An Analysis of the Change in Discussions on Social Media with Bitcoin Price

Andrew Burnie
The Alan Turing Institute and University College London
aburnie@turing.ac.uk

Emine Yilmaz
The Alan Turing Institute and University College London
emine.yilmaz@ucl.ac.uk

ABSTRACT

We develop a new approach to temporalizing word2vec-based topic modelling that determines which topics on social media vary with shifts in the phases of a time series to understand potential interactions. This is particularly relevant for the highly volatile bitcoin price with its distinct four phases across 2017-18. We statistically test which words change in frequency between the different stages and compare four word2vec models to assess their consistency in relating connected words in weighted, undirected graphs. For words that fall in frequency when prices shift from rising to falling, all eight topics are identified with the four approaches; for words rising in frequency, three out of the five topics remain constant. These topics are intuitive and match with actual events in the news.

KEYWORDS

Content Analysis; Social Media; Bitcoin; Word2Vec; Topic Modelling; Graph Theory

ACM Reference Format:
Andrew Burnie and Emine Yilmaz. 2019. An Analysis of the Change in Discussions on Social Media with Bitcoin Price. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331304

1 INTRODUCTION

We develop a methodology and compare four practical implementations of word2vec to examine how social media discussions vary with different phases in a bitcoin price time series. This helps elucidate potential interactions between changes on social media and the time series. The framework is applied to the 2017-18 bitcoin price series as it shifted across distinct stages, including rising (by twenty-fold), falling (by 70%) and relatively stable prices (Figure 1).

In previous literature topics were identified by looking at the complete text. Different topics over time were then matched with changes to the bitcoin price series [16]. Topic models can be trained across the data either ignoring time [16] or incorporating it into the learning algorithm [5, 27].

Our main contribution is to ask what words changed in frequency with shifts in the bitcoin price series and then use word2vec-based clustering of text to identify which topics were associated with these words. This focusses topic generation on themes whose popularity varied with changes in price.

After identifying stages in the bitcoin price (Section 2.1) and processing the text (Section 2.2 to 2.3), non-parametric statistics (Section 2.4) are applied to delineate words that changed statistically significantly in frequency across stages in the price series. The significant risers and fallers are placed on separate undirected graphs where each edge has a weight measuring how similar the context was in which the two words were used [13]. The weight is the cosine similarity between the word2vec-derived numeric vectors [19, 20]. A threshold is applied to enable the inferring of distinct topics (Section 2.5).

We also compare four different word2vec architectures in deploying this framework. Using a neural network trained to predict the current word using its context, the Continuous Bag-of-Words model (CBOW), is evaluated against training to predict the context using the current word, the continuous Skip-gram (SG) model [19]. Computational complexity being mitigated through the original approach of Hierarchical Softmax (HS) [19] is assessed against the alternative Negative Sampling (NEG) [20].

Evaluation metrics are developed that consider the number and quality of the topics identified (Section 3.3). Experimentation applies these to compare different practical implementations of the framework (Section 3.4). The optimal approach is applied on the phasic shift with the most data available to enable qualitative evaluation (Section 3.5).

Word2vec representations have previously been clustered using k-means [14, 16]. Applying thresholds to weighted, undirected graphs is preferred in obviating the need to select the number of topics and in allowing for polysemy.

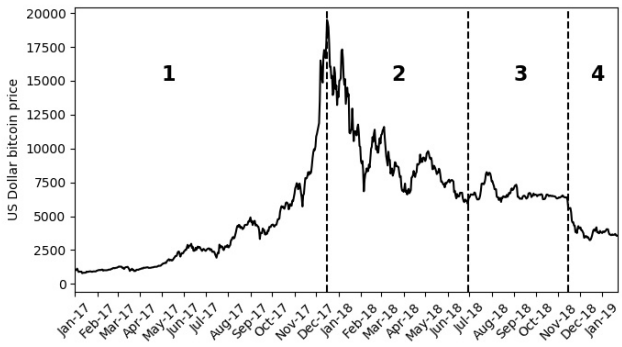

Figure 1: The Four Stages of the US Dollar Bitcoin Price

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331304

2 THE MODEL AND METHODOLOGY

2.1 Phasic Shifts in Bitcoin Price

Daily US Dollar prices from Blockchain Luxembourg S.A. [6] from 1 January 2017 to 22 January 2019 split into four phases (Figure 1):

(1) Stage 1: Prices rose 1854% 1 January to 16 December 2017.
(2) Stage 2: Prices fell 70% until 29 June 2018, in a series of cycles.
(3) Stage 3: Prices were relatively stable until 15 November 2018, varying between 30%-42% of the 16 December 2017 peak.
(4) Stage 4: Prices fell to 55% of the previous low (29 June 2018). The final price was 60% that on 29 June 2018.

2.2 Choice of Social Media Source

The 'Bitcoin' subreddit was selected because moderators act to ensure that the 'primary topic is Bitcoin' [24] and it has a larger userbase than bitcointalk.org [15, 16]. On 1 February 2019 at 19:14 UTC, there were 900 users online in bitcointalk.org [4] compared to 4100 on the Reddit forum [24].

Twitter has been used [1, 9] but lacks moderators and has an estimated 10 million likely fake accounts being created per week to tweet artificial opinions [21]. Lamon et al [17] found Reddit posts outperformed both Twitter data and news headlines in predicting the bitcoin price.

2.3 Text Preparation

Reddit submissions text was extracted from the Pushshift API [2]. Submissions were more relevant than comments, which were prone to irrelevant discussion topics (https://www.reddit.com/r/Bitcoin/comments/9svjcp/10_years_ago_today_2008_oct_31/). Uninformative text was removed: automated submissions ('rBitcoinMod' and 'crypto_bot'); blank, repetitive or removed submissions; text that did not relate to words (including punctuation, URLs, HTML tags, social media handles, non-ASCII and text of more than 50 consecutive word characters); and stopwords (provided by NLTK Version 3.3). Words relating to the same concept were standardised with all text placed into lower case, lemmatised using NLTK's 'WordNetLemmatizer' and stemmed using 'SnowballStemmer' [12]. Words and their common abbreviations were equalised: BTC and XBT were converted into 'bitcoin'; '$', 'usd', 'dollar(s)' and 'us dollar(s)' into 'dollar_marker_symbol'; and 'ln' and 'lightning network' into 'ln'. Words in 100 or less submissions were removed.

2.4 Delineating Words that Changed in Frequency

Daily word frequency was the proportion of submissions on a day that contained that word. For each consecutive pair of stages in the price series (Section 2.1), two-sided Wilcoxon Rank Sum Tests were applied to determine which words had changed statistically significantly in daily frequency using a p-value cut-off of 1%. This was Bonferroni-corrected [18] to mitigate against the higher risk of false positives associated with the large number of tests.

2.5 Topic Modelling

Word2vec models were trained ('gensim' Version 3.5.0 [26]) using the default hyperparameter values suggested by gensim [26], except that the number of noise words drawn (in the case of negative sampling) and iterations were increased to 20 to reflect the limited dataset size [19, 20]. Words with a total frequency below 100 were excluded. These were trained using text from all submissions from 00:00 1 January 2017 to before 00:00 23 January 2019 (UTC).

The trained models were applied to words that had risen and fallen significantly in frequency across consecutive pairs of stages. Word2vec assigned to each word a vector of 100 continuous-scaled numbers. Each word was placed into a graph ('NetworkX' Version 2.2 [22]) where the weight of each edge corresponded to the cosine similarity between the words' vectors. A threshold was applied to remove the edges with the lowest cosine similarities. Topics were identified as groups of more than one word that were connected with each other and not connected with words outside the group.

Different practical implementations of this topic modelling methodology are compared in experiments detailed subsequently (Section 3).

Table 1: Datasets

Stage	Days	Submissions	Submissions per Day	Risers	Fallers
All	752	338415	450.02	N/A	N/A
1	349	181327	519.56	N/A	N/A
2	195	101110	518.51	129	586
3	139	38706	278.46	83	40
4	69	17272	250.32	63	8

3 EXPERIMENTS

3.1 Datasets

Table 1 shows a decline in the number of words that statistically significantly rose and fell over time as successive stages had fewer associated days and submissions, and so less data available. This was exacerbated by a decline in Reddit activity. Over 500 submissions per day were being posted on average in Stages 1 and 2, the periods when prices were most volatile. This fell 46% as prices stabilised (Stage 3) and a further 10% in Stage 4.

3.2 Model Variants

Experimentation compares different practical implementations of the discussed framework (Section 2). This includes four word2vec approaches (CBOW, SG, HS and NEG) and the following percentile thresholds applied to the graph: 90, 95, 99, 99.90, 99.95 and 99.99. A pre-trained model was not used in comparison as these were developed for words without stemming and lemmatisation [10].

3.3 Evaluation Metrics

A 'group' here refers to when there are two or more words that are connected by edges. The words within each group generated should be similar to each other and dissimilar with words outside the group. The median cosine similarity between words within the same group ('INTRA') and between words in a group and words outside ('INTER') were calculated. These are of the same scale and so INTER was deducted from INTRA to provide a measure of the quality of the groups generated. Using just this quality metric resulted in only

Table 2: Grouping 586 Words Falling from Stages 1 to 2

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.90	0.6863	0.1939	73	35.95
SG, HS	99.90	0.5946	0.0360	67	37.43
CBOW, NEG	99.90	0.6626	0.0206	62	39.81
CBOW, HS	99.90	0.5742	0.002178	62	35.46

Table 3: Grouping 129 Words Rising from Stages 1 to 2

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.6139	0.1865	19	8.12
SG, HS	99.00	0.6815	0.0790	19	11.45
CBOW, NEG	99.00	0.5068	0.0573	15	6.74
CBOW, HS	99.00	0.4806	0.0297	17	7.67

Table 4: Grouping 40 Words Falling from Stages 2 to 3

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	95.00	0.5893	0.1696	7	2.94
SG, HS	95.00	0.6013	0.0530	8	4.39
CBOW, NEG	95.00	0.5879	0.0351	6	3.32
CBOW, HS	95.00	0.5476	0.0120	7	3.75

one or two groups being generated with the exception of the words that fell from Stages 1 to 2.

The more groups, the more potential, distinct topics that can be interpreted from them, and so the quality metric was multiplied by the number of groups generated (Equation 1), resulting in an objective evaluation score ('EVAL'). Models with a negative INTRA or INTER score were excluded. This meant that the same percentage increase in either quality or number of groups had the same impact on the evaluation metric.

$$EVAL = (INTRA - INTER) \times \text{Number of Groups} \quad (1)$$

3.4 Evaluating Model Variants

Examining Stages 1 to 2 (Tables 2 and 3), the optimal threshold (by EVAL), the number of groups and the evaluation score were similar for a given table. There was no consistent tendency for one word2vec architecture to outperform all others. The exception was that, for the risers from Stages 1 to 2 (Table 3), the evaluation score for SG with HS was 41% higher (11.45) compared with the second highest value (8.12). This trend continued for Stages 2 to 3 (Tables 4 and 5) and Stages 3 to 4 (Table 6). The results for the words falling from Stages 3 to 4 are not shown as there were only eight words and only one group was generated across the models compared. The more words available for extracting topics, the more groups were generated and the higher the optimal threshold.

3.5 Topic Modelling

The shift from Stage 1 (rising) to Stage 2 (falling) prices had the most associated data (Table 1). The optimal model for falling words was CBOW with NEG (Table 2) and for rising was SG with HS (Table 3). The largest groups (with more than three words) are displayed

Table 5: Grouping 83 Words Rising from Stages 2 to 3

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.5514	0.2154	8	2.69
SG, HS	99.00	0.5881	0.0895	10	4.99
CBOW, NEG	99.00	0.5517	0.0509	12	6.01
CBOW, HS	99.00	0.4932	0.0286	12	5.58

Table 6: Grouping 63 Words Rising from Stages 3 to 4

Model	Threshold	INTRA	INTER	Groups	EVAL
SG, NEG	99.00	0.6918	0.2246	8	3.74
SG, HS	99.00	0.7712	0.0607	9	6.39
CBOW, NEG	99.00	0.8036	0.0317	8	6.18
CBOW, HS	99.00	0.7005	0.0060	8	5.56

in Figures 2 and 3 using a force-embedded algorithm [8] to display the graph for each group.

The topics generated by the different approaches were similar, despite CBOW predicting current words using their context and SG using words to predict their context [19]. The optimal model identified eight topics in the fallers which when the other three model variants were examined were constant. For risers, the results were again constant for the SG with NEG and with HS, but no 'ICO' topic could be found for CBOW with NEG and no 'Startup' topic could be found for CBOW with HS.

The largest groups of rising words clustered around five topics (Figure 2) centred on 'East Asia', 'Competition', 'Startup', 'ICO' and the 'Lightning Network'. Regarding East Asia, Japanese Coincheck and South Korean Bithumb were both subject to investigations and hacks [23], whilst 'giant' could refer to large Japanese firms entering partnerships with exchanges to accept bitcoin [25]. Bitcoin competitors that became more discussed included Tron ('trx'), Stellar, EOS ('eo'), Cardano, Ripple ('rippl', 'xrp') and Verge ('verg'). The 'Startup' topic focussed on incubators ('incub'), the Silicon Valley ('silicon', 'valley' and 'bay'), investment ('angel') and founders. There was also growing interest in Initial Coin Offerings ('ico') and the Lightning Network.

The largest groups of falling words clustered around eight topics (Figure 3) which reflected a notable fall in discussion around how Bitcoin works. This involved topics covering 'Wallet', 'Transfer', 'Exchanges', 'Password' and 'Posts'. In response to escalating confirmation times (topic 'Confirmation'), a split (topic 'Fork') emerged between Bitcoin Unlimited ('bu' and 'unlimit'), for a larger blocksize limit (topic 'Blocksize'), and Segregated Witness (SegWit), for moving information off network [7]. This involved protests such as the User-Activated Soft Fork ('uasf') Bitcoin Improvement Proposal 148 ('bip148') [11] and the abortive compromise SegWit2x [3]. Results show how interest in this debate declined after SegWit was implemented and Bitcoin Cash was forked, both on 1 August 2017 [7].

4 CONCLUSIONS

In this paper, we show the value of first identifying the words that changed with phasic shifts in a time series and then using

word2vec to extract which topics they relate to. Experimentation demonstrates its effectiveness in generating interpretable groups reflecting distinct topics, and that similar topics are generated by the four different word2vec approaches both quantitatively and qualitatively. For fallers (Stages 1 to 2), all eight topics were represented; for risers, three out of five topics were constant. This temporalizes word2vec-based topic modelling and is widely applicable to associating changes in social media discussions with external events.

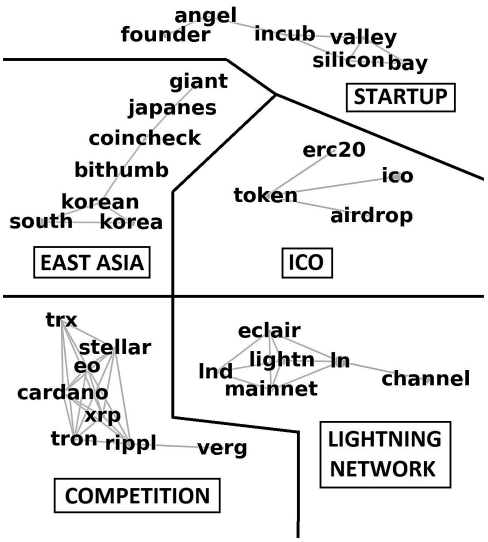

Figure 2: Groups Rising in Frequency from Stages 1 to 2.

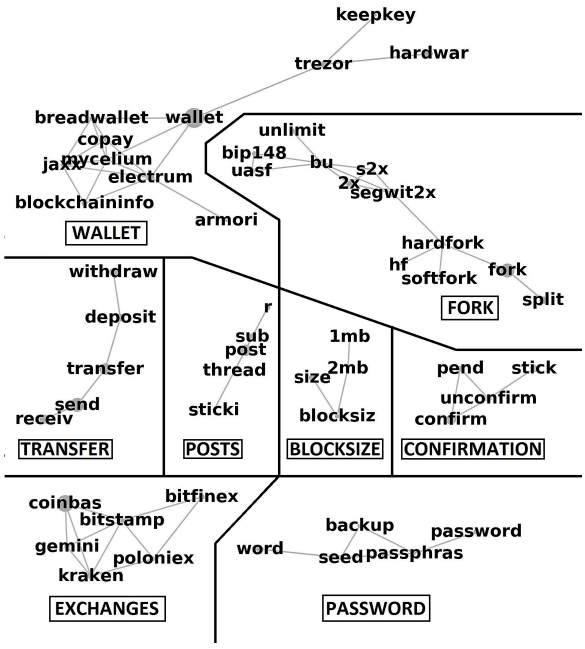

Figure 3: Groups Falling in Frequency from Stages 1 to 2.

ACKNOWLEDGMENTS

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and Turing award number TU/C/000028. This project was partially funded by the EPSRC Fellowship titled "Task Based Information Retrieval", grant reference number EP/P024289/1.

REFERENCES

[1] Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. 2018. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review* 1, 3 (2018).
[2] Jason Michael Baumgartner. 2018. Pushshift API. https://github.com/pushshift/api.
[3] Mike Belshe. 2017. [Bitcoin-segwit2x] Segwit2x Final Steps. https://lists.linuxfoundation.org/pipermail/bitcoin-segwit2x/2017-November/000685.html
[4] bitcointalk.org. 2019. Statistics Center. https://bitcointalk.org/index.php?action=stats.
[5] David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, Pittsburgh, Pennsylvania, 113–120.
[6] Blockchain Luxembourg S.A. 2018. Blockchain Charts & Statistics API. https://www.blockchain.com/api/charts_api.
[7] Cointelegraph. 2018. What is Bitcoin Cash? *Cointelegraph* (Jan. 2018).
[8] Thomas M. J. Fruchterman and Edward M. Reingold. 1991. Graph drawing by force-directed placement. *Software: Practice and Experience* 21, 11 (1991), 1129–1164.
[9] David Garcia and Frank Schweitzer. 2015. Social signals and algorithmic trading of Bitcoin. *Royal Society Open Science* 2, 9 (2015).
[10] Google. 2013. word2vec. https://code.google.com/archive/p/word2vec/.
[11] Alyssa Hertig. 2017. Bitcoin UASF Proposal Quietly Activates - to Little Effect. *CoinDesk* (Aug. 2017).
[12] Anjali Ganesh Jivani. 2011. A Comparative Study of Stemming Algorithms. *International Journal of Computer Technology and Applications* 2, 6 (2011), 1930–1938.
[13] Andrei Kashcha. 2019. Exploring word2vec embeddings as a graph of nearest neighbors: anvaka/word2vec-graph. https://github.com/anvaka/word2vec-graph
[14] Han Kyul Kim, Hyunjoong Kim, and Sungzoon Cho. 2017. Bag-of-concepts: Comprehending document representation through clustering words in distributed representation. *Neurocomputing* 266 (Nov. 2017), 336–352.
[15] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLOS ONE* 11, 8 (August 2016).
[16] Young Bin Kim, Jurim Lee, Nuri Park, Jaegul Choo, Jong-Hyun Kim, and Chang Hun Kim. 2017. When Bitcoin encounters information in an online forum: using text mining to analyse user opinions and predict value fluctuation. *PLOS ONE* 12, 5 (May 2017).
[17] Connor Lamon, Eric Nielsen, and Eric Redondo. 2018. Cryptocurrency Price Change Prediction Using News and Social Media Sentiment. http://cs230.stanford.edu/files_winter_2018/projects/6929537.pdf
[18] John H. McDonald. 2014. *Handbook of Biological Statistics* (3 ed.). Sparky House, Baltimore, Maryland, U.S.A.
[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (Jan. 2013). arXiv: 1301.3781.
[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.
[21] Steven Musil. 2018. Twitter ramps up effort to combat abusive bots, trolls. *CNET* (June 2018).
[22] NetworkX. 2019. Software for complex networks. https://networkx.github.io/.
[23] Stephen O'Neal. 2018. From Coincheck to Bithumb: 2018's Largest Security Breaches So Far. *Cointelegraph* (June 2018).
[24] r/Bitcoin. 2019. *Reddit* (Feb. 2019). https://www.reddit.com/r/Bitcoin/
[25] Sujha Sundararajan. 2018. Japanese Electronics Retail Giant Launches Bitcoin Payments. *CoinDesk* (Jan. 2018).
[26] Radim Řehůřek. 2019. models.word2vec - Word2vec embeddings. https://radimrehurek.com/gensim/models/word2vec.html
[27] Zhuofeng Wu, Cheng Li, Zhe Zhao, Fei Wu, and Qiaozhu Mei. 2018. Identify Shifts of Word Semantics Through Bayesian Surprise. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 825–834. https://doi.org/10.1145/3209978.3210040