coefficients on these are then averaged over all time periods, and their significance can be tested using standard t-test. This approach is ideal for testing and evaluating linear asset pricing models in a cross-sectional dimension.

For the methodical representation of the regressions run in this paper, an example of a model with factors for the cryptocurrency market and sentiment, looks as follows:

$$E\left[\tilde{r}_i\right] - r_f = \gamma_0 + \gamma_{MKT}\beta_{i,MKT} + \gamma_{SENT}\beta_{i,SENT} \tag{1}$$

Thus, as the betas are not directly observable, they need to be estimated first, this is the first-pass regressions. So, we estimate betas for each asset $I = 1, .., N$, following the argument of Cochrane [Cochrane, 2009] by conducting a full sample time series regression:

$$\tilde{r}_{t,i} - \tilde{r}_{t,f} = \alpha_i + \beta_{iMKT}\left(\tilde{r}_{t,MKT} - \tilde{r}_{t,f}\right) + \beta_{i,SENT}\left(\tilde{r}_{t,SENT} - \tilde{r}_{t,f}\right) + \tilde{\varepsilon}_{t,i} \tag{2}$$

With $\tilde{\varepsilon}_{t,i}$ being a zero-mean error term. Saving the estimated beta coefficients from these regressions then brings on the next step, the second-pass regressions. For the second-pass regression we run cross-sectional regressions for each time period $t = 1, \ldots, T$, to estimate consistent standard errors in the presence of cross-sectional correlation.

$$\tilde{r}_{t,i} - \tilde{r}_{t,f} = \gamma_{t,0} + \gamma_{t,MKT}\hat{\beta}_{i,MKT} + \gamma_{t,SENT}\hat{\beta}_{i,SENT} + \nu_{t,i} \tag{3}$$

With $\nu_{t,i}$ being a zero-mean error term and the set of independent variables are the estimated betas from the first-pass regression. These T cross-sectional regressions result in a time series of estimates for the gamma coefficients. Assuming that each time period is independent, such that $\{\hat{\gamma}_{t,0}, \hat{\gamma}_{t,MKT}, \hat{\gamma}_{t,SMB}\}$ constitute an IID time series. We can then form estimates using the time series average of each coefficient, these are what we define as the risk premiums.

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=1}^{T}\hat{\gamma}_{t,j}, \;\; j = \{\hat{\gamma}_{t,0}, \hat{\gamma}_{t,MKT}, \hat{\gamma}_{t,SENT}\} \tag{4}$$

The standard errors can then be computed in a similar way, using the standard deviations from the cross-sectional

regression estimates to generate the sampling errors for the estimates, remembering to use $\frac{1}{T^2}$ as we are finding standard errors of sample means, $\frac{\sigma^2}{T}$.

$$\sigma^2\left(\hat{\gamma}_j\right) = \frac{1}{T^2}\sum_{t=1}^{T}\left[\hat{\gamma}_{t,j} - \hat{\gamma}_j\right]^2,\ j = \{\hat{\gamma}_{t,0}, \hat{\gamma}_{t,MKT}, \hat{\gamma}_{t,SENT}\} \tag{5}$$

Extracting these mean that we have the necessities for testing and inference. Testing can be conducted using a standard t-test with $T - K$ degrees of freedom, K being the amount of factors in the model. Described above is the standard Fama-MacBeth two-pass cross-sectional regression methodology, however one main flaw occurs which we seek to handle. The flaw being the introduction of the Errors-In-Variables problem [Shanken, 2015], stemming from the fact that we use estimated beta values as independent variables in the cross-sectional regressions. Fama and MacBeth suggested using portfolios instead of individual assets to handle this problem, as it can be expected that some of the individual noise can be mitigated as it is 'averaged-out', reducing the measurement error in the estimated betas ($\hat{\beta}_{i,MKT}, \hat{\beta}_{i,SMB}$). However, as we are in a case where our computational restrictions do not allow us to construct portfolios of a size large enough to assume that the noise is simply averaged-out, we seek to handle this problem actively. This is done through a correction provided by Shanken [Shanken, 2015], wherein we compute the errors-in-variables corrected covariance matrix related to the risk factors, utilizing the Fama-MacBeth estimation and t-by-t-cross-sectional approach, with c being a scaling factor:

$$Var_{EIV}\left[\hat{\gamma}^f\right] = \frac{(1+c)\left(T \cdot Var\left[\hat{\gamma}^f\right] - Var\left[f\right]\right) + Var\left[f\right]}{T}, \quad c = \hat{\gamma}^f \cdot Var\left[f\right]\hat{\gamma}^f \tag{6}$$

$VAR\left[\hat{\gamma}^f\right]$ denotes the part of the covariance matrix of the risk prices related to the risk factors, excluding the constant. $Var\left[f\right]$ denotes the variances of the factors themselves. $Var\left[\hat{\gamma}^f\right]$ thus denotes the error-in-variables corrected covariance matrix related to the risk factors. The empirical analysis will report both the standard Fama-MacBeth standard errors and t-statistcs, along with the Shanken corrected values.

4 Empirical Analysis

In the following section, we conduct cross-sectional asset pricing tests on our sentiment-based cryptocurrency return predictors, to see if they can be spanned by a single or two factors. The analysis is conducted as inspired by Fama and French, and follows the procedures specified in the Methodology section. Initially we test a single factor model, with the factor being the coin market returns. We find that the cryptocurrency CAPM fails in accounting for the excess returns of the specified strategies. Following this result, we test two-factor models, utilizing factors accounting for market returns, size, momentum as control for our proposed sentiment factor. We find that sentiment is priced in all cases when running the regression on sorted portfolios and that no model is supports the single factor CAPM.

To examine the cross-sectional relationship of the returns of sentiment-based investing strategy, we utilize portfolio

sorting of the cryptocurrencies. We construct portfolios based on each of the previously defined sentiment factors and analyze the resulting returns. Portfolios are constructed using breakpoints determined by evenly spaced percentiles, we opt for quartiles, therein resulting in four sentiment portfolios. This choice is made to accommodate the trade-off between the number of assets in each portfolio and the cross-sectional variation in the expected returns that can be identified using our sentiment sorting factors. Smaller portfolios will likely have noisy returns with a potentially high amount of idiosyncratic risk, while larger portfolios reduce the cross-sectional variability in portfolio returns, potentially hindering the detection of cross-sectional relationships. With twelve coins at our disposal, we opt for the construction of four portfolios, keeping in mind that with only three currencies in each portfolio, we are heavily exposed to the former of the risks. The formation of the portfolios is done by allocating each cryptocurrency into the appertaining portfolio according to the sorting factor.

Each day (week), each currency is sorted into its relevant portfolio according to the day's quartile breakpoints calculated from the sentiment factor. The first portfolio then contains the currencies with the lowest level of the sorting factor, and the fourth portfolio contains the highest. We then calculate value-weighted returns for each portfolio, weighted according to market capitalization. The returns are weighted according to the market capitalization of the previous day, as to mimic the use of such an investment strategy. Specifically, in time t you observe the sentiment and invest in a portfolio weighted by the current-day market capitalization, which means that you obtain this value-weighted return the next day.

$$r_{k,t} = \frac{\sum_{i=1}^{N_{k,t}} r_{i,t} \cdot MCAP_{i,t-1}}{\sum_{i=1}^{N_{k,t}} MCAP_{i,t-1}} \tag{7}$$

From these portfolios, we construct a zero-cost long-short portfolio (spread portfolio), taking a long position in the fourth portfolio and a short position in the first portfolio. To assess a potential cross-sectional relationship between future cryptocurrency returns and our different sentiment sorting factors, we test the significance of the average portfolio returns by regressing on a constant. Return statistics of these portfolios can be seen in Table 1 below, and descriptive statistics can be found in Appendix Table A.5.

Table 1: Portfolio returns

*Stars indicate degrees of significance. * indicates significance on a 10% significance level against a two-sided alternative hypothesis. ** indicates significance on a 5% significance level against a two-sided alternative hypothesis. *** indicates significance on a 1% significance level against a two-sided alternative hypothesis.*

	Quartile portfolios				
	P1	P2	P3	P4	Spread
Share Positive	Low			High	
$Mean$	0.0025^{**}	0.0028^{**}	0.0031^{**}	0.0053^{***}	0.0029^{**}
$t\,(Mean)$	(2.20)	(2.35)	(2.08)	(3.44)	(1.97)
Share Negative	Low			High	
$Mean$	0.0053^{***}	0.0015	0.0033^{**}	0.0027^{**}	-0.0026^{*}
$t\,(Mean)$	(3.38)	(1.13)	(2.5)	(2.37)	(-1.68)
Mom. Share Positive	Low			High	
$Mean$	0.0020	0.0031^{**}	0.0038^{*}	0.0056^{***}	0.0036^{**}
$t\,(Mean)$	(1.57)	(2.45)	(2.78)	(3.39)	(2.05)
Mom. Share Negative	Low			High	
$Mean$	0.0023^{*}	0.0040^{***}	0.0018	0.0055^{***}	0.0032^{*}
$t\,(Mean)$	(1.76)	(3.13)	(1.43)	(3.27)	(1.84)

As the result shows, all of the spread portfolios are significant at a 10% significance level, with the two factors based on positive sentiment also significant on a 5% significance level. The significant spread ranges from -0.0026 to 0.0036. For further empirical analysis, we will focus on the "Mom. Share Positive" factor, as this factor has shown the largest spread returns, whilst also having the largest t-statistic of the spread portfolios. The spread portfolios returns will be refereed to as the $SENT$ factor. The positive and significant result indicates that the current values of the sentiment sorting factor is informative about future cross-sectional differences in returns. Further, there can be seen monotonicity in the average return relationships, which moreover indicate an ability in our sentiment sorting factor(s) to explain cross-section of returns.

We likewise test the same sorting scheme for weekly grouped data; this leads to contrasting results. None of the sentiment sorted portfolios has significant results when testing. The results are shown in Appendix Table A.6. This indicates that the value of knowing the sentiment is heavily dependent on at what time the market atmosphere is studied as the information depreciates rather quickly. Due to the insignificant results, we will not continue with weekly data for the remainder of the analysis.

Following these initial results, we seek to establish whether the observed patterns and significance of the average portfolio returns can be traced to cross-sectional variation in portfolio sensitivities to systematic risk factors, as inspired by [Fama & French, 1992] on traditional assets and [Liu *et al.* , 2019] on the cryptocurrency market. We construct a number of different factors to include in this analysis. The cryptocurrency market excess returns factor, CMKT, as alluded to in section 2, is constructed as the difference between the value-weighted return of all cryptocurrencies and the risk-free rate. We measure size using the market capitalization, and momentum as the rolling sum of returns from the previous 7 days[5]. Results can be found in the appendix. The size, CSMB, and momentum, CMOM, factors are constructed

[5]Analysis has also been conducted for 3, 5 and 7 day horizons. 7 days has been chosen as this generates the largest long-short spread in the data, following the procedure seen in [Liu *et al.* , 2019]

in the same manner as Fama & French and Jegadeesh & Titman [Fama & French, 1992][Jegadeesh & Titman, 1993]. For the size (momentum) factor, the currencies are split into three groups according to market capitalization (momentum); bottom 30 percent (Small), middle 40 percent (Mid), and top 30 percent (Big). We then calculate the appertaining value-weighted returns for each portfolio, and define the size (momentum) factor, as the return difference between Small and Big portfolios.

As an initial test we run the single factor cryptocurrency CAPM.

$$\tilde{r}_{t,i} - \tilde{r}_{t,f} = \alpha_i + \beta_{iMKT} \left(\tilde{r}_{t,MKT} - \tilde{r}_{t,f} \right) + \tilde{\varepsilon}_{t,i}$$

The results in the Panel A of Table 2, from running the cryptocurrency CAPM, show that the model does actually live up to some of the expectations. First of all, the constant, γ_0, is both economically small and insignificant, which is in line with the standard CAPM assumptions. Secondly, whilst γ_{MKT} is insignificant, it does retain a positive sign, indicating the positive risk-reward trade-off that the CAPM implies. This being said, the market gamma is not significant, especially after the Shanken correction, where the t-statistic reduces from 1.40 to 1.15, and the model can be seen to only explain approximately 36% of the variation in the excess return on the four sentiment sorted portfolios. There is also no monotonic relationship to be seen in the first stage time series betas, where we would have like to observe a positive relationship between the betas (risk) and the expected returns from Table 1. The sentiment strategy does not seem to be exposed to the coin market returns. Thus, we conclude that the market excess returns factor is likely not the sole source of risk relevant to investors; thus, the cryptocurrency CAPM cannot be considered a sufficient model. The appertaining pricing error plot can be found in Appendix Figure A.4.

We now add a sentiment factor to test the effect it has on the market factor from the previous regression. Results of the two-factor model are depicted in Table 2. Starting with panel B of the table with MKT and $SENT$ factors. When studying the time series results on the left side, we do not see a pattern of monotonicity for either in α or β_{MKT}. β_{SENT} has a clear increase from P1 to P4. In the second stage, we see that the properties of γ_0 remains in favor of CAPM, yet the inclusion of SENT has further reduced the significance of the market excess returns. This is a clear divergence from the CAPM hypothesis of MKT being the only priced risk in the market. On the other hand, the price of risk is significant for the SENT factor and of economic magnitude. Together the two factors explain 99.36% of the variation, which is impressing, but it is essential to underline that the size is likely attributed to the low number of portfolios when running a multi-factor model. In Panel C, we keep the sentiment factor and replace the market factor with the momentum factor to see if sentiment remains priced. The left-hand side of the table remains troubling, with non-monotonic relations regarding α and β_{MOM}, but monotonicity remains for β_{SENT}. As for the cross-sectional part, γ_0 is insignificant, and sentiment is still priced when including the momentum portfolio's excess returns. Specifically the momentum factor is quite large economically, but fails to be statistically significant. The model achieves a slightly lower R^2 than the first two-factor model. Appendix Table A.7 contains the concurrent results when running a two-factor model with a size factor

and a sentiment factor.

Worth noting is that the γ_{SENT} becomes insignificant, in all tested relations, once the standard errors are corrected for the errors in variables problem. This indicates that our Fama-Macbeth regressions suffer from errors in variables, implying the potential for our estimated risk premiums to be downward biased, and the intercepts to be upwards biased.

Table 2: Two-factor cryptocurrency

These tables show results from running a Fama-MacBeth two-pass cross-sectional regression on the cryptocurrency factor models. Panel A is the cryptocurrency CAPM/MKT factor, panel B is MKT and SENT and panel C is CMOM & SENT. Test assets are 4 sentiment sorted portfolios. The right table reports estimated coefficients for both intercept and market beta for time-series regression on the entire sample. Newey and West [Newey & West, 1987] standard errors are in parentheses, with lag chosen as $T^{\frac{1}{4}}$ [Wooldridge, 2006]. R^2 is reported in percent. The left table shows the results from the cross-sectional regressions, coefficients are reported alongside standard-errors and t-statistics in parantheses, and Shanken [Shanken, 2015] corrected reported in hard brackets. The cross-sectionalR^2 is also reported.

Panel A:

	P1	P2	P3	P4
α	−0.0011	−0.0001	0.0007	0.0025
	(−0.79)	(−0.08)	(0.48)	(1.42)
β_{MKT}	0.0087	−0.0271	0.0431	0.0424
	(0.23)	(−0.81)	(1.13)	(0.98)
R^2 (%)	0.30	3.00	7.00	4.00

	γ_0	γ_{MKT}	R^2 (%)
Coefficient	0.00001	0.02695	
s.e	(0.001)	(0.019)	
	[0.001]	[0.024]	36.42
t-stat	(0.012)	(1.40)	
	[0.012]	[1.15]	

Panel B:

	P1	P2	P3	P4
α	0.0001	−0.0002	0.0005	0.0001
	(0.06)	(−0.12)	(0.38)	(0.06)
β_{MKT}	0.0198	−0.0276	0.0418	0.0198
	(0.59)	(−0.82)	(1.10)	(0.59)
β_{SENT}	−0.3277	0.0144	0.0387	0.6723
	(−4.76)	(0.89)	(0.81)	(9.76)
R^2 (%)	19.04	0.07	0.32	49.76

	γ_0	γ_{MKT}	γ_{SENT}	R^2 (%)
Coefficient	0.00002	0.00730	0.00349	
s.e	(0.001)	(0.020)	(0.002)	
	[0.001]	[0.021]	[0.002]	99.36
t-stat	(0.0194)	(0.3582)	(1.9937)	
	[0.0194]	[0.3515]	[1.4054]	

Panel C:

	P1	P2	P3	P4
α	−0.0007	0.0000	−0.0004	−0.0007
	(−0.55)	(−0.01)	(−0.34)	(−0.55)
β_{MOM}	0.1220	−0.0161	0.1426	0.1220
	(2.34)	(−0.47)	(2.33)	(2.34)
β_{SENT}	−0.3547	0.0178	0.0073	0.6453
	(−7.54)	(0.99)	(0.23)	(13.71)
R^2 (%)	21.30	0.08	3.14	51.16

	γ_0	γ_{CMOM}	γ_{SENT}	R^2 (%)
Coefficient	−0.00011	0.00325	0.00348	
s.e	(0.001)	(0.008)	(0.002)	
	[0.001]	[0.008]	[0.002]	98.79
t-stat	(−0.0935)	(0.4011)	(1.9916)	
	[−0.0935]	[0.3924]	[1.4141]	

Pricing error plots for each of these models can be found in Appendix Figure A.4 A.5A.6A.7. These gauge the performance of each model, as they plot the realized average excess returns of the sentiment portfolios against the model implied excess returns - as a graphical assessment of each model's ability to reproduce the cross-sectional variation in excess returns. All two-factor models display values along the 45°, indicating that the model's variation in excess returns are almost identical to the realized average excess returns. This does not hold true for the cryptocurrency CAPM. These relations are also implied by the R^2 values found in the cross-sectional regressions.

In summation, the results show that the sentiment factor derived from Reddit.com user comments is significantly positively priced in the cryptocurrency market, and the cryptocurrency CAPM is rejected.

5 Limitations and extensions

Whilst the empirical analysis does result in interesting and significant results, it is of considerable importance to emphasize some of the potential underlying pitfalls that could be impacting the results. This short section accentuates some of the limitations of the analysis along with potential future extensions that could handle these. The limitations regarding the analysis reside mainly in the sentiment data being used. As we seek to determine the sentiment of investors by utilizing Reddit.com comments, it limits the size of our data notably. We are restricted to only being able to use cryptocurrencies which are actively discussed on the website, and even though the website is the 13th most visited website worldwide, the fact that the cryptocurrency markets are still very much in a developing state, means that it is only in 2019 that we have all 12 currencies in the portfolios. This limitation comes with a further complication regarding the comparison of absolute amounts of positive and negative comments. The analysis uses only definitions of sentiment determined relatively within each currency, meaning that the absolute amounts of comments are not considered. In essence, this means that, for e.g., the "Share Positive" definition, a coin with two positive comments, and a single negative would be classified as having a more positive atmosphere than say a coin with 2000 positive comments and 1200 negative comments. Ignoring the potential value of amounts of comments has thus been a forced limitation because of the formerly stated problem of actively discussed coins on the website. Descriptive statistics of the sentiment data, accentuating the size differences in the discussion of currencies, can be found in Appendix Table A.3.A further potential flaw in our dataset is also the fact that the cryptocurrencies are selected from the size of their market capitalization, which means that there is a possibility of survivorship bias as we only handle coins that are traded today.. [Liu *et al.* , 2019] alleviated this potential bias by conducting their analysis on all cryptocurrencies available from Coinmarketcap.com, which include both active and defunct cryptocurrencies. However, our introduction of investor sentiment prevents us from doing the same. Such problems are difficult to handle as of now, but future extensions could be made into this field once the cryptocurrency market has saturated, and more currencies can be used to retrieve sentiment data. Alternatively, another source of sentiment could be Twitter.com, which has also been shown to be of use when conducting sentiment analysis on cryptocurrencies [Galeshchuk *et al.* , 2018], or even perhaps a combination of Reddit and Twitter sentiment in order to enlarge the data foundation; this, however, comes with associated weighting problems. It could also be of interest to conduct a similar analysis using other types of Natural Language Processors, such as the TextBlob or Pattern libraries for Python. With a larger dataset, an interesting extension would be to examine this field with different definitions of investor sentiment and incorporating the potential value of absolute amounts of comments to establish better definitions of sentiment or even an alternative for a Size factor.

Another characteristic of our data that could have been informative to look into is the possibility of herding behavior

in the cryptocurrency market. Herding entails investor decisions on individual assets based on the action of others in the market, or even basing their investment decision on a particular asset on the performance of another larger asset, which in our case could be Bitcoin [Vidal-Tomas *et al.* , 2019], [Bouri *et al.* , 2019], [Murray Leclair, 2018]. Herding could have an influence in that it has been shown to lead to pricing instability resulting in excess volatility, miss-pricing, bubble formation, and market crashes[Hwang & Salmon, 2004]. Likewise, overconfidence could also be a dominating factor in our data set which we do not take into account, as the cryptocurrency market's degree of volatility and trading activity behavior, especially in the beginning of 2018, has been shown to have a high degree of similarity to the behavior observed during the Dot-Com bubble, following the same psychologically-based market cycle [Tran, 2019]. Investor overconfidence has been shown to affect market properties such as market volatility, price distortion, and trading volume [Yeh & Yang, 2011].

6 Conclusion

This paper studies the effects of social media sentiment in theoretic pricing of cryptocurrency. The sentiments factor is constructed by categorizing social media comments as positive, neutral, or negative using a lexicon-based method. We display significant empirical results in both portfolio sorting and cross-sectional asset pricing test using excess returns of sentiment sorted portfolios. In particular, we construct a sentiment factor that relates the current share of positive comments to the past seven days. We find that this factor is positively priced in the market for all tested models. Furthermore, results implicate that the CAPM is rejected in all tested cases. We underline that the sentiment risk prices cease to be significant when using Shanken corrected standard errors.

It is paramount to remark that results are based on a low number of assets due to the difficulty in constructing the data set. Here we propose an extension to be an expansion of the data set by including more assets and sentiment factors to be constructed using numerous sources of social media data and more sophisticated natural language processors.

References

[Baker & Wurgler, 2000] Baker, Malcolm, & Wurgler, Jeffrey. 2000. The equity share in new issues and aggregate stock returns. *the Journal of Finance*, **55**(5), 2219–2257.

[Baker & Wurgler, 2006] Baker, Malcolm, & Wurgler, Jeffrey. 2006. Investor sentiment and the crossâsection of stock returns. *The journal of Finance*, **61**(4), 1645–1680.

[Borri & Shakhnov, 2019] Borri, Nicola, & Shakhnov, Kirill. 2019. The cross-section of cryptocurrency returns. *Available at SSRN 3241485*.

[Bouoiyour & Selmi, 2015] Bouoiyour, Jamal, & Selmi, Refk. 2015. What does Bitcoin look like? *Annals of Economics and Finance*, **16**(2), 449–492.

[Bouri *et al.* , 2019] Bouri, Elie, Gupta, Rangan, & Roubaud, David. 2019. Herding behaviour in cryptocurrencies. *Finance Research Letters*, **29**, 216–221.

[Cochrane, 2009] Cochrane, John H. 2009. *Asset pricing: Revised edition*. Princeton university press.

[Fama & French, n.d.] Fama, Eugene F, & French, Kenneth R. Common risk factors in the returns on stocks and bonds.

[Fama & French, 1992] Fama, Eugene F, & French, Kenneth R. 1992. The cross-section of expected stock returns. *the Journal of Finance*, **47**(2), 427–465.

[Fama & MacBeth, 1973] Fama, Eugene F., & MacBeth, James D. 1973. Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, **81**(3), 607–636.

[Galeshchuk *et al.* , 2018] Galeshchuk, Svitlana, Vasylchyshyn, Oleksandra, & Krysovatyy, Andriy. 2018. Bitcoin Response to Twitter Sentiments.

[Garcia & Schweitzer, 2015] Garcia, David, & Schweitzer, Frank. 2015. Social signals and algorithmic trading of Bitcoin. *Royal Society open science*, **2**(9), 150288.

[Gregoriou, 2019] Gregoriou, Andros. 2019. Cryptocurrencies and asset pricing. *Applied Economics Letters*, **26**(12), 995–998.

[Hutto & Gilbert, 2015] Hutto, C. J., & Gilbert, Eric. 2015. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*.

[Hwang & Salmon, 2004] Hwang, Soosung, & Salmon, Mark. 2004. Market stress and herding. *Journal of Empirical Finance*, **11**(4), 585–616.

[Jegadeesh & Titman, 1993] Jegadeesh, Narasimhan, & Titman, Sheridan. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, **48**(1), 65–91.

[Kim *et al.* , 2016] Kim, Young Bin, Kim, Jun Gi, Kim, Wook, Im, Jae Ho, Kim, Tae Hyeong, Kang, Shin Jin, & Kim, Chang Hun. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, **11**(8).

[Kristoufek, 2013] Kristoufek, Ladislav. 2013. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, **3**(1), 3415.

[Kristoufek, 2015] Kristoufek, Ladislav. 2015. What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PloS one*, **10**(4).

[Li *et al.* , 2019] Li, Tianyu Ray, Chamrajnagar, Anup, Fong, Xander, Rizik, Nicholas, & Fu, Feng. 2019. Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. *Frontiers in Physics*, **7**, 98.

[Ling & Zhu, 2019] Ling, Aifan, & Zhu, Zhikai. 2019. Network Hypes and Asset Prices of Cryptocurrencies: Empirical Evidence Based on Google-Attention Approach. *Available at SSRN 3424155*.

[Liu & Tsyvinski, 2018] Liu, Yukun, & Tsyvinski, Aleh. 2018. *Risks and returns of cryptocurrency*. Tech. rept.

[Liu *et al.* , 2019] Liu, Yukun, Tsyvinski, Aleh, & Wu, Xi. 2019. *Common risk factors in cryptocurrency*. Tech. rept.

[Murray Leclair, 2018] Murray Leclair, Emmanuel. 2018. *Herding in the cryptocurrency market*.

[Neal & Wheatley, 1998] Neal, Robert, & Wheatley, Simon. 1998. Do Measures of Investor Sentiment Predict Returns? *Journal of Financial and Quantitative Analysis*, **33**(Dec.), 523–547.

[Newey & West, 1987] Newey, Whitney K., & West, Kenneth D. 1987. Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, **28**(3), 777–787.

[Phillips & Gorse, 2017] Phillips, Ross C., & Gorse, Denise. 2017. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. IEEE.

[Shanken, 2015] Shanken, Jay. 2015. On the Estimation of Beta-Pricing Models. *Rev Financ Stud*, **5**(1), 1–33.

[Shiller, 2000] Shiller, Robert C. 2000. Irrational exuberance. *Philosophy and Public Policy Quarterly*, **20**(1), 18–23.

[Shuyo, 2010] Shuyo, Nakatani. 2010. *Language Detection Library for Java*.

[Tran, 2019] Tran, Hai Yen. 2019. Overconfidence Test in Cryptocurrencies Markets Using VAR Analysis. *Available at SSRN 3416394*.

[Vidal-Tomas *et al.* , 2019] Vidal-Tomas, David, Ibanez, Ana M., & Farinos, Jose E. 2019. Herding in the cryptocurrency market: CSSD and CSAD approaches. *Finance Research Letters*, **30**, 181–186.

[Wooldridge, 2006] Wooldridge, Jeffrey M. 2006. Introduction to econometrics: A modern approach. *Michigan State University. USA*, 390.

[Yeh & Yang, 2011] Yeh, Chia-Hsuan, & Yang, Chun-Yi. 2011. Examining the Effects of Tradersâ Overconfidence on Market Behavior.

Appendix A:

Table A.1: Statistical properties of daily cryptocurrency return series.

	Mean	Std	Skewness	Kurtosis
BTC	0.19%	0.04	0.05	5.33
BCH	0.28%	0.09	1.66	9.81
Cardona	0.43%	0.10	5.77	65.93
dogecoin	0.43%	0.10	5.77	65.93
EOS	0.26%	0.08	1.03	4.98
ETH	0.53%	0.07	0.27	14.00
LTC	0.26%	0.08	1.03	4.98
XRP	0.35%	0.08	7.54	148.43
Monero	0.21%	0.06	1.80	16.90
BNB	0.98%	0.09	2.94	23.49
IOTA	0.21%	0.08	0.90	5.56
TEZOS	0.03%	0.06	0.72	3.45

Figure A.1: Standardized level of comments and price for Bitcoin

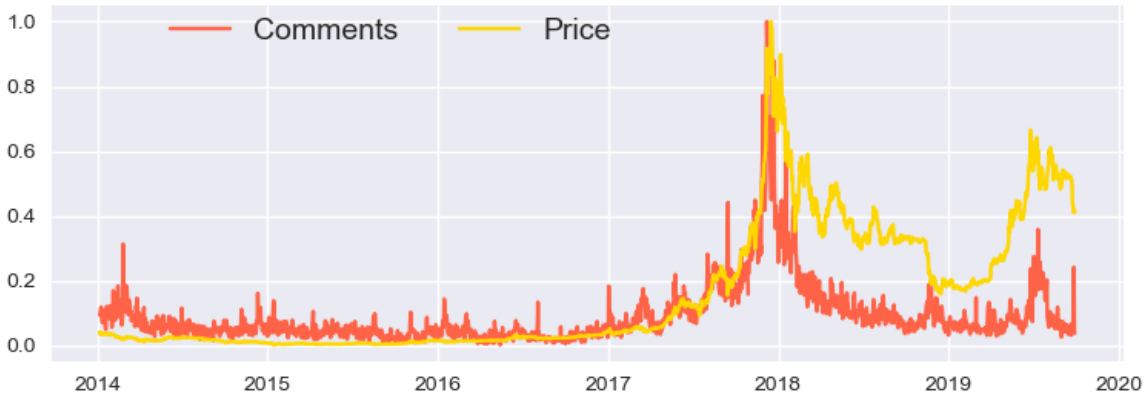

Table A.2: Descriptive statistics of financial cryptocurrency data. Number of coins shows the active number of coin in the sample. Market Cap Index is the value weighted cryptocurrency index. Market Cap Coin is on a individual coin basis.

Year	Number of Coins	Market Cap Index (mil) Mean	Median	Market Cap Coin (mil) Mean	Median
2014	4	1,590.78	1,543.87	1,437.00	144.45
2015	5	806.23	760.89	733.64	88.95
2016	7	1,703.90	1,775.43	1,703.90	208.93
2017	8	11,605.03	10,376.03	11,132.07	2,897.05
2018	11	21,570.55	19,049.89	20,383.37	5,699.08
2019	12	15,222.95	16,261.92	15,222.95	3,550.71
Full	12	8,749.90	6,075.73	8,435.49	1,552.990

Figure A.2: Cummulative net excess returns for cryptocurrency strategies

The plot depicts all considered strategies for the small minus big size portfolio, seven days momentum, (MOM), share positive, (Sh. Pos), is a long-short strategy going long in a portfolio containing the most positive coin according to sentiment and short in the least, share negative is a long-short strategy going long in the least negative and short in the most negative. Momentum share is the current share of positive or negative comments divided by the five day momentum in positive or negative share.

Figure A.3: Distribution of positive, negative and neutral comments in the sample.

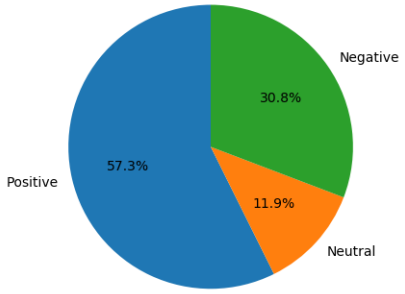

Table A.3: Sentiment data descriptives

The table contains the average daily amount of comments and the maximum amount of comments on a particular day, for each respective coin.

Cryptocurrency	Average daily amount of comments	Maximum amount of comments
BTC	1,678	13,891
BCH	297	11,382
Cardona	36	572
dogecoin	106	1,821
EOS	102	1,058
ETH	574	6,824
LTC	77	2,547
XRP	238	8,026
Monero	101	893
BNB	46	786
IOTA	156	6,595
TEZOS	260	1,590

Table A.4: Momentum portfolios

Table contains the long short spread for 3, 5, and 7 days momentum portfolios. The spread portfolio takes a long position in the third portfolio, P3, and a short position in the first portfolio, P1.

MOM3	P1	P2	P3	Spread
beta	0.0017	0.0013	0.0064	0.0046
se	0.0012	0.0011	0.0017	0.0017
t-values	1.4466	1.2490	3.8131	2.6875
p-values	0.1482	0.2118	0.0001	0.0073

MOM5	P1	P2	P3	Spread
beta	0.0007	0.0017	0.0067	0.0061
se	0.0012	0.0011	0.0016	0.0017
t-values	0.5602	1.5704	4.1671	3.5788
p-values	0.5754	0.1165	0.0000	0.0004

MOM7	P1	P2	P3	Spread
beta	0.0007	0.0019	0.0077	0.0070
se	0.0012	0.0011	0.0016	0.0017
t-values	0.5612	1.6801	4.8244	4.2570
p-values	0.5747	0.0931	0.0000	0.0000

Table A.5: Descriptive statistics of sentiment portfolios
The spread portfolio takes a long position in the fourth portfolio, P4, and a short position in the first portfolio, P1.

Share Positive	P1	P2	P3	P4	Spread
Mean (%)	0.24	0.28	0.30	0.53	0.29
Std	0.05	0.05	0.07	0.07	0.07
Skewness	0.92	1.10	2.45	1.85	1.87
Kurtosis	10.35	8.72	22.10	16.22	22.59

Share Negative	P1	P2	P3	P4	Spread
Mean (%)	0.53	0.15	0.33	0.27	-0.25
Std	0.07	0.06	0.06	0.05	0.07
Skewness	1.94	1.17	2.75	1.09	1.78
Kurtosis	17.03	11.07	26.84	12.35	21.06

Mom. Share Positive	P1	P2	P3	P4	Spread
Mean (%)	0.20	0.31	0.38	0.56	0.35
Std	0.06	0.06	0.06	0.08	0.08
Skewness	1.41	0.89	2.21	5.48	-4.12
Kurtosis	10.11	7.98	23.52	92.40	77.69

Mom. Share Negative	P1	P2	P3	P4	Spread
Mean (%)	0.23	0.39	0.18	0.55	0.32
Std	0.06	0.06	0.06	0.08	0.08
Skewness	1.48	1.34	0.88	5.60	-4.38
Kurtosis	10.32	9.43	11.94	88.05	77.35

Table A.6: Portfolio returns of weekly data
The spread portfolio takes a long position in the fourth portfolio, P4, and a short position in the first portfolio, P1.

	Quartile portfolios				
	P1	P2	P3	P4	Spread
Share Positive	Low			High	
$Mean$	0.0291^{***}	0.0319^{**}	0.0245^{*}	0.0234^{**}	-0.0058
$t\,(Mean)$	(3.26)	(2.26)	(1.86)	(2.05)	(-0.58)
Share Negative	Low			High	
$Mean$	0.0356^{***}	0.0248^{***}	0.0232^{**}	0.0318^{***}	-0.0038
$t\,(Mean)$	(2.81)	(2.06)	(1.67)	(3.48)	(-0.34)
Mom. Share Positive	Low			High	
$Mean$	0.0307	0.0206^{**}	0.0224^{**}	0.0442^{***}	0.0135
$t\,(Mean)$	(1.61)	(2.11)	(2.05)	(2.85)	(0.61)
Mom. Share Negative	Low			High	
$Mean$	0.0285	0.0110	0.0226^{**}	0.0481^{***}	0.0196
$t\,(Mean)$	(1.43)	(1.25)	(1.99)	(3.20)	(0.86)

Table A.7: Two-factor model results

	P1	P2	P3	P4
α	-0.0009	-0.0010	-0.0008	-0.0009
	(-0.75)	(-0.90)	(-0.61)	(-0.75)
β_{SENT}	0.2298	0.2174	0.2829	0.2298
	(6.99)	(5.51)	(4.54)	(6.99)
β_{SMB}	-0.3213	0.0201	0.0467	0.6787
	(-4.62)	(1.20)	(0.95)	(9.76)
R^2 (%)	24.98	5.70	8.63	53.45

	γ_0	γ_{SENT}	γ_{SMB}	R^2 (%)
Coefficient	-0.00187	0.00816	0.00356	
s.e	(0.005)	(0.021)	(0.002)	
	[0.005]	[0.021]	[0.002]	99.95
t-stat	(-0.3664)	(0.3900)	(2.0350)	
	$[-0.3664]$	[0.3855]	[1.4408]	

Figure A.4: CAPM pricing-error plot

Pricing errors for the cryptocurrency CAPM. Test assets are 4 sentiment sorted portfolios. Errors are deviations from the 45-degree line.

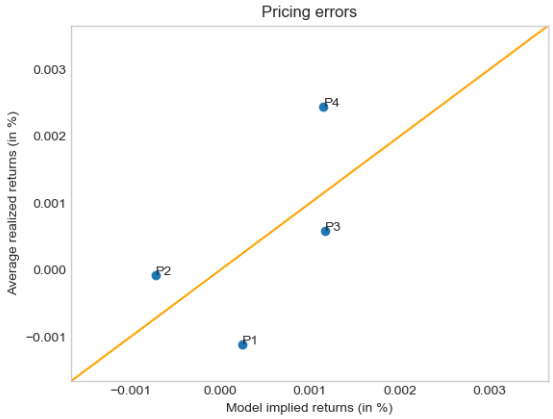

Figure A.5: Market and Sentiment two-factor model pricing-error plot

Pricing errors for the cryptocurrency Market and Sentiment two-factor model. Test assets are 4 sentiment sorted portfolios. Errors are deviations from the 45-degree line.

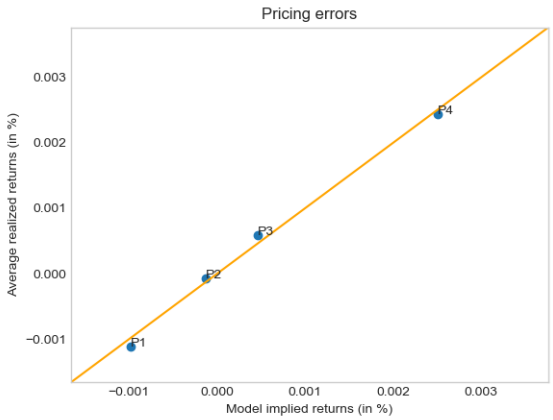

Figure A.6: Momentum and Sentiment two-factor model pricing-error plot

Pricing errors for the cryptocurrency Momentum and Sentiment two-factor model. Test assets are 4 sentiment sorted portfolios. Errors are deviations from the 45-degree line.

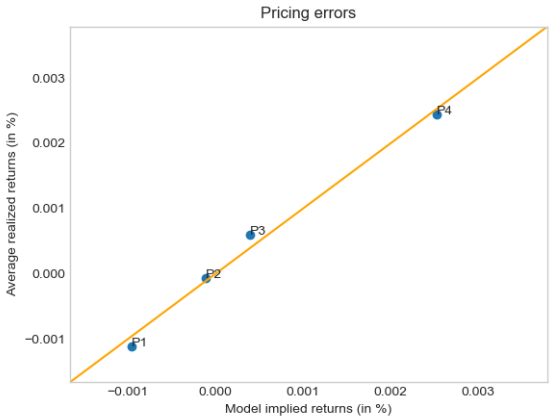

Figure A.7: Size and Sentiment two-factor model pricing-error plot

Pricing errors for the cryptocurrency Size and Sentiment two-factor model. Test assets are 4 sentiment sorted portfolios. Errors are deviations from the 45-degree line.