Your Sentiment Matters

A Machine Learning Approach for Predicting Regime Changes in the Cryptocurrency Market

Parra-Moyano, José; Partida, Daniel; Gessl, Moritz

Document Version
Final published version

Published in:
Proceedings of the 56th Annual Hawaii International Conference on System Sciences

DOI:
10125/102744

Publication date:
2023

License
CC BY-NC-ND

Citation for published version (APA):
Parra-Moyano, J., Partida, D., & Gessl, M. (2023). Your Sentiment Matters: A Machine Learning Approach for Predicting Regime Changes in the Cryptocurrency Market. In T. X. Bui (Ed.), *Proceedings of the 56th Annual Hawaii International Conference on System Sciences* (pp. 920-929). Hawaii International Conference on System Sciences (HICSS). https://doi.org/10125/102744

Your Sentiment Matters: A Machine Learning Approach for Predicting Regime Changes in the Cryptocurrency Market

José Parra-Moyano
Copenhagen Business School
jpm.digi@cbs.dk

Daniel Partida
University of Zurich
daniel.partida@bf.uzh.ch

Moritz Gessl
Moonpass
moritzfgessl@gmail.com

Abstract

Research suggests that a significant number of those investing in cryptocurrencies do not follow what we might call rational, profit-maximizing behavior. We also know that with the progressive lowering of entry barriers to online trading platforms, an increasing number of inexperienced investors are investing in cryptocurrencies. Increasingly, the behavior of investors contradicts the predictions made by traditional financial models and challenges the assumptions on which such models have previously relied when anticipating returns on cryptocurrency investments. To overcome this issue we develop a random forest model which we train with features stemming from a sentiment analysis performed on data generated by cryptocurrency enthusiasts using Twitter, Google Trends, and Reddit. Our findings show that such features have an important role to play in capturing the behavior of cryptocurrency investors and increase our model's ability to anticipate regime changes in the cryptocurrency market. Our model outperforms the predictive ability of the Log-Periodic Power Law model—currently, the model most widely-used to predict regime changes in financial markets. These results imply that scholars and practitioners aiming to understand and predict the development of cryptocurrency markets stand to benefit from analyzing social media data generated by cryptocurrency enthusiasts.

Keywords: Bitcoin, Cyptocurrencies, LPPL, Machine Learning, Sentiment Analysis

1. Introduction

Cryptocurrencies have sparked widespread interest among investors of all kinds. Unsurprisingly, scholars and practitioners aiming to better understand the price development of cryptocurrencies have used sophisticated financial models to predict returns on cryptocurrency investments (Albrecht et al., 2019). One of the key components driving positive and negative returns on investments in financial markets are regime changes (i.e., sharp changes in the tendency of the market). Anticipating regime changes is an effective way to improve investment decision-making and enable better management of risks.

Although scholars and practitioners have previously had success predicting returns on cryptocurrency investments (Geuder et al., 2019), anticipating regime changes in the cryptocurrency market remains a challenge (Rane and Dhage, 2019). This challenge prevails even for some of the most sophisticated models, such as the Log-Periodic Power Law (LPPL) model (Sornette et al., 1996) which has often and consistently predicted regime changes in traditional financial markets (Ghosh et al., 2021). The result is that cryptocurrency investors are often surprised by such regime changes.

One explanation for the challenges faced by scholars and practitioners trying to anticipate regime changes in the cryptocurrency market focuses on behavioral differences between investors in traditional assets and investors in cryptocurrencies. Researchers have identified differences in motivation between some members of the two groups (Mattke et al., 2021), as well as different reactions on the part of traditional and cryptocurrency investors confronted with the same information (Keller and Scholz, 2019). Another explanation for these challenges is the lowering of entry barriers to investment in cryptocurrencies, which is enabling many inexperienced investors to "try their luck" in the cryptocurrency market. The investment strategies of some such novice investors have been observed as irrational or inexistent (Hasso et al., 2019).

Recently, it has been proposed that the characteristics of cryptocurrency investors could be detected and accounted for by performing sentiment analysis of user-generated data (Wołk, 2020). However, such analyses cannot be incorporated into a model like the LPPL model. This is because, when attempting to predict regime changes, the LPPL model cannot account for information other than previous prices of the asset in question. This limitation means that the LPPL model cannot effectively anticipate regime changes in cryptocurrency markets. This challenge has been the motivation for our paper, in which we pose the following research question: *How can we develop a model that anticipates regime changes in the cryptocurrency market while accounting for the specific characteristics of cryptocurrency investors?*

To answer this research question, we built an ensemble random forest machine learning model, which we trained using a variety of features, including features stemming from a sentiment analysis conducted on user-generated data from Twitter, Google Trends, and Reddit. We compared the performance of our model with the performance of the LPPL model using a forward-chain, out-of-sample dataset on the cryptocurrencies Bitcoin and Luna. The results showed that our model outperformed the LPPL model in terms of accuracy, recall, and precision. Furthermore, we measured how the variables stemming from the sentiment analysis of user-generated data contributed to increases in the model's accuracy, recall, and precision. The results revealed how these variables played a pivotal role in improving the performance of the model. In all, our results showcase how the analysis of social media data generated by cryptocurrency users can help scholars and practitioners to better predict the development of cryptocurrency markets.

The rest of this paper is structured as follows. In Section 2, we review the literature on financial modeling for prediction and relevant theory on cryptocurrency investment behavior. In Section 3, we explain how a machine learning approach enabled us to predict regime changes and allowed for the incorporation of features stemming from the sentiment analysis performed on data generated by cryptocurrency investors. In Section 4, we compare the results of the models and, in Section 5, we discuss the implications and limitations of our results. Finally, in Section 6, we present a brief conclusion.

2. Literature Review

2.1. Modelling Financial Markets as Complex Systems to Predict Asset Returns

Historically, it has been assumed that the price of securities reflects all of the information available in the market (Fama, 1970). According to this hypothesis, termed the efficient-market hypothesis (EMH), the availability of new information leads to instant changes in the behavior of investors, altering the prices (and, therefore, the returns) of investments in the market. Hence, an acceptance of the EMH implies an acceptance that consistently outperforming the market is theoretically impossible (Keller and Scholz, 2019).

However, the EHM is highly contested, and there is plenty of evidence demonstrating how market anomalies and information asymmetries enable investors to consistently outperform the market (Ball et al., 2016; Grossman and Stiglitz, 1980). Likewise, the development of algorithms and sophisticated data analysis techniques has led to exclusive insights that can be used as signals of high fit (Clemons and Thatcher, 1997), giving some investors an advantageous position in the market.

In this regard, scholars and practitioners from different fields have used various techniques and approaches in an attempt to anticipate the behavior of financial markets. Following the 2007 financial crisis, a number of theories emerged aiming to anticipate the development of financial markets. Among the most prominent of these was complex system theory (Farmer et al., 2012), which is based on the notion that the state of a system is determined by the complex interactions of the heterogeneous agents that comprise the system (Eisenhardt and Bhatia, 2017). The theory tries to model the interactions of agents in the market and their (heterogeneous) reactions to new information, such as financial market indicators or price developments.

One such model is the LPPL model. The LPPL model is based on the notion that, prior to crashes, the mean function of a stock index price time series is characterized by a power law decorated with log-periodic oscillations which, in turn, leads to a market crash (Geraskin and Fantazzini, 2013). This model is able to predict regime changes in the long run and has reinforced the view of the stock market as a self-organizing cooperative system. This model has also been applied in other contexts and has often successfully predicted regime changes in asset returns (Barrau and Douady, 2022).

Given the numerous bubbles, so far, in the price of cryptocurrencies, it is not surprising that the LPPL

model has already been used to predict regime changes in the cryptocurrency market (Wheatley et al., 2019; Shu et al., 2021). However, the model has been deemed "not as applicable" to the cryptocurrency market as it is to traditional financial markets, largely because the "specific decision rule being used by cryptocurrency investors (which can potentially include human judgement and biases), increases the risk of the LPPL model predicting false positives or missing regime changes in this specific market," (Wheatley et al., 2019). Fundamentally, the strength of the LPPL model lies in anticipating long-term rather than short- or mid-term regime changes (Geuder et al., 2019). Hence, the LPPL model falls short when dealing with the short-term volatility of the cryptocurrency market.

2.2. On the Differences Between Traditional Investors and Cryptocurrency Investors

The literature has identified significant differences in the behavior of investors in the cryptocurrency market and investors in traditional markets (Mattke et al., 2021). While anticipated profits are the main (or only) factor influencing the decisions of those investors driven by financial ambition, it has been shown that some cryptocurrency investors invest because they support the ideology of a particular cryptocurrency. For these investors, profit expectancy is not a necessary condition of currency investment. At the same time, the low cost of entry into cryptocurrency markets and the profusion of online trading platforms that ease investments in any type of asset (Fink, 2021) have increased the number of inexperienced investors using irrational strategies—or no strategies at all—in the cryptocurrency investment (Low and Marsh, 2019).

The fact that profit maximization may not be the main investment rationale for all investors in the cryptocurrency market, coupled with the fact that some investors follow no strategy at all, increases the complexity of cryptocurrency systems. The presence of irrational (as one could define the traders investing in cryptocurrencies on an ideological, rather than financially rational, basis) or inexperienced investors results in the unpredictability of asset price developments and creates additional risk for rational, experienced investors (De Long et al., 1990). Such noise can result in prices diverging from fundamental values. This noise challenges the essential principles of rational behavior in times of uncertainty, which form the basis of decisions made by investors when designing their portfolios (Tversky and Kahneman, 1979).

One potential strategy for addressing the presence of irrational or inexperienced cryptocurrency investors may lie in the field of behavioral finance, which uses models of human psychology to study the motivations of individual investors (Shiller, 2003). Many of these models can be empirically tested using sentiment analysis on user-generated data. The analysis of user-generated data to explain the development of financial markets is commonplace in traditional finance. In fact, recent studies in the field of behavioral finance (Bollen et al., 2011; Li et al., 2018) show that investment decisions are strongly affected by individuals' emotional impulses, which are related to the individuals' moods (Albrecht et al., 2019). Such findings support the idea of incorporating sentiment analysis on user-generated data into existing models of financial prediction. In this vein, the use of news reports (Feuerriegel and Prendinger, 2016), message boards (Antweiler and Frank, 2004; Das and Chen, 2007), and Twitter or Google data (Oliveira et al., 2016; Ho et al., 2017; Pant et al., 2018; Kraaijeveld and De Smedt, 2020; Bing et al., 2014; Ranco et al., 2015) has proven very useful when addressing the particularities of investors in financial markets.

Hence, it is not surprising that the analysis of user-generated data from social networks has also proven to be a valuable source of information for generating features that address the behavior of cryptocurrency investors. A sentiment analysis conducted on a corpus of 144,492 tweets by 522 ventures has served to provide empirical evidence that positive language and a high, yet steady, level of interactivity among the community of investors are linked to higher token valuation (Albrecht et al., 2019). Similarly, sentiment analysis of user-generated data has enabled the anticipation of token prices of Initial Coin Offerings (Chanson et al., 2020), and there is evidence to show that sentiment analysis of user-generated data is useful in explaining the trading behavior of cryptocurrency investors (Keller and Scholz, 2019). Advanced social media sentiment analysis has also proven useful for making short-term cryptocurrency price predictions (Wołk, 2020).

In sum, sentiment analysis of user-generated data may help us to understand the development of the cryptocurrency market and enable us to account for the specific investment rationale of cryptocurrency investors.

2.3. Using Sentiment Analysis Techniques for Regime Change Prediction

On the basis that sentiment analysis performed on user-generated data could reveal the particularities of cryptocurrency investors, it seems reasonable

to incorporate features stemming from sentiment analysis into models for predicting regime changes in cryptocurrency markets. However, the LPPL model considers only past price development when predicting future returns, which precludes the use of features stemming from sentiment analysis.

This poses a challenge for scholars and practitioners alike, as the LPPL model cannot address the specifics of cryptocurrency investors. While this affects the use of the LPPL model in all kinds of markets, the limitation is a particular issue when the model is applied to the cryptocurrency market due to the presence of relatively large groups of irrational or inexperienced investors. This problem motivated our (aforementioned) research question.

3. Models and Data

3.1. An Ensemble Machine Learning Model to Predict Regime Changes

To develop a model to predict regime changes while allowing for the incorporation of features other than past asset prices, we turned to the machine learning literature. Machine learning models have been extensively used to predict returns on investments in assets of all kinds (Ma et al., 2021). In fact, predictive analytics—making use of machine learning and network analysis—has become a cornerstone of modern organizations in recent years (Sarlin and Mezei, 2020). In this vein, machine learning techniques have already been used to predict Bitcoin prices (Chowdhury et al., 2020). Nonetheless, we could find no machine learning model that incorporates both sentiment analysis of user-generated data and macro features and technical indicators to consistently predict regime changes in the cryptocurrency market. We aimed to develop a model that achieved this goal.

To this end, we used a random forest estimator as an ensemble model to regress the returns of investments on cryptocurrencies. The main advantages of choosing a random forest regressor are its ability to handle big data with numerous features, clearly specify the importance of particular features, and reduce variance to prevent overfitting, which is achieved by taking the aggregation of multiple decision tree outputs (Xia Liu et al., 2021; Domingos, 2012; Zhang and Ram, 2020). As ensemble models can incorporate a multitude of features, such as those stemming from sentiment analysis, and can be used to model complex systems and predict binary outcomes such as regime changes, we deemed this type of model adequate for our purpose.

Specifically, we used a Classification and Regression Tree (CART), to which we applied a random forest method to reduce the instability of the tree. This method is based on decision trees that learn how to best split the observations of a dataset into ever-decreasing subsets of data in order to predict a target value (Rokach and Maimon, 2005). The target value can be continuous (e.g., returns on asset investments), or categorical (e.g., regime changes). Hence, this model not only allows us to predict regime changes but also to predict asset returns. We deem this a strength of the model, as it has the mathematical foundation to make holistic predictions about the cryptocurrency market. Hence, it will provide investors with more information than would a model that only predicts regime changes (as would the LPPL model), or only asset returns (as would the majority of machine learning models applied to cryptocurrencies).

For the regression problem (i.e., when regressing asset returns), the resulting tree is the one that achieves maximum variance reduction using Mean Square Error when estimating the asset return for a particular week (Y_i), which is given by this formula:

$$MSE = \frac{1}{n}\sum_{i=1} n(Y_i - \hat{Y}_i)^2$$

For the classification problem (i.e., when regressing labels that stand for the presence and type of regime changes), the resulting tree is the one that minimizes the Gini Impurity. The Gini Impurity measures the probability of incorrectly classifying a data point (i.e., of assigning the wrong label to a data point), and is given by this formula:

$$Gini\ Impurity = \sum_{i=1} Cf_i(1 - f_i)$$

The feature f stands for the frequency of label i at a node, and C is the number of unique labels. In our case, there are three possible labels, namely a positive regime change, a negative regime change, and no regime change. For the estimation we used the Python library *sci-kit earn* (Pedregosa et al., 2011).

3.2. Data

We trained our model using a range of features. These features came in different datasets containing macro features and technical indicators. For the macro features, we specifically considered four data inputs from the Federal Reserve Economic Data: the expected inflation, the consumption index, the money supply M2 in the form of cash, and checking deposits, such as saving deposits, money market funds, and certificates of deposits. The money supply M2 is closely observed

as an indicator of future inflation and also as a proxy for central bank monetary policy. We also considered other macro features such as the exchange rates of the EUR/USD, the GBP/USD, and the RMB/USD, as well as the commodity prices of gold, silver, copper, and oil. This data can be downloaded in multiple formats from the webpage of the St. Louis Fed.

In terms of stock market data, we considered three different indices to account for the correlation between the cryptocurrency market and the stock market. These indices were the NASDAQ Composite, the S&P 500, and the VIX, which helps us to capture the expected volatility of the stock market. Furthermore, we include all the technical indicators provided by the Python's Technical Analysis Library (Lopez-Padial and Harris, 2018). These included, for example, the relative strength index, stochastic relative strength index, and rate of change among the momentum indicators; the money flow index, negative volume index, and on-balance volume among the volume indicators; Bolliger Bands, Donchian Channel, and annualized volatility among the volatility indicators; and Ichimoku clouds, moving average convergence divergence, and exponential moving averages among the trend indicators.

To incorporate the sentiment analysis and try to capture the specific characteristics of cryptocurrency investors, we performed a sentiment analysis on data gathered from Google Trends, Reddit, and Tweeter. Specifically, we searched for the Google Trends "Bitcoin", "LUNA", "gm" and "wagmi". The latter two of these acronyms are cryptocurrency jargon, "gm" standing for "good morning" and "wagmi" for "we are all gonna make it". For the Reddit data, we searched for similar terms in the subreddits r/terraluna and r/Bitcoin. For the Twitter data we used tweets containing the same terms, and, thanks to the Twitter Annotations API, we also considered tweets relating to these terms and referring to – yet, not including the specific terms - "Bitcoin" and "Luna".

To perform the sentiment analysis, we used the dictionary-based sentiment model VADER (Wilksch and Abramova, 2022) to label the sentiment score of our data. We refined our analysis using the transformer NLP architectures BERT and RoBERTa (Hartmann et al., 2022), taking into account semantic lexicon from social media that provides a sentiment score for the analyzed user-generated data. This sentiment analysis served as a proxy against which to measure the social effect in the cryptocurrency market. We used the resulting sentiment score as an additional feature in the final random forest model. Our complete dataset is available upon request.

3.3. Out of Sample Testing

We began by using our model to anticipate the weekly (7-day) returns of Bitcoin and Luna, respectively. To avoid overfitting due to in-sample training, we used a forward-chain approach that enabled us to train the models using only past data and to predict the returns of the subsequent week in an out-of-sample way. This approach was iterative in the sense that, once it had performed an out-of-sample prediction for the future week, it incorporated the data of that week to retrain the model and make a prediction for the subsequent week. This enabled us to avoid in-sample prediction and to ensure the validity of our results. Furthermore, the returns regressor was updated to predict returns from out-of-sample using this forward-chain philosophy.

In order to improve the performance, we ran a randomized grid search as part of the hyperparameter tuning necessary to calibrate the model. We use the resulting model to predict the weekly (7-day) returns on investments in Bitcoin (starting from the first week of 2015), and the returns on investments in Luna (starting from the last July week of 2019).

Once the weekly, out-of-sample returns had been predicted, we adapted our model to anticipate regime changes. This meant that, rather than using our model to predict a random continuous feature (the returns), we used it to solve a classification problem. This resulted in a model that, as an output, could predict three possible scenarios: namely, a draw-up of more than 10% in weekly returns (what we consider a positive regime change), a draw-down of more than 10% in weekly returns (what we consider a negative regime change), and a scenario for all return changes between -10% and +10%, which we consider to be a scenario involving no regime change. By undertaking this exercise, we were able to use the ensemble random forest method to predict regime changes, just as the LPPL model does.

Using these parameters of regime change, we have observed 87 regime changes in Bitcoin since the beginning of 2015, of which 29 have been positive regime changes (i.e., sharp increases in the price of Bitcoin), and 58 have been negative regime changes (i.e., sharp decreases in the price of Bitcoin). Using these parameters we have also observed 299 weeks in which we determined that no regime change had happened. For Luna, we have observed 61 regime changes since the last week of July 2019, out of which 27 were positive regime changes (i.e., sharp increases) and 34 were negative regime changes (i.e., sharp decreases). We also identified 88 weeks in which no regime change occurred for Luna. Finally, we

used the LPPL model to predict regime changes and asset returns in Bitcoin and Luna, employing the same forward-chain approach we had used with our model.

Overall, these activities provided us with a weekly out-of-sample asset return prediction for Luna and Bitcoin from our model, as well as two 7-day predictions about a regime change in Bitcoin and Luna (one from the LPPL model and one from our model).

4. Results

4.1. Results of Predicting the Regime Changes

To address the classification problem (i.e., the problem of regime change prediction), we can compare the confusion matrices of the LPPL model with those resulting from our model. This enables us to compare the accuracy (i.e., the number of correctly predicted regime changes divided by the total number of predicted regime changes), the recall (i.e., the number of correctly predicted regime changes divided by the total number of correctly predicted regime changes plus the number of states without a regime change that were predicted as a regime change), and the precision (i.e., the ratio between the number of regime changes correctly classified and the total number of predicted regime changes). The results of these metrics for each of the models (the LPPL model and our machine learning model, which we refer to as "ML"), can be found in Table 1. We report these results for both Bitcoin and Luna which, in the table, we refer to as "assets".

Table 1. Metrics Comparison for the Regime Changes.

Asset	**Bitcoin**		**Luna**	
Model	**LPPL**	**ML**	**LPPL**	**ML**
Accuracy	80.3%	86.2%	61.7%	82.1%
Recall	38.6%	72.1%	31.1%	80.2%
Precision	34.9%	83.8%.	32.8%	77.2%

When comparing the metrics of the two models we observe that, for both Bitcoin and Luna, the ML model consistently outperformed the LPPL model. This result implies that the ML model enables us to correctly predict the development of the Bitcoin market for 12 weeks more than when using the LPPL model. Similarly, the increase in accuracy implies that using the ML model enables us to correctly assess the development of the Luna market for 22 weeks more than when using the LPPL model. When comparing the recall metrics for Bitcoin, we observed that the ML model enabled us to predict twice as many actual regime changes as did the LPPL model. Furthermore, when comparing the precision metrics for Bitcoin, we observed that using the ML model we were able to predict more than twice as many regime changes than when using the LPPL model. As the strength of the LPPL model is its ability to anticipate regime changes, we consider these results highly encouraging.

4.2. Contribution of the Sentiment Analysis for Predicting Regime Changes

In order to identify the contribution of the variables stemming from the sentiment analysis performed on the user-generated data, we compared the accuracy, recall, and precision of the models for both currencies with and without such variables. We observed that, when including the variables stemming from the sentiment analysis, the accuracy for the Bitcoin ML model increased by 7.7%, the recall increased by 3.2%, and the precision increased by 6.2%. Likewise, the accuracy for the Luna ML model increased by 6.3%, the recall increased by 7.7%, and the precision increased by 6.1%. It is important to note that these increases were in absolute terms, meaning that they represented absolute increases in the performance of the model (and not improvements relative to the previous performances, which would be much higher).

4.3. Results of Predicting the Returns

Since the LPPL model cannot anticipate return changes, we cannot compare our results with the results of the LPPL model. However, we hereafter present the results of our model when predicting returns in Bitcoin and Luna in order to provide a complete overview of our model's performance. Specifically, we report the R^2, Mean Absolute Error (MAE), and Mean Squared Error (MSE) of our model. Table 2 provides the results of these metrics. Given that current state-of-the-art for the monthly return forecasting R^2 value of random forest models of asset pricing (for all the stocks in the market), lies at around 38% (Gu et al., 2020), and for the MAE they lie at around 9% (Nagula and Alexakis, 2022) we consider our results in this realm a significant contribution.

Table 2. Metrics Comparison for the Returns.

Asset	**Bitcoin**	**Luna**
R^2	41.1%	38.6%
MAE	5.41%	12.4%
MSE	0.65%.	3.74%

Figure 1 shows the weekly (out-of-sample) fitted and historical returns of investing in Bitcoin and Luna for the period between November 28th, 2021, and May 29th,

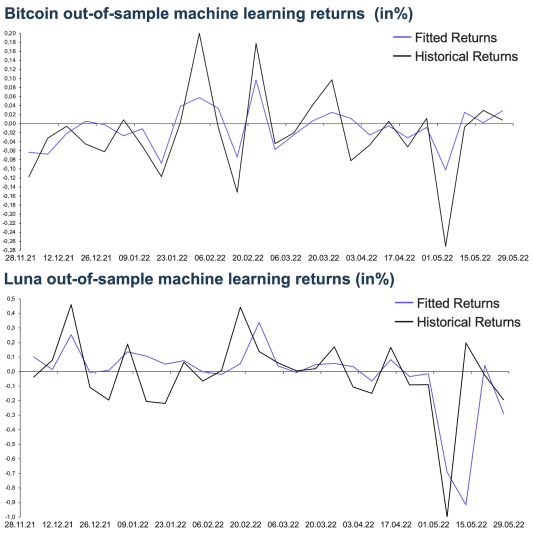

Figure 1. Weekly returns for Bitcoin and Luna using the ML model.

2022. We can observe that the direction of the fitted (predicted) asset return line and that of the historical asset return line coincide in the majority of the cases. More encouragingly, our model is able to grasp the sharp drop in returns that occurred in the first week of May 2022 for both Bitcoin and Luna. The fact that our model only makes out-of-sample predictions means it was able to anticipate the return decrease and, therefore, give a clear signal of Bitcoin and Luna selling before the crash actually happened.

5. Discussion

In this paper, we have considered how to address the limitations faced by scholars and practitioners in their attempts to anticipate regime changes in the cryptocurrency market. We used the LPPL model as a baseline model for our analysis. We continued by developing a supervised machine learning ensemble model using a random forest algorithm to predict both regime changes and weekly returns on Bitcoin and Luna. Our machine learning model has enabled us to add features stemming from a sentiment analysis performed on Google Trends, Twitter, and Reddit data—features that the LPPL model cannot incorporate. Our model outperformed the LPPL model in terms of accuracy, precision, and recall when predicting regime changes in Bitcoin and Luna, and offered consistent predictions on the returns of these cryptocurrencies, being able to anticipate the crash of these two cryptocurrencies that occurred in the first week of May 2022.

Together, these results enable us to answer our research question by stating that the use of a random forest ensemble model including macro features, technical indicators, and features stemming from a sentiment analysis performed on user-generated data can outperform the predictions of the LPPL model for Bitcoin and Luna. By comparing the performance of our model with and without the features stemming from the sentiment analysis, we can state that analyzing data generated by cryptocurrency investors in social platforms can yield useful information for identifying the specific characteristics of cryptocurrency investors and can, therefore, help to better predict the development of cryptocurrency markets.

5.1. Implications for Scholars

Our results show that performing a sentiment analysis on user-generated data can play a pivotal role in improving the prediction of regime changes and asset returns in the cryptocurrency market. We suggest that it is the use of variables stemming from the sentiment analysis performed on user-generated data that enables the model to anticipate the behavior of irrational (Mattke et al., 2021) or inexperienced (Low and Marsh, 2019) cryptocurrency investors. Hence, when assumptions about rational behavior under conditions of uncertainty are not fulfilled, and traditional models are, therefore, challenged (Tversky and Kahneman, 1979), incorporating variables stemming from sentiment analysis of user-generated data may result in better prediction models.

This finding is in line with behavioral finance literature which uses models of human psychology to study the motivations of individual investors (Shiller, 2003). Furthermore, our results build on work by previous scholars employing user-generated data to address the behavioral particularities of investors groups and subgroups (Bollen et al., 2011; Li et al., 2018; Oliveira et al., 2016; Sprenger et al., 2014; Nofer and Hinz, 2015; Ho et al., 2017). Our work joins these previous studies in demonstrating the importance of such data when anticipating regime changes in the cryptocurrency market.

Our results also emphasize the virtues of machine learning models and provide an example of how such models—when trained using appropriate features—can outperform structured models rooted in complex system theory, such as the LPPL model. This opens the possibility of combining machine learning models with insights from the behavioral finance literature and models of human psychology to solve unstructured and unsolved problems.

5.2. Implications for Practitioners

Practitioners could use a model such as ours to model the market and try to prevent large losses due to unforeseen regime changes. For highly volatile markets, such as the cryptocurrency market, this could generate interesting results. Furthermore, practitioners could use our model to attempt to capitalize on positive regime changes and, eventually, outperform the market by reaping higher returns. An increasing number of individuals rely solely on the income they generate from cryptocurrencies (Liedel, 2018)—as is the case with workers quitting their jobs to earn cryptocurrencies by playing Axie Infinity Mataverse Videogames (Kellers, 2022). Models such as ours may become increasingly important in helping such individuals in predicting short-term regime changes in cryptocurrencies.

Given the entry costs of investing in all kind of assets (not only cryptocurrencies) is decreasing (Low and Marsh, 2019), we anticipate that models involving sentiment analysis will play a crucial role in anticipating asset returns. In all cases, the limitations of the model referred hereafter need to be carefully taken into account.

5.3. Limitations and Future Work

Our results are encouraging and contribute to the literature by showing how machine learning models—when trained using adequate features—can outperform the most sophisticated finance models in predicting regime changes. Yet, our results are not without limitations.

First, we do not experiment with our model to find out whether adding features stemming from a sentiment analysis is of particular advantage when making predictions about the cryptocurrency market. It could be that adding such features to a machine learning model like ours is equally helpful both for traditional and cryptocurrency markets. If this were the case, other sources of information would have to be taken into account in order to identify the characteristics of cryptocurrency investors.

Second, our method of defining regime changes is rather rudimentary (defining a positive regime change as a return's increase above 10%, and a negative regime change as a return's decrease of more than 10%). This leaves an intermediary scenario of returns between 10% and -10% which we consider to be a scenario involving no regime change. Refining this labeling by including more states and breaking down the intermediary state would be valuable and result in more useful predictions. Incorporating volatility metrics into the definition of regime change could also be of value.

Third, there is a component of "self-fulfilling prophecy" that, while ubiquitous to all predictive models in finance, implies that, at some point, the advantages of using a model like ours would vanish.

Fourth, it is unclear whether our model would perform similarly for currencies other than Bitcoin and Luna. Future research should address these shortcomings. Furthermore, if the role of institutional investors in the cryptocurrency space increases, the importance of anticipating the behavior of irrational or inexperienced investors would decrease, as would the contribution of the variables stemming from user-generated data.

Fifth, our model provides indications as to the likelihood of future regime changes but does not represent an investment strategy in itself. Investment strategies that incorporate the insights of this paper would need to be developed in order to operationalize the applicability of the model.

6. Conclusion

The fact that a significant proportion of investors in cryptocurrencies do not follow what is considered a rational strategy poses a challenge for scholars and practitioners aiming to anticipate regime changes in this type of asset. We have developed a random forest model that helps to anticipate such regime changes by incorporating features from the analysis of user-generated data from Google Trends, Twitter, and Reddit. Our model is the first to address this issue and highlight the contribution of such variables in anticipating such regime changes for the cryptocurrencies Bitcoin and Luna. These results present several avenues for future research, particularly in relation to the increasing number of inexperienced and—from a traditional point of view—irrational investors in cryptocurrencies, and (due to decreasing entry costs), traditional financial assets too.

References

Albrecht, S., Lutz, B., & Neumann, D. (2019). How sentiment impacts the success of blockchain startups - an analysis of social media data and initial coin offerings. *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019, Hawaii.*, 4545–4556.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259–1294.

Ball, R., Gerakos, J., Linnainmaa, J. T., & Nikolaev, V. (2016). Accruals, cash flows, and operating profitability in the cross section of stock returns. *Journal of Financial Economics*, *121*(1), 28–45.

Barrau, T., & Douady, R. (2022). Predictions of market returns. In *Artificial intelligence for financial markets* (pp. 59–81). Springer.

Bing, L., Chan, K. C., & Ou, C. (2014). Public sentiment analysis in twitter data for prediction of a company's stock price movements. *2014 IEEE 11th International Conference on e-Business Engineering*, 232–239.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1–8.

Chanson, M., Martens, N., & Wortmann, F. (2020). The role of user-generated content in blockchain-based decentralized finance. *Proceedings of the 28th European Conference on Information Systems, Online.*

Chowdhury, R., Rahman, M. A., Rahman, M. S., & Mahdy, M. (2020). An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. *Physica A: Statistical Mechanics and its Applications*, *551*, 124569.

Clemons, E. K., & Thatcher, M. E. (1997). Evaluating alternative information regiines in the private health insurance industry: Managing the social cost of private information. *Journal of Management Information Systems*, *14*(2), 9–31.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375–1388.

De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, *98*(4), 703–738.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Eisenhardt, K. M., & Bhatia, M. M. (2017). Organizational complexity and computation. *The Blackwell companion to organizations*, 442–466.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, *25*(2), 383–417.

Farmer, J. D., Gallegati, M., Hommes, C., Kirman, A., Ormerod, P., Cincotti, S., Sanchez, A., & Helbing, D. (2012). A complex systems approach to constructing better models for managing financial markets and the economy. *The European Physical Journal Special Topics*, *214*(1), 295–324.

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, *90*, 65–74.

Fink, C. (2021). Why millennials gravitate to new brands in online investing. *Journal of Brand Strategy*, *9*(4), 401–407.

Geraskin, P., & Fantazzini, D. (2013). Everything you always wanted to know about log-periodic power laws for bubble modeling but were afraid to ask. *The European Journal of Finance*, *19*(5), 366–391.

Geuder, J., Kinateder, H., & Wagner, N. F. (2019). Cryptocurrencies as financial bubbles: The case of bitcoin. *Finance Research Letters*, *31*.

Ghosh, B., Kenourgios, D., Francis, A., & Bhattacharyya, S. (2021). How well the log periodic power law works in an emerging stock market? *Applied Economics Letters*, *28*(14), 1174–1180.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2022). More than a feeling: Accuracy and application of sentiment analysis. *SSRN Electronic Journal*.

Hasso, T., Pelster, M., & Breitmayer, B. (2019). Who trades cryptocurrencies, how do they trade it, and how do they perform? evidence from brokerage accounts. *Journal of Behavioral and Experimental Finance*, *23*, 64–74.

Ho, C.-S., Damien, P., Gu, B., & Konana, P. (2017). The time-varying nature of social media sentiments in modeling stock returns. *Decision Support Systems*, *101*, 69–81.

Keller, A., & Scholz, M. (2019). Trading on cryptocurrency markets: Analyzing the behavior of bitcoin investors. *Proceedings of the Fortieth International Conference on Information Systems, Munich.*

Kellers, L. (2022). *How we can predict the next financial crisis*. Forkast. https : / / forkast . news / axie - infinity - fans - play - their - way - payday - game - token-prices-soar

Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188.

Li, T., van Dalen, J., & van Rees, P. J. (2018). More than just noise? examining the information content of stock microblogs on financial markets. *Journal of Information Technology*, *33*(1), 50–69.

Liedel, D. A. (2018). The taxation of bitcoin: How the irs views cryptocurrencies. *Drake L. Rev.*, *66*, 107.

Lopez-Padial, D., & Harris, C. (2018). *Python technical analysis* (tech. rep.). https://technical-analysis-library-in-python.readthedocs.io/en/latest/

Low, R., & Marsh, T. (2019). Cryptocurrency and blockchains: Retail to institutional. *The Journal of Investing*, *29*(1), 18–30.

Ma, Y., Han, R., & Wang, W. (2021). Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, *165*, 113973.

Mattke, J., Maier, C., Reis, L., & Weitzel, T. (2021). Bitcoin investment: A mixed methods study of investment motivations. *European Journal of Information Systems*, *30*(3), 261–285.

Nagula, P. K., & Alexakis, C. (2022). A new hybrid machine learning model for predicting the bitcoin (btc-usd) price. *Journal of Behavioral and Experimental Finance*, 100741.

Nofer, M., & Hinz, O. (2015). Using twitter to predict the stock market. *Business & Information Systems Engineering*, *57*(4), 229–242.

Oliveira, N., Cortez, P., & Areal, N. (2016). Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, *85*, 62–73.

Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, 128–132.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 2825–2830.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. *PloS one*, *10*(9), e0138441.

Rane, P. V., & Dhage, S. N. (2019). Systematic erudition of bitcoin price prediction using machine learning techniques. *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, *65*(2), 594–598.

Rokach, L., & Maimon, O. (2005). Decision trees. In *Data mining and knowledge discovery handbook* (pp. 165–192). Springer.

Sarlin, P., & Mezei, J. (2020). Introduction to the minitrack on machine learning and predictive analytics in accounting, finance and management.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, *17*(1), 83–104.

Shu, M., Song, R., & Zhu, W. (2021). The 2021 bitcoin bubbles and crashes—detection and classification. *Stats*, *4*(4), 950–970.

Sornette, D., Johansen, A., & Bouchaud, J.-P. (1996). Stock market crashes, precursors and replicas. *Journal de Physique I*, *6*(1), 167–175.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, *20*(5), 926–957.

Tversky, A., & Kahneman, D. (1979). An analysis of decision under risk. *Econometrica*, *47*(2), 263–292.

Wheatley, S., Sornette, D., Huber, T., Reppen, M., & Gantner, R. N. (2019). Are bitcoin bubbles predictable? combining a generalized metcalfe's law and the log-periodic power law singularity model. *Royal Society open science*, *6*(6), 180538.

Wilksch, M. V., & Abramova, O. (2022). The predictive power of social media sentiment for short-term stock movements. *Proceedings of the 17th International Conference on Wirtschaftsinformatik, Nürnberg*, 1092–1100.

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

Xia Liu, A., Li, Y., & Xu, S. X. (2021). Assessing the unacquainted: Inferred reviewer personality and review helpfulness. *MIS Quarterly*, *45*(3).

Zhang, W., & Ram, S. (2020). A comprehensive analysis of triggers and risk factors for asthma based on machine learning and large heterogeneous data sources. *MIS Quarterly*, *44*(1).