Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using A Gradient Boosting Tree Model

Tianyu "Ray" Li, Anup Chamrajnagar, Xander Fong, Nick Rizik
Math 106/87: Professor Feng Fu

ABSTRACT

In this paper, we analyze Twitter signals as a medium for user sentiment to predict the price fluctuations of a small-cap alternative cryptocurrency called *ZClassic*. We extracted tweets on an hourly basis for a period of 3.5 weeks, classifying each tweet as positive, neutral, or negative. We then compiled these tweets into an hourly sentiment index, creating an unweighted and weighted index, with the latter giving larger weight to retweets. These two indices, alongside the raw summations of positive, neutral, and negative sentiment were juxtaposed to ~ 400 data points of hourly pricing data to train an Extreme Gradient Boosting Regression Tree Model. Price predictions produced from this model were compared to historical price data, with the resulting predictions having a 0.81 correlation with the testing data. Our model's predictive data yielded statistical significance at the $p < 0.0001$ level. Our model is the first academic proof of concept that social media platforms such as Twitter can serve as powerful social signals for predicting price movements in the highly speculative alternative cryptocurrency, or "alt-coin", market.

METHODS

We began by researching different alternative cryptocurrencies to decide which would be best suited within the confines of our analysis. We decided to choose ZClassic (ZCL), a private, decentralized, fast, open-source community driven virtual currency, as the primary target of our academic focus. The technological nature of the ZClassic cryptocurrency lends itself to a high level of predictability via tweet analysis, as it is set to "hard fork" into Bitcoin Private on February 28th, 2018. A hardfork is a major change to blockchain protocol which makes previously invalid blocks or transactions valid [19]. Typically, in the time leading up to the "hard fork," a coin's price reflects investor outlook on the success of the resulting fork.

To collect the tweets, we decided to base our program in RStudio, using open-sourced rtweet package [20], which accesses Twitter's REST and stream APIs. We were able to use the rtweet package to retrieve, from each of the last seven days, searching from midnight backwards, tweets that had the terms "ZClassic," "ZCL," and "BTCP" while simultaneously eliminating repeat tweets. This collection process was repeated 3 times over the course of three and a half weeks to provide sufficient data for our analysis. In the end, we garnered a final data set of 130,000 unique tweets.

We then created an algorithm to classify each tweet as positive, negative, or neutral sentiment using natural language processing. The dictionary, primarily sourced from the Python package "Textblob", that assigns impactful words and phrases a polarity value (e.g., "top" and "not great" have values of 0.5 and −0.4, respectively), which we view as sentiment

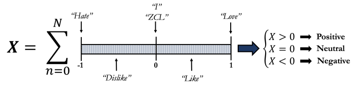

Figure 1: Summation of polarity values to Sentiment Conversion (***N*** = *total words)*

For model selection, we employed 10-fold cross validation on 589 data points to choose an optimal model framework among linear regression, logistic regression, polynomial regression, exponential regression, tree model, and support vector machine regression. A tree model called the Extreme Gradient Boosting Regression (also known as XGBoost [21]), exhibited the smallest loss, or inaccuracy, and was thus chosen to train the model on our data. The XGBoost model, as well as other tree-based models, is particularly suited for applications on our data for the following reasons:

1. Tree models are not sensitive to the arithmetic range of the data and features. Thus, we do not need to normalize the data and possibly prevent loss due to normalization.
2. Tree models are by far the most scalable machine learning model due to their construction processes –simply adding more children nodes to the pre-existing tree nodes will update the tree and allow our strategy to continue to accurately predict price as our collection of price and tweet data increases into the future. It also makes the model adaptable for currencies with larger daily tweet volumes.
3. On the abstract level, the tree model is a rule-based learning method which, unlike a traditional regression learning method, has more potential to unveil insightful relationships between features.

For completeness, we sketch the key ideas behind XGBoost as follows. Let us define:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F},$$

- $\varphi(x_i)$ is our predicting function
- $\hat{y}_i$ is the prediction from our model for the i-th observation
- Each f representing a tree in our regression tree forest, F .

Algorithm 1: Exact greedy algorithm for split finding [21] used in our price prediction model.

```
Input: I, instance set of current node
Input: d, feature dimension
gain ← 0
G ← Σ_{i∈I} g_i, H ← Σ_{i∈I} h_i
for k = 1 to m do
    G_L ← 0, H_L ← 0
    for j in sorted (I, by x_jk) do
        G_L ← G_L + g_j, H_L ← H_L + h_j
        G_R ← G − G_L, H_R ← H − H_L
        score ← max(score, G_L^2/(H_L+λ) + G_R^2/(H_R+λ) − G^2/(H+λ))
    end
end
Output: Split with max score
```

Our goal is to minimize the objective function L , defined below:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k),$$

$$\Omega(f) = \gamma \mathcal{T} + \frac{1}{2}\lambda \|w\|$$

		Algorithm Sentiment Prediction Positive	Neutral	Negative
Manual Sentiment Decision	Positive	79%	21%	0%
	Neutral	34%	51%	15%
	Negative	25%	0%	75%

Figure 2: Validation analysis of algorithm sentiment prediction by manual inspection.

To minimize the above objective function, we employed a greedy Algorithm (1) to create our regression tree forest F as originally implemented in Ref. [21].
One-third of the 589 data points is separated as the testing data, and the remainder is used as the training set as we built our Extreme Gradient Boosting Regression model. The model also tests different lead-lag on the range of [0, 1, 2, 3, 4, 5 hours] since we do not know how quickly the public would react to the market update or the social media sentiment. Based on the testing result, we decided that there is a 3-hour lag effect between social media information and price effects.

RESULTS

Sentiment	Tweets
Positive	"Bitcoin Private $btcp is coming! better hop in $zcl to grab some before the snapshot on the 28th! 15 days is a times for a crypto to be going up." "#ZCL going to give Grandma her retirement back!" "If you haven't got $ZCL yet, now is the time to grab some."
Neutral	"Do you plan on selling your $zcl before the fork or holding on to it for the $btcp?" "don't miss out registering on Binance, before they close registration again." "RT @cryptobr4in: When everyone realizes your get FREE #BitcoinPrivate $BTCP if you buy $ZCL #CryptoCurrency #FreeCoins ..."
Negative	"RT@Crypto.Bitlord: Bitcoin private is just another scam coin knock off like bitcoin cash trying to ride the satoshi's original vision." "2017 end - most thought 2018 would b the year crypto would explode. I said it b4 n will say it again - gonna b a bad year." "What a joke this $zcl game is. Good news, drops like a rock. #cryptocurrency ..."

Figure 3: Examples of Tweets with positive, neutral and negative sentiment classifications in our dataset.

Having set the sentiment classification algorithms in place, we decided to train our model using six different features: Pure Positive Sentiment, Pure Negative Sentiment, Neutral Sentiment, An Unweighted Sentiment index, A Weighted Sentiment Index, and Hourly Trading Volume. These six features proved to be varied enough to train the model effectively on a variety of different trading points and resulted in the best and most accurate overall correlation with the testing data (as shown in Figure 5). The detailed co-plots of the different features versus the price curve over the study period is shown in Figure 4.

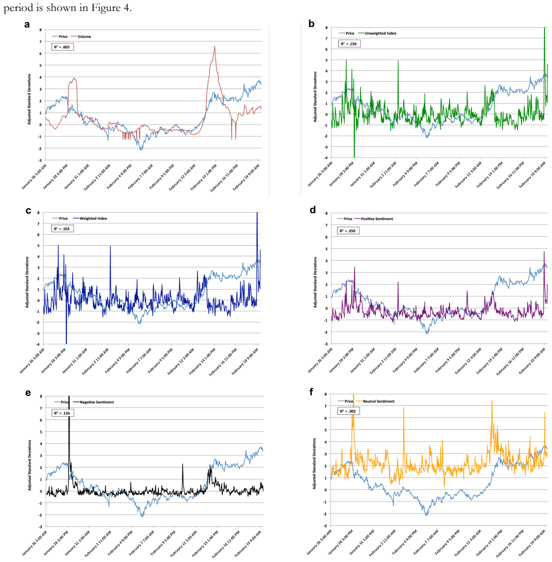

Figure 4: Shown are the price fluctuations versus our chosen six features, respectively, over the time period under consideration. (a) Price vs. Volume, (b) Price vs. Unweighted Index, (c) Price vs. Weighted Index, (d) Price vs. Pos. Sentiment, (e) Price vs. Neg. Sentiment, and (f) Price vs. Neutral Sentiment.

	Correlation Coefficients
Trading Volume	0.605
Neutral	0.302
Positive	0.250
Negative	0.156
Unweighted Index	0.150
Weighted Index	0.103

Figure 5: 5-factor correlation coefficients between the chosen feature and the price data, respectively.

RESULTS CONTINUED

Figure 6: Comparison of model prediction and actual price data. Panel (a) plots the fitted price curve obtained from the training price data and the predicted price curve with respect to the testing data. Panel (b) details the model prediction price data as compared to the testing real price data.

In testing our model, we were able to produce price data that strongly reflected the actual fluctuations (see Fig. 2). In particular, it is significant that our model achieved a Pearson correlation of .806 when tested against the actual test data, yielding a statistical significance at the $p<0.0001$ level. As such, our model provides a viable method to predict price fluctuations, and also serves as a proof of concept that statistical analyses using Twitter sentiment can also be used to analyze price fluctuations in additional cryptocurrencies.

It is also interesting to note that despite the similar directionality between the price model and actual price fluctuations, there appears to be a price gap between the two of around $30 (see Fig. 2b). One possible explanation to this gap is the discrepancies between the training and testing data (as summarized in Table 4). First, it is important to note that the model was trained on data that primarily exhibited a negative trend (see Table 4). As such, it is possible that the model became more desensitized to positive stimuli, and more sensitive to negative stimuli. In the testing data, however, the model was exposed to ~ 3% decrease in positive stimuli and ~ 0.5% increase in negative stimuli (Table 4). The number of average tweets per hour also increased by ~ 15% (Table 4). As such, it is possible that the model reacted to the change in these factors by exhibiting a slightly lower price expectation than what the actual market reflected. However, the overall directionality and correlation within the model remained strong, suggesting that if the model were also trained on data that exhibited positive trends, a more accurate set of predictions would have resulted.

	Positive	Negative	Neutral	Total
Average Hourly Tweets in Training Data	73.7	18.8	75.2	167.7
Percentage of Total	43.9%	11.2%	44.8%	-
Average Hourly Tweets in Testing Data	82.4	23.5	95.5	201.4
Percentage of Total	40.9%	11.7%	47.4%	-

Figure 7: Discrepancies of Twitter sentiments between testing and training data.

CONCLUSION

In conclusion, our results suggest that by analyzing Twitter sentiment and trading volume, an Extreme Gradient Boosting Regression Tree Model serves as a viable means of predicting price fluctuations within the ZClassic cryptocurrency market. As such, given the complete lack of research within this academic sphere, our model serves as a proof of concept that social media platforms such as twitter can be used to capture investor sentiment, and that this sentiment is an early signal to future price fluctuations in alternative cryptocurrencies. Of particular interest is seeing whether this approach produces similarly strong results when applied to other alternative cryptocurrencies such as ZCash and Bitcoin Private. However, this discovery sheds light to the possibility of arbitrage opportunities that utilize social media platform sentiment to predict future cryptocurrency prices.

Our pricing model could be further improved by factoring in other social media platforms or data, such as Google Search results, Facebook posts, and Reddit Posts. Moreover, the dictionary that we have used in our model could be also be aided by adding investment-specific terms that indicate positive and negative sentiment such as "bull" and "bear" respectively. As seen from our manual vs. algorithm cross-analysis, the algorithm's greatest weakness is in classifying tweets that should otherwise be characterized as "negative" as "positive." After careful review it is evident that such inaccurate characterizations are due to the algorithm's inability to detect sarcasm, a pervasive language schema in popular social media platforms. As such, further research to enhance our algorithm to detect sarcasm would increase the reliability of the sentiment analysis, and thereby potentially improve the accuracy of our prediction to retail driven price changes.

REFERENCES

[1] Greenberg, A. Crypto Currency. https://www.forbes.com/forbes/2011/0509/technology-psilocybin-bitcoins-gavi andresen-crypto- currency.html, April 2011 (accessed 30 April 2018).
[2] Allison, I. If Banks Want Benefits of Blockchains, They Must Go Permissionless. https://www.ibtimes.co.uk/nick-szabo-if banks-want- benefits-blockchains-they-must-go-permissionless-1518874/, September 2015 (accessed 30 April 2018).
[3] Nakamoto,S.Bitcoin:APeer-to PeerElectronicCashSystem.https://bitcoin.org/bitcoin.pdf,October2008(accessed30April2018).
[4] Davis,J.TheCrypto Currency:Bitcoinanditsmysteriousinventor,TheNewYorker.https: //www.newyorker.com/magazine/2011/10/10/the- crypto-currency, October 2014 (accessed 30 April 2018).
[5] ElBahrawy A, Alessandretti L, Kandler A, Pastor-Satorras R, Baronchelli A. Evolutionary dynamics of the cryptocurrency market. Royal Society Open Science. 2017 Nov 1;4(11):170623.
[6] ChartfromCoinMarketCap:https://coinmarketcap.com/currencies/zclassic/,(accessed30April2018).
[7] Cornish,C.GrowingNumberofCryptocurrenciesSparkConcerns,FinancialTimes.https://www.ft.com/content/a6b90a8c f4b7-11e7-8715- e94187b3017e, January 2018 (accessed 30 April 2018).
[8] Parino F, Gauvin L, Beiro MG. Analysis of the Bitcoin blockchain: Socio-economic factors behind the adoption. arXiv preprint arXiv:1804.07657. 2018 April 20
[9] Macedo, J.M. Are We in a cryptocurrency bubble? A comparison with the 2000 dotcom bubble. FreeCodeCamp.com.https://medium.freecodecamp.org/are-we-in-a-cryptocurrency-bubble-a-comparison-with-the-2000 dotcom-bubble- a463d8dd8d8b/, 29 November 2017 (accessed 30 April 2018).
[10] BarivieraAF.TheinefficiencyofBitcoinrevisited:Adynamicapproach.EconomicsLetters.2017Dec1;161:1-4.
[11] KrafftPM,DellaPennaN,PentlandA.AnExperimentalStudyofCryptocurrencyMarketDynamics.arXivpreprintarXiv:1801.05831.2018 Jan 17.
[12] GuoT,Antulov-FantulinN.Predictingshort-termBitcoinpricefluctuationsfrombuyandsellorders.arXivpreprintarXiv:1802.04065.2018 Feb 12.
[13] Leinz, K. "Who Is Buying Bitcoin? This Charts Reveals the Answer—Money." Time, 24 Jan. 2018, time.com/money/5116904/who-is-buying-bitcoin/ (accessed on 30 April 2018).
[14] Meucci,A.'P'Versus'Q':DifferencesandCommonalitiesbetweentheTwoAreasofQuantitativeFinance(January22,2011).GARPRisk Professional, pp. 47-50, February 2011. https.//ssrn.com/abstract=1717163
[15] GarciaD,SchweitzerF.SocialsignalsandalgorithmictradingofBitcoin.RoyalSocietyopenscience.2015Sep1;2(9):150288.
[16] KimYB,KimJG,KimW,ImJH,KimTH,KangSJ,KimCH.Predictingfluctuationsincryptocurrencytransactionsbasedonusercomments and replies. PloS one. 2016 Aug 17;11(8):e0161197.
[17] PhillipsRC,GorseD.Predictingcryptocurrencypricebubblesusingsocialmediadataandepidemicmodelling.InComputationalIntelligence (SSCI), 2017 IEEE Symposium Series on 2017 Nov 27 (pp. 1-7). IEEE.
[18] Williams-Grut, Oscar. The cryptocurrency market is now doing the same daily volume as the New York Stock Exchange. Business Insider. 20 December 2017.
[19] Hayes,A.'HardFork.'Investopedia,20Mar.2017,www.investopedia.com/terms/h/hard-fork.asp(accessedon30April2018)
[20] M.W.Kearney,rtweet:CollectingTwitterData,2017.https://cran.r-project.org/package=rtweet.
[21] Chen T, Guestrin C. XGboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIG KDD International conference on Knowledge Discovery and Data Mining 2016 Aug 13 (pp. 785-794). ACM.