Bloggers and Bitcoin Prices: A Textual Machine Learning Analysis

Eric Ghysels
UNC Chapel Hill

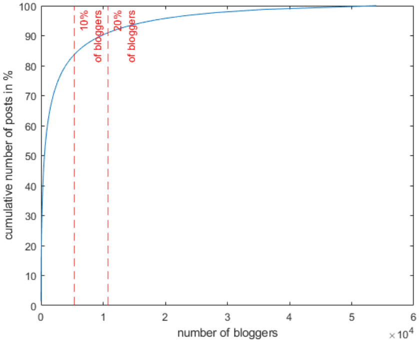
Joint with
Giang Nguyen (Penn State)
Donghwa Shin (UNC Chapel Hill)
Zhe Wang (UNC Chapel Hill)

October 27, 2020

Motivation

- Absence of traditional information intermediaries in the cryptocurrency market
  - In the stock market, sell-side analysts produce earning forecasts and recommend stocks based on both public (such as financial statements) and private information.
- Given the absence of reporting requirements in the cryptocurrency market, it is not clear
  - Whether private information held by individuals have any values
  - How to extract private information of the individuals
- We study a discussion forum specialized in the cryptocurrency market where anonymous individuals freely discuss their views using a state-of-the-art textual machine learning technique.

Growing popularity of BitcoinTalk initiated by Satoshi Nakamoto



- The forum has experienced a significant growth since its inception (at the same time as Bitcoin). More than 200,000 blogs per year recently.
- Importance of active bloggers
  - 10% (20%) of bloggers write 84% (91%) of overall posts.
  - The discussion forum is dominated by a small fraction of active bloggers.

Preview of empirical findings

- A traditional dictionary-based model is not useful to predict future returns.
- A machine learning-based model does not show predictability for daily aggregated posts.
- Importance of individual blogger-level modeling
  - Individual bloggers appear to have different writing styles (based on Jaccard distance)
  - Individual bloggers exhibit heterogeneity in predictability.
  - Interesting to observe that posts which get more comments from other bloggers exhibit poorer performance. $\rightarrow$ Implies the importance of understanding how the bloggers interact.

Literature

- Wisdom of Crowds (relatively new field in finance and economics)
  - Chen, De, Hu, and Hwang (RFS 2014): Study the predictability of stock opinion transmitted through social media (Seeking Alpha)
  - Budescu and Chen (MS 2015): Focus on how to aggregate dispersed opinions using weighted-average scheme.
  - Da and Huang (MS 2020): Study how individuals use public and private information in earning forecasts and its implication on predictability of group forecast. Encouraging individuals to use more of their private information increases the predictability of the group forecast.
- Unlike the previous literature, extracting the private information is much more challenging from unstructured data and there is no publicly available information in our study.
  - We overcome this barrier by using a state-of-the-art textual machine learning technique.

Data

- BitcoinTalk.org
  - One of the oldest and the most famous online discussion forums
  - Forum where people freely express their views on the prospect of the Bitcoin price
  - We choose posts that contain the keywords, bitcoin or BTC, to exclude the posts that are irrelevant for the prediction of Bitcoin price.
- Kaiko
  - We obtain the prices of Bitcoin in USD in 11 reputable cryptocurrency exchanges.
  - We construct the volume-weighted average bitcoin price across these exchange.
  - Based on the volume-weighted average bitcoin price, we compute the returns for various horizons. (5 minutes - 90 days)

Dictionary-based approach

- Tone Measure based on Dictionary (Harvard psychosocial dictionary, Loughran-McDonald sentiment word lists)
  - "Bag of Words"
  - Tone of the article: weight of negative words (proportional or tf.idf)
  - Return-predictability based on the calculated tone
- Dictionary-dependent
- One dictionary for all (topics, authors, et al...)

Machine learning (ML)-based approach

- Tone Measure based on Machine Learning (Ke, Kelly & Xiu, KKX)
  - Sentiment word counts in article i follow a mixture multinomial distribution:

    $$d_{i,[S]} \sim Multinomial(s_i, p_i O_+ + (1 - p_i) O_-)$$

    Sentiment topics O_+/O_- describes the expected word frequencies in a maximally positive/negative sentiment article.
  - Estimate $O = [O_+ \ O_-]$ using a two-topic model:

    $$\mathbb{E}\tilde{D} = OW$$

    $\tilde{D}$ is the set of sentiment-charged words. W is the sentiment score matrix.

Machine learning (ML)-based approach

- Tone Measure based on Machine Learning (Ke, Kelly & Xiu, KKX)
  - Construct the sentiment-charged words set S (used to estimate $\tilde{D}$):

$$S = \begin{bmatrix} & Article\ 1 & Article\ 2 & \dots & Article\ n \\ word\ 1 & f_{1,1} & f_{1,2} & \dots & f_{1,n} \\ word\ 2 & f_{2,1} & f_{2,2} & \dots & f_{2,n} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ word\ j & f_{j,1} & f_{j,2} & \dots & f_{j,n} \end{bmatrix}$$

  - Construct the sentiment score:

$$W = \begin{bmatrix} Article\ 1 & Article\ 2 & \dots & Article\ n \\ p_1 & p_2 & \dots & p_n \\ 1 - p_1 & 1 - p_2 & \dots & 1 - p_n \end{bmatrix}$$

where $p_i = \dfrac{rank\ of\ return(i)\ in\ all\ returns}{n}$

Machine learning (ML)-based approach

- Tone Measure based on Machine Learning (Ke, Kelly & Xiu, KKX)
  - Construct S with selection:

    $$\hat{S} = \{j : f_j \geqslant 1/2 + \alpha,\ or\ f_j \leq 1/2 - \alpha\} \cap \{j : k_j \geq \kappa\}$$

    where:
    f_j is the frequency with which word j co-occurs with a positive return:

    $$f_j = \frac{\#\ ariticles\ including\ word\ j\ AND\ having\ sgn(return) = 1}{\#\ articles\ including\ word\ j}$$

    k_j is the count of articles including word j (the denominator in f_j), and restrict the analysis to words for which $k_j > \kappa$.
  - α and κ are hyper-parameters to be tuned.

Machine learning (ML)-based approach

- Tone Measure based on Machine Learning (Ke, Kelly & Xiu, KKX)
  - Scoring new articles through MLE with a penalty term:

$$\hat{p} = \arg\max_{p \in [0,1]} \{ \hat{s}^{-1} \sum_{j=1}^{\hat{s}} d_j log(p\hat{O}_{+,j} + (1-p)\hat{O}_{-,j}) + \lambda log(p(1-p)) \}$$

  - $\hat{s}$ is the total count of words from $\hat{S}$ in the new article
  - λ is a hyper-parameter to be tuned.
  - Do not rely on specific dictionary
  - Topic/author-specific

Sample construction

- Testing period: 2017, 2018, 2019
- Training period: from 2014 to the beginning of testing year
- Sample: all posts containing the key words "btc" or "bitcoin" (case-insensitive)
- Return: 1/7/30/90 days since the time when the blog is published
- Drop blogs that do not have essential keywords (<5%)
- Top 10 bloggers are the most quoted bloggers during our sample period

Predictability (Spearman rank correlation) of daily aggregate posts

	1 day return	7 day return
Full	-0.018	-0.022
Top Decile	-0.053	-0.050
Bottom Decile	-0.034	-0.007

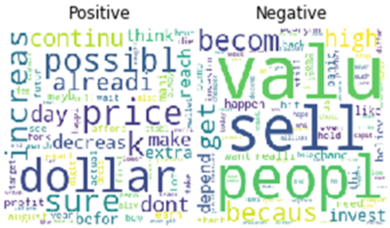
- KKX constructed based on daily aggregate posts does not show predictability.

Summary statistics - Individual bloggers

				# Words				
Blogger	# blogs (1)	Start date (2)	End date (3)	Average (4)	Stdev (5)	25 percentile (6)	Median (7)	75 percentile (8)
B1	978	16-Jul-2017	28-Jul-2019	74.7	68.3	23	52	120
B2	1024	26-Oct-2013	31-Jul-2019	57.6	46.9	27	46	72.25
B3	511	28-Jan-2014	31-Jul-2019	83.3	75.2	33	59	110
B4	371	16-Dec-2013	2-Aug-2019	41.6	53.4	15	28	48.5
B5	525	28-Sep-2014	1-Aug-2019	48.5	31.2	27	41	66
B6	596	8-Oct-2013	15-Jul-2019	58.6	61.8	20	39	74
B7	408	4-Dec-2013	25-Jul-2019	66.8	49.7	40	58	79
B8	242	9-Jan-2017	26-Jun-2019	72.3	95.4	21	40	87.5
B9	110	29-May-2016	15-Jul-2019	42.3	48.9	12	23	46
B10	184	9-Nov-2013	20-Jun-2019	55.4	44.7	25	42	69

- The average (median) # of words range from 41.6 (28) to 83.3 (59).
- Individual bloggers seem to have different writing styles in terms of average (median) # words.
- The length of blog posts are much shorter than newspaper articles or other social media posts (such as Seeking Alpha)

Word cloud



- An example of a word cloud of a blogger constructed based on KKX.
- Is there any difference between the word clouds of the different bloggers?

Comparison of Writing Styles: Jaccard Index

- Similarity between two word sets measured using Jaccard index:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- Positive words: $O_+(i) > O_-(i)$
- Negative words: $O_+(i) < O_-(i)$
- Writing style comparison: Jaccard index of the positive/negative word sets between two individuals

Different writing styles of individual bloggers

Panel A: Positive words

Blogger	B1	B2	B3	B4	B5	B6	B7	B8	B8	B10
B1	1									
B2	0.101	1								
B3	0.122	0.172	1							
B4	0.116	0.135	0	1						
B5	0.118	0.139	0.158	0	1					
B6	0.124	0.153	0.178	0	0.158	1				
B7	0.113	0.148	0.178	0.162	0	0	1			
B8	0.101	0.114	0.149	0.177	0	0	0	1		
B9	0.045	0.056	0.057	0.064	0	0	0	0	1	
B10	0.087	0.125	0.126	0.148	0	0	0	0	0	1

- The above table presents the Jaccard index between word clouds of 10 bloggers. Jaccard index measures the similarity of word clouds.
- Individual bloggers appear to have different writing styles. Consistent with the poor performance of aggregate posts.

Different writing styles of individual bloggers

Panel B: Negative words

Blogger	B1	B2	B3	B4	B5	B6	B7	B8	B8	B10
B1	1									
B2	0.117	1								
B3	0.122	0.179	1							
B4	0.125	0.137	0	1						
B5	0.138	0.159	0.156	0	1					
B6	0.134	0.178	0.159	0	0.153	1				
B7	0.124	0.155	0.168	0.149	0	0	1			
B8	0.135	0.152	0.155	0.177	0	0	0	1		
B9	0.076	0.071	0.079	0.092	0	0	0	0	1	
B10	0.117	0.122	0.140	0.134	0	0	0	0	0	1

Predictability of individual bloggers

Blogger	1 day return	7 day return	30 day return	90 day return	# obs
B1	0.038	0.110(***)	0.168(***)	0.060(*)	868
B2	0.037	0.136(***)	0.105(**)	0.134(***)	512
B3	0.007	-0.033	0.037	0.200(***)	326
B4	0.042	0.034	0.032	0.101(*)	294
B5	0.029	-0.033	-0.103(*)	0.058	272
B6	0.106(*)	-0.033	-0.003	0.002	253
B7	-0.079	0.093	0.057	0.067	170
B8	-0.093	0.030	0.015	0.116	137
B9	-0.118	-0.035	0.430(***)	0.400(***)	66
B10	0.037	-0.078	-0.126	0.016	53

- Spearman Rank Correlation is presented.
- Heterogeneous predictability across different bloggers and horizons.

Horse race (KKX vs Dictionary-based approach)

Blogger		1 day			7 day	
	QPS_{KKX}	$QPS_{Dictionary}$	ΔError(KKX-Dic)	QPS_{KKX}	$QPS_{Dictionary}$	ΔError(KKX-Dic)
B1	0.592	0.701	-0.059(***)	0.692	0.708	-0.008
B2	0.500	0.528	-0.014(**)	0.499	0.513	-0.007
B3	0.501	0.557	-0.028(**)	0.535	0.609	-0.037(**)
B4	0.499	0.566	-0.033(*)	0.505	0.618	-0.056(***)
B5	0.500	0.558	-0.029(*)	0.496	0.547	-0.026(**)

Blogger		30 day			90 day	
	QPS_{KKX}	$QPS_{Dictionary}$	ΔError(KKX-Dic)	QPS_{KKX}	$QPS_{Dictionary}$	ΔError(KKX-Dic)
B1	0.763	0.823	-0.030(***)	0.844	0.931	-0.044(***)
B2	0.491	0.540	-0.024(***)	0.505	0.55	-0.022(***)
B3	0.591	0.668	-0.039(***)	0.680	0.76	-0.040(***)
B4	0.488	0.540	-0.026(*)	0.389	0.397	-0.004
B5	0.494	0.499	-0.003	0.506	0.537	-0.015

- $QPS = \frac{1}{T}\sum 2 * (\hat{p} - p)^2$, quadratic probability score (Brier, 1950). Range [0,2]. 0 is perfect accuracy.
- ΔError: Diebold-Mariano test
- KKX produces more accurate forecast than the traditional dictionary-based approach.

Predictability and attentions

	In Top Decile			Not In Top Decile		
Blogger	Correlation	p-value	# Blogs	Correlation	p-value	# Blogs
B1	0.040	0.377	497	0.142(***)	0.005	386
B2	0.050	0.624	99	0.160(***)	0.001	413
B3	0.021	0.807	135	-0.090	0.213	192
B4	-0.015	0.884	100	0.086	0.231	207
B5	-0.077	0.537	67	-0.020	0.777	205
B6	-0.127	0.366	53	-0.003	0.963	201
B7	-0.029	0.860	40	0.136	0.122	130
B8	-0.080	0.480	80	0.190	0.157	58
B9	-0.021	0.932	29	-0.028	0.852	76
B10	0.452	0.260	8	-0.110	0.470	45

- 7 day return prediction
- It is puzzling to observe that when a blogger got more attention measured by the number of comments (replies), the predictability seems to be worse.
- Why? Need to further understand the feedback by other bloggers.
  - When do people leave comments?
  - Is there any asymmetry in leaving the comments?

Combine Probability Forecast

- Aggregation of probability forecasts from distinct sources with different information set (Ranjan and Gneiting, 2010).

$$p_t = H_{\alpha,\beta}(\sum w_i p_{i,t})$$

  - p_t is the combined forecast
  - $p_{i,t}$ is forecast from source i
  - $H_{\alpha,\beta}$ is a cumulative β distribution
  - $\sum w_i = 1$
  - Estimate parameters (w_i, α, β) by maximizing the following log likelihood function:

$$l(w_i, \alpha, \beta) = \sum y_t * log(p_t) + (1 - y_t) * log(1 - p_t)$$

where y_t=1 if the future return is positive and 0 otherwise.

Combine KKX with Return-Driven Model

Blogger	7 day				
	QPS_{KKX}	QPS_{Ret}	Combined	ΔError(Com-KKX)	ΔError(Com-Ret)
B1	0.692	0.708	0.510	-0.091(***)	-0.099(***)
B2	0.499	0.513	0.464	-0.018(*)	-0.025(**)
B3	0.535	0.609	0.432	-0.052(***)	-0.089(***)
B4	0.505	0.618	0.703	0.099(***)	0.042
B5	0.496	0.547	0.495	-0.001	-0.026(*)

Blogger	30 day				
	QPS_{KKX}	QPS_{Ret}	Combined	ΔError(Com-KKX)	ΔError(Com-Ret)
B1	0.844	0.931	0.474	-0.185(***)	-0.228(***)
B2	0.505	0.550	0.423	-0.041	-0.063(**)
B3	0.680	0.760	0.473	-0.103(**)	-0.143(***)
B4	0.389	0.397	0.692	0.152(***)	0.148(***)
B5	0.506	0.537	0.488	-0.009	-0.025

- Combine KKX with return-driven model improves the accuracy.

Conclusion

- A traditional dictionary-based model is not useful to predict future returns.
- A ML-based model does not show predictability for daily aggregated posts.
- Importance of individual blogger-level modeling
  - Individual bloggers appear to have different writing styles.
  - Individual bloggers exhibit heterogeneity in predictability.
  - Interesting to observe that posts which get more comments from other bloggers exhibit poorer performance.
- Future works
  - The feedbacks by other bloggers.
  - Can we aggregate the outcomes of the individual models to construct a better predictor? (Wisdom of Crowds)