# Studying the Correlation between Online User Comments and Bitcoin Fluctuations using Latent Dirichlet Allocation

TOM JANSSEN GROESBEEK, Radboud University

Ever since the sudden rise in value, Bitcoin and similar cryptocurrencies have been an interesting topic for discussion. Nowadays numerous forums exist on which discussions related to these currencies are held. One such forum is the website named Reddit. Almost every cryptocurrency has its own subreddit with people speculating on its price. The current study uses Latent Dirichlet Allocation in order to analyse daily Bitcoin subreddit discussions on the topics that are discussed. Then Spearman's correlation coefficients are computed to find out if specific topics correlate to the weekly rise or fall of the Bitcoin price.

Additional Key Words and Phrases: Topic Analysis, Text Mining, Bitcoin, LDA, TF-IDF, Reddit, Correlation Analysis

## 1 INTRODUCTION

Some people see an important future for cryptocurrencies and the blockchain technology behind them, while others believe that we are dealing with a new bubble like the dot-com bubble[1]. While it is still to see if the latter is true, the cryptocurrencies, especially Bitcoin, are gaining in popularity and more people are investing in them[2]. Cryptocurrencies function as digital "money" for users to perform transactions with. They are called "crypto" currencies as they make use of cryptography to protect against attempts of tampering or forgery of transactions made with these currencies [14]. But cryptocurrencies also make an interesting investment, because of the profits that are possible due to sudden fluctuations[3].

As people are investing their money, they also start to share information on social media. One such medium is Reddit, a online source for "what is new and popular on the web"[4]. Almost every cryptocurrency has its own so called subreddit, a seperate forum, on which users can discuss and share information with each other.

Current study sets out to broaden the available literature on the relation between social media and Bitcoin fluctuations. Most studies focus on using known Bitcoin forums (e.g. [9], [8]), while studies related to predicting financial data often make use of Twitter data (e.g. [1], [15]). The subreddit named 'BitCoinMarket', which as the names states is dedicated to Bitcoin, shows increasing number of subscribers since the start of 2014[5] and contains a well moderated daily discussion. This is why this study will analyse the posts of these daily discussions in order to investigate any correlations with each week's closing Bitcoin price. The main research question of this study is: Do the topics discussed on Reddit correlate with the fluctuations of the Bitcoin value?

The remainder of this paper is structured as follows. The next section will provide a summary of the related work. Section 3 will then elaborate on the data that has been collected for this project and the topic modelling technique used. Then section 4 presents the results of the experiments and we conclude this paper with a discussion and conclusion in section 5. Additional figures, data, or important links can be found under Appendices.

## 2 RELATED WORK

Research on predicting financial data using online chatter is nothing new. Take for example the work done by Asur and Huberman, who demonstrate how Twitter content can be used to predict box-office revenues for movies [1]. They saw social media as a form of collective wisdom and hypothesized that when "a movie is well talked about, it will also be well-watched". Their work consists out of making a linear regression model using the tweet-rate, number of tweets per hour, of a movie. Using this model they showed, among other things, that there is a correlation between the attention a movie gets and its future ranking. Similar research was performed by Joshi et al. [6] who used film critics' reviews from several online sources to predict opening weekend revenues. And Gruhl et al. who set out to demonstrate that spikes in book sales can be predicted based on postings in, for example, blogs [3]. Tsapeli et al. were even able to show that social media data do not only correlate with stock market returns, but also influence them [15]. They proposed a causal inference framework for time-series that they then use to show that the sentiment of tweets does have a causal impact on stock market prices of four big technological companies.

So with the rise in popularity of Bitcoin and other cryptocurrencies which in turn led to the existance of online chatter and communities. It is no surprise that new studies focus on the possibilities of social media to predict the fluctuation of cryptocurrencies. The study performed by Kim et al. is a good example [9]. Their study focusses on extracting keywords of interest from user comments posted on Bitcoin online forums. In turn they developed a model based on deep learning to predict the Bitcoin transaction count and price. Similar work in 2016 was already performed by Kim et al. [8]. However, the interest of this work was to determine how the opinions of online community users are related to cryptocurrency price and transaction fluctuations. Noteworthy is that their research was not focussed solely on Bitcoin, but on several cryptocurrencies.

Other work on cryptocurrency and related online postings is done by Linton et al. [10]. Who made use of dynamic topic modelling (DTM) to study the evolution of topics found on popular cryptocurrency message boards. They then also investigate the relation between these topics and cryptocurrency events as well as the predictive power of DTM. Their work provides the steps to construct such a model and also shows that the number of topics are a good indicator of an event. Another known topic modelling technique is Latent Dirichlet Allocation (LDA) by Blei et al. [2]. According to Linton et al. the aspect of time is lost when using LDA, because it

[1] https://www.bloomberg.com/gadfly/articles/2017-11-27/bitcoin-crash-may-take-a-while-longer

[2] https://www.cnbc.com/2017/11/28/bitcoin-surpasses-10000-for-the-first-time.html

[3] https://www.forbes.com/sites/rogeraitken/2017/11/15/cryptocurrency-cfd-trading-consumers-warned-over-high-investing-risks/#3d0de0e844b9

[4] https://www.reddit.com/wiki/faq

[5] http://redditmetrics.com/r/BitcoinMarkets

assumes that all documents are exchangeable. Which makes event detection pointless. The DTM technique analyses documents in discrete time slices and makes the assumption that the topics in these documents evolve over time.

Overall there is quite some literature on topic modelling used on social media data (e.g. [18], [11], [13]). As well as research on sentiment analysis performed on online postings (e.g. [5], [12], [1]). Current work sets out to broaden the available literature on the relation between financial data and social media postings. In order to do so, Reddit is examined instead of popular choices like Twitter or Facebook. Topic modelling will be performed on user posts to see if the underlying topics have any correlation to the fluctuations of the Bitcoin price.

## 3 METHODS

### 3.1 Data

The current study collected data from the subreddit named 'BitCoinMarket' and its daily discussions. This subreddit is focussed on the bitcoin price and contains discussions on trading and price speculations. Other subreddits on other cryptocurrencies, like Ethereum and Omg, were also examined. However, Bitcoin is one of the oldest and well known cryptocurrencies out there. This results in a larger amount of subscribers to the related subreddit and also results in a more stable daily discussion over time. Other cryptocurrency subreddits are still being formed and it is still unclear on what should be discussed or not.

After analyzing the increase of the Bitcoin price per year and the amount of subscribers to the BitCoinMarket subreddit, it was clear that most activity started around the end of 2013. The actual gathering of data was started around the end of 2017 and therefore it was decided to gather Reddit comments and Bitcoin closing prices between November 30th 2013 and November 30th 2017.

### 3.2 Preprocessing

Preprocessing the reddit comments was an important step in this study. As there are hardly any limitations on who can comment on Reddit, the comments can contain a lot of meaningless information. So the first step in preprocessing was to remove website links from all the comments. While these could lead to interesting sources of information, a lot of these links were actually spam and did not contribute to a meaningful discussion on Reddit. Futhermore, punctuation was removed in order to ensure that duplicate words were not confused for being different because one of them contained some form of punctuation. Like when a word is placed at the end of the sentence and thus contains a period as final character. Next, newlines were removed from the comment to turn the comment into one big line of words. Finally, the entire comment was made lowercase. This should again ensure that duplicate words were not seen as two seperate terms.

After the first round of cleaning, the comment were passed to another preprocessing function. This function made sure to tokenize the words inside the comment and also remove any stop words. A stop word list was manually crafted as it turned out that the known stop word list of Python was to short. Normally, it is the case that during preprocessing words are also stemmed. However, when looking at the preprocessed comments after stemming, terms like "Google" and other important potential topic words were affected by the stemming process. As Jurafski and Martin note: "We don't want to stem, say the word Illustrator to illustrate, since the capitalized form Illustrator tends to refer to the software package" [7]. This is why for this study it was decided not to stem the terms during preprocessing. However, only during the last week of experimenting it was noted that lemmatisation would maybe have improved results. Because the experiments were already performed and it was too late to do any preprocessing, this step was also left out.

A final step in preprocessing was to group the comments per week. The comments were collected from the daily discussions placed on BitCoinMarket and so were marked with a date. However, as the daily discussions are never closed, comments could still be made after a certain day had already passed. It could be the case that certain comments were made weeks after the original day. Which in turn could lead to misleading correlations. Therefore, the date corresponding to each daily discussion was converted to a timestamp. The comment creation dates, also converted to timestamps, were then compared to that day's timestamp to decide if it was made during the same week or afterwards. Next, after all comments were grouped, only the most informative information was stored. This was: the comment creation date, comment id, author, stripped comment (one big line of words), and the cleaned comment (tokenized and stop words removed).

The Bitcoin closing prices did not need any sophisticated preprocessing. The site Coindesk offered an excell sheet with all the daily closing prices from 2013 till 2017. the prices only needed to be grouped in a similar fashion as the Reddit comments and the average weekly Bitcoin closing price had to be computed.

### 3.3 Hierarchical Clustering

In order to work with the LDA algorithm, certain parameters had to be initialized. One of these parameters was the expected number of topics. As the data ranges from november 2013 till november 2017, it was hard to say how many topics could be expected. That is where hierarchical clustering comes in handy.[6] In order to get an idea of which weekly discussions are quite similar, the Ward clustering algorithm [16] was used in combination with TFIDF and the Cosine Similarity.

The first step was to create a TFIDf matrix of all the weekly comments. The TFIDF score helps to show how important a word is in a document as it's document frequency helps to increase the score, but the frequency among the other documents decreases it[17]. Each document here is one week of comments.

The next step was to compute a distance matrix by computing the cosine similarity between every document by making use of the TFIDF matrix. This distance matrix is then passed along to the Ward hierarchical clustering algorithm to cluster the different weekly comments. The different clusters are depicted in a dendrogram in Fig 1. Unfortunately, the dendrogram is to large for this paper to be readable. But it is easy to see that the number of clusters ranges from 2 to more than 10. So the dendrogram helped to decide on reasonable values for the number of topics parameter of the LDA.

[6] http://brandonrose.org/clustering

In the end it was decided to vary this parameter by assigning it a value in the range from 2 to 10.

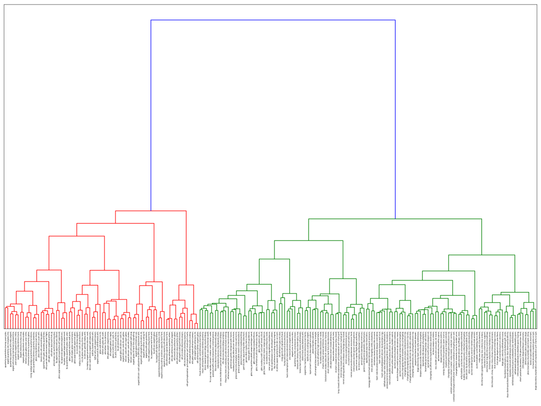

Fig. 1. Dendrogram.

### 3.4 Topic Modelling

The current study makes use of the LDA topic modelling algorithm in order to investigate the topic distribution among the Reddit comments. The algorithm has a few parameters which need to be initialized before it can be used. These are:

- num_topics: this is the number of expected topics and for this study it is varied between 2 and 10.
- passes: this paramter was set to 20 and not varied because of time constraints.

The algorithm also requires a corpus on which it performs the topic modelling. After running it on the preprocessed comments it became clear that the topic words returned still contained many meaningless words. In order to solve this problem, TFIDF was performed on the comments. The comments were then reduced to either the top 5 or top 10 scoring words. These two numbers where chosen based on the average comment length, which was approximately 16 words. If a comment was shorter than either 5 or 10 words, all its words were included. These decisions resulted in several experiment settings that in turn resulted in 18 different experiments. Dynamic topic modelling was also mentioned. However, this method was not freely available on the internet and because of time constraints it was decided not to implement this method.

### 3.5 Correlation Analysis

Before any correlation analysis could be performed, the first priority was to turn the topic counts and Bitcoin fluctuations into vectors of the same length. Two kinds of vectors were created: binary and non-binary. The binary vectors were created as many of the topics were not mentioned during the specified time period. This means that a lot of counts were equal to zero.

The resulting binary vectors for the topic counts contain zeros for the weeks that the specific topic was not mentioned and ones for the weeks that they were actually mentioned. Even if the actual topic count for that week was very low. The binary vector for the Bitcoin prices was computed in a similar way. It contained zeros for the weeks that the Bitcoin price was either lower than the previous week or similar to the previous week. It contained ones for the weeks that the Bitcoin price was higher than the previous week.

The non-binary vectors were created in a different way. For the Reddit comments this means that resulting vector contained the actual topic count, which was obtained by taking the product of the LDA probability distribution for that topic for each week and the amount of comments that week. The Bitcoin price vector was created by taking the actual price difference between the previous and current week.

In order to compute the correlation, two correlations methods were examined: Pearson's and Spearman's correlation coefficients. Pearson's assumes that the two variables are normally distributed and have a linear relationship[4]. This was tested for both the binary and non-binary vectors of both the topic counts and Bitcoin price differences. From these tests it was concluded that no vector could have been drawn from a normal distribuition, so Pearson's method was discarded. The assumptions of Spearman's is that the variables must at least be ordinal. Luckily, the Python library concerning Spearman's correlation will ensure that any variable is turned ordinal before correlation coefficients are computed.

This means that for this study the Spearman's correlation coefficients between the resulting vectors was computed. The correlation coefficients can take a value between -1 and 1. With positive values corresponding to positive correlations and negative values corresponding to negative values. Futhermore, correlations above 0.3 and or lower than -0.3 are considered as correlations of medium strenght and above 0.5 and lower than -0.5 as strong correaltions. A lot of the resulting topic count vectors still contained a lot of zero counts as the topics were simply not mentioned that often. As no solution was found to deal with this problem, the correlations were still measured on these zero counts. But the resulting correlation coefficients were simply ignored.

## 4 RESULTS

In total 18 different experiments were run. However, as mentioned above, most of the topic vectors contained over 50% of zero counts. Mainly as the topics were only mentioned during some weeks of the selected time period. As to avoid mentioning strong correlations between the Bitcoin price rise and fall and these vectors, it was decided to ignore the correlations of those vectors. The results obtained from the binary vectors were also ignored, as the resulting correlations were either not strong enough or not significant. So the remaining of this paper is dedicated to the non-binary vectors.

Furthermore, a strong positive correlation is defined as >=0.5 and a strong negative correlation is defined as <=-0.5. Unfortunately, none of the computed correlations could therefore be considered strong. The correlation analysis did however return some correlations of medium strength. Because of space restrictions only the strongest correlations is shown. Table 1 depicts these correlations. These results were obtained with comments containing the top 10 TFIDF words and setting the number of topics parameter to 4. The p-values in the table indicates if the correlation is significant or not. This is the case when the p-value is lower than 0.05. Which is the case for the first three topics return by the ldamodel, but not for the last topic. But this topic did also contain more than 50% zero count.

The strongest correlation is -0.42 and thus a negative one. This correlation was found for topic 3. The corresponding scatterplot is depicted in Fig 2. The top 10 words belonging to this topic can be read in Table 2. The interested reader should consult the Appendices for the remaining correlation coefficients, scatterplots, and topic terms per experiment setting.

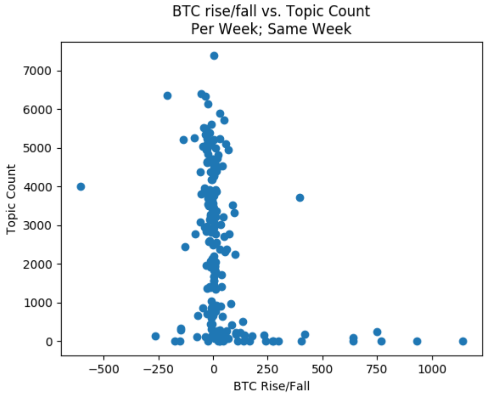

Fig. 2. TFIDF=10; num_topics=4; Same Week; topic=3

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3385 | 5.6897e-07 |
| 1 | 0.3163 | 3.2592e-06 |
| 2 | -0.4233 | 1.8819e-10 |
| 3 | 0.0510* | 0.4645 |

Table 1: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=4.

| topic | terms |
|---|---|
| 0 | bitcoin;like;btc;im;price;dont;think;time;good;get; |
| 1 | bitcoin;btc;like;im;price;dont;think;people;time;get; |
| 2 | bitcoin;price;like;im;think;dont;people;btc;time;market; |
| 3 | bch;btc;bitcoin;like;think;bcash;im;dont;going;get; |

Table 2: Top 10 topic words TFIDF=10, num_of_topics=4.

## 5 DISCUSSION AND CONCLUSION

As mentioned in the previous section, the obtained results did not contain any strong correlations (either positive or negative). But the results did contain some correlations of medium strength. The strongest one was -0.4233 and also significant. This correlation, which is negative, can be interpreted that a weekly rise in Bitcoin price correlates with a decrease in Topic mentions that same week. The latter as the correlation was found when comparing the weekly topic count to that seem weeks Bitcoin price rise or fall. Unfortunately, this correlation is only of medium strength.

The other correlations showed similar or even weaker correlations and therefore no strong correlation has been found between any of the topics returned by the LDA algorithm and the Bitcoin price rise or fall. The current work does include many limitations. First of all, comments are preprocessed in such a way that only unigrams remain. No other types of N-grams are considered. Also no lemmatisation or stemming is performed during preprocessing. While there was a clear reason as to why to ignore stemming. More sophisticated ways of stemming could have been implemented.

Futhermore, the comments are grouped per week and are only compared with the average Bitcoin price of either the same or the next week. Other ways of grouping or price fluctuations are not experimented with. Also the LDA algorithm is used on the entire dataset ranging from November 2013 till November 2017. This resulted in many zero counts for certain topics, which in turn was simply ignored.

Also other subreddits dedicated to Bitcoin could have been considered to increase the number of comments. Currently only the BitCoinMarket subreddit is used and it could very well be that the topics discussed on this subreddit show no correlation at all, but topics discussed on other subreddits do. Finally, the topics returned by the LDA model look quite similar most of the time. More preprocessing could have been performed in order to improve the differences between the topics. Most importantly, as the subreddit was dedicated to Bitcoin and the study was interested in finding a correlation between the Bitcoin price and the subreddit topics. It seems that making sure to remove words like "Bitcoin" and "btc" could possibly be one way of improving the returned topics. Mainly, as most topics will surely mention Bitcoin.

It is therefore suggested that future work should tackle these limitations in order to see if these were the cause of the poor correlation coefficients. Aside from the above mentioned limitations it is also suggested that more recent comments are examined, as it was not until the start of 2017 that the Bitcoin price and subreddit subscriber amount saw an extreme increase in numbers. It could be that the amount of comments gathered from November 2013 till November 2017 is too low for any strong and significant correlations to be found.

Finally, while research related to social media and its predictive power on Bitcoin or cryptocurrency price fluctuations does exist. Still too little research is focussed on the predictive power of Reddit in specific. Reddit does form a perfect place for discussion as no limitations on words is placed and people do not have to be friends to read each others comments.

Current work did not manage to answer the proposed research question as it did not find any strong correlation between the posts made on the subreddit and the price fluctuations of the Bitcoin price. It did, however, manage to open ways for future work to study the specific relations between user posts made on Reddit and the fluctuations of the Bitcoin price. As this work was limited in multiple aspects, it is expected that proposed improvements to the current work will discover more profound relations and results.

## REFERENCES

[1] S. Asur and B. A. Huberman. 2010. Predicting the Future with Social Media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent*

*Agent Technology*, Vol. 1. 492–499. https://doi.org/10.1109/WI-IAT.2010.63

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (01 2003), 993–1022.

[3] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2005. The Predictive Power of Online Chatter. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*. ACM, New York, NY, USA, 78–87. https://doi.org/10.1145/1081870.1081883

[4] Jan Hauke and Tomasz Kossowski. 2011. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaestiones Geographicae* 30 (05 2011).

[5] Shu Huang, Wei Peng, Jingxuan Li, and Dongwon Lee. 2013. Sentiment and Topic Analysis on Social Media: A Multi-task Multi-label Classification Approach. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13)*. ACM, New York, NY, USA, 172–181. https://doi.org/10.1145/2464464.2464512

[6] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. 2010. Movie Reviews and Revenues: An Experiment in Text Regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 293–296. http://dl.acm.org/citation.cfm?id=1857999.1858037

[7] Daniel Jurafsky and James H. Martin. [n. d.]. *Speech and Language Processing*. Pearson Education.

[8] 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 (08 2016), 1–17. https://doi.org/10.1371/journal.pone.0161197

[9] 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 (05 2017), 1–14. https://doi.org/10.1371/journal.pone.0177630

[10] M. Linton, E. G. S. Teo, E. Bommes, C. Y. Chen, and Wolfgang Karl Härdle. 2017. *Dynamic Topic Modelling for Cryptocurrency Community Forums*. Springer Berlin Heidelberg, Berlin, Heidelberg, 355–372. https://doi.org/10.1007/978-3-662-54486-0_18

[11] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. 2007. Topic and Role Discovery in Social Networks. *Journal of Artificial Intelligence Research* 30 (10 2007), 249–272.

[12] Gilad Mishne and Natalie Glance. 2006. Predicting Movie Sales from Blogger Sentiment. *American Association for Artificial Intelligence* (2006).

[13] Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating Topic Models for Social Media User Recommendation. In *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 101–102. https://doi.org/10.1145/1963192.1963244

[14] Rasheed Sabar. 2017. What Are Cryptocurrencies? *Ellington Management Group* (2017).

[15] Fani Tsapeli, Mirco Musolesi, and Peter Tino. 2017. Non-parametric causality detection: An application to social media and financial data. *Physica A: Statistical Mechanics and its Applications* 483, Supplement C (2017), 139 – 155. https://doi.org/10.1016/j.physa.2017.04.101

[16] Joe H. Jr. Ward. 1963. Hierarchical Grouping to Optimize an Objective Function. *J. Amer. Statist. Assoc.* 58 (03 1963), 236–244.

[17] ChengXiang Zhai and Sean Massung. [n. d.]. *Text Data Management and Analysis*. ACM.

[18] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. *Comparing Twitter and Traditional Media Using Topic Models*. Springer Berlin Heidelberg, Berlin, Heidelberg, 338–349. https://doi.org/10.1007/978-3-642-20161-5_34

# Appendices

## Appendix A

. The tables in this section contain the top 10 words per topic for every different experiment setting. Each topic number is associated with those terms. It should be noted that the topic terms do differ per experiment setting. The different experiment settings that influence the topic terms are:

- comments fed to the LDAmodel contain either top 5 TFIDF words or top 10 TFIDF words
- num_of_topic parameter LDA is varied from 2 till 10

| topic | terms |
|---|---|
| 0 | bitcoin;btc;like;good;lol;im;thanks;right;buy;dont; |
| 1 | bitcoin;thanks;btc;like;price;im;good;think;dont;buy; |

Table A1: Top 10 topic words TFIDF=5, num_of_topics=2.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;thanks;like;good;price;im;dont;think;buy; |
| 1 | bch;fork;bitcoin;btc;im;lol;good;segwit;think;right; |
| 2 | bitcoin;btc;like;thanks;im;price;good;think;buy;dont; |

Table A2: Top 10 topic words TFIDF=5, num_of_topics=3.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;thanks;price;good;like;im;short;dont;think; |
| 1 | bitcoin;btc;like;im;lol;good;thanks;buy;dont;right; |
| 2 | bitcoin;like;thanks;btc;im;good;price;time;right;think; |
| 3 | bitcoin;thanks;btc;price;im;like;good;buy;think;dont; |

Table A3: Top 10 topic words TFIDF=5, num_of_topics=4.

| topic | terms |
|---|---|
| 0 | bitcoin;thanks;btc;like;good;price;im;think;buy;dont; |
| 1 | bid;bubble;news;gox;thanks;market;btc;bitcoin;31st;price; |
| 2 | whalebearpig;triggerd;cbi;mmm;bested;china;ponzi;defenseless;317320;umangalaiii; |
| 3 | securityprivacy;bancor;350450;cheezits;btsx;quartz;1888;ostrich;undies;antbleed; |
| 4 | bitcoin;btc;like;im;thanks;good;price;dont;buy;think; |

Table A4: Top 10 topic words TFIDF=5, num_of_topics=5.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;thanks;like;price;good;buy;im;think;dont; |
| 1 | bitcoin;btc;like;im;going;good;price;think;right;dont; |
| 2 | bitcoin;btc;right;thanks;see;dont;price;sell;short;people; |
| 3 | bitcoin;btc;thanks;good;like;price;im;short;dont;time; |
| 4 | bitcoin;thanks;im;price;like;btc;good;think;buy;dont; |
| 5 | bitcoin;btc;like;lol;good;im;buy;think;thanks;time; |

Table A5: Top 10 topic words TFIDF=5, num_of_topics=6.

| topic | terms |
|---|---|
| 0 | news;bitcoin;btc;etf;bu;good;gox;market;im;like; |
| 1 | gox;mtgox;bitcurex;5pm;dean;withdrawals;coins;goxbtc;news;bitcoin; |
| 2 | bitcoin;thanks;like;btc;price;im;good;think;dont;buy; |
| 3 | stripe;nfc;tokenization;applepay;braintree;1400btc;967;becuz;offexchange;unix; |
| 4 | bitcoin;btc;im;good;thanks;like;right;think;price;lol; |
| 5 | price;thanks;bitcoin;coins;btc;bubble;news;uchangetip;market;like; |
| 6 | collected;china;uchangetip;christmas;im;thanks;bitcoin;coins;btc;volume; |

Table A6: Top 10 topic words TFIDF=5, num_of_topics=7.

| topic | terms |
|---|---|
| 0 | btg;tether;bch;8k;cme;thanksgiving;usdt;7500;8000;diamond; |
| 1 | bitcoin;btc;good;im;price;thanks;like;buy;dont;right; |
| 2 | fork;b2x;short;rbf;hearn;bitcoin;ddos;like;im;360; |
| 3 | fork;btg;2x;alts;bitcoin;b2x;like;7k;eth;btc; |
| 4 | collected;uchangetip;bubble;ghash;cycle;auction;btc;bubblewatch;bitcoin;coins; |
| 5 | bitcoin;btc;thanks;like;im;good;price;think;buy;dont; |
| 6 | bitcoin;btc;thanks;price;like;im;good;think;dont;time; |
| 7 | btc;bitcoin;right;sell;bch;people;buy;dont;thanks;value; |

Table A7: Top 10 topic words TFIDF=5, num_of_topics=8.

| topic | terms |
|---|---|
| 0 | bitcoin;like;good;thanks;dont;time;im;price;buy;right; |
| 1 | bitcoin;btc;like;thanks;im;see;dont;good;right;buy; |
| 2 | 385;bearwhale;newegg;bitcoin;wall;whalebearpig;cbi;cyprus;coins;benefit; |
| 3 | ghash;ghashio;dean;hudghe;popocorn;lucas;dio;payscreen;2339;crass; |
| 4 | bcc;grammar;errors;wouldve;bittrex;viabtc;information;block;right;deposits; |
| 5 | bitcoin;btc;like;thanks;im;buy;price;good;thats;dont; |
| 6 | bitcoin;btc;good;thanks;think;im;dont;thats;like;right; |
| 7 | bitcoin;thanks;price;btc;like;im;good;think;dont;buy; |
| 8 | bitcoin;btc;like;im;good;lol;thanks;right;think;dont; |

Table A8: Top 10 topic words TFIDF=5, num_of_topics=9.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;like;im;good;thanks;think;price;lol;dont; |
| 1 | bitcoin;thanks;like;btc;price;im;good;buy;think;dont; |
| 2 | mc;mv;r3;hearnia;gabi;jd;st;firstmark;beanzflooz;fssrbf; |
| 3 | stampjacked;workstations;daisies;outofpocket;wiley;markdowns;worsealthough;buttstamp;27680;bitmessage; |
| 4 | btc;bitcoin;good;thanks;etf;im;buy;channel;news;gdax; |
| 5 | bitcoin;bch;like;im;btc;good;thanks;time;think;get; |
| 6 | gdax;icos;gbtc;3k;bubble;like;bitcoin;goldman;mean;segwit2x; |
| 7 | bch;bitcoin;bcc;btc;right;see;thanks;fork;lol;sell; |
| 8 | dish;millibitcoins150;600;coffee;570;595;cup;moralagent;562;417; |
| 9 | eb;donut;548;btu;pitchfork;speckled;quickbooks;emergent;920950;anniversay; |

Table A9: Top 10 topic words TFIDF=5, num_of_topics=10.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;like;im;dont;price;think;time;people;get; |
| 1 | bitcoin;price;like;im;think;dont;people;time;btc;good; |

Table A10: Top 10 topic words TFIDF=10, num_of_topics=2.

| topic | terms |
|---|---|
| 0 | bitcoin;price;like;im;think;dont;people;market;btc;see; |
| 1 | bitcoin;btc;like;im;dont;think;price;people;time;get; |
| 2 | bitcoin;price;like;im;dont;think;btc;time;people;good; |

Table A11: Top 10 topic words TFIDF=10, num_of_topics=3.

| topic | terms |
|---|---|
| 0 | bitcoin;like;btc;im;price;dont;think;time;good;get; |
| 1 | bitcoin;btc;like;im;price;dont;think;people;time;get; |
| 2 | bitcoin;price;like;im;think;dont;people;btc;time;market; |
| 3 | bch;btc;bitcoin;like;think;bcash;im;dont;going;get; |

Table A12: Top 10 topic words TFIDF=10, num_of_topics=4.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;like;im;price;dont;think;people;time;get; |
| 1 | bitcoin;price;like;think;im;dont;people;coins;see;time; |
| 2 | bitcoin;like;price;im;think;dont;people;china;buy;market; |
| 3 | bitcoin;like;price;im;dont;think;time;long;btc;good; |
| 4 | synapse;like;bitfinex;price;think;im;dont;bfx;btc;synapsepay; |

Table A13: Top 10 topic words TFIDF=10, num_of_topics=5.

| topic | terms |
|---|---|
| 0 | synapse;synapsepay;bfxcoin;bitcoin;ltc;btc;adjective;price;dont;litecoin; |
| 1 | bitcoin;price;like;im;dont;think;time;market;good;people; |
| 2 | bitcoin;like;btc;im;dont;price;think;time;good;people; |
| 3 | bitcoin;like;price;im;btc;think;dont;people;time;buy; |
| 4 | bch;btc;bcc;bitcoin;block;right;dont;fork;get;like; |
| 5 | bitcoin;btc;like;im;price;dont;think;people;time;good; |

Table A14: Top 10 topic words TFIDF=10, num_of_topics=6.

| topic | terms |
|---|---|
| 0 | bitcoin;btc;like;price;im;think;dont;people;time;going; |
| 1 | bitcoin;price;like;im;dont;think;btc;time;good;see; |
| 2 | china;news;chinese;exchanges;bitcoin;ban;banning;caixin;bounce;think; |
| 3 | like;lsd;paypal;bitcoin;price;halving;people;dont;thats;im; |
| 4 | bitcoin;im;like;btc;think;dont;price;buy;people;time; |
| 5 | bitcoin;like;btc;im;dont;think;price;people;time;get; |
| 6 | bitcoin;btc;price;like;dont;think;im;people;see;get; |

Table A15: Top 10 topic words TFIDF=10, num_of_topics=7.

| topic | terms |
|---|---|
| 0 | bitcoin;price;like;im;think;dont;time;good;see;btc; |
| 1 | 300;mmm;330;350;bitcoin;volume;openbazaar;325;like;time; |
| 2 | bitcoin;btc;like;im;price;dont;think;people;time;going; |
| 3 | yobo;nucleus;zar;classice;megaflag;97000;33399;xorg;showstopper;mitsubishi; |
| 4 | bitcoin;like;price;im;think;dont;btc;people;see;time; |
| 5 | like;bitcoin;btc;im;dont;price;think;people;thats;time; |
| 6 | ghash;ghashio;stability;bitcoin;berryfarmerconcious;prom;dissertation;like;price;im; |
| 7 | bitcoin;like;think;price;btc;time;dont;thats;buy;im; |

Table A16: Top 10 topic words TFIDF=10, num_of_topics=8.

| topic | terms |
|---|---|
| 0 | bid;auction;winning;ltc;bidding;syndicate;bidders;secondmarket;bitcurex;news; |
| 1 | bitcoin;price;like;im;dont;think;btc;time;good;see; |
| 2 | btc;bcc;bitcoin;dont;price;im;think;like;get;good; |
| 3 | bitcoin;price;like;think;im;dont;btc;people;market;get; |
| 4 | like;im;dont;think;people;price;time;thats;see;good; |
| 5 | bitcoin;btc;like;im;dont;price;think;people;time;get; |
| 6 | etf;approved;approval;bu;decision;sec;bitcoin;im;price;dont; |
| 7 | gox;submissionmessageim;daa;goxbtc;bcrash;clashic;violation;bch;btc;txmal; |
| 8 | thanksgiving;btg;tether;diamond;8000;8k;cme;8500;8300;8100; |

Table A17: Top 10 topic words TFIDF=10, num_of_topics=9.

| topic | terms |
|---|---|
| 0 | bitcoin;like;im;price;dont;think;people;btc;see;good; |
| 1 | bitcoin;btc;like;im;price;dont;think;time;people;get; |
| 2 | bitcoin;im;think;btc;like;right;price;time;people;thats; |
| 3 | bitcoin;price;like;im;think;dont;btc;people;time;buy; |
| 4 | bip141;2250;bitcoin;matthias;90bn;2420;like;1051;squozen;650k; |
| 5 | bitcoin;price;im;money;dont;thats;think;like;right;get; |
| 6 | bitcoin;bch;btc;like;im;think;time;dont;price;people; |
| 7 | bitcoin;price;like;btc;dont;im;think;people;time;good; |
| 8 | bitcoin;like;btc;im;dont;think;price;good;people;buy; |
| 9 | bitcoin;price;like;im;btc;think;dont;time;see;long; |

Table A18: Top 10 topic words TFIDF=10, num_of_topics=10.

## Appendix B

. This section contains all the correlation coefficients and corresponding p-values. A p-value lower than 0.01% indicates a significant correlation. The correlation was computed using Spearmans' rho and a '*' indicates that the corresponding topic count vector contained over 50% zero count. Each table is followed by a caption that contains two numbers. The TFIDF number stands for the number of words each comment could have during preprocessing. In more detail: in order to filter out meaningless words, TFIDF-scoring was used and either a comment was reduced to the top 5 TFIDF words or the top 10 TFIDF words. Next the num_topics number indicates the LDA parameter. So during experiments this number was varied ranging from 2 till 10. Finally, the topic counts per week were either correlated against the BTC price rise or fall during that same week or against the next week.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3743* | 2.5491e-08 |
| 1 | -0.2231 | 0.0012 |

Table B1: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=2.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.01164 | 0.8675 |
| 1 | 0.1936* | 0.0051 |
| 2 | 0.2285 | 0.0009 |

Table B2: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=3.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.02614 | 0.7078 |
| 1 | 0.3208 | 2.3042e-06 |
| 2 | 0.0116 | 0.8684 |
| 3 | -0.1616 | 0.0197 |

Table B3: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=4.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1817 | 0.0086 |
| 1 | -0.1455 | 0.0360 |
| 2 | -0.0400 | 0.5657 |
| 3 | 0.0161 | 0.8172 |
| 4 | 0.3260 | 1.5422e-06 |

Table B4: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=5.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0312* | 0.6547 |
| 1 | 0.0235* | 0.7364 |
| 2 | 0.1092* | 0.1164 |
| 3 | 0.2163 | 0.0017 |
| 4 | -0.2234 | 0.0012 |
| 5 | 0.2960* | 1.4158e-05 |

Table B5: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=6.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0507* | 0.4670 |
| 1 | 0.0054* | 0.9379 |
| 2 | -0.2731 | 6.5569e-05 |
| 3 | -0.0104* | 0.8812 |
| 4 | 0.3662 | 5.3448e-08 |
| 5 | -0.1112* | 0.1098 |
| 6 | -0.0352* | 0.6138 |

Table B6: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=7.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0075* | 0.9147 |
| 1 | 0.1602* | 0.0208 |
| 2 | 0.0785* | 0.2595 |
| 3 | -0.0429* | 0.5388 |
| 4 | 0.0137* | 0.8440 |
| 5 | 0.0925 | 0.1837 |
| 6 | 0.0163 | 0.8153 |
| 7 | 0.0287* | 0.6807 |

Table B7: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=8.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0271* | 0.6973 |
| 1 | nan* | nan |
| 2 | -0.1298* | 0.0616 |
| 3 | -0.0747* | 0.2838 |
| 4 | -0.0373* | 0.5930 |
| 5 | 0.1089* | 0.1174 |
| 6 | nan* | nan |
| 7 | -0.1969* | 0.0044 |
| 8 | 0.2224* | 0.0012 |

Table B8: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=9.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3597 | 9.4736e-08 |
| 1 | -0.2887 | 2.3535e-05 |
| 2 | -0.0308* | 0.6583 |
| 3 | -0.1036* | 0.1364 |
| 4 | 0.0733* | 0.2925 |
| 5 | -0.0415* | 0.5522 |
| 6 | -0.1805* | 0.0091 |
| 7 | 0.1739* | 0.0120 |
| 8 | 0.101* | 0.1457 |
| 9 | 0.0845* | 0.2247 |

Table B9: Topic Count vs. Next Week BTC Price Correlation. TFIDF=5; num_topics=10.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3432 | 3.8700e-07 |
| 1 | -0.2865 | 2.7272e-05 |

Table B10: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=2.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1317* | 0.0580 |
| 1 | 0.3326 | 9.2137e-07 |
| 2 | -0.1611 | 0.0201 |

Table B11: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=3.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.1641 | 0.0179 |
| 1 | 0.3293 | 1.1940e-06 |
| 2 | -0.3164 | 3.2198e-06 |
| 3 | -0.0352* | 0.6132 |

Table B12: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=4.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.1498 | 0.0307 |
| 1 | -0.1991* | 0.0039 |
| 2 | -0.2161 | 0.0017 |
| 3 | -0.0262 | 0.7072 |
| 4 | 0.0133* | 0.8490 |

Table B13: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=5.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0591* | 0.3962 |
| 1 | -0.1272 | 0.0670 |
| 2 | -0.0258* | 0.7110 |
| 3 | -0.0501 | 0.4727 |
| 4 | -0.0311* | 0.6557 |
| 5 | 0.3329 | 8.9934e-07 |

Table B14: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=6.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0170* | 0.8070 |
| 1 | -0.0480 | 0.4914 |
| 2 | 0.0283* | 0.6845 |
| 3 | -0.0330* | 0.6361 |
| 4 | -0.1110* | 0.1106 |
| 5 | 0.1808 | 0.0090 |
| 6 | 0.1054* | 0.1299 |

Table B15: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=7.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1013 | 0.1455 |
| 1 | -0.0018* | 0.9793 |
| 2 | 0.1995 | 0.0039 |
| 3 | -0.0128* | 0.8545 |
| 4 | -0.1273 | 0.0668 |
| 5 | -0.0008* | 0.9910 |
| 6 | -0.0399* | 0.5670 |
| 7 | nan* | nan |

Table B16: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=8.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1702* | 0.0140 |
| 1 | -0.1597 | 0.0212 |
| 2 | 0.0537* | 0.4408 |
| 3 | -0.1807* | 0.0090 |
| 4 | -0.0054* | 0.9384 |
| 5 | 0.1838 | 0.0079 |
| 6 | 0.0848* | 0.2235 |
| 7 | -0.0648* | 0.3527 |
| 8 | -0.0140* | 0.8409 |

Table B17: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=9.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0106* | 0.8787 |
| 1 | 0.1841 | 0.0078 |
| 2 | nan* | nan |
| 3 | -0.1750 | 0.0115 |
| 4 | 0.0382* | 0.5836 |
| 5 | nan* | nan |
| 6 | 0.3569* | 1.2078e-07 |
| 7 | 0.1558* | 0.0246 |
| 8 | 0.1662* | 0.0165 |
| 9 | 0.1321 | 0.0571 |

Table B18: Topic Count vs. Next Week BTC Price Correlation. TFIDF=10; num_topics=10.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3805* | 1.4329e-08 |
| 1 | -0.2624 | 0.0001 |

Table B19: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=2.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0435 | 0.5327 |
| 1 | 0.0393* | 0.5727 |
| 2 | 0.2024 | 0.0034 |

Table B20: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=3.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0262 | 0.7068 |
| 1 | 0.3122 | 4.4119e-06 |
| 2 | 0.0435* | 0.5332 |
| 3 | -0.0809 | 0.2457 |

Table B21: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=4.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1662 | 0.0164 |
| 1 | 0.0594* | 0.3938 |
| 2 | -0.0049* | 0.9444 |
| 3 | -0.0200 | 0.7742 |
| 4 | 0.3263 | 1.5094e-06 |

Table B22: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=5.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0351* | 0.6150 |
| 1 | 0.0846* | 0.2246 |
| 2 | 0.0876* | 0.2084 |
| 3 | 0.2129 | 0.0020 |
| 4 | -0.3007 | 1.0163e-05 |
| 5 | 0.3301* | 1.1217e-06 |

Table B23: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=6.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0696* | 0.3181 |
| 1 | -0.0861* | 0.2162 |
| 2 | -0.2702 | 7.9220e-05 |
| 3 | 0.0480* | 0.4912 |
| 4 | 0.3450 | 3.3500e-07 |
| 5 | -0.0827* | 0.2348 |
| 6 | 0.0427* | 0.5396 |

Table B24: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=7.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0653* | 0.3485 |
| 1 | 0.0949* | 0.1726 |
| 2 | -0.0102* | 0.8833 |
| 3 | 0.0031* | 0.9644 |
| 4 | -0.1842* | 0.0077 |
| 5 | 0.1715 | 0.0132 |
| 6 | -0.0963 | 0.1665 |
| 7 | 0.0599 | 0.3900 |

Table B25: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=8.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0318* | 0.6482 |
| 1 | nan* | nan |
| 2 | -0.0868* | 0.2125 |
| 3 | -0.0862* | 0.2155 |
| 4 | -0.0341* | 0.6252 |
| 5 | 0.0438* | 0.5295 |
| 6 | nan* | nan |
| 7 | -0.1933 | 0.0052 |
| 8 | 0.2626 | 0.0001 |

Table B26: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=9.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3919 | 4.7972e-09 |
| 1 | -0.3178 | 2.8983e-06 |
| 2 | -0.0767* | 0.2710 |
| 3 | -0.0851* | 0.2218 |
| 4 | 0.1740* | 0.0120 |
| 5 | 0.1914* | 0.0056 |
| 6 | 0.0421* | 0.5460 |
| 7 | 0.1977* | 0.0042 |
| 8 | 0.0732* | 0.2932 |
| 9 | -0.0624* | 0.3707 |

Table B27: Topic Count vs. Same Week BTC Price Correlation. TFIDF=5; num_topics=10.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3996 | 2.2455e-09 |
| 1 | -0.3733 | 2.7887e-08 |

Table B28: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=2.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0687* | 0.3241 |
| 1 | 0.3218 | 2.1330e-06 |
| 2 | -0.1526 | 0.0277 |

Table B29: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=3.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.3385 | 5.6897e-07 |
| 1 | 0.3163 | 3.2592e-06 |
| 2 | -0.4233 | 1.8819e-10 |
| 3 | 0.0510* | 0.4645 |

Table B30: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=4.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.2208 | 0.0013 |
| 1 | -0.2589* | 0.0002 |
| 2 | -0.2792 | 4.4353e-05 |
| 3 | 0.0336 | 0.6302 |
| 4 | 0.0408 | 0.5584 |

Table B31: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=5.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | 0.0340* | 0.6259 |
| 1 | -0.1074 | 0.1225 |
| 2 | 0.0438* | 0.5303 |
| 3 | -0.0550 | 0.4305 |
| 4 | 0.0495* | 0.4777 |
| 5 | 0.3130 | 4.1627e-06 |

Table B32: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=6.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.2428* | 0.0004 |
| 1 | 0.0446 | 0.5223 |
| 2 | -0.0097* | 0.8896 |
| 3 | -0.0081* | 0.9077 |
| 4 | -0.0508* | 0.4664 |
| 5 | 0.1438 | 0.0382 |
| 6 | -0.0055 | 0.9368 |

Table B33: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=7.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0175 | 0.8022 |
| 1 | -0.0404* | 0.5628 |
| 2 | 0.0938 | 0.1780 |
| 3 | 0.0085* | 0.9035 |
| 4 | -0.0421 | 0.5460 |
| 5 | 0.0340* | 0.6268 |
| 6 | -0.0961* | 0.1675 |
| 7 | nan* | nan |

Table B34: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=8.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.0361* | 0.6051 |
| 1 | -0.0866 | 0.2135 |
| 2 | 0.0294* | 0.6736 |
| 3 | -0.1668* | 0.0160 |
| 4 | -0.0531* | 0.4465 |
| 5 | 0.1627 | 0.0189 |
| 6 | 0.1335* | 0.0546 |
| 7 | 0.0021* | 0.9765 |
| 8 | 0.1348* | 0.0523 |

Table B35: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=9.

| topic | corrcoef | p-value |
|---|---|---|
| 0 | -0.1063* | 0.1263 |
| 1 | 0.2158 | 0.0017 |
| 2 | nan* | nan |
| 3 | -0.1678 | 0.0154 |
| 4 | -0.1696* | 0.0143 |
| 5 | nan* | nan |
| 6 | 0.3605* | 8.8602e-08 |
| 7 | 0.1559* | 0.0245 |
| 8 | 0.2381* | 0.0005 |
| 9 | 0.1426 | 0.0399 |

Table B36: Topic Count vs. Same Week BTC Price Correlation. TFIDF=10; num_topics=10.

## Appendix C

. This section contains graphs depicting the topic counts per week for every topic per experiment setting. However, only the topic counts for the experiment settings with the most interesting correlation coefficients are shown.

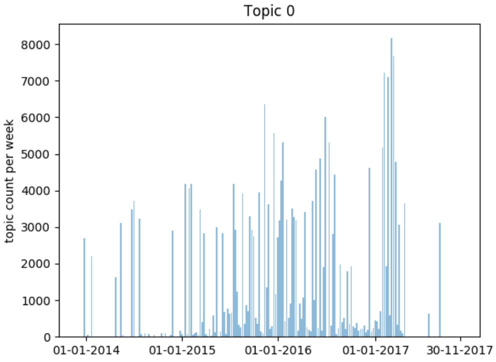

Fig. 3. TFIDF=5; num_topics=4;

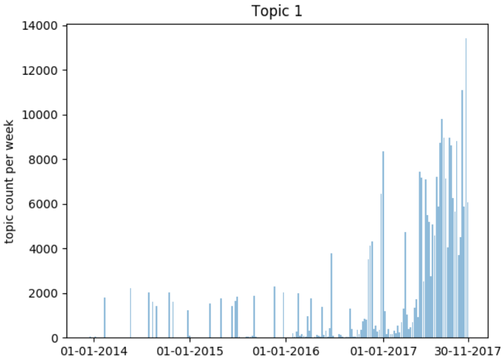

Fig. 4. TFIDF=5; num_topics=4;

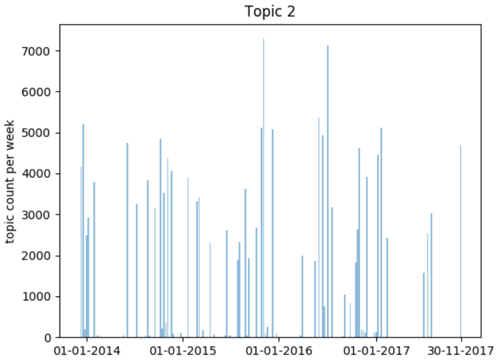

Fig. 5. TFIDF=5; num_topics=4;

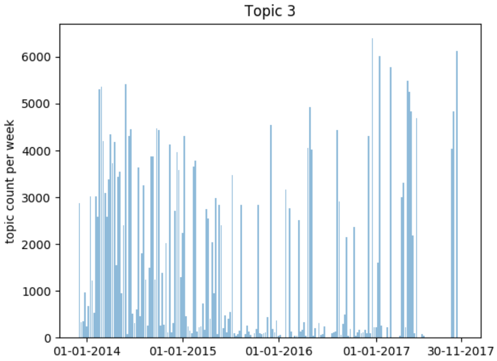

Fig. 6. TFIDF=5; num_topics=4;

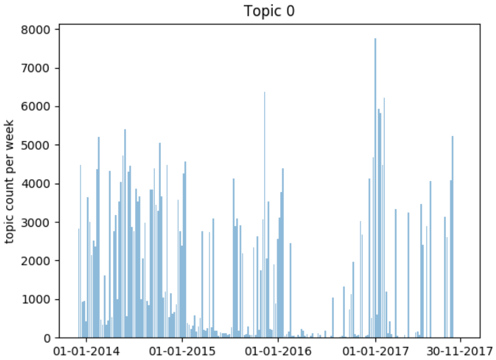

Fig. 7. TFIDF=5; num_topics=5;

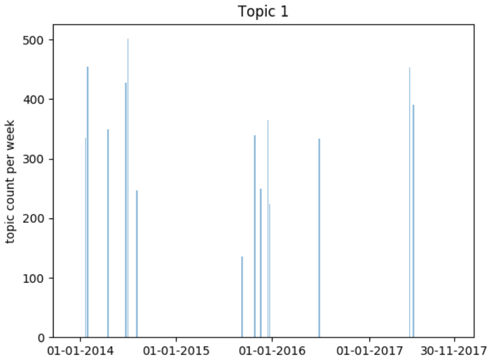

Fig. 8. TFIDF=5; num_topics=5;

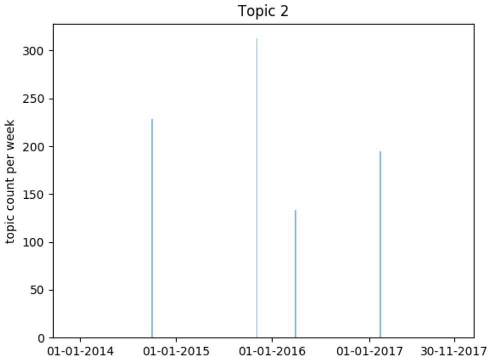

Fig. 9. TFIDF=5; num_topics=5;

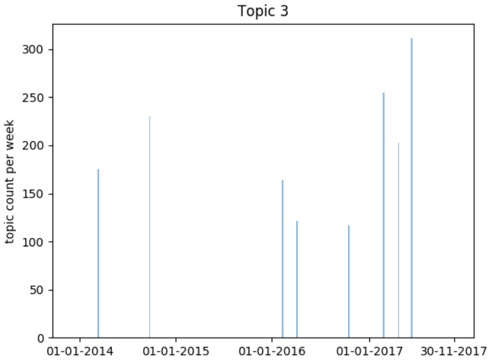

Fig. 10. TFIDF=5; num_topics=5;

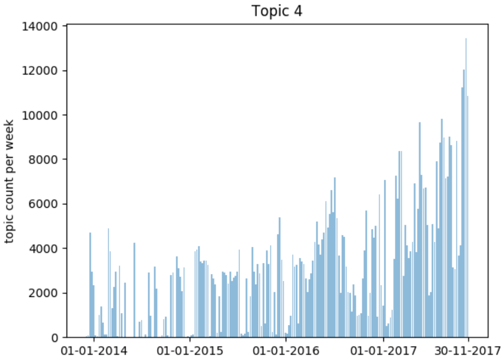

Fig. 11. TFIDF=5; num_topics=5;

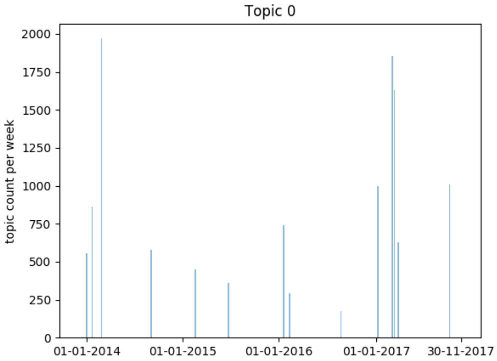

Fig. 12. TFIDF=5; num_topics=7;

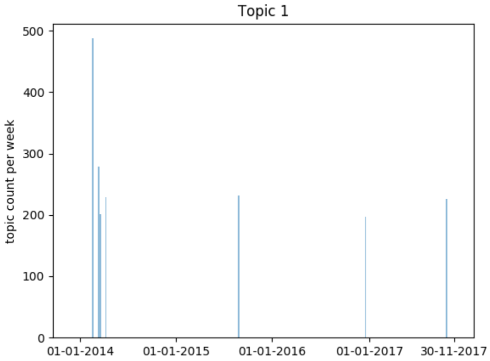

Fig. 13. TFIDF=5; num_topics=7;

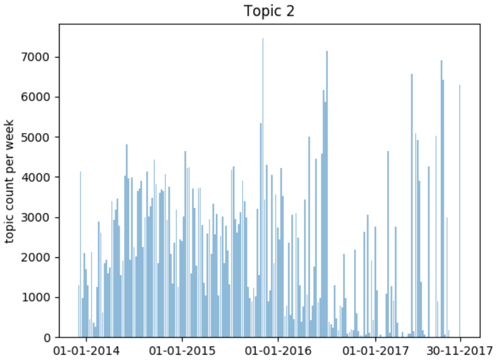

Fig. 14. TFIDF=5; num_topics=7;

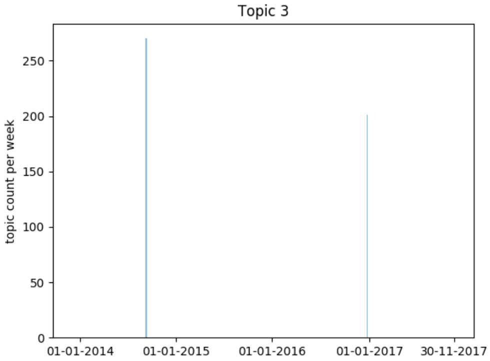

Fig. 15. TFIDF=5; num_topics=7;

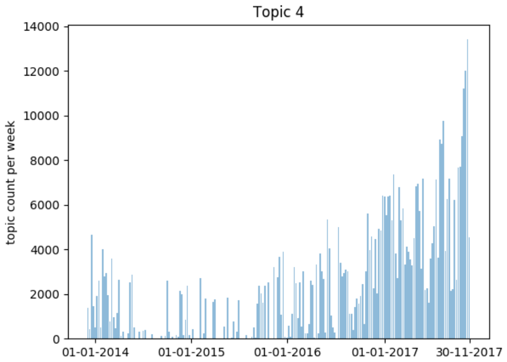

Fig. 16. TFIDF=5; num_topics=7;

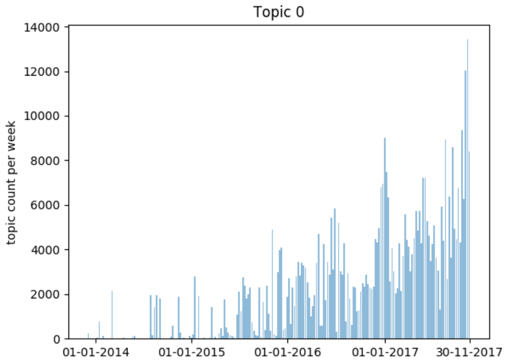

Fig. 19. TFIDF=5; num_topics=10;

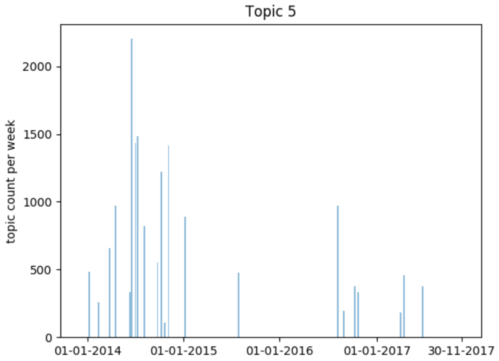

Fig. 17. TFIDF=5; num_topics=7;

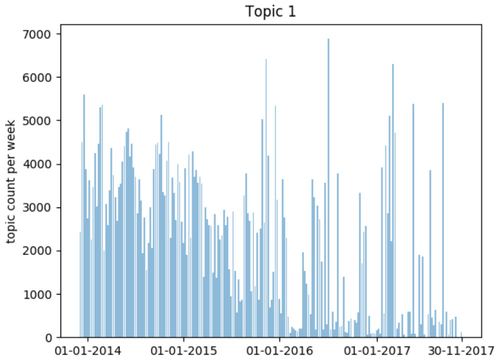

Fig. 20. TFIDF=5; num_topics=10;

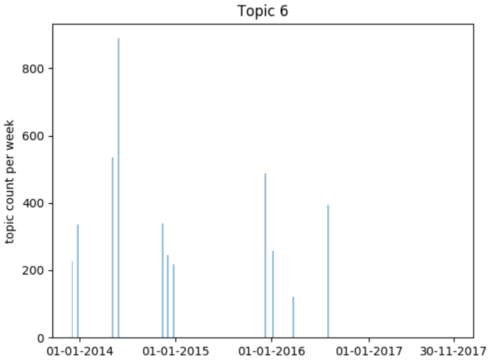

Fig. 18. TFIDF=5; num_topics=7;

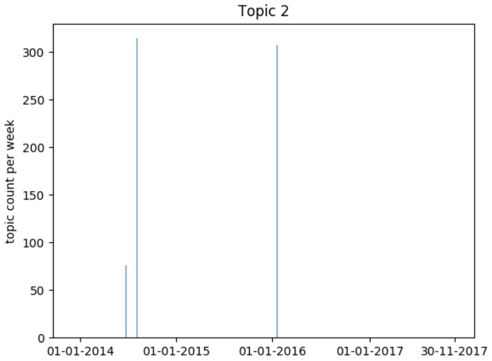

Fig. 21. TFIDF=5; num_topics=10;

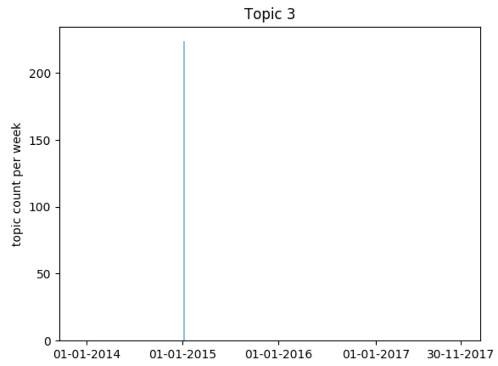

Fig. 22. TFIDF=5; num_topics=10;

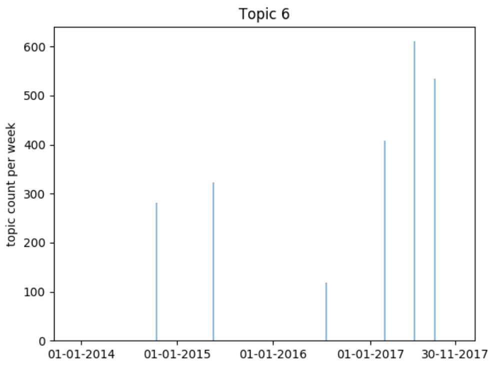

Fig. 25. TFIDF=5; num_topics=10;

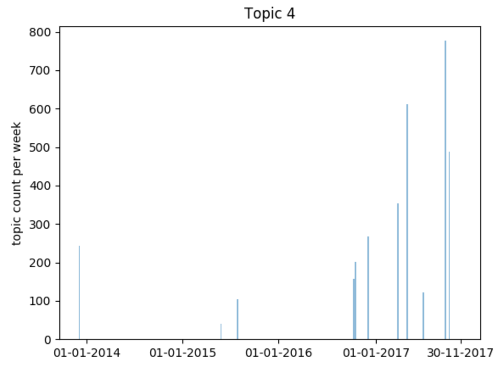

Fig. 23. TFIDF=5; num_topics=10;

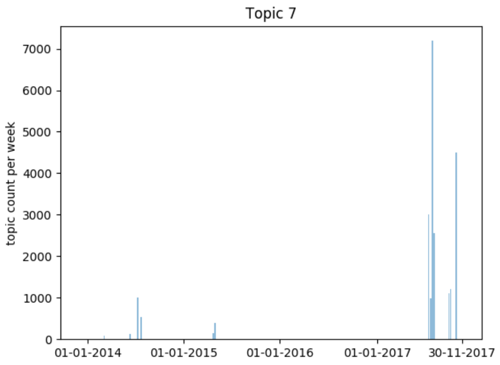

Fig. 26. TFIDF=5; num_topics=10;

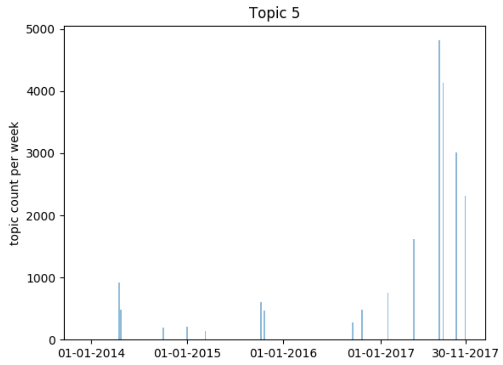

Fig. 24. TFIDF=5; num_topics=10;

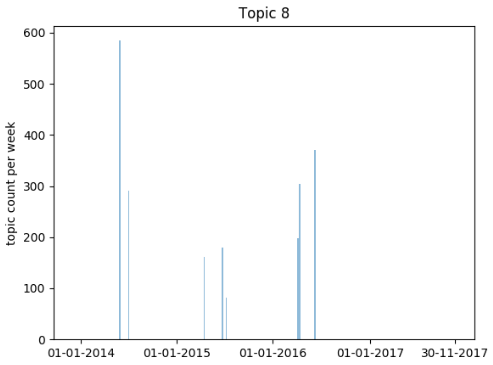

Fig. 27. TFIDF=5; num_topics=10;

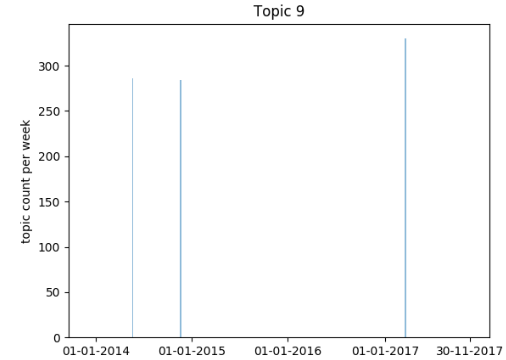

Fig. 28. TFIDF=5; num_topics=10;

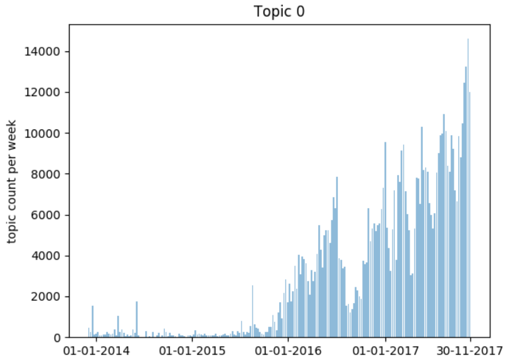

Fig. 29. TFIDF=10; num_topics=2;

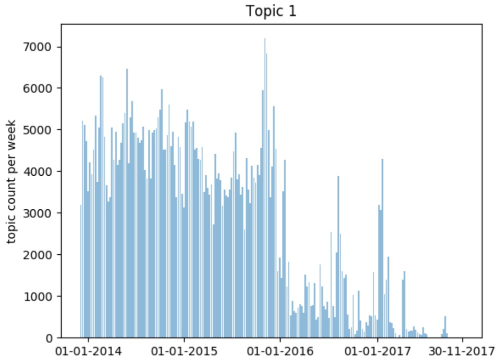

Fig. 30. TFIDF=10; num_topics=2;

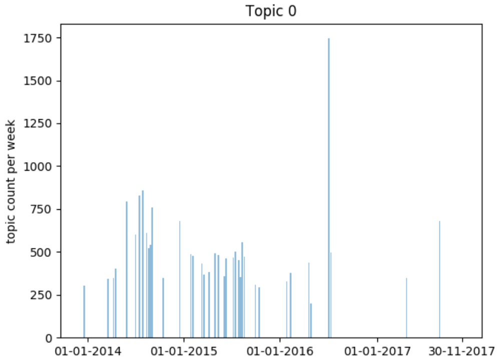

Fig. 31. TFIDF=10; num_topics=3;

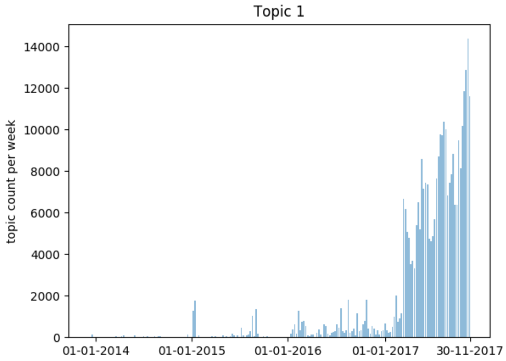

Fig. 32. TFIDF=10; num_topics=3;

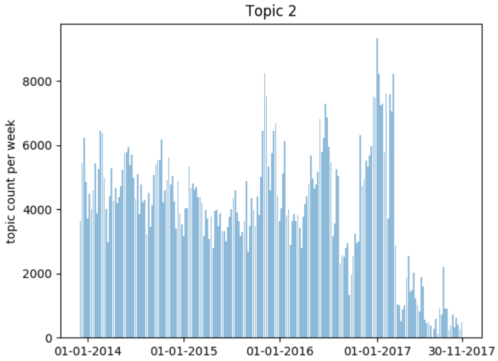

Fig. 33. TFIDF=10; num_topics=3;

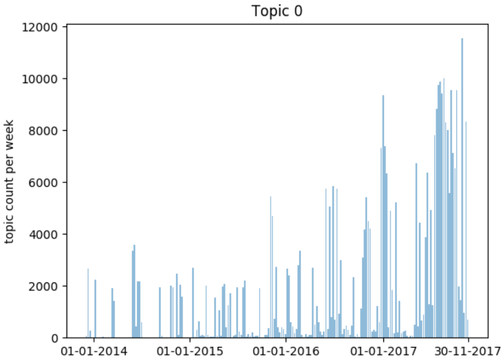

Fig. 34. TFIDF=10; num_topics=4;

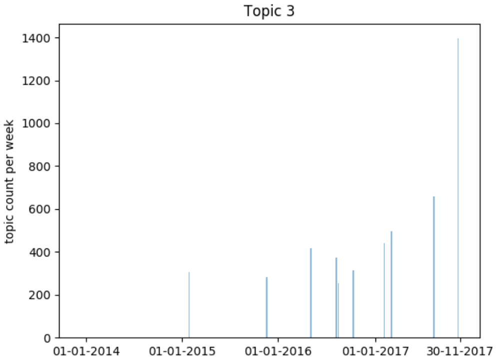

Fig. 37. TFIDF=10; num_topics=4;

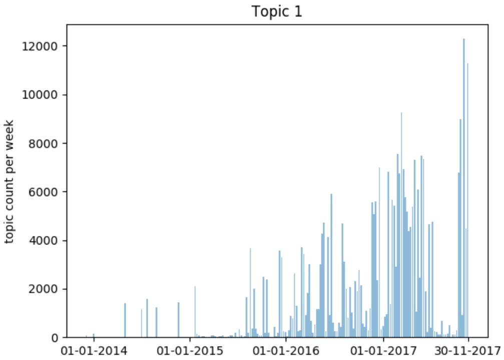

Fig. 35. TFIDF=10; num_topics=4;

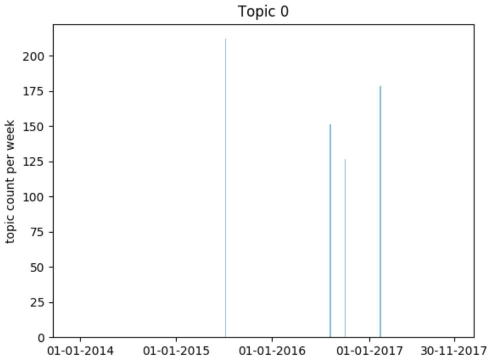

Fig. 38. TFIDF=10; num_topics=6;

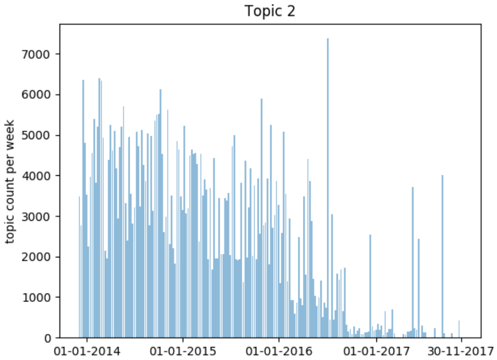

Fig. 36. TFIDF=10; num_topics=4;

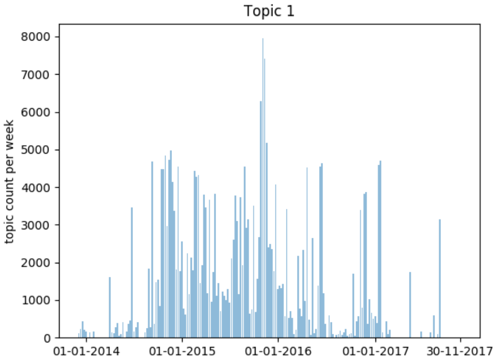

Fig. 39. TFIDF=10; num_topics=6;

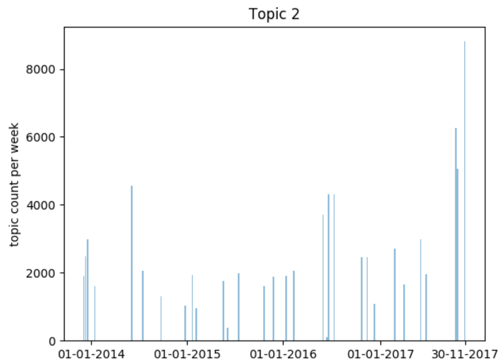

Fig. 40. TFIDF=10; num_topics=6;

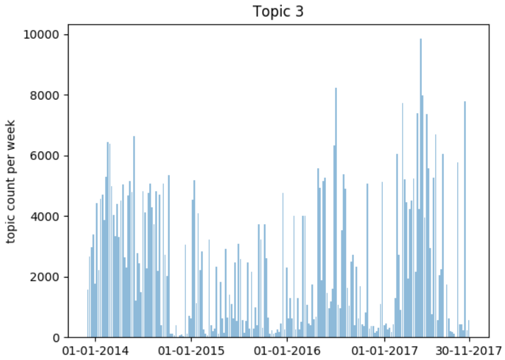

Fig. 41. TFIDF=10; num_topics=6;

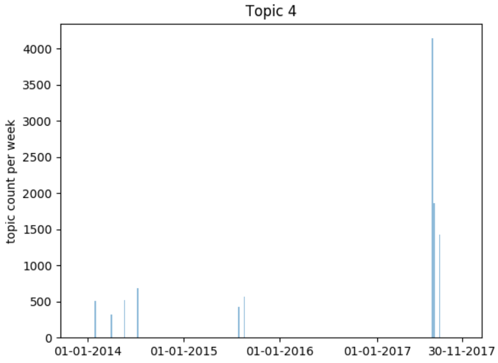

Fig. 42. TFIDF=10; num_topics=6;

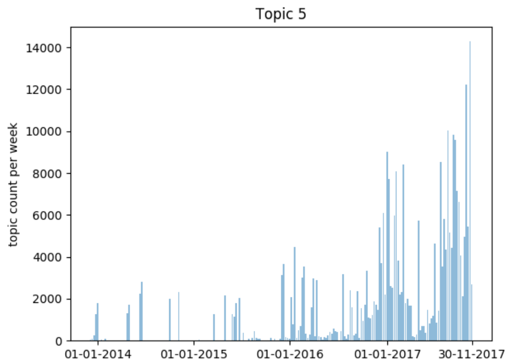

Fig. 43. TFIDF=10; num_topics=6;

## Appendix D

. This section contains scatter plots of the topic counts versus the BTC price rise and fall. Again, only the scatter plots of the experiment settings with the most interesting correlation coefficients are shown. 'Next Week' indicates that the topic counts were compared to the rise and fall of BTC price of the next week. 'Same Week' means the BTC price of the same week as the topic count.

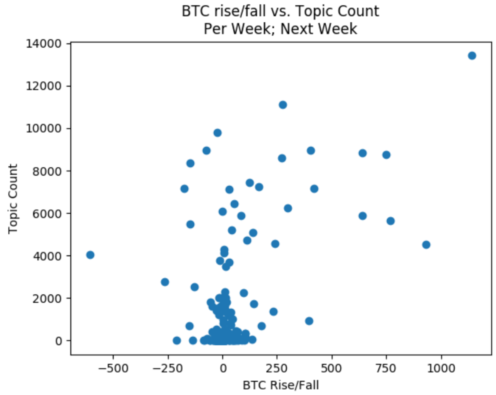

Fig. 44. TFIDF=5; num_topics=4; Next Week; topic=2

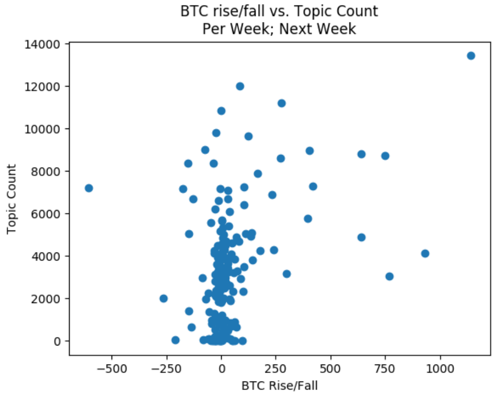

Fig. 45. TFIDF=5; num_topics=5; Next Week; topic=5

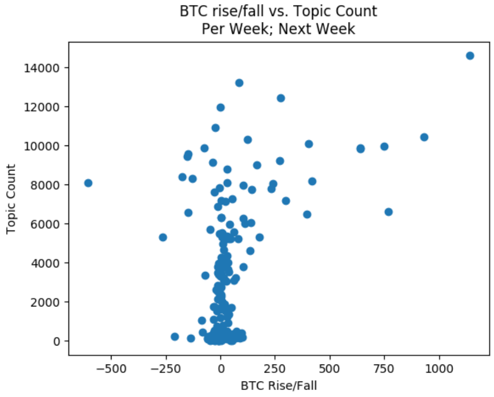

Fig. 48. TFIDF=10; num_topics=2; Next Week; topic=1

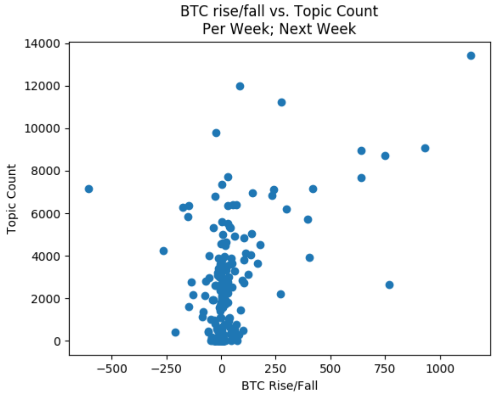

Fig. 46. TFIDF=5; num_topics=7; Next Week; topic=5

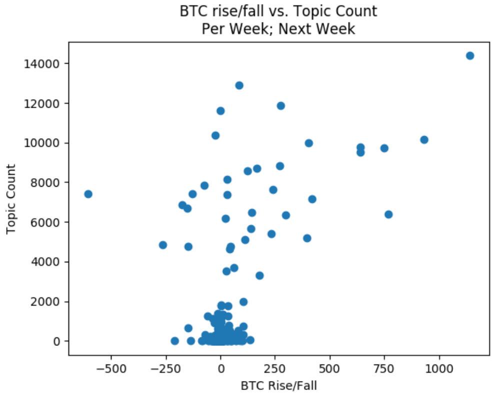

Fig. 49. TFIDF=10; num_topics=3; Next Week; topic=2

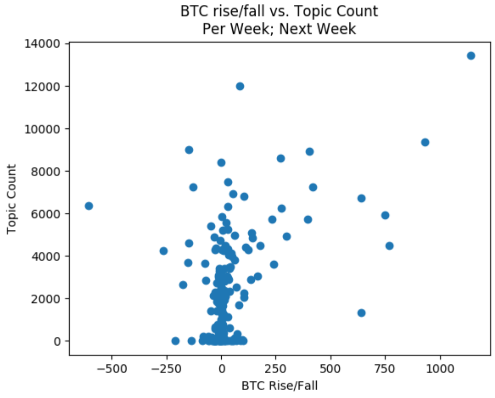

Fig. 47. TFIDF=5; num_topics=10; Next Week; topic=1

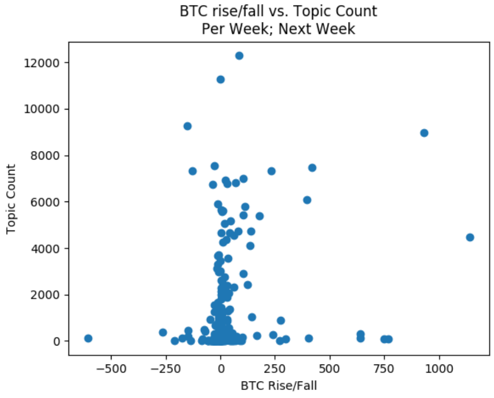

Fig. 50. TFIDF=10; num_topics=4; Next Week; topic=2

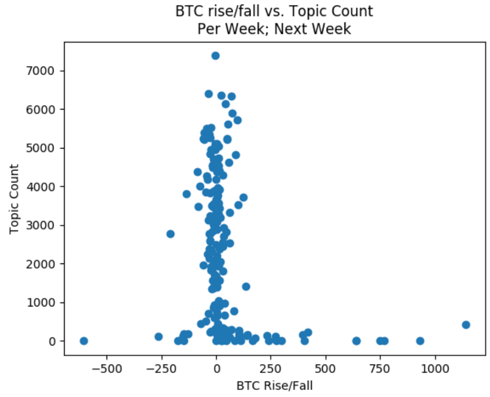

Fig. 51. TFIDF=10; num_topics=4; Next Week; topic=3

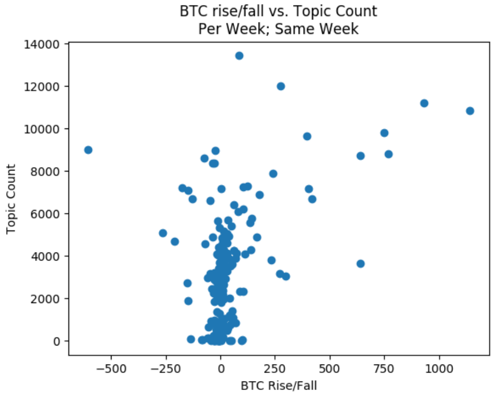

Fig. 54. TFIDF=5; num_topics=5; Same Week; topic=5

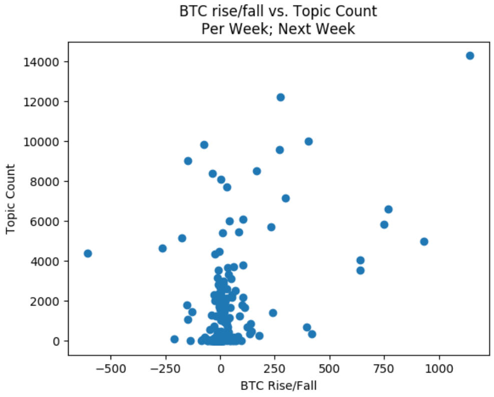

Fig. 52. TFIDF=10; num_topics=6; Next Week; topic=6

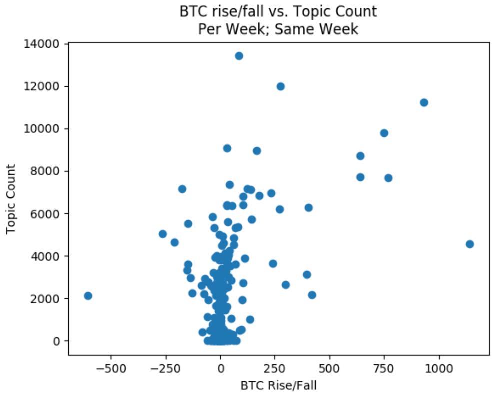

Fig. 55. TFIDF=5; num_topics=7; Same Week; topic=5

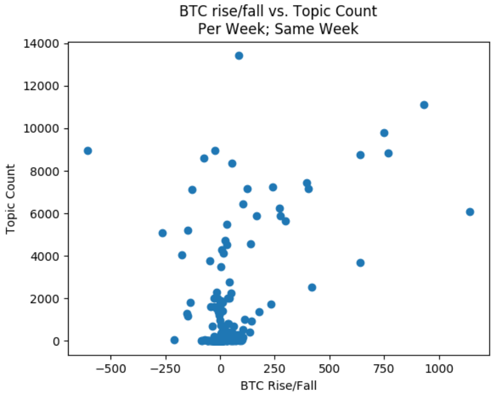

Fig. 53. TFIDF=5; num_topics=4; Same Week; topic=2

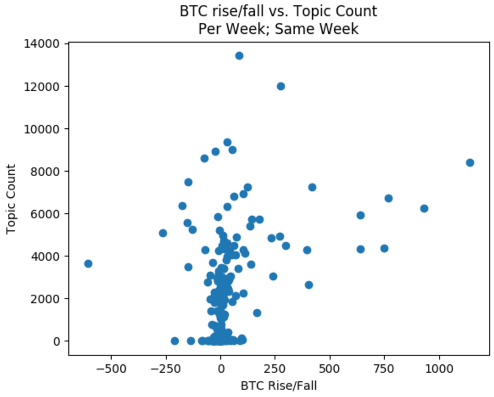

Fig. 56. TFIDF=5; num_topics=10; Same Week; topic=1

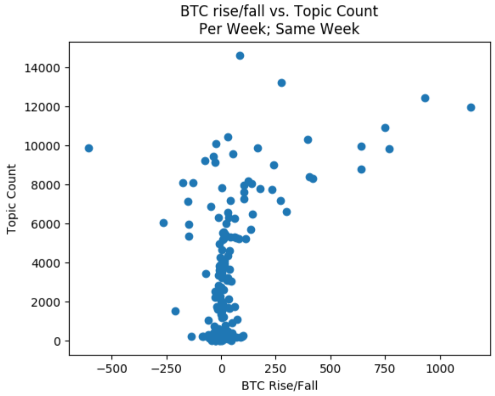

Fig. 57. TFIDF=10; num_topics=2; Same Week; topic=1

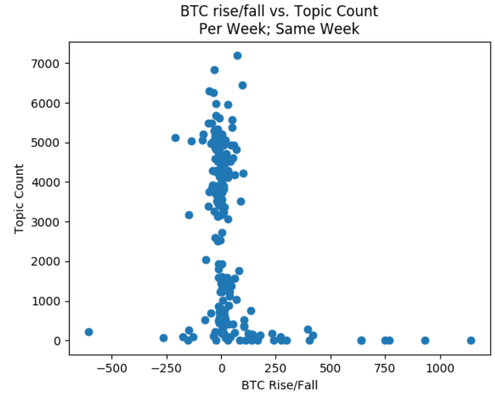

Fig. 58. TFIDF=10; num_topics=2; Same Week; topic=2

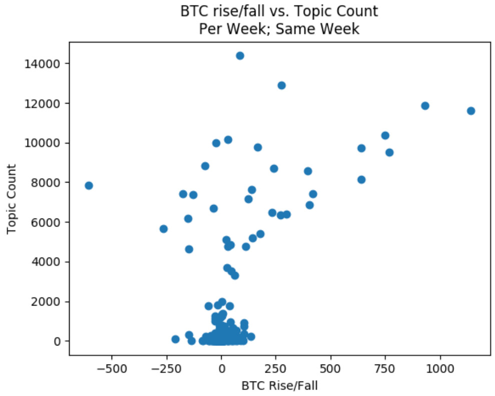

Fig. 59. TFIDF=10; num_topics=3; Same Week; topic=2

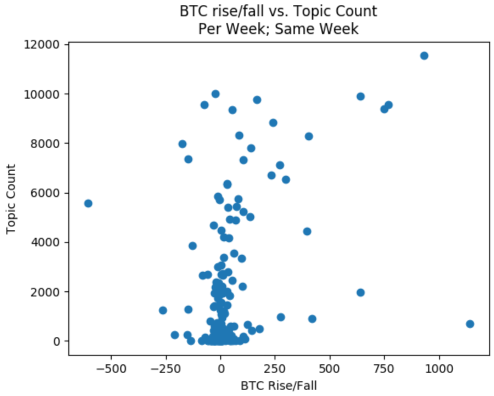

Fig. 60. TFIDF=10; num_topics=4; Same Week; topic=1

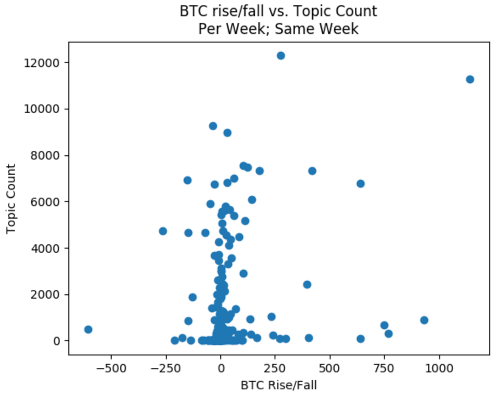

Fig. 61. TFIDF=10; num_topics=4; Same Week; topic=2

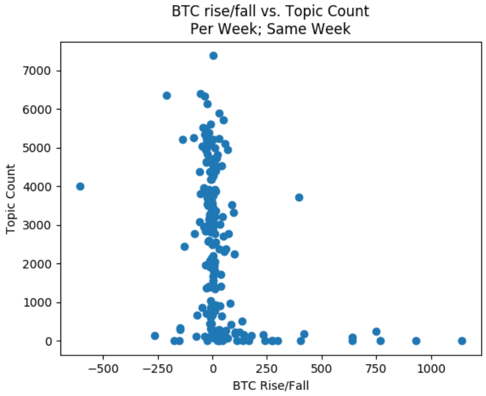

Fig. 62. TFIDF=10; num_topics=4; Same Week; topic=3

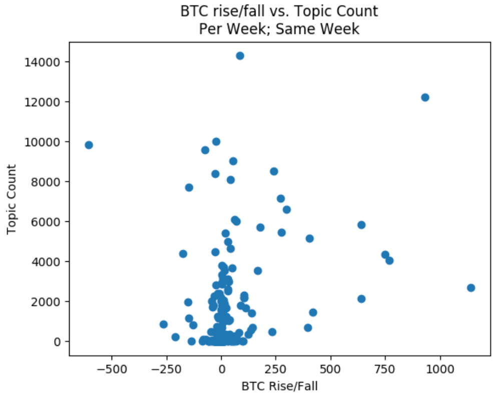

Fig. 63. TFIDF=10; num_topics=6; Same Week; topic=6

## Appendix E

. This section contains some informative plots on the BTC price from 30-11-2013 till 30-11-2017 and also on the subreddit 'BitCoinMarket'.

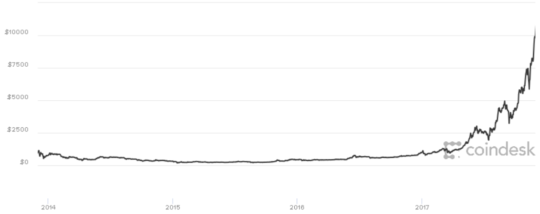

Fig. 64. Bitcoin price from 30-11-2013 till 2018.

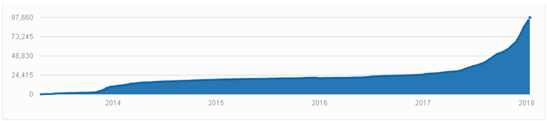

Fig. 65. Subreddit 'BitCoinMarkets' total subscribers from 2013 till 2018.

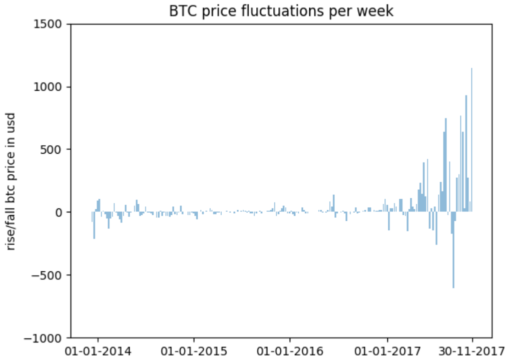

Fig. 66. BTC price rise and fall per week from 30-11-2017 till 30-11-2017

## Appendix F

. This section contains the dendrogram that helped to decide on reasonable values for the num_of_topics parameter for the LDA algorithm. Unfortunately, the dendrogram contains too many clusters so the different cluster terms are not readable.

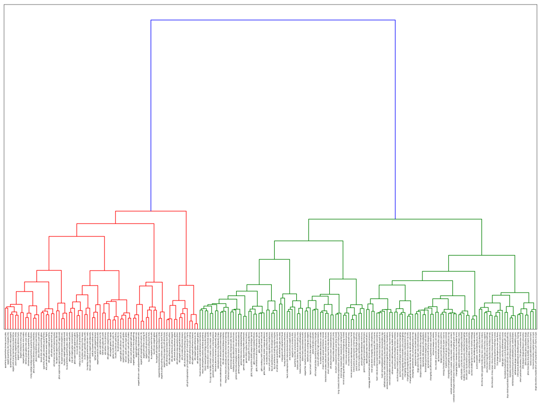

Fig. 67. Dendrogram.