ANALYZING CRYPTOCURRENCY GROUPS USING TOPIC MODELING ON TWITTER POSTS

Bruno Nascimento Rubio

Tampere University
Faculty of Natural Sciences
Computational Big Data Analytics
M.Sc. thesis
Supervisor: Jaakko Peltonen
May 2019

Tampere University
Faculty of Natural Sciences
Computational Big Data Analytics
Bruno Nascimento Rubio: Analyzing Cryptocurrency Groups Using Topic Modeling on Twitter Posts
M.Sc. thesis, 67 pages, 5 index and 22 appendix pages
May 2019

Cryptocurrencies are decentralized digital coins that use cryptographic protocols to provide more secure financial transactions. The world witnessed an impressive rise in the prices of these assets in the last few years, which stimulated a great interest regarding them. This thesis shifts the focus from the most notorious one, bitcoin, and from price aspects to concentrate on other cryptocoins and their technical features. A total of 25 cryptocoins were selected and then divided into 3 groups representing fundamental characteristics: Faster transactions, Smart Contracts and Privacy. Then, daily comments about these cryptocoins on Twitter were collected for 4 months. The main objective was to check whether the categorization fits well for each group, detect the prominent themes under discussion and perform a prediction task to see which ones may be discussed again in the future. Topic modeling, specifically Latent Dirichlet Allocation (LDA), was utilized to process the text data in order to find the topics that best represented each one of the groups. Coherence measures were applied to discover the optimal number of topics, which were later grouped into themes. Daily average probability distributions for topics, or topic weights, were treated as a time series data along with their theme representations. With that, it was possible to forecast theme weights using ARIMA and check the predictive ability of each theme by comparing mean squared error (MSE) of ARIMA and Naive methods. Overall, the cryptocoins seemed to be well represented since in every group there is at least one topic that directly refers to the meaning of the group. However, none of the previously mentioned topics was the most important in any of the groups. Faster transactions and Smart Contracts ended up being similar groups, having a Financial topic-group as the most notable theme, a similar organization of their remaining ones and low predictive ability, while the Privacy currency-group had different results, with a Mixed topic-group as the best-positioned theme and slightly better forecasting results.

Keywords: cryptocurrency, altcoins, twitter, topic modeling, latent Dirichlet allocation, coherence measures, time series data, faster transactions, smart contracts, privacy

ACKNOWLEDGMENT

I would like to express my gratitude to the teachers of Tampere University for the lectures that allowed me to greatly improve my knowledge of not only statistics and computer science but also of Finnish culture and society. I extend my thanks to my fellow colleagues, remembering how our group studies were essential to overcome the difficulties.

I also would like to thank my thesis supervisor Prof. Jaakko Peltonen, whose insightful remarks were a valuable contribution since the earliest stages of the research.

My sincere thanks to Fernando Pradella, who first introduced me to this fascinating world of cryptocurrencies. Last but not least, a special thanks to my beloved friends and family for being always present and supportive.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The creation and diffusion of virtual currencies, also known as cryptocurrencies or cryptocoins, is currently one of the most discussed subjects in the media. The main one, bitcoin (symbolized by ₿), drew the attention of the whole world due to the exponential rise of its price: in mid-2009 1 unit of bitcoin was valued US$ 0.003, but at the end of 2017 it reached more than US$ 17,000 [1].

Aside from news coverage about the fluctuation of bitcoin price, articles about the misuse of cryptocoins are fairly common. While illicit practices do occur, emphasizing those generates a wrong perception that cryptocoins were created just to fulfill the needs of criminal activities, such as money laundering or tax evasion. General awareness of cryptocurrencies is then divided between those who perceive it as a speculative asset to achieve wealth and those who do not consider having it at all, given the association with unlawful schemes. Both standpoints fail to understand the reasons why virtual currencies were developed.

While price volatility and illegal occurrences are undeniably the main driving force that puts bitcoin and other cryptocoins in the spotlight, it is also necessary to explore other aspects that surround them. For example, the mechanism behind nearly all cryptocurrencies, the blockchain, represents a groundbreaking technology that is changing how financial systems work. It eliminates the need for a middleman on transactions and largely facilitates the lengthy process of fund transfers across the globe, among other innovations.

Following the focus of the media coverage, academic research on cryptocurrencies also has heavy emphasis on price. Most of the studies with a statistical approach are focused on price prediction, especially of bitcoin. However, this thesis takes an unconventional approach to the subject: it does not utilize price as a basis for statistical analysis, nor focuses on bitcoin. Instead, the research is designed based on the characteristics that drove the creation of other cryptocoins and studies them through text analysis on social media.

In this thesis, selected cryptocoins were divided into groups that represent features shared among all of them. These features are improvements in areas in which bitcoin underperforms. From that, related comments made on social media were collected during a certain period of time. Twitter was the chosen platform for this work, not only because of the accessibility of its data but

also due to its larger user base, allowing collection of opinions from a broader audience and not exclusively from enthusiasts of cryptocurrency.

This large quantity of text data was broken down into small collections of words that represents specific topics. This task was performed by the usage of a topic modeling technique called Latent Dirichlet Allocation (LDA). Thus, the main objectives of this study are: (1) understanding if the defined groups are either well fitted according to the topics or if there are characteristics that may suggest a new group category, (2) observing which are the underlying themes discussed in each one of the groups and how they interact themselves and (3) performing a prediction task of which are the themes most likely to appear in the future.

This thesis is organized in the following way. Chapter 2 offers a general view on bitcoin, briefly describes the “altcoins” concept and each one of the selected cryptocurrency groups, presents Twitter social media and explains topic modeling, LDA, coherence measures and time series data. Chapter 3 reviews important academic literature pertinent to this work. Chapter 4 concerns the research methods. Data collection and preprocessing steps are detailed, along with data modeling and handling of time series data. Chapter 5 discusses and analyzes the findings. Finally, Chapter 6 provides a conclusion of this thesis and suggests potential future work.

2. BACKGROUND

2.1 General view on bitcoin

Bitcoin is an open-source peer-to-peer (P2P) virtual currency and the first decentralized payment system, at the same time [2]. Until the invention of bitcoin, in 2008, online transactions had always required a third trusted intermediary. For example, if the individual A wanted to send 500 monetary units (m.u.) to individual B, it would depend on third services such as PayPal, Visa and MasterCard. Intermediaries like PayPal keep a record of the customers account balances. When individual A sends 500 m.u. to B, PayPal debits the amount from A's account and credits the same quantity to B's account. These intermediaries were necessary since, without them, a digital currency could be spent twice.

It becomes clearer if one imagines a hypothetical situation in which there is no balance registries, no intermediaries and the digital currency is nothing more than a digital file, which can be sent freely from a computer to another one, just like a PDF file: individual A could send 500 m.u. simply attaching the money file in a message. However, just like it occurs in an e-mail, attaching a file does not remove this file from the original computer. Therefore, individual A would keep a copy of the file and could easily send it again to individuals C, D, E and so on. This problem is known as "double-spending" and, until the advent of bitcoin, having a third trusted intermediary recording the transactions was the unique way to prevent a digital currency to be spent twice [3].

With bitcoin, for the first time ever, the double-spending problem was solved with no need of a third; the bitcoin system solved this problem by distributing the transaction historical registry to every single system users through a P2P web. All transactions that are made in bitcoin are registered in a sort of public ledger named blockchain, which is a large public database and contains the history of all transactions carried out. New transactions are always verified in the blockchain to ensure that the same bitcoins have not been previously expended, preventing the double-spending. The verification process are conducted by the miners, which are the individuals who ensures the authenticity of information by providing computational power, often treated as hash power, in order to solve mathematical problems required to update the blockchain. The miners receive back a reward: a little amount of bitcoin. Therefore, this process has been called mining because of the comparison with the mining process of precious metals.

With the blockchain technology, the global P2P web, consisting of millions of users, becomes itself the intermediary. Of note, it is important to emphasize that transactions in the bitcoin web are not denominated in dollar or euro. Instead, they are denominated in bitcoins. Therefore, the bitcoin system is, in addition to a decentralized payment system, a virtual currency. Its price does not derive from the gold or from some governmental decree. It derives from the value that people who buy and sell bitcoin attribute to it.

2.2 Altcoins

The name “altcoin” is an abbreviation of the extended name “alternative coin”, and the term altcoin may refer to any cryptocurrency other than bitcoin. All altcoins were modeled on bitcoin and, thus, they are technologically quite similar: most of them are generated through "mining", have a limited and known amount of supply, and a blockchain to record transactions.

The first one, Namecoin, came out in 2011, 2 years after the creation of the first block on the bitcoin blockchain. Until 2013, there were only 7 altcoins in the market. Nevertheless, the rise of bitcoin price, in that year, made the creation of other cryptocurrencies spread throughout the world. In March 2018, there were around 1,700 active cryptocurrencies [4]. While some of them were created just in an attempt to reproduce the success of bitcoin, others actually seek to improve some of the bitcoin deficiencies.

It is possible to state that many altcoins have a strong resemblance with bitcoin, meaning that they have almost the same encryption protocol that bitcoin uses, just with small differences. On the other hand, lots of altcoins bring along a variety of new interesting features in their protocol, of which, for purposes of this thesis, are considered three: 1) Smart Contracts; 2) high privacy level and/or anonymity; 3) faster transactions.

Smart Contracts

Smart Contracts, just like the traditional ones, are used to form an agreement between two or more parties, containing all rights, obligations and specifications. However, whereas a typical contract describes the conditions of an accord, an intelligent contract is conceived through cryptographic code. This means that issues inherent in a contract, in the broad sense of the term, in the case of smart contracts are defined and validated partially or entirely automatically.

This technology was conceived initially in 1997, by the computer scientist and cryptographer Nick Szabo [5], but it became possible only with the creation of the blockchain. He realized that ledgers could be used for Smart Contracts.

In this arrangement, contracts can be transferred into computing language, stored and reproduced in the framework and supervised by the computational system that runs the blockchain. Cryptocurrencies which prioritize the smart contract feature attempt to replace restrictive language of bitcoin by a language that enables developers to compose their own projects. In other words, these cryptocurrencies allow developers to design their own intelligent contracts. The language is 'Turing-complete' [6], implying it is able to riddle complex mathematical questions using computational tools.

A contract of a simple financial nature, for example, could define that a given total value of 10 thousand dollars must be reserved and paid to someone in 5 monthly installments of 2 thousand dollars each. In the traditional system, one could think of a paper containing the signature of the two or more parties involved, in order to signal that they are all in accordance with what was written above. Notably, the text should contain the rules and conditions that guide how, when, by whom and to whom the money should be periodically sent.

In a Smart Contract, everything would be done entirely digitally instead, from the stages of sending and receiving the initial amount until the signature of the parties deciding that the payments were made under the terms defined. With a pair of cryptographic keys, it is possible not only to sign digital money transactions that are addressed to the target portfolio in one of the next blocks generated by the network but also to define less obvious and more complex rules for the movement of certain funds. Similarly to what any crypto-wallet application does, with the use of a private key, it becomes possible to interact with snippets of code stored in a blockchain. Blockchain allows the establishment of consensus not only in decentralized financial protocols, such as bitcoin, but its use in a number of other non-financial use cases also becomes feasible. Therefore, just like bitcoin is a kind of "money without banks”, Smart Contracts can be described as agreements without the need for a judge, arbitration chamber or any other intermediary since the rules can be preprogrammed to self-execute [7].

High privacy level / Anonymity

Bitcoin is commonly perceived as a digital asset with high level of anonymity. However, this is inaccurate considering it is one of the most traceable coins since all transactions are registered in the blockchain and the historical record is public. Thus, for example, if individual A transfers a number of bitcoins to B, this transaction is for sure going to leave traces. It is true the blockchain registers neither the name of who made the transaction nor the name of whom it was sent for. Instead, the numbers of the public keys of the sender and the addressee are registered along with a variety of information of the parties, like IP and etc. Therefore, full anonymity is not guaranteed in any bitcoin transaction.

Trying to fill this gap that the bitcoin protocol leaves, many altcoins were designed to promote more privacy or even complete anonymity to users. These currencies are also suitable as value reserves since they are safe and hold values that can be traded without major commitments. The strategies for providing privacy vary from coin to coin, and, here, just some of them will be briefly described.

Monero, the major cryptocoin which emphasizes privacy, attempts to ensure confidentiality to the transactions by using ring signatures: pieces of signature of different people are collected[1] and a new signature is created so that it becomes impossible to know who is the sender because any of the people whose signatures were collected would be a possible candidate. To hide the receiver, Monero uses a stealth address, which creates a new send address not publicly associated with the actual address of the receiver [8].

Zcash, another cryptocurrency which promotes privacy, uses the technology of zero-knowledge proof: a verifier can prove that a statement is valid without knowing the information. When a user transmits his transaction to the miners, they are able to verify whether the transaction is valid or not based on a string of information that could only be generated if that user had the corresponding private key. This way, the Zcash system hides who sends, who receives and the amount sent [9].

[1] If the ring size value is 5, there are 5 people - the larger this number, the more anonymous, but the more expensive the transaction rate.

Verge Currency is another cryptocoin and allows users to choose between public or private transactions. For that, it uses the technologies Wraith protocol and TOR: the former allows to transact anonymously, with the help of some "invisible addresses" that do not let the transactions be associated with the real public addresses. The latter hides the IP of the sender, so it is not stamped on Verge blockchain. Nonetheless, public transactions are still available if the user shuts down the Wraith protocol [10].

Faster transactions

The speed of transaction of a cryptocurrency depends on the balance between the demand for transactions and computational power (known as "hash power") available for resolving the mathematical problems, the size of the block and the limit of transactions per second that the specific blockchain allows.

In the bitcoin case, when there is sufficient offer of hash power, the transactions linger about 10 minutes. It means that a new bitcoin block is created every 10 minutes. As the bitcoin protocol requires 6 confirmations of a transaction[2] to be considered an irreversible transaction, 1 hour is necessary to make a transaction for good within the bitcoin blockchain. However, more than once, the enthusiasm with this cryptocurrency has led many people to suddenly buy it, which enormously increased the demand for transactions with no correspondent increase of the hash power. As a result, the transactions were delayed more than 1 hour and the transaction fees were rising constantly. Bitcoin developers doubled the block size from 1MB to 2MB by implementing new rules to the protocol (Segregated Witness or just SegWit) in an attempt to solve this problem [11].

Even so, one-hour transactions are still a breakthrough in the international transaction area, since the current bank system works only in commercial hours and, sometimes, delays a week to confirm the transaction.

However, some cryptocurrency developers saw in this bitcoin weakness a window of opportunity to make money and, therefore, developed blockchains with faster transactions and

[2] Confirming a transaction means verifying the information of the last block one more time, so errors are prevented.

lower fees. Among many cryptocurrencies, this study is going to cite three examples: 1) Ripple, 2) Stellar, and 3) Litecoin.

Ripple was thought to work with banks and not to compete with them. Its goal is to modernize the current bank system with fast (a few seconds to confirm a transaction), cheap and secure transactions. Many banks have already adopted Ripple to transact, which made the value of this cryptocurrency rise significantly [12].

Stellar is intended to work as an intermediary exchange between two different coins, and not necessarily cryptocoins. In other words, within the Stellar system, one can deposit bitcoin and another can receive in US dollar, or euro, British pounds, etc., or vice versa [13]. The developers bet that the fast transactions (which are confirmed in a few seconds) and the low-cost transaction fee (0.00001 Stellar, equivalent to 0.0000019 USD in March 2018 [4]) will lead people to prefer to use this system rather than the current bank system.

Litecoin is developed by the same group of developers of bitcoin. It has a bitcoin-based blockchain, with some changes. It takes only two and a half minutes to make a new block [14]. This means that the transactions in the Litecoin system are confirmed way faster than in the bitcoin system, which is considered to be a great advantage of this cryptocoin over bitcoin.

2.3 Twitter

Twitter is a social network that enables users to interact among themselves by posting messages (known as “tweets”) with 280 characters maximum, also allowing uploads of photos and videos. According to the latest data (first quarter of 2018), it has reached an average of 335 million monthly active users [15], making it one of the most important platforms of human interaction of the 21st century so far.

The spectrum of Twitter users can be considered very democratic: it goes from regular people to popular celebrities and even heads of state. The platform became a powerful channel for companies to communicate with their clients, for media outlets to publish news headlines in the form of text or video and even for government agencies to make announcements.

This social network does not offer a tool to build groups or communities of any kind. Instead, its dynamics are based on the following: each user can follow any other account and can

be followed by other users as well. All the tweets posted by a user are shown at the feed of their followers. That way, each person can be considered the curator of the content seen on their feed, by following everything that suits their interests.

The most important feature of Twitter is “retweeting”, which is the literal replication of a tweet referencing the user who wrote it. A tweet from a user can be retweeted by its own or any other account. It is an efficient method of spreading a message outside of the range of followers. Tweets that are retweeted a large number of times are either from well-known people (i.e. celebrities or world leaders) or from common users who were able to create messages that “went viral”[3].

Twitter is responsible for popularizing the hashtag mechanism within the social networks. A hashtag is created when the symbol # is associated with a word or a phrase. It is then used as keywords that create hyperlinks that direct the user to a new page, where all the posts that mentioned that hashtag are listed. Some names of the selected cryptocoins for this work are also common words, such as Verge, Stellar or Ark. In that sense, using the hashtags to look for comments about them on Twitter minimizes the risk of getting text data that are just using that word in a context not related to cryptocurrency.

The platform also does a ranking of all hashtags and meaningful expressions. This rank is called Trending Topics (TT) and shows up to 20 most popular terms that have been posted in a certain period. Any user can see the current TT, tailored by country or worldwide. Of note, there is no point in tweeting the same term many times with the sole intention of entering it on the TT, as this depends on the spread between different users and other variables calculated by Twitter algorithms. Twitter recently allowed advertisements on TT, in which a company, organization or government can promote a term to appear on top of it. For those cases, the promoted term is explicitly displayed as advertising, to avoid being confused with the high ranked terms.

Together hashtags and retweets play an important role in the dissemination of content and ideas across the network. The more relevant the theme is for the community that makes use of social networks, the more posts and sharing will occur, leading to greater visibility. With the

[3] Expression defining any content that is quickly spread on social networks due to being shared a large number of times by its users.

democratization of information, communication and immersion of knowledge, social networks became a space for interactivity, coexistence and debate. Nonetheless, if in one hand the online environment might become a representation of collective intelligence, on the other hand, it is easily manipulated by fake accounts and bots, which may lead a spread of false information that has the potential of influencing the population to believe in something without any basis in reality. In the context of this thesis, it is crucial to make a thorough screening of the collected data in order to identify whether the author is a real person or not.

2.4 Topic Modeling

Managing an extensive amount of text data in order to extract meaning from it is a challenging task. When the collection of documents reaches thousands (if not millions) of entries, it is not possible to manually try to highlight important information and an automated procedure is required to do the work. One motivating example is corruption investigation: it generally involves a large number of items to be examined, going from e-mails from politicians to executives of large companies, to bills and contracts from a multitude of enterprises. Another example could be a journalist exploring all the New York Times articles published during the 1990s to discover what was most discussed in that time. For both cases, the underlying information in all documents can be figured out with the implementation of natural language processing (NLP) and machine learning (ML) techniques.

One of the resources in NLP and ML is the bag-of-words model. It has a simple concept, understandable from its own name: every word in a document is put in a list but the order of the words is ignored, so that the collection of these lists is treated as an abstract “bag” containing all the words with its respective frequency considering all documents. Then, each document is transformed into an array of numbers that represent the overall frequencies, in a process called vectorization [16]. Figure 1 below illustrates the entire procedure with a small example.

D_1: "Mike and Amy travelled to Asia."
D_2: "Mike likes Japanese food. Amy also likes Japanese food."

Lists:

["Mike", "and", "Amy", "travelled", "to", "Asia"],
["Mike", "likes", "Japanese", "food", "Amy", "also", "likes", "Japanese", "food"]

Bag-of-words: *10 unique words*

{
"Mike": 2, "and": 1, "Amy": 2, "travelled": 1, "to": 1, "Asia": 1,
"likes": 2, "Japanese": 2, "food": 2, "also": 1
}

Vectorization:

D_1: [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
D_2: [1, 0, 1, 0, 0, 0, 2, 2, 2, 1]

Figure 1: Example of bag-of-words model application in documents *D*

Another resource in NLP and ML is topic modeling, which was the one chosen for this thesis. A topic model is a kind of statistical model for the discovery and clustering of the abstract topics in a document collection and thus offering a way to get a corpus-level view of major themes without having to be supervised. In topic models, the input data is an assortment of documents where each one is a collection of words alongside with an input integer *k* that represents the number of topics to be found. The model outputs:

- on a topic-level: all the words associated with each topic, and their proportions in the topic;
- on a document-level: every topic that is expressed by that document, and their proportions.

There are several different topic models, such as Explicit Semantic Analysis, Hierarchical Dirichlet Process or Non-negative Matrix Factorization. This thesis focuses on just one: Latent Dirichlet Allocation (LDA).

Latent Dirichlet Allocation (LDA)

Blei et al [17] define LDA as a generative probabilistic model for text data, in which a document represents random mixtures over a set of topics, each one corresponding to a multinomial distribution over words. The obtained posterior distribution should answer the questions: (1) how likely each word is to appear in each topic; (2) with which probability each

topic appears in each of the texts of the dataset. The result of fitting an LDA model into a text set is an estimate of the values of these parameters.

For a better understanding, the generative process for LDA will be described in a reverse way, starting with documents and working towards finding the topics in it. The model assumes that new documents are created in phases: i) deciding the amount of words in the document; ii) picking a topic blend for the document over an arrangement of topics. For instance, 30% topic t_1, 20% topic t_2 and 50% topic t_3; iii) creating the words in the document. For each word, firstly choosing a topic based on the aforementioned multinomial distribution of the document (namely, 30%, 20% and 50%). Then, by picking a word according to the topics multinomial distribution of words. Finally, this process must be repeated until reaching the number of words determined in the first phase.

In conclusion, LDA works with probability distributions instead of word frequencies strictly. Unlike the bag-of-words model that focuses on the most frequent words in a document, LDA offers a more global approach by concentrating on the word distribution throughout the topics.

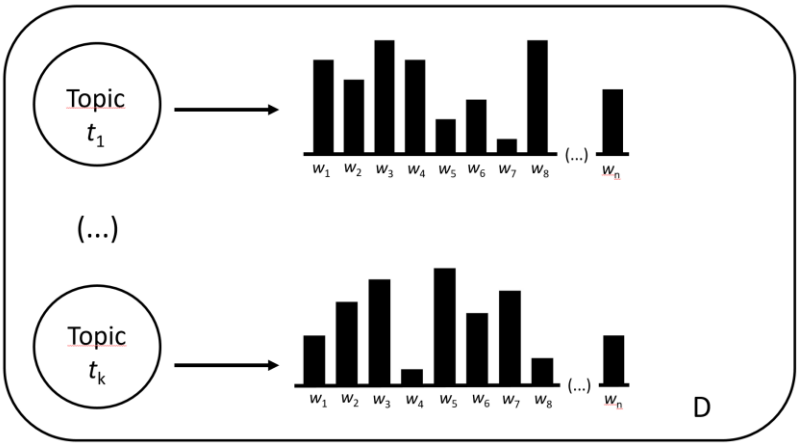


Figure 2: Distributions of the words *w* across topics *t* in document *D*

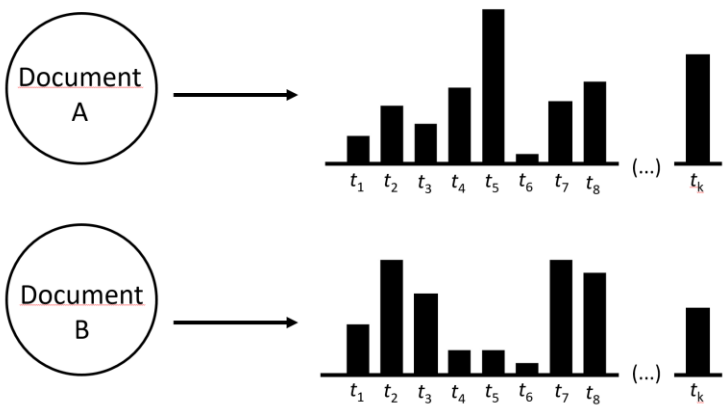


Figure 3: Distributions of the topics *t* across different documents

The figures above give an overview of LDA outputs. In Figure 2, every topic *t* in document *D* contains a number *n* of possible words *w*, each topic showing different distributions over the words, represented by the vertical bars. Words with higher probabilities are most representative for the topic and vice-versa. Similarly, Figure 3 shows that both documents have a number *k* of topics, each document with different distributions over the topics (vertical bars). Topics with higher probabilities are most representative for the document and vice-versa.

The framework for an LDA model is represented in the following figure:

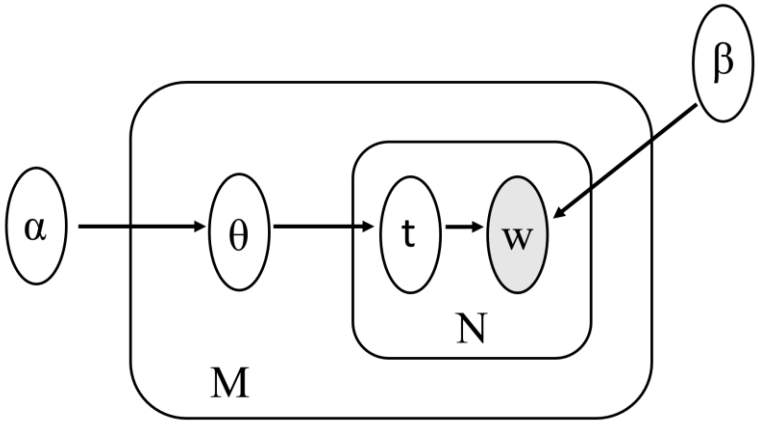


Figure 4: Plate notation of LDA model

The plate notation illustrated in Figure 4 is a compact form to represent in a visual manner the affiliation and the subjection among the model parameters. The outer rectangle *M* represents the aggregate number of documents inside the corpus while the inner rectangle *N* denotes the number of word occurrences in a document. The variables in the figure are described as follows:

- α is the parameter of the uniform Dirichlet prior on the document topic distributions: on one hand, a high α suggests that each document is probably going to contain a blend of the greater part of the topics and not only a couple specifically. On the other hand, a low α is indicative that each document will probably contain only a couple of topics;
- β is the parameter of the uniform Dirichlet prior on the topic word distribution: high β shows that every topic is going to hold a blend of most words, whereas low β indicates that every topic may comprise a blend of only a couple of words.
- θ is the topic distribution for each document;
- *t* the topic that is used to generate each word *w*, and, hence, every single document is a blend of these topics.

A study from Darling [18] addressed the problem of learning the posterior distribution of the variables in the equation:

$$p(\theta, \boldsymbol{t}|\boldsymbol{w}, \alpha, \beta) = \frac{p(\theta, \boldsymbol{t}, \boldsymbol{w}|\alpha, \beta)}{p(\boldsymbol{w}|\alpha, \beta)} \quad (1)$$

Given that this equation is incredibly hard to work with (Darling states that $p(\boldsymbol{w}|\alpha, \beta)$ cannot even be computed), an alternative method can be applied such as the Gibbs Sampling algorithm[4]. Similarly with the aforementioned description for the generative process, it runs LDA backward from the document level to recognize the topics that created the corpus, in order to find the probability of the latter. The process is explained as follows:

1. Randomly assign a word, one by one, in every document to one of the *k* topics;
2. For every single document D:
    - Consider that all topic assignments aside from the present one are right.
    - Calculate 2 probabilities:
        i. Based on words in the document (topic distribution θ) that are currently assigned to topic *t:* $p(t|\theta)$, for each topic *t*.

[4] Gibbs Sampling is an algorithm part of the Markov Chain Monte Carlo (MCMC) method that generates a sample from the distribution and stimulates it to converge to the posterior distribution of interest. It is commonly applied when the direct approach to the distribution is not feasible [19].

ii. Based on all assignments of words to topics over all documents, compute the probability that a particular word *w* will be chosen from topic *t*: $p(w|t)$, for each possible word *w* and for each topic *t*.

- Multiply those 2 probabilities and assign w a new topic based on the result: $p(t|\theta)p(w|t)$

3. Repeat step 2 a considerable amount of times. In the end, it will possibly achieve a stage in which the assignments are satisfactory.

Therefore, considering the parameters α and β, as well as the meanings of rectangles M and N in plate notation, the joint distribution of θ, $\boldsymbol{t}$ and $\boldsymbol{w}$ is given by:

$$p(\theta, \boldsymbol{t}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(t_n|\theta)p(w_n|t_n, \beta) \quad (2)$$

where $p(\theta|\alpha)$ occurs because LDA considers that the document is issued by incorporating the distribution of topics. The following steps recall the work of Blei et al, in which the marginal distribution of a document is obtained by integrating over θ and summing over *t*:

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{t_n} p(t_n|\theta)p(w_n|t_n, \beta)\right)d\theta \quad (3)$$

Finally, the probability of a corpus D is achieved by applying the product of the marginal probabilities of the documents:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{t_{dn}} p(t_{dn}|\theta_d)p(w_{dn}|t_{dn}, \beta)\right)d\theta_d \quad (4)$$

Coherence Measures

Choosing the ideal number of topics *k* is essential to produce good results, as a low *k* generates few and quite general topics, in the same way a high *k* may result in repetitive topics that should have been combined. One way to perform this task is simply testing the results given by the model over multiple choices of *k* and choosing the one that is best suited for the purposes of the research. Although possible, this method is merely based on human judgment. And therefore,

the chosen model would not be chosen based on non-subjective arguments. The alternative is a more scientific method, which allows evaluation of the topics, both qualitatively and quantitatively: through coherence measures.

Coherence measures have been designed in order to qualitatively gauge the coherence of a topic, which is essentially based on the distributional hypothesis of linguistics. Fundamentally, it is presumed that words with similar meaning usually appear in similar circumstances. Thus, coherent topics are the ones in which the majority of words - or even all of them - are somehow associated. In this context, the goal is to design attain measures that have a high correlation with human topic ranking data [20].

There are multiple topic coherence metrics, each one with different approaches, such as Perplexity, UMass or UCI. This thesis will implement two of them, UMass and C_V, which will be detailed as follows.

UMass was first introduced by Mimno et al [21] as an intrinsic method for measuring coherence, meaning that it does not utilize any external source other than the original corpus to evaluate the topic models. It is based on the similarity scores of words $w_1, \dots, w_n$ grouped in pairs, expressed by

$$coherence(W) = \sum_{(w_i, w_j) \in W} score(w_i, w_j) \tag{5}$$

where W is a list of top-k words for some k in a topic. This generalization also applies to UCI, which is, opposed to UMass, an extrinsic method [20].

Letting $D(w_i)$ denote the number of documents D containing the word w_i and letting $D(w_i, w_j)$ denote the count of documents D having the pairwise w_i, w_j, the score function for UMass is defined as:

$$score(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)} \tag{6}$$

In contrast, the study of Röder et al [22] employs a systematic approach that explores existing methods utilized by other coherence measures and summarizes them in just one metric,

called C_V. Its framework, summarized in the study from Syed and Spruit [23], is based on four steps:

i. Segmentation: pairs of subsets S are created from the top n words w in a topic, represented by W, having every single word paired with another one. Thus, for $W = \{W_1, \ldots, W_n\}$ and S_i as a subset of word $W' \in W$ paired with word $W^* \in W$, all subsets S can be defined as $S = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\}$.

ii. Probability Estimation: both single words probabilities and joint probabilities of word pairs are estimated using Boolean document calculation, namely the number of single words or word pairs over the total number of documents. Since it does not consider the distances between words in the documents, a sliding window is integrated into the process and moves over the documents one word token per step. The new virtual documents that are created at every movement of the sliding window are used to calculate the word probabilities.

iii. Confirmation Measure (ϕ): it is calculated for all single pairs $S_i = (W', W^*)$ to determine how W^* supports W' based on how similar W^* and W' are to all words in W. A way to illustrate the semantic support of the words in W is to represent W' and W^* as vectors $\vec{u} = \vec{v}(W')$ and $\vec{w} = \vec{v}(W^*)$. But first it is necessary to calculate the association between single words w_i and w_j using normalized pointwise mutual information (NPMI), given by

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \,.\, P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \right)^{\gamma} \tag{7}$$

where γ is used to add more weight to high NPMI scores and ϵ is added to inhibit logarithm of zero. Then, the context vectors can be created by joining them to every word in W using the sum of NPMI values as follows:

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{8}$$

Next, the last step to get the confirmation measure ϕ of a pair S_i is by calculating the cosine vector similarity $\phi_{S_i}(\vec{u}, \vec{w})$ of the vectors inside S_i:

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \, . \, w_i}{\|\vec{u}\|_2 \, . \, \|\vec{w}\|_2} \tag{9}$$

iv. Aggregation: all individual confirmation measures ϕ are aggregated by the arithmetic mean into one final coherence score.

While for UMass an indicator of a good *k* number of topics is a low coherence score value, the opposite holds for Cv. Nevertheless, the optimal *k* is not necessarily pointed by the lowest or highest score but instead, it is usually the one marked after a steep growth of topic coherence before it reaches a plateau. Choosing the coefficient after that may result in the aforementioned problem of picking a high *k*.

2.5 Time Series Data

A time series data is an ordered sequence of observations X_t, usually with equal intervals of time t but that is not strictly necessary. Such data can be seen in a vast range of areas, practically everything that can be covered with statistical knowledge, such as business, medicine, social and natural sciences, among many others. Figure 5 depicts an example, exhibiting bitcoin price variation (in USD) from Jan/2017 to Dec/2018 [1].

Figure 5: Time series data for bitcoin price (USD) for 2017 and 2018

Basic objectives of analyzing a time series include: describing features and patterns present in the data, understanding how past behaviors may affect next observations, establishing control standards for future values by the adjustment of certain parameters and forecasting future data [24], which is the focus of this thesis. A time series can be either univariate or multivariate, depending on the number of variables it contains. In the context of the research, the output from the LDA model first generates a multivariate time series that is later transformed to univariate. Forecasting will advance using the latter.

Below are listed a few important features that can be present in time series data:

- Trend: an overall or continuous increase or decrease of variable values across time;
- Seasonality: a pattern that is systematically repeated along regular intervals, denoting a specific behavior at specific times;
- Cycle: any periodic variation, not necessarily tied with a specific season;
- Outliers: observations that are drastically different from the other ones.

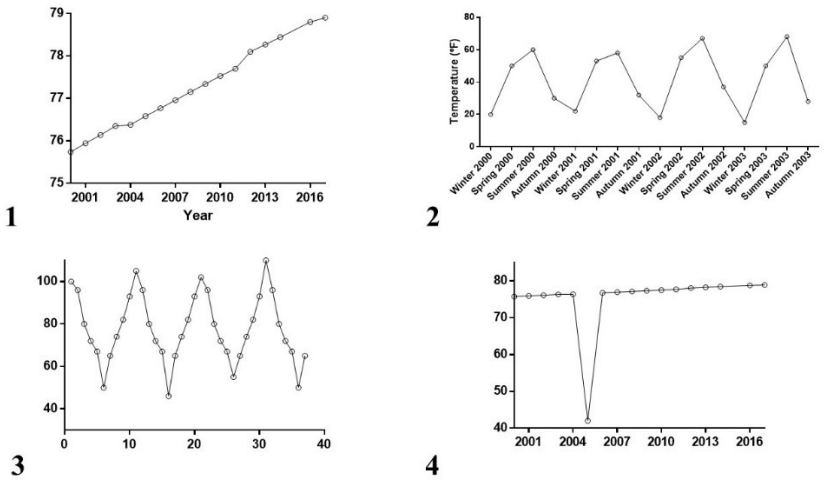


Figure 6: Charts exhibiting examples of time series data features: (1) trend, (2) seasonality, (3) cycle and (4) outlier

One fundamental concept of time series phenomena is stationarity, which denotes any time series that has neither trends nor seasonality or periodicity. In mathematical terms, means that the properties of the time series $\{X_t, t = 0, \pm 1, \dots\}$ are similar to other $\{X_{t+h}, t = 0, \pm 1, \dots\}$ for any input lag h. The exploratory analysis of a time series data involves testing stationarity of the underlying phenomenon and, in the absence of it, apply techniques to remove trends and seasonal components to make it stationary, so it can be used in the forecasting models.

There are plenty of different prediction models for a time series, such as Vector Autoregression or Simple Exponential Smoothing. The simplest one is the Naive method, which basically consists in assuming that the predicted value is the same as the last observation. Considering $h = 1$, commonly known as 1-step-ahead, it is denoted by:

$$\hat{Y}_{t+1} = Y_t \tag{10}$$

A more robust method, the one chosen for this thesis, is the Autoregressive Integrated Moving Average (ARIMA), which is a generalization of the Autoregressive Moving Average (ARMA) method that incorporates non-stationary series [24]. ARMA(p,q) models consist of two parts, an autoregressive part (AR) and a moving average part (MA). The AR(p) part corresponds to modeling future observations from the prior time steps, satisfying the equation

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t \tag{11}$$

where X_t is a stationary time series, $\phi_1, \ldots, \phi_p$ are fixed coefficients, Z_t is independent of X_t and is also a White Noise sequence, meaning that it has normal distribution with zero mean and constant variance σ^2. It can be simply denoted as $X_t \sim N(0, \sigma^2)$.

The MA(q) part treats the error term as a linear combination of error terms happening at the same time and at several past times [25], satisfying the equation for a stationary series

$$X_t = \sum_{k=0}^{q} \theta_k Z_{t-k} \qquad Z_t \sim N(0, \sigma^2) \tag{12}$$

where $\theta_0 = 1$ is usually assumed.

Therefore, an ARMA(p,q) model can be written as

$$X_t = \left(\sum_{i=1}^{p} \phi_i X_{t-i} + Z_t \right) + \left(\sum_{k=0}^{q} \theta_k Z_{t-k} \right) \qquad Z_t \sim N(0, \sigma^2) \tag{13}$$

or more concisely as

$$\phi(B) X_t = \theta(B) Z_t \tag{14}$$

where $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$, $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$ and B is the backshift operator ($B^j X_t = X_{t-j}$, $B^j Z_t = Z_{t-j}, j = 0, \pm 1, \ldots$).

ARIMA models, in turn, are utilized in situations in which there is evidence that the data are nonstationary. The AR component of the ARIMA model designates that the evolutionary variable of interest is returned in its proper lagged values, just as in the ARMA model. The MA component also follows the description used for ARMA. The integrated part (I), employed to remove nonstationarity, designates that the data values have been supplanted by the variation among their previous and actual values, a procedure that might have been rendered multiple times [26]. An ARIMA(p,d,q) model satisfies the equation

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t \qquad Z_t \sim N(0, \sigma^2) \tag{15}$$

where, if the difference iteration part d equals zero, the process becomes ARMA(p,q) again.

Evaluation of forecast results can be done through mean squared error (MSE), regardless of the method applied. MSE is calculated through the average of squared forecast errors, each of which is simply the squared difference between the actual value, extracted from the test data, and the predicted value obtained from the training data.

3. LITERATURE REVIEW

This chapter presents related studies that make use of topic modeling or other NLP and ML techniques to develop research on cryptocurrencies. Papers about this subject have been published in an increasing amount as the monetary value of these crypto assets increases and draws more attention of the public. Many statistical works focus on bitcoin price prediction as it was the first to be launched, the one which has the most active web community and the greatest market capitalization. For instance, the research from Blau [27] tried to establish a connection between bitcoin's price volatility and speculative trading. Meanwhile, Jang et al [28] utilized information on bitcoin's blockchain to create a statistical model for price prediction utilizing Bayesian neural networks. Other similar examples of bitcoin studies are abundant.

Narrowing the scope to research that used text data, one common NLP application is Sentiment Analysis, which aims to identify the feeling that users have about some entity of interest (a specific product, company, place, targeted group of people etc.) based on the content available on the Web or digitalized from a physical source. The approach is capable of recognizing subjectivity and the underlying opinion in texts. The outcome does not show what the content expresses, but rather the predominant emotion in it: positive, negative or neutral [29]. The work developed by Matta et al [30] made use of sentiment analysis to model a correlation among Twitter, Google Trends and the fluctuations in the bitcoin market. Even though data was gathered for just one coin and only spanning 60 days, the total amount of retrieved tweets was roughly 50% of the total collected for this thesis. There is no mention about the handling of retweets and bot posting if there was any at all. It was then observed that peaks expressing positive sentiment towards bitcoin, both in Twitter and Google Trends, have a strong correlation with posterior peaks in the bitcoin price for a few days. Consequently, the platforms may be used as predictors of bitcoin price fluctuations.

Twitter and Sentiment Analysis were also used by Stenqvist and Lönnö [31] for predicting bitcoin price. Similarly to the paper of Matta et al, the work from Stenqvist and Lönnö tries to find a correlation between volatility and social media sentiment but in a different fashion: disregarding macro trends and only analyzing feelings on a short term basis, that is, within intervals ranging from 5 minutes to 4 hours. A simple prediction of price increase or decrease according to sentiment indicated in those frequencies is then shifted forward in 4 different timestamps to indicate how the

coin behaved. A total of 2.27 million tweets were collected just in one month, showing how much bitcoin dominates the cryptocurrency discussions. After a cleaning process to mitigate the effect of bots, around 55% of this total was removed. The authors concluded that, for some sentiment frequency/shift counts, the prediction accuracy is around 50%, indicating that there is neither positive nor negative correlation between sentiments and bitcoin price fluctuations. On the other hand, for other sentiment frequency/shift counts, the accuracy increases, suggesting a positive correlation. Also, the accuracy is higher 1 hour after posting the sentiment and the effect on the price is seen around 4 hours later.

A multidisciplinary approach based on social network data was taken by Laskowski and Kim [32], utilizing NLP techniques in order to better understand variations in price and in other trends regarding the cryptocurrency market. Text data were gathered from Twitter and from Internet Relay Chat (IRC) for a six months period, regarding discussions about bitcoin. IRC organizes the discussions into topics in which the subjects are followed by a hash symbol, similarly to Twitter hashtags. Although the total amount of entries were not disclosed, it was stated that the number of messages retrieved from Twitter surpassed the total collected from IRC by a large margin. Daily bitcoin USD price and volume of transactions were collected for the same period. The processing pipeline was conducted using an open source for NLP called GATE (General Architecture for Text Engineering) along with a pre-packaged application for Twitter analysis. It was found that Twitter post volume is positively correlated with the bitcoin volume transactions, whereas the total IRC post volume in the channel “#bitcoin-pricetalk” is also positively correlated with bitcoin price and transaction volume. Conversely, some negative correlation was found between the daily amount of discussions in the channel “#Bitcoin-assets” and the bitcoin price. The conclusion points that, even with a lower volume of messages, IRC data may be very much valuable for this kind of analysis, as they complement Twitter data. A limitation was noted that also became clear prior to the start of data collection for this thesis: real-time analysis of data coming from different sources is limited by computational power.

Related work that expands the cryptocurrency spectrum outside bitcoin is more limited. One example is the article published by Kim et al [33] that proposes a method to predict fluctuations in the prices and amount of transactions of not only bitcoin but also of Ethereum and Ripple. Litecoin was considered but ultimately discarded because its online community was not considered active enough, despite its great market cap. Data was collected from dedicated message

boards for the three cryptocoins under analysis, corresponding to post texts, the time stamps and the number of replies and views of each post but excluding all repeated sentences in replies referring to the previous post. Techniques to filter spam data were applied, for instance, removing any post of more than two sentences that appeared more than five times a day. Time frames for the data collection were different for every coin, justified by particularities of each one, and retrieval of price data and number of transactions corresponded to each period. Once the entire data had been crawled, a Sentiment Analysis algorithm was used to classify every single comment into a range that varies from very negative to negative and positive until very positive. This classification was further compared to the price and to the number of transaction fluctuations of each crypto for formulating the predictive model. Very positive/positive posts, positive replies and the number of topics proved to be significantly correlated with bitcoin price fluctuations, in such a way that the maximum accuracy (79.57%) was obtained with a 6-day lag. Alternatively, very negative/negative and positive posts showed a significant association with Ethereum price variation, and the accuracy was the highest (71.823%) with a 6-day lag. Ripple price fluctuations, in turn, ended up being significantly associated with very negative posts and negative replies, being the maximum accuracy (71.76%) with a 7-day lag. In conclusion, the prediction accuracy varied around 8% among different cryptocurrencies, so that the predictions regarding bitcoin were the most precise and the ones regarding Ripple were the least precise. This is attributable to the difference in the market capitalizations and number of daily comments and replies. In addition, comments proved to be useful to predict the number of transactions of all coins and they estimated that user opinions might be used to predict fluctuation in six or seven days.

Even scarcer are statistical researches that focus neither on bitcoin nor on price. In the research by Linton et al [34], data regarding bitcoin and altcoins were extracted from popular online message boards, such as bitcointalk.org, and analyzed through topic modeling and text mining in an attempt to identify new trends, economic matters and legal or illegal schemes regarding cryptocurrencies across the time. To do so, they collected thread ids, post ids, usernames, time stamps, post titles, post texts, quotes of other posts and links of each post and stored them in the database. Thread and post ids were used to who is posting and which thread they belong; usernames were necessary to make an association between a post and an agent; time stamps were collected to make time slices; post titles and texts were used as source for LDA; links and quotes were useful for determining how one post is related to another. A total of approximately $15x10^6$

posts, $5x10^5$ threads and 200 subforums were collected from November 22nd, 2009 to August 6th, 2016. The results were compared to what actually occurred to test the power of the analysis. The article identifies interesting words that are associated with the historical events that happened in the year in which the data collection was made. For example, the results highlighted the relevance of words like CPU, difficult and mining in 2009, when the bitcoin mining difficulty augmented rapidly. The authors also identify that words related to fraud, pyramids schemes and other semantically related words appeared in a high volume. They speculate that the reason for the high occurrence frequency of these words may be the lack of governmental regulation concerning bitcoin and other cryptocurrencies. Additionally, they affirm that the biggest event in bitcoin history to that year (2016) is the insolvency of the MtGox bitcoin exchange that happened in 2014. They found that the word probabilities are usually flat probabilities along the time and have peaks during events.

4. RESEARCH METHODS

4.1 Data collection

The cryptocoins analyzed in this thesis were selected to fit into one of the groups (explained in Chapter 2) that focus on deficiencies of bitcoin: faster transactions, smart contracts and privacy levels. Upon investigation on websites specialized on cryptocurrency news and analysis, 57 cryptocoins were first selected.

Even though market cap is not a feature to be explored in this thesis, it was taken into consideration to narrow down the list of the first selected cryptocoins. A very low value of daily volume and global market cap is a strong indication that the coin has low awareness. Therefore, there may not be a reasonable amount of people commenting about it on Twitter.

After considering the eligibility of only those cryptocoins that performed in the top 100 position of global market cap on the website https://coinmarketcap.com/ [4] in March 2018, the final list was narrowed down to 25 coins, as follows:

FASTER TRANSACTIONS	SMART CONTRACTS	PRIVACY
Ripple (XRP)	Ark (ARK)	Bytecoin (BCN)
NEM (XEM)	Cardano (ADA)	Komodo (KMD)
Steem (STEEM)	ChainLink (LINK)`	MaidSafeCoin (MAID)
Bitcoin Cash (BCH)	EOS (EOS)	Monero (XMR)
Litecoin (LTC)	Ethereum (ETH)	PIVX (PIVX)
Dash (DASH)	Ethereum Classic (ETC)	Verge (XVG)
Bitshares (BTS)	NEO (NEO)	Zcash (ZEC)
IOTA (MIOTA)	Qtum (QTUM)	ZCoin (XZC)
Stellar (XLM)		

Table 1: Groups of cryptocoins separated according to three main features

By using Twitter API access, a Python script was written to extract tweets that made any mention to the selected cryptocoins. Due to the restrictions that prevent retrieval of messages older than two weeks, the script was executed to get all the tweets within a period of a day. The process was repeated from May 2018 to August 2018 resulting in 4 months of daily data.

4.2 Cleaning steps

Text data retrieved from Twitter cannot be used in their original form. There are multiple repetitions of the same information and a lot of action from automated accounts (called “bots”). The former does not add any significance to this work and the latter is not on the scope of the research since it is seeking comments only from real people on social media. The cleaning steps required during the preprocessing stage consist of:

1) Removal of repeated messages;
2) Mitigation of the effects caused by bot accounts that generate a large amount of meaningless data;
3) Protection of any relevant information that could be wrongly removed during step (2);
4) Elimination of tweets unrelated to cryptocurrency.

Retweets

One noticeable characteristic of each file with daily data was the high number of retweets, which forms roughly half of the total number of tweets. It is pointless to analyze if it has been published by bots or real people because either way it consists of unnecessary repetition of the same information. Consequently, it should be removed.

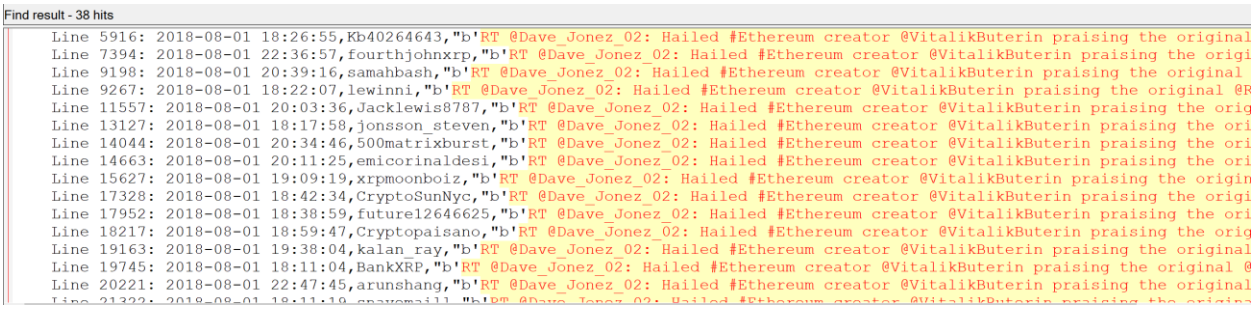


Figure 7: Example of a single post retweeted multiple times

It is important to point out that the elimination of all retweets does not mean the removal of the message entirely. Given the nature of the retweeting feature, the original tweet is likely to be retrieved on the same day or on the day before. In addition, since all retweets are marked with “RT”, the process for eliminating them is fairly easy.

Bot accounts

Tweets generated by bot accounts are abundant in every file and appear in a wide variety of forms. A common type are bots that keep track of a particular coin and post variations on its price within a certain period of time. However, the most predominant ones are those which post advertisement-type of messages. Usually the same text is posted multiple times across many different accounts on the platform. Figure 8 below illustrates one occurrence:

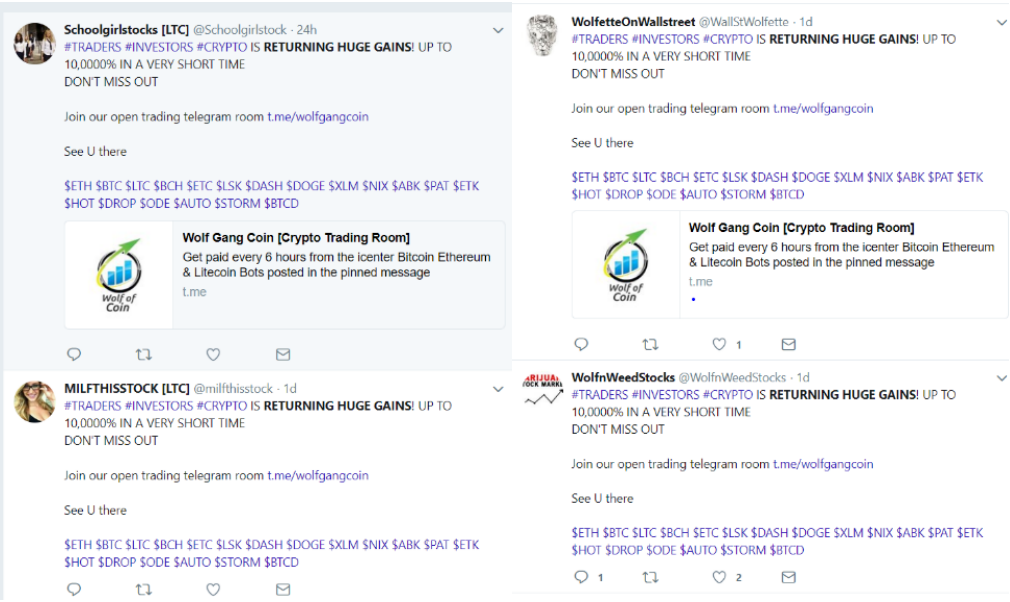


Figure 8: Same comment posted by different accounts, illustrating actions of bots on Twitter

Bot identification is a challenging task because it is not always clear if an account is in fact automated or belongs to a real person. There are some common characteristics among bots, such as a high number of posts per day (hundreds or even thousands) or weird usernames that seem randomly generated. However, these are mere indications to of an account being a bot, not concrete proof. Since developing advanced techniques to identify bot accounts would take a long time and, was not the purpose of this work, an alternative method was implemented to solve the issue.

After the dismissal of all retweets, the remaining ones were alphabetically ordered and each one was compared to the following one. In case both of them were 75% similar (threshold arbitrarily defined), the two users that posted it were flagged. At the end of this process, all tweets from the flagged accounts were discarded. Approximately 2,500 accounts were ignored per daily file, representing 41% of total users retrieved for this research.

Relevant accounts

Upon inspection of the list of flagged users, it was noticed that some accounts from real people were being flagged in the process. They were easily distinguishable because these users are prominent members of the cryptocurrency community. It was then discovered that another type of bot action is simply replicating (not by retweeting it) the message of one of these important users as one of its own. With that, some real persons' accounts ended up being flagged by the procedure previously described.

To prevent the removal of relevant data, a bypass list of all well-known usernames were created, meaning that all tweets from these accounts were preserved. The list contained around 100 different users, built according multiple sources on the media, such as the article "19 Bitcoin Accounts You Should Follow on Twitter" published by Fortune Magazine [35], the article "The 100 Most Influential People in Crypto - 2018 Edition - " by Crypto Weekly newsletter [36], among others [37][38]. All official accounts of each one of the selected cryptocoins were also included, if they existed.

Unrelated tweets

The last part of the cleaning was implemented just after the first results. Some of the created topics for the cryptocurrency groups presented words that hardly have any connection to the subject, such as "ice cream", "chef", among others that apparently were from an African language ("chigumba", "mnangagwa", "chamisa" etc). After investigation of the tweets, it was discovered that some of the coins names and acronyms were shared with terms used in situations unrelated to cryptocurrency, as following:

- ZEC (Zcash) was a hashtag used prior to the Zimbabwe elections;
- BCN (Bytecoin) is an acronym for an online community called "The Best Chef Network";
- Ripple is also part of an ice cream flavor called "Raspberry Ripple".

There was no way to prevent the retrieval of this unwanted data given that the situations mentioned above were unknown before the start of this thesis. All tweets related to those were removed.

After the last step of the cleaning phase, all remaining tweets were eligible to be used. Figure 9 shows the breakdown of the total number of retrieved tweets and highlights the great amount of useless data generated in the social media. It is possible to reaffirm that retweets make up for around 50% of all collected tweets, while the sum of valid tweets to be used on this work was less than 20% of the total.

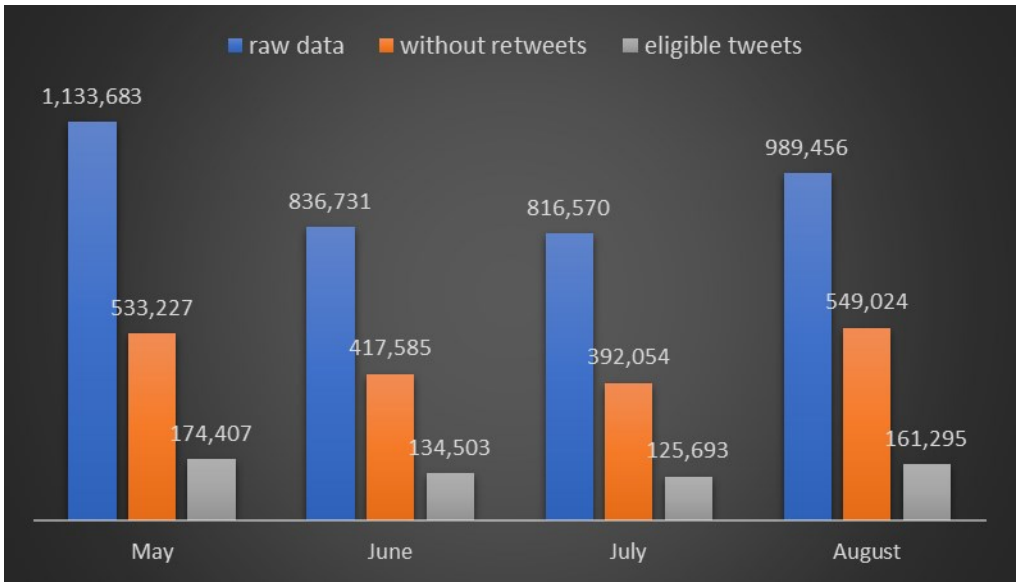


Figure 9: Monthly totals of retrieved tweets

4.3 Preprocessing

All eligible tweets were stored in a single file. Along with the message text, each tweet contained information about the date and time of the posting and which cryptocurrency group it belongs to. The text data required further transformation to be appropriately used in the models. The first one was the tokenization process: punctuation was removed and each tweet was transformed into a list of words. Next, the lemmatization process was performed: all words were converted to their root form (for example: “studies”, “studying” and any other variation were stored as “study”), avoiding unnecessary repetition of the inflectional forms of a term.

The last part was the removal of stop words, which usually consists of function words that allow to maintain a cohesive phrase structure and do not have meaning by themselves (i.e. “the”, “at”, “as”). The Natural Language Toolkit (NLTK) used to process the data has a library with a preset list of stop words. However, it needed to be expanded to include more terms that should not

end up appearing in the topics: any references to cryptocurrency names and acronyms (including those not in the scope of the research), people usually referenced in the discussions and popular cryptocurrency exchanges[5]. These words do not add any significance to the research: cryptocurrency names are redundant information and it is not of interest to explore names of individuals discussing them or the exchanges where they are traded. Such terms may appear very frequently within the messages, affecting the model and possibly resulting in low-quality topics if not removed.

Another addition to the stop word list are words with low frequency considering the entirety of the data. Infrequent words are not evenly distributed among the topics. As a result, some topics show more relevant words than others, affecting the quality of the model. The threshold for infrequency was defined as a word frequency lower than the 50% of the mean of the total amount of words per total unique words, calculated separately for each cryptocurrency group. As seen in Table 2, only 10% to 15% of total unique words were eligible to be used in the model.

	Privacy	Faster transactions	Smart Contracts
All words (A)	628,266	2,715,660	3,469,310
Unique words (B)	24,436	58,709	62,942
Minimum frequency – (A/B) * 0.5	13	23	27
Low frequency unique words	20,704	52,570	56,531
Eligible words (% of total)	3,732 (15.3%)	6,139 (10.5%)	6,411 (10.2%)

Table 2: Breakdown of word quantities for each group

4.4 Data modeling

After the preprocessing steps, the text data was ready to be fed into the LDA models. As explained in Chapter 2, text data is one of the two required inputs, the other one being the integer *k* defining the number of topics. To find the best *k* value for an optimal model, coherence measures C_V and UMass were calculated for all LDA models for each cryptocurrency group using *k* values from 2 to 40, with step size 4. As pointed out in section 2.4, an optimal *k* is usually the last value before topic coherence reaches a plateau. In the case of both C_V and UMass indicating approximate

[5] It is worth mentioning that the actual word "exchange" was not added to the stop word list.

k values, the selected *k* will be an average of the results. However, if both measures differ greatly, *k* will be chosen from either C_V or UMass according to the thesis author's judgment.

The execution of the modeling task was conducted using the Machine Learning for Language Toolkit (MALLET), a Java-based application for topic modeling and other statistical analysis using text data [39]. The text input was converted into a .mallet file, one for each cryptocurrency group. The process then output, per group:

- *k* number of topics with its respective word associations for each one;
- probability distributions over all words within each topic;
- probability distributions over the *k* topics for each tweet.

Further analysis of the topics proceeded considering the top 7 words with highest probabilities (referred to as “weights” from here on) for each topic. These words should show if the categorization of the cryptocoins within the groups fit accordingly, at least for one topic. That is, if the group category was reflected as a common topic within the group corresponding to the characteristic that was used to create the group, or, if there are characteristics of one group shown in others, suggesting a potential interference of one or more coins that were misplaced. Then, every one of the topics was interpreted in regard to find a broader meaning for it, which may be shared with more than one topic. With that, topics could be manually grouped into topic groups represented by just a keyword, which this thesis names as “theme”. Summarization allowed by themes greatly helped to better understand the results of the topic modeling.

4.5 Analysis of the time series

Considering that all tweets were stored with its respective date and time of its posting, the output of the LDA model containing the topic weights for each tweet can be seen as a multivariate time series data, with *k* variables. Each topic was represented in a scatter plot, having multiple data points (topic weights) for every timestamp, according to the number of messages posted in the same period. While this allows the topics to be analyzed individually within the timeframe of the data collection, a more general comprehension of how they relate to each other is compromised. To address this issue, the weights needed to be aggregated in order to have a single data point displayed for each aggregate interval of timestamps (here, one day), enabling visualization of all topics in a single line chart. For that, the average topic weight was calculated, in a simple manner:

by adding all the topic weights for a timestamp then dividing it by the total number of entries for the same period.

Even the average topic weights chart can be difficult to understand if the *k* number of topics is too large (for instance, higher than 15): many topics may present close weight values, creating a lot of overlapping. Consequently, data can be aggregated again according to the themes defined for each topic. This not only helps for a more cohesive view but also in clearly identifying which theme is more prevalent in each one of the groups' discussions. Gathering multiple topics under one theme may give a false impression that this particular theme has a better overall weight than another one that contains fewer topics. Therefore, it is only possible to point out the prevailing theme in each one of the groups through the comparison of themes aggregated topic weights.

Calculation of theme weights was done by simply summing the topic weights assembled by day and theme it belongs to, per cryptocurrency group. Moreover, the average theme weight was computed using the same procedure applied to the average topic weight. Finally, the results were plotted in a line chart, one for each group.

Furthermore, a prediction task for every theme was performed using the average theme weight values where the aim was 1-step ahead prediction. It is important to point that 4 months of daily data is considered a quite short period for a conventional time series forecasting, thus not presenting any kind of seasonality. In case of a clear trend, data should be transformed in order to remove it. Then, the first 2/3 of the data for a single theme was separated as a training set for an ARIMA model and the remaining 1/3 was used to test the accuracy of the predicted results.

In order to determine which *p, d, q* values are the most suitable for the ARIMA(p,q,d) model, grid search was performed combining multiple values for each parameter: ranging from 0 to 10 for the lag values (p) and from 0 to 5 for both difference iterations (d) and residual errors (q). The chosen parameters were based on the model that presented the lowest mean squared error (MSE). This process was first repeated for the remaining themes in the same group and then replicated to the other two groups. Consequently, predictions for every theme within a group could be done using several different ARIMA models.

Although simple, Naive prediction was also implemented, again for 1-step ahead. This was mainly done to compare the MSE values for both Naive and ARIMA and detect the prediction

ability of the latter model. Finally, a single line chart was made combining the real data points with the predicted ones using ARIMA, for all average theme weights within a group. This allowed an overall view of the forecasting results.

4.6 Additional tests

Later experiments were conducted subsequently either to find another perspective from the obtained results or to further investigate certain phenomena present in the data. For the latter, the time series data were plotted in a more detailed timescale, on an hourly basis, in order to check the observed topic activity and evaluate if it is possible to get a better understanding of the noisiness seen on a daily basis. As for the former, data was again modeled in LDA, but only with 8 topics as output, which is the same number of themes that resulted from the aggregation of topics for each group (except Privacy, with 9). The aim is to detect whether there is any equivalency between the new topics and the obtained themes in the first run. Then, all the previous steps explained in sections 4.4 and 4.5, excluding the aggregation into themes, were replicated.

5. FINDINGS AND DISCUSSION

In this Chapter, results for the three groups are presented and discussed in the following order: 1) Coherence measures Cv and UMass and the chosen *k* number of topics; 2) Topics generated by the LDA model; 3) Aggregation of topics into themes; 4) Scatterplots for daily topic weights; 5) Line charts for average of daily topic weights; 6) Line charts for average of daily theme weights; 7) ARIMA predicted values for themes with MSE comparison versus the Naive predictive method and an overview of actual and predicted data together; and finally 8) Results for additional experiments.

Coherence Measures

Figure 10 shows that both Faster transactions and Smart Contracts C_V results indicated the same pattern: a rapid increase of the score values for the first number of topics, followed by a plateau behavior for the higher ones. Privacy scores, however, just kept increasing, except for a small negative variation between 25 and 30 topics. UMass results shown in Figure 11 indicate a continuous decrease in score values for all the groups.

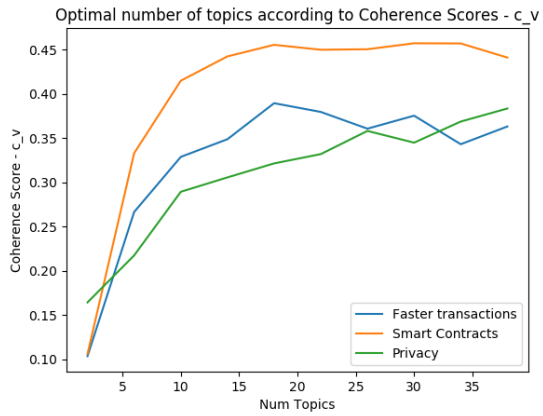


Figure 10: Optimal number of topics according to C_V results

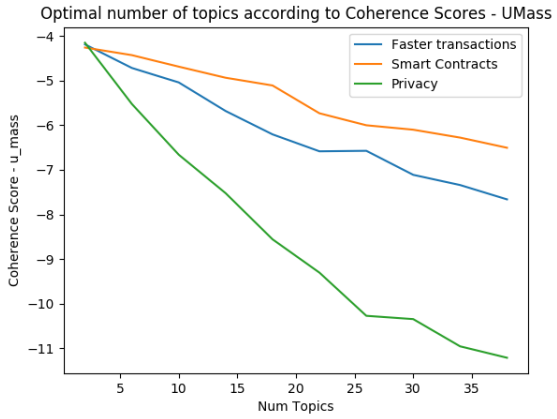


Figure 11: Optimal number of topics according to UMass results

While UMass scores are inconclusive, indicating that the optimal *k* number of topics is the highest possible, C_V results give a better perspective about the best possible coefficient. A quick growth in topic coherence is shown until reaching 18 topics for Faster transactions and Smart Contracts, keeping almost stable for higher numbers. Therefore, both groups can assume $k = 18$.

As for Privacy, C_V scores did not reach a plateau for any number of topics, making it difficult to draw conclusions solely based on it. However, $k = 18$ can also be for this group, given that it is still a number with high topic coherence and to make it cohesive with the other groups.

Topics

The 18 topics per group generated by the LDA model are shown in Figures 12, 13 and 14. Every chart represents one topic, exhibiting its respective 7 most prominent words in a top-bottom perspective, with the horizontal bars indicating how representative each word is, according to its weight value.

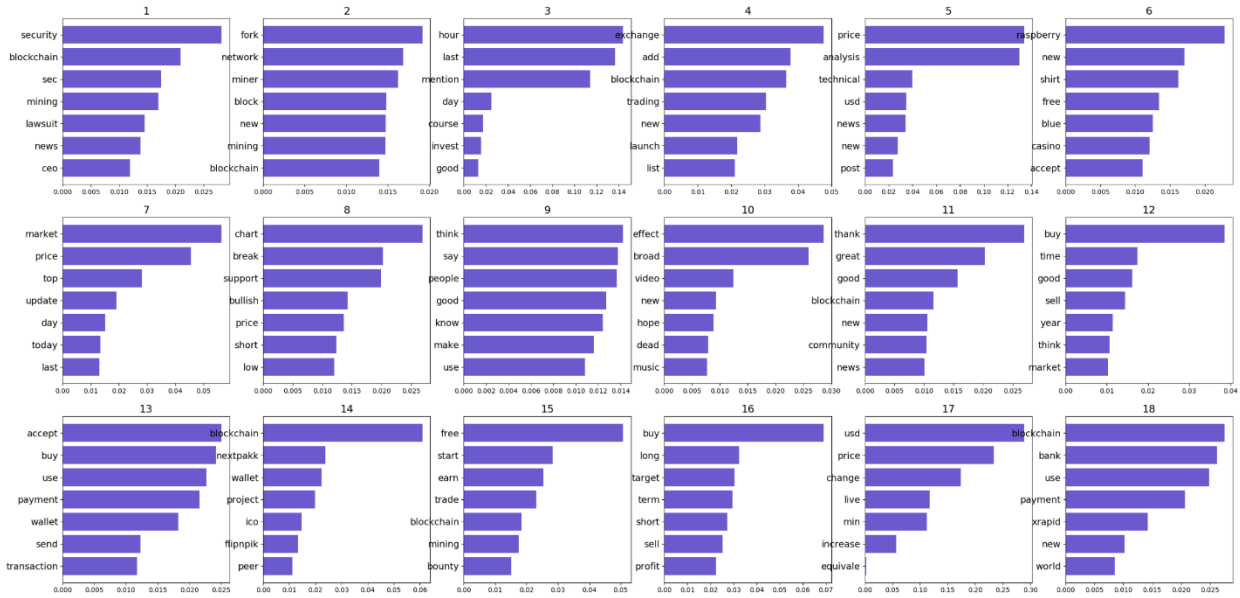


Figure 12: Topics for Faster transactions

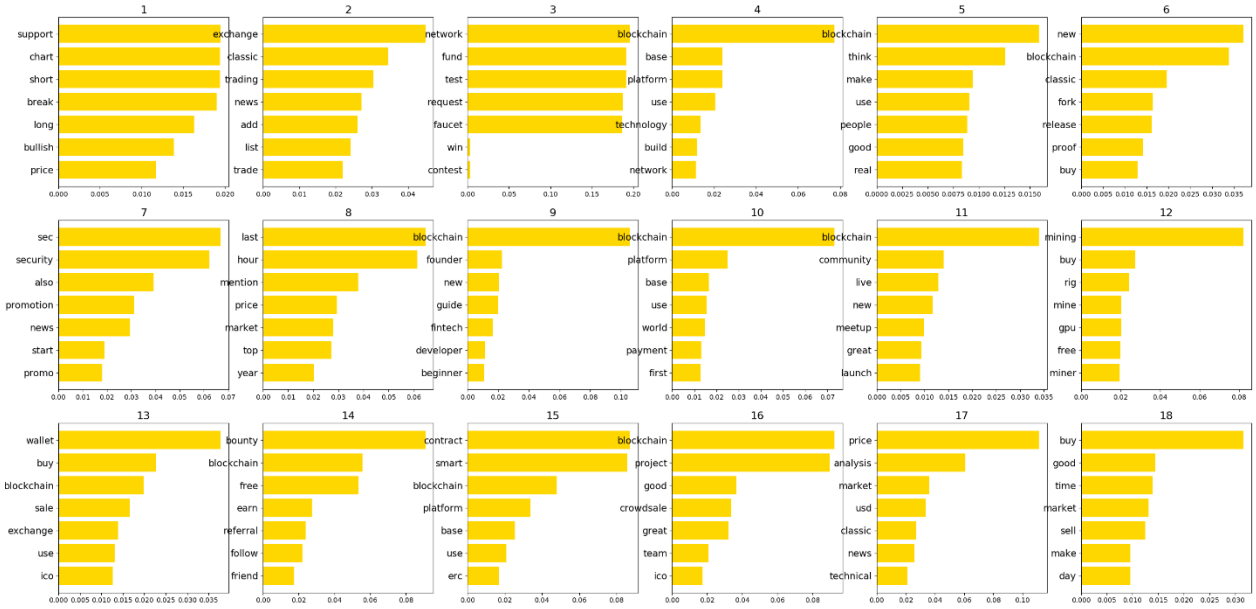


Figure 13: Topics for Smart Contracts

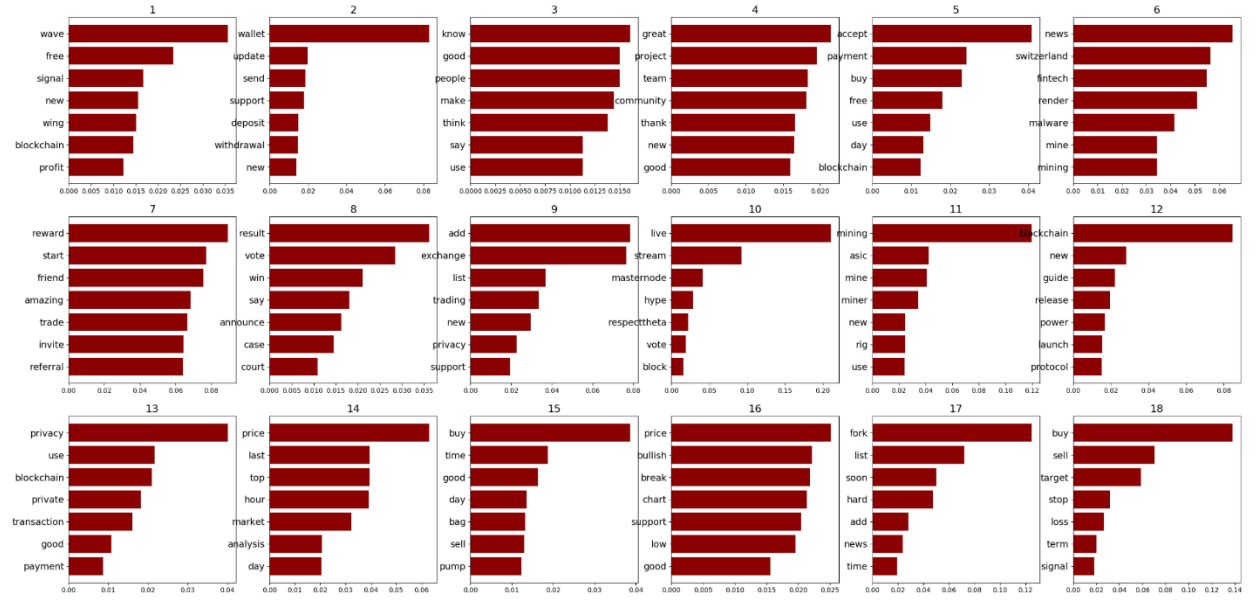


Figure 14: Topics for Privacy

Topic results not only show particularities of each group but also reveal some characteristics shared among them. The most noticeable one is the high quantity of topics presenting technical (“mining”, “block”, “network”, “platform”) and financial (“price, market, “bank”, “payment”) terms. Despite this being expected since the very nature of cryptocurrency discussion is a mixture of economics and technology, it is worth mentioning the data is reflecting the assumption. Which one, technical or financial, is more prevalent will be discussed later.

Another important feature is the presence of at least one topic that corresponds to the intended semantic categorization of each group. This is demonstrated in topic 15 for Smart Contracts and in topic 13 for Privacy, both having defining keywords in the top position: "smart" and "contract" for the former and "privacy" for the latter group. Although Faster transactions also has the topic 18 as one of this kind, the only word that refers to the group's intended meaning is "xrapid"[6], which appears in the 5th position and relates to only one cryptocoin, Ripple (XRP). Another candidate topic could have been the number 13, but it just contains the term "transaction" in 7th place and no other direct or similar reference to "faster". As such, it is not clear if the other cryptocoins aside from XRP can be considered as a part of the same group.

The term "blockchain" is the most popular among the three groups: it has the highest frequency of all words, appearing 21 times in total, 9 of which in the 1st position. This highlights a strong interest in this technology and it may indicate the pervasiveness of technical over financial discussions, yet to be confirmed by theme weight analysis. Some topics display very specific terms in the cryptocurrency environment, such as "ico"[7] or "erc"[8], as well as common expressions used for enthusiasts ("respecttheta", "flipnpik") but most probably unknown for the general public.

Themes

Aside from Technical and Financial, other themes can be inferred according to the prevailing semantic field among the topics, as follows:

- Group related: topics that relate directly to the group categorization;
- Emotional: topics having words expressing judgment or feeling, either positive or negative. Examples: "good", "great", "bad", "think", "dead";
- Marketing: topics having words relating to information broadcast. Examples: "live", "video", "broad", "stream", "promotion";
- Mixed: topics in which three or more of the aforementioned themes could be identified among the words.

[6] System that provides liquidity solution to banks using XRP platform to greatly reduce the cost of transactions and eliminate delays in payments worldwide [40].
[7] Initial Coin Offering, a fundraising activity for the development and launch of a new cryptocoin. It is the cryptocurrency equivalent of the Initial Public Offering (IPO) for stock exchanges.
[8] Ethereum Request for Comment or simply ERC-20, a smart contract protocol based on Ethereum blockchain [41].

If two prevalent themes are identified within a single topic, both are considered one double theme, such as Financial/Marketing, Technical/Emotional or Marketing/Emotional.

Final aggregation into themes are displayed in Tables 3, 4 and 5 for the three groups. Topics for Faster transactions and Smart Contracts were assembled into 8 themes while for Privacy the topics were assembled to 9 themes. Every group has one Group Related theme (highlighted in each table), as mentioned in the previous section, and also Financial, Technical and Mixed, in addition to 3 double themes for each. Faster transactions amasses half of its topics just in Financial and Technical, having 6 and 3 topics in those, respectively. Smart Contracts, in its turn, has an even higher concentration in the same themes, with 7 topics in Financial and 5 topics in Technical, leaving the rest of the remaining themes with just 1 topic for each. Finally, Privacy has a more diverse aggregation, having all the main themes in which topics are less concentrated than the other groups.

Faster transactions	
topic words	*theme*
blockchain, bank, use, payment, xrapid, new, world	group related
market, price, top, update, day, today, last chart, break, support, bullish, price, short, low buy, time, good, sell, year, think, market accept, buy, use, payment, wallet, send, transaction buy, long, target, term, short, sell, profit usd, price, change, live, min, increase, equivale	financial
security, blockchain, sec, mining, lawsuit, news, ceo fork, network, miner, block, new, mining, blockchain blockchain, nextpakk, wallet, project, ico, flipnpik, peer	technical
think, say, people, good, know, make, use thank, great, good, blockchain, new, community, news	emotional
hour, last, mention, day, course, invest, good raspberry, new, shirt, free, blue, casino, accept	mixed
exchange, add, blockchain, trading, new, launch, list free, start, earn, trade, blockchain, mining, bounty	financial/ technical
price, analysis, technical, usd, news, new, post	financial/ marketing
effect, broad, video, new, hope, dead, music	marketing/ emotional

Table 3: Themes for Faster transactions

Smart Contracts	
topic words	*theme*
contract, smart, blockchain, platform, base, use, erc	group related
support, chart, short, break, long, bullish, price exchange, classic, trading, news, add, list, trade last, hour, mention, price, market, top, year wallet, buy, blockchain, sale, exchange, use, ico bounty, blockchain, free, earn, referral, follow, friend price, analysis, market, usd, classic, news, technical buy, good, time, market, sell, make, day	financial
blockchain, base, platform, use, technology, build, network new, blockchain, classic, fork, release, proof, buy blockchain, founder, new, guide, fintech, developer, beginner blockchain, platform, base, use, world, payment, first mining, buy, rig, mine, gpu, free, miner	technical
sec, security, also, promotion, news, start, promo	marketing
blockchain, project, good, crowdsale, great, team, ico	mixed
network, fund, test, request, faucet, win, contest	financial/ technical
blockchain, think, make, use, people, good, real	technical/ emotional
blockchain, community, live, new, meetup, great, launch	technical/ marketing

Table 4: Themes for Smart Contracts

Privacy	
topic words	*theme*
privacy, use, blockchain, private, transaction, good, payment	group related
accept, payment, buy, free, use, day, blockchain price, last, top, hour, market, analysis, day price, bullish, break, chart, support, low, good buy, sell, target, stop, loss, term, signal	financial
news, switzerland, fintech, render, malware, mine, mining mining, asic, mine, miner, new, rig, use blockchain, new, guide, release, power, launch, protocol	technical
great, project, team, community, thank, new, good	emotional
live, stream, masternode, hype, respecttheta, vote, block	marketing
wave, free, signal, new, wing, blockchain, profit know, good, people, make, think, say, use result, vote, win, say, announce, case, court add, exchange, list, trading, new, privacy, support	mixed
reward, start, friend, amazing, trade, invite, referral buy, time, good, day, bag, sell, pump	financial/ emotional
wallet, update, send, support, deposit, withdrawal, new	financial/ technical
fork, list, soon, hard, add, news, time	technical/ marketing

Table 5: Themes for Privacy

Daily topic weights

Topic weights for every day from May 1, 2018, to August 30, 2018, are shown in the scatter plots for all groups in Figures 15, 16 and 17. Each colored dot depicts a single topic weight value for one document, meaning that every tweet has an exact 18 data points (1 per topic), that represent it, regardless of its value. Thus, all the charts have the same number of dots. Even with some of them looking emptier than others do, that just implies a high number of data points concentrated in one vertical region.

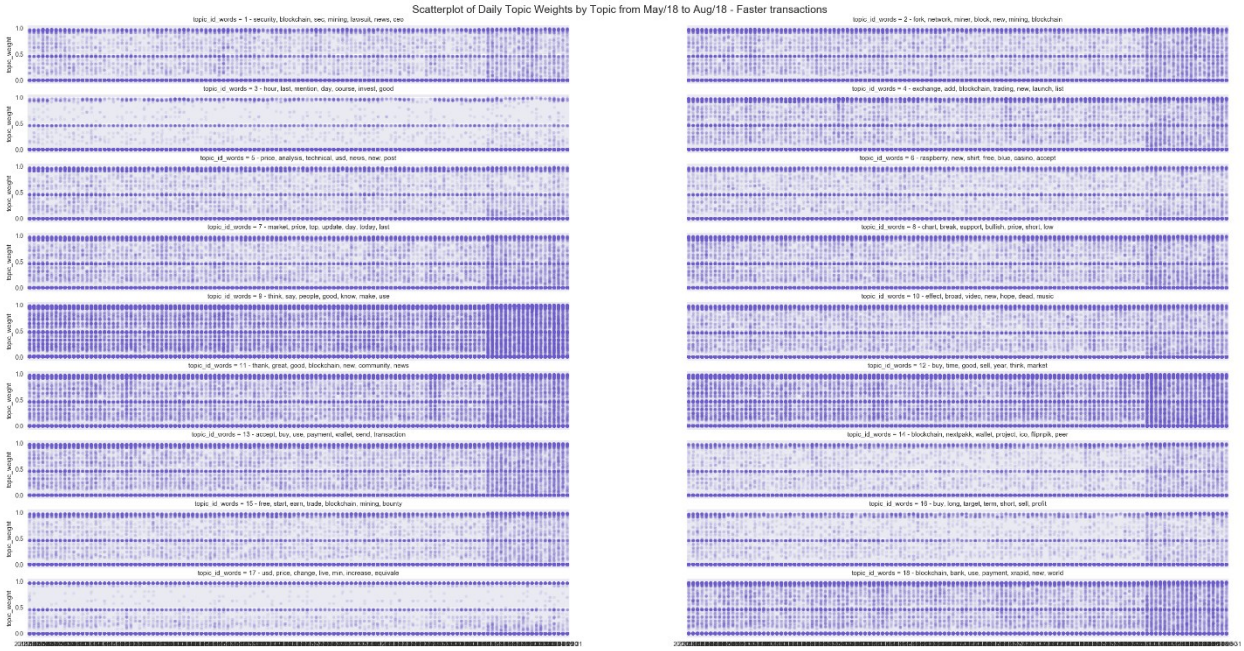


Figure 15: Scatterplot of daily topic weights by topic from May 2018 to August 2018 for Faster transactions

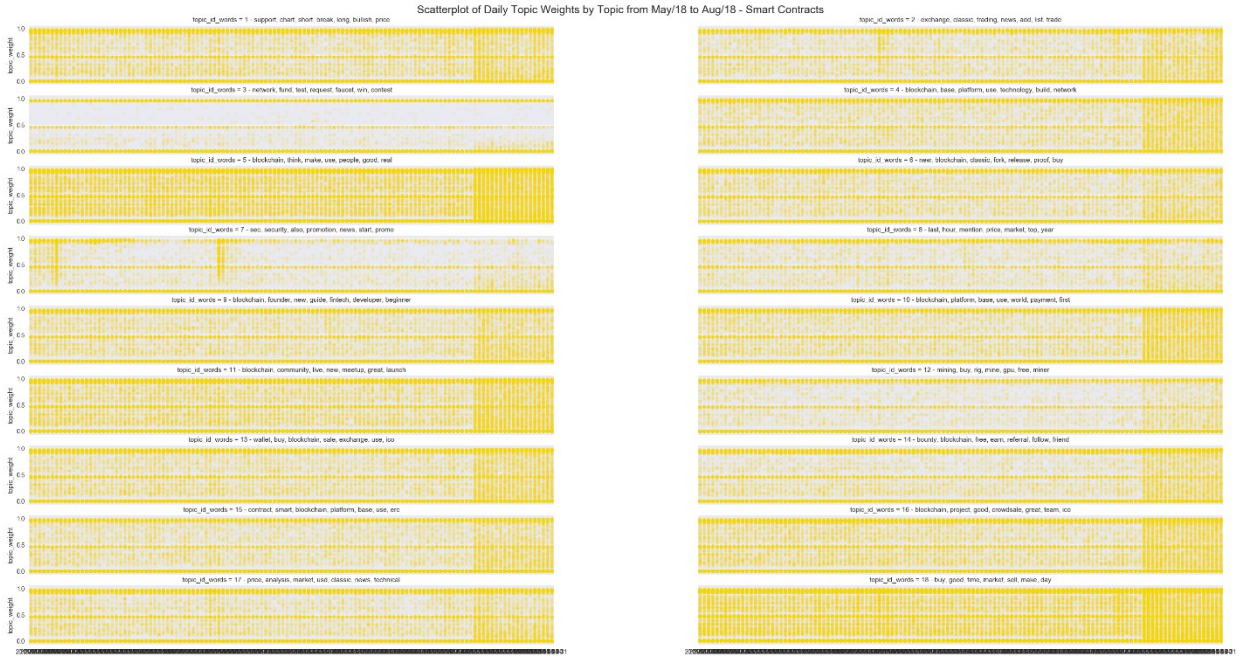


Figure 16: Scatterplot of daily topic weights by topic from May 2018 to August 2018 for Smart Contracts

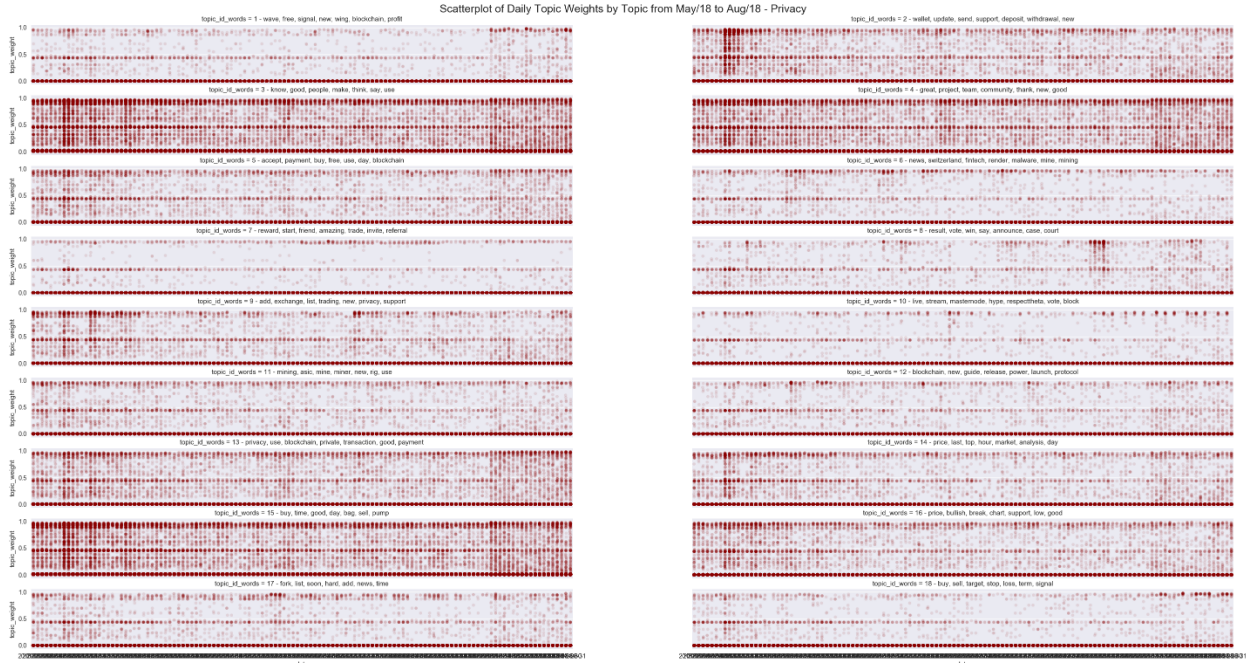


Figure 17: Scatterplot of daily topic weights by topic from May 2018 to August 2018 for Privacy

The three groups display a pattern of clustering data points near 0.0, 0.5 and 1.0 topic weight values in most of its charts. Another notable common feature is a higher concentration of

points in the last 20 days, present in the majority of the topics, although this merely reflects the sudden increase of the number of tweets retrieved during the final days of the data collection, whose causes remain unclear.

It is possible to observe that some topics present a better dispersion of points, instead of just concentrating them around 0.0, 0.5 and 1.0 values. That is the case for the topics number 9 and 12 for Faster transactions, 5 and 18 for Smart Contracts and 3, 4 and 15 for Privacy. There are different levels of point dispersion for the remaining topics but overall differences are subtle, making it difficult to draw any other conclusion. Even for the aforementioned topics with higher weight, making an accurate statement about which one is the most representative for the group is not feasible. Consequently, going forward with this analysis requires the calculation of average topic weights.

Average of daily topic weights

The daily average of topic weights depicted in Figures 18, 19 and 20 allow a global view of all the topics within a group and how they interact among themselves. A certain pattern is seen for the three groups: topics with higher average weight are usually very noisy, with significant variations in its value happening in short periods of time, while most of the topics with lower scores do not vary as much and are clustered in the same region of weight values. A few of the latter present a few significant upsurges, having a higher score for only one day or up to five days. Normally these occurrences would be characterized as outliers, but this denomination does not fit well in the context of the research. Instead, these sudden increases are more likely to illustrate events that happened for one or more cryptocoins, which triggered a discussion represented in the topic.

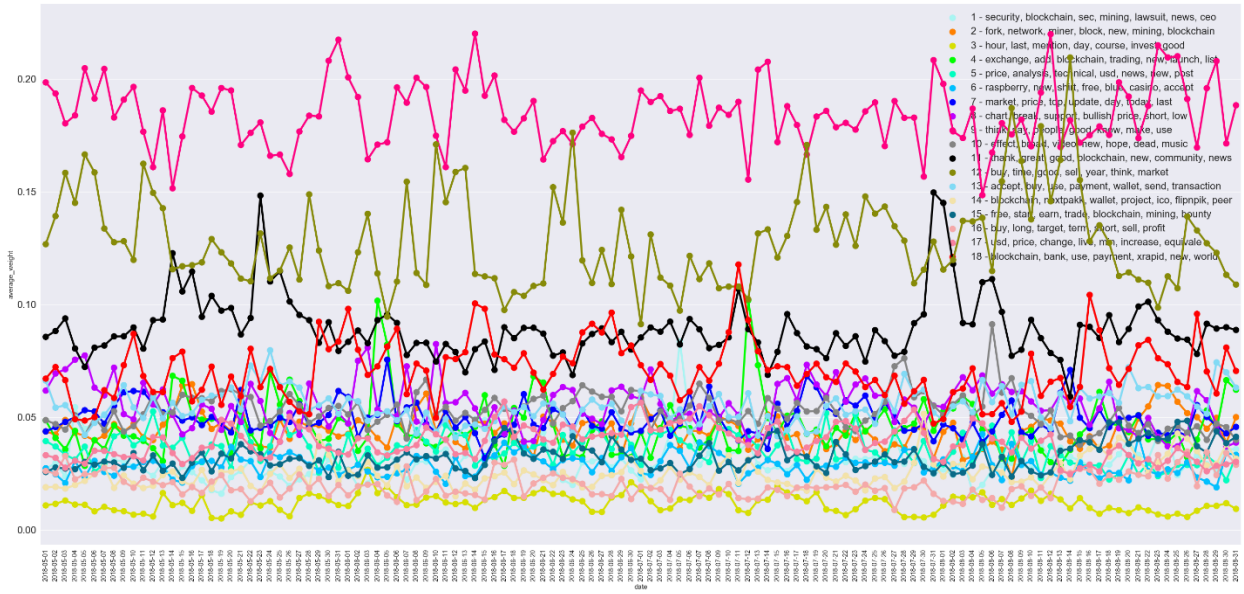


Figure 18: Daily average of weights per topic from May 2018 to August 2018 for Faster transactions

The overall strongest topic for Faster transactions is number 9, followed by topics 12 and 11. There were five major surges for topic 12 that made it be positioned very close to topic 9, even surpassing it some days. A similar situation happened with topics 12 and 11, with the latter settled below the former, except for the days it spiked. The rest of the topics are clustered between 0.00 and 0.10 weights, with occasional minor spikes for some of them.

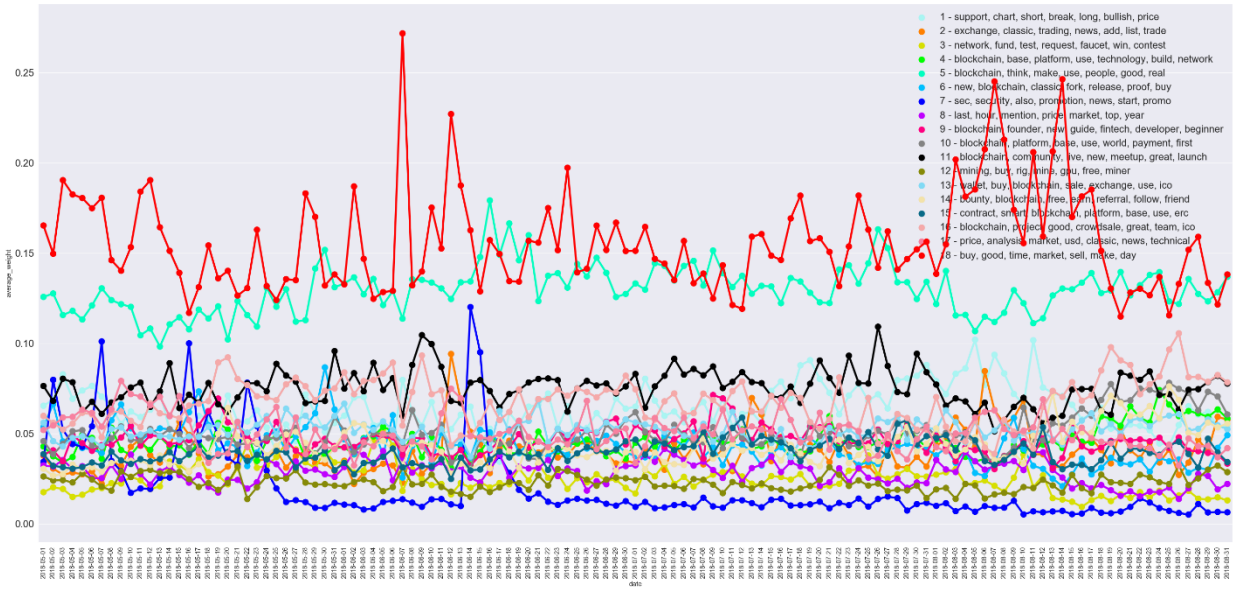


Figure 19: Daily average of weights per topic from May 2018 to August 2018 for Smart Contracts

The two strongest topics for Smart Contracts, numbers 18 and 5, stand apart from the topics clustered in the lower region of weights. Even having the strongest overall weight, topic 18 presents considerable spikes throughout the entire timeframe (the major one happened in June 7), causing to exchange the 1st position multiple times with topic 5. During the first half of August, topic 18 has a significant increase in its weight, isolating itself in the first position. In contrast with having the lowest general weight of the group, topic 5 shows a large spike in June 14, positioning it very close with the top two topics.

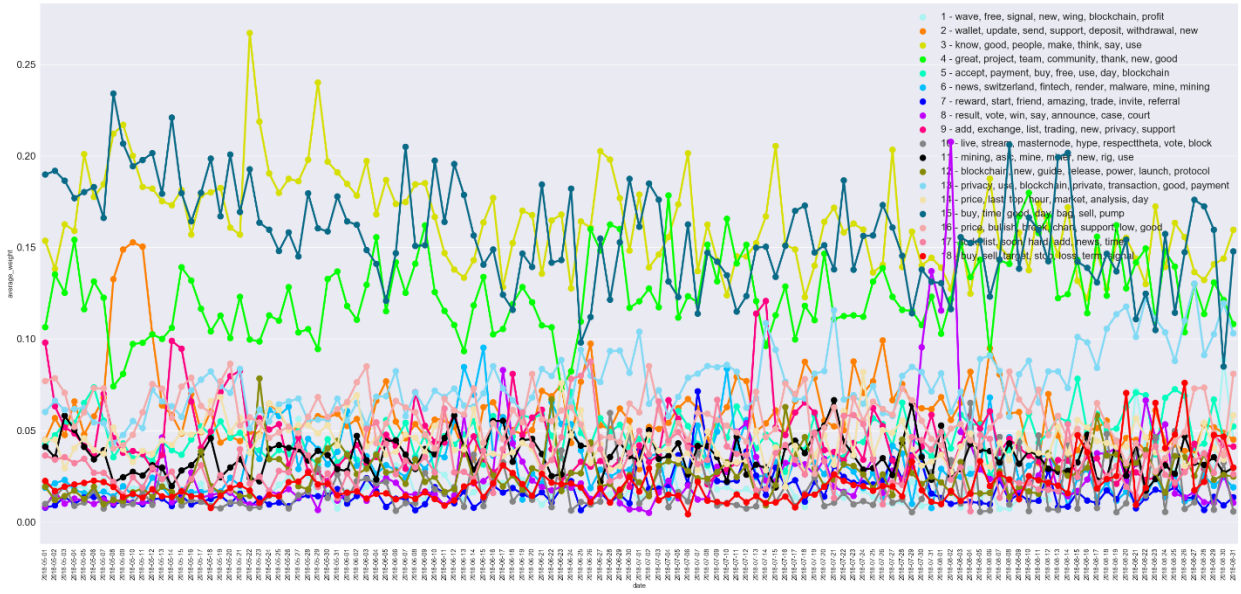


Figure 20: Daily average of weights per topic from May 2018 to August 2018 for Privacy

Privacy was considered the more diverse group regarding theme distribution and this is also reflected in the average topic weights. There is not one clear topic in the top position, given that topics with the highest overall weight, numbers 3 and 15, present intense variation, both in quantity and weight range, which results in a constant interchange of their positions throughout the four months. In addition, topic 4 reduces the gap with the top-positioned topics at the beginning of June and, excluding sporadic score decreases, could sustain its position for the remaining months, even achieving the highest weight value for the day in a few occasions. A similar situation happened with topic 13: it was positioned in the cluster region of lower weight topics for most of the time until it started an uptrend in the second half of August and placed itself close to the other top 3 topics during the final days of the timeframe. Lastly, two huge spikes are noted for two topics with low scores, the first from May 08 to May 11 for topic 2 and the second between July 31 and August 02 for topic 8.

Of note, topics that are included in the Group Related theme were chosen to be representative of the thesis author's intended semantic meaning for the cryptocurrency group. Therefore, it is interesting to check their weights before moving to the analysis of themes in order to get a clearer view without the influence of aggregations. In other words, if this particular topic does not have high weight when compared to other individual topics, it surely will also have low

weight when compared to other themes, since it is an individual topic and the rest may be multiple topics assembled together. Conversely, if this topic has a high weight, its relative rank among themes depends directly on how the other topics were aggregated.

Overall, these topics have neither achieved significant weight values to be amidst higher positions nor having the lowest weights of the group. The highest weight can be seen for Privacy, of which the topic 13, as aforementioned, ended higher weighted due to a weight growth during the last days. Likewise, topic 18 for Faster transactions reached the 4th place among higher weights, but there is a considerable gap between it and the top 2 positions. It is in the higher area of the low weight topic cluster, moving away from it at least five times. Ultimately, topic 15 for Smart Contracts did not show higher weights, appearing in the middle region of the cluster, without any meaningful variation in its weight values.

Average of daily theme weights

The aggregated view in themes displayed in Figures 21, 22 and 23, as expected, reduced the overall noise seen in individual topics, in different levels across the groups. A few general features can be listed: contrary to the suggestion brought due to the high frequency of the term “blockchain”, no group had the Technical theme as the one with the highest weight values. In addition, themes with low weights are also clustered, at least for Faster transactions and Smart Contracts. Lastly, in line with the result of the individual topic analysis, the Group Related theme did not have high weight in any of the groups.

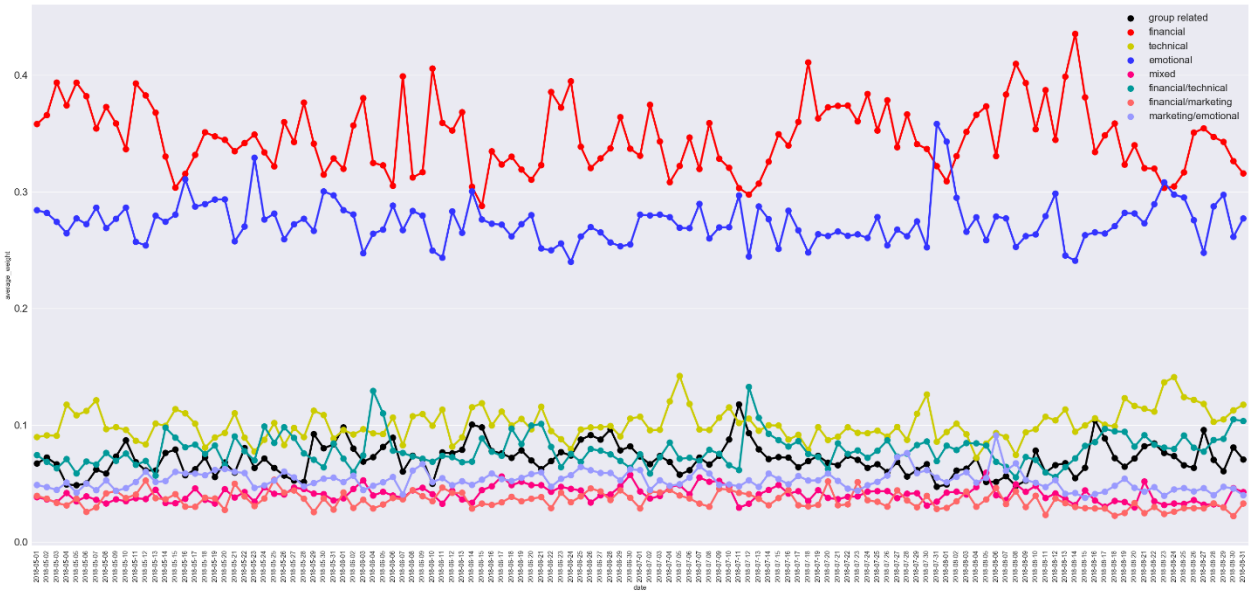


Figure 21: Daily average of weights per theme from May 2018 to August 2018 for Faster transactions

Financial was the most prominent theme for Faster transactions, followed by Emotional. While Financial keeps itself distant from Emotional for most of the time, the latter closes the gap on some occasions, even surpassing it for three days (July 31, August 01 and August 23 by a small margin). Both are isolated from Technical in 3rd place as well as from the rest of the themes with lower scores. It is not clear which one has the overall highest weight value after Technical considering that the weight differences amidst them are fairly small, and the constant exchange of positions gives an impression they are "blended" together. Consequently, none of them seems more relevant than others.

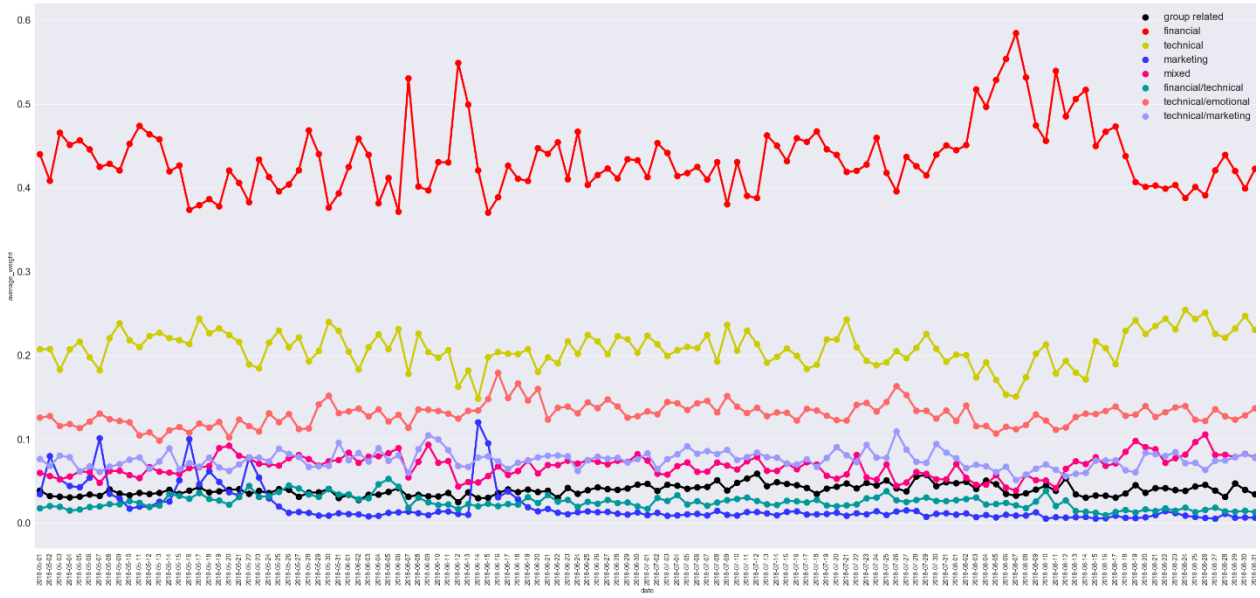


Figure 22: Daily average of weights per theme from May 2018 to August 2018 for Smart Contracts

Financial appears as the most prominent theme for Smart Contracts, with a significant distance between it and Technical, which is in 2nd place. It is relevant to point that, out of the three double themes (Financial/Technical, Technical/Emotional, and Technical/Marketing) in the group, Financial is in just one and Technical is in all of them. If the double themes were incorporated to the respective single themes their weight would be added to the weight of the respective single theme, and the space between 1st and 2nd place would be narrowed. Moreover, the appearance of Financial/Technical in 3rd place shows how the two themes stand out in this group. Score differences for the other themes are clearer than the ones in Faster transactions: they are less intertwined and have less daily variation. The theme with the general lowest score, Marketing, presents some spikes before mid-June that approximately place it to the 3rd place in terms of overall weight.

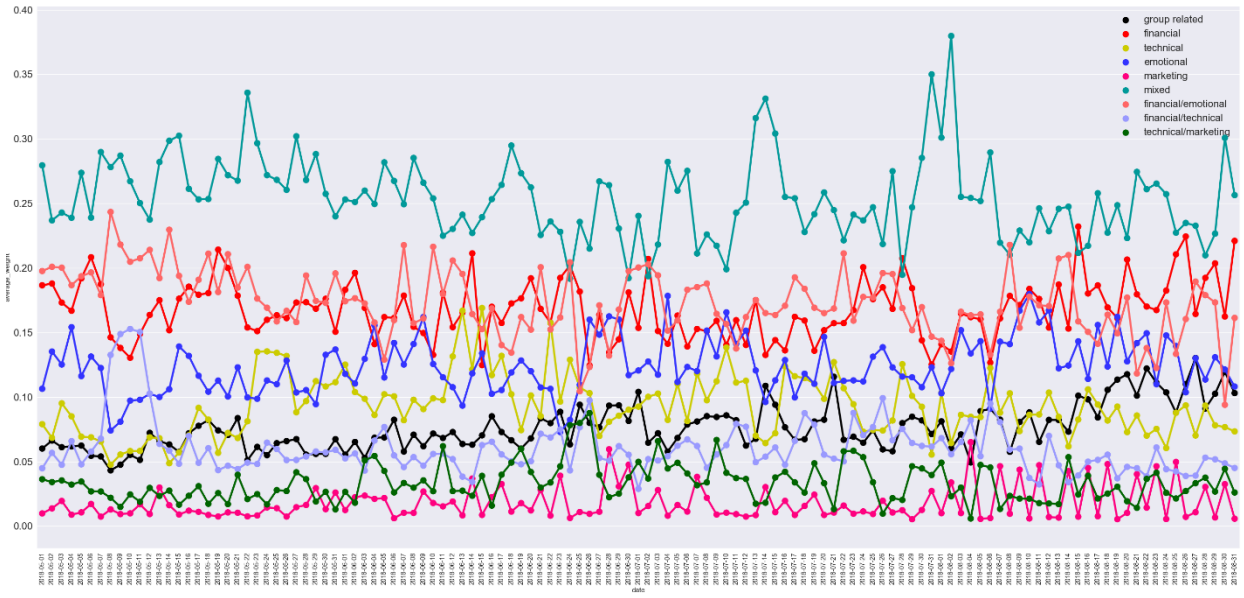


Figure 23: Daily average of weights per theme from May 2018 to August 2018 for Privacy

Privacy is the only group that does not have a cluster of themes with a low score, having more visible distances between each one. The theme with the highest weight is Mixed, followed by Financial and Financial/Emotional, both in 2nd place. If these two were merged, the 1st position would probably switch over Financial and Mixed. Emotional is the next highest-weighted theme, followed closely by Technical and Group Related. In general, the aggregation into themes did not reduce the noise as much as happened with the other groups. Nevertheless, the diversity of themes in this cryptocurrency group is reflected in a more uniform distribution of the weight values.

Average theme weight forecasting

The outcome of ARIMA prediction for all themes is shown in Figures 24, 25 and 26 for each group, along with Tables 6, 7 and 8 displaying a comparison of MSE values between ARIMA and Naive methods in order to evaluate the predictive ability of the former model. A high ratio of ARIMA MSE compared to Naive MSE (shown here as a percentage of the Naive MSE value) indicates a low level of predictability of the data, meaning that the ARIMA predicted curve may resemble one generated by Naive model. Aside from these cases, forecasting events that ended up creating spikes in the test data have proven itself very difficult for the model. Accordingly, many of the predicted curves do not show them.

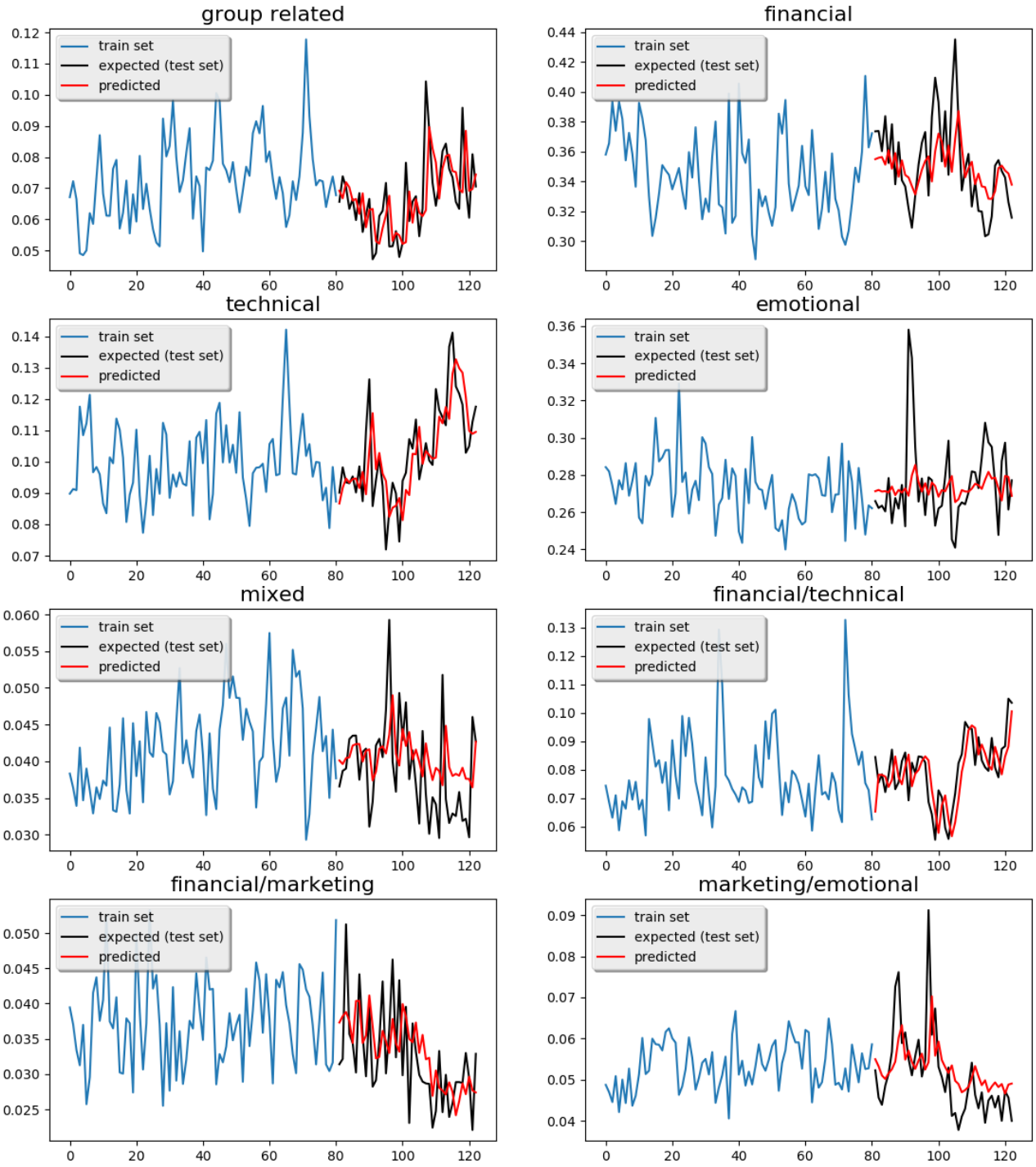


Figure 24: Predicted theme weight values using ARIMA for Faster transactions

Faster transactions				
Themes	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
Group Related	0.000185	(4,1,0)	0.000133	72%
Financial	0.000733	(1,0,0)	0.000627	86%
Technical	0.000162	(2,1,0)	0.000133	82%
Emotional	0.000650	(0,0,1)	0.000490	75%
Mixed	0.000064	(1,0,0)	0.000043	68%
Financial/Technical	0.000086	(1,1,0)	0.000079	92%
Financial/Marketing	0.000084	(6,1,0)	0.000032	38%
Marketing/Emotional	0.000095	(2,0,0)	0.000081	85%

Table 6: MSE comparison between Naive and ARIMA predictions for Faster transactions

Results for Faster transactions show that Financial/Technical, having the highest MSE ratio of the group, was the theme where ARIMA prediction is closest in performance to Naive. Both Emotional and Marketing/Emotional predictions did not present the large spikes seen in the test data, thus failing in foresee these events. Generally, the forecast for the remaining themes presented the same pattern of the test set but with fewer variations.

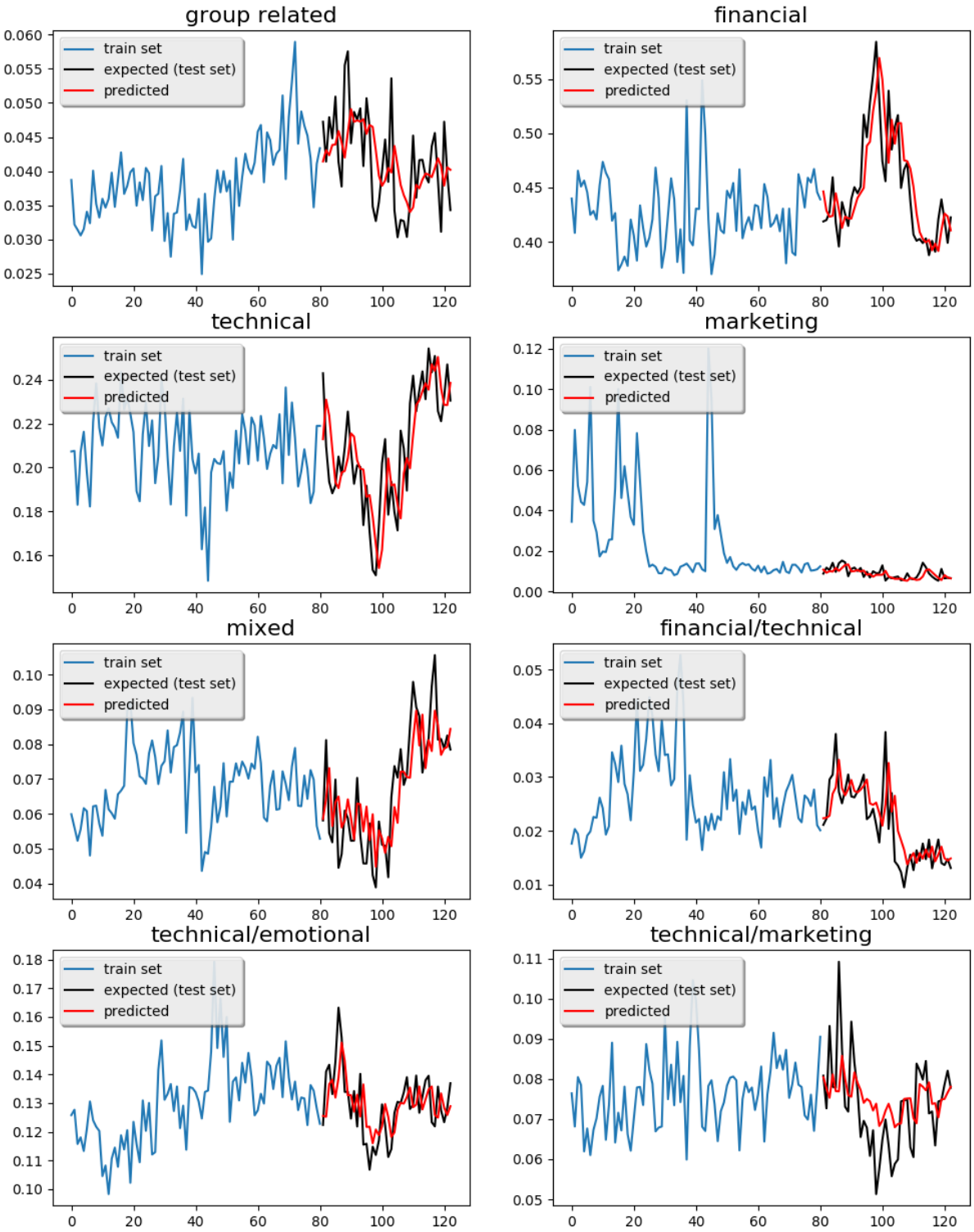


Figure 25: Predicted theme weight values using ARIMA for Smart Contracts

Smart Contracts				
Themes	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
Group Related	0.000060	(1,0,1)	0.000042	70%
Financial	0.000975	(2,1,0)	0.000947	97%
Technical	0.000356	(2,1,0)	0.000352	99%
Marketing	0.000009	(0,1,1)	0.000007	78%
Mixed	0.000154	(8,0,1)	0.000104	68%
Financial/Technical	0.000030	(1,1,1)	0.000025	83%
Technical/Emotional	0.000108	(1,0,0)	0.000087	80%
Technical/Marketing	0.000124	(1,0,0)	0.000100	81%

Table 7: MSE comparison between Naive and ARIMA predictions for Smart Contracts

Financial and Technical were the two themes with the lowest predictability in Smart Contracts, having almost the same MSE values for ARIMA and Naive. The forecast for remaining themes follows the same behavior described for the previous group: a similar pattern of the test set with fewer variations.

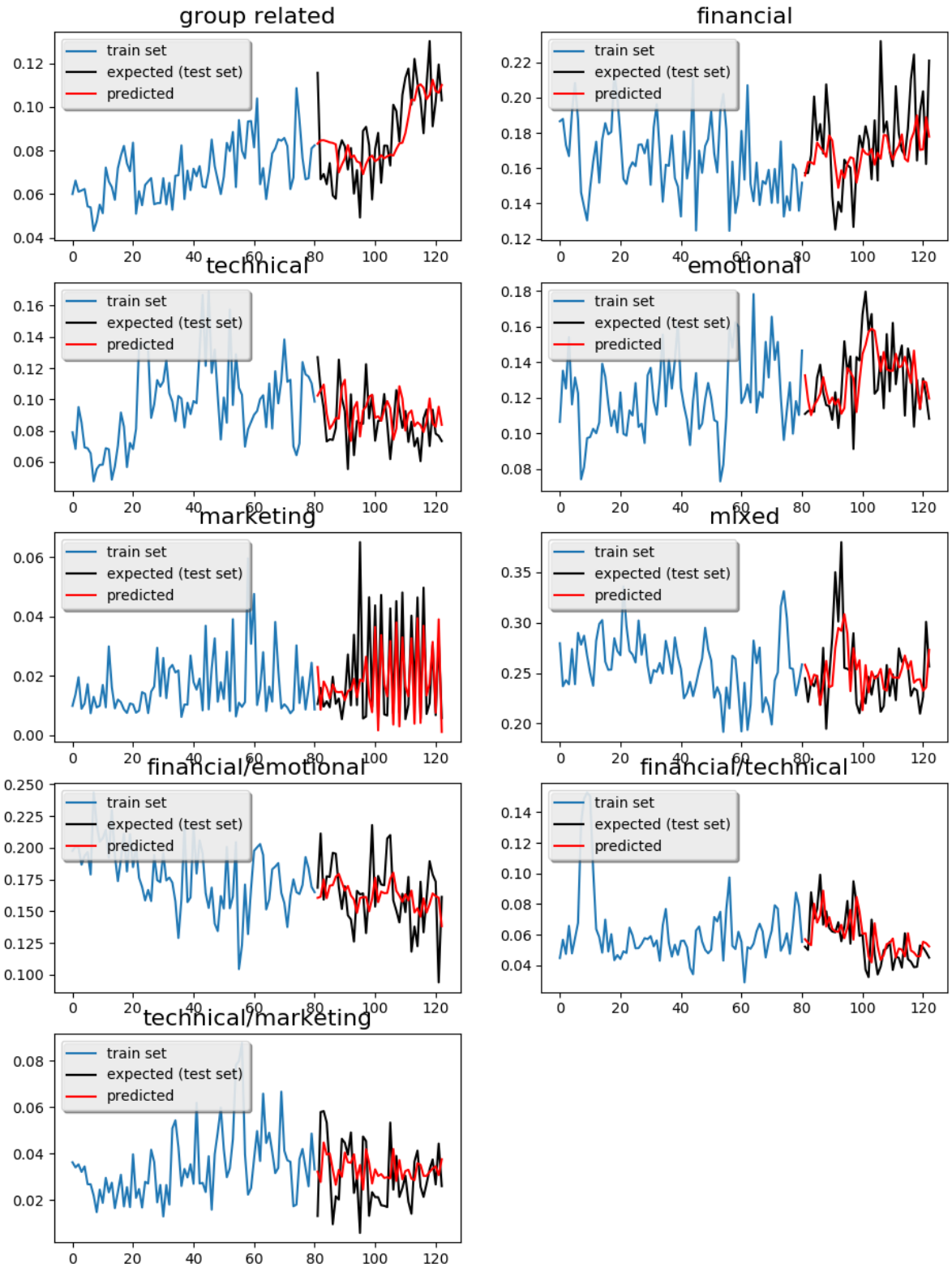


Figure 26: Predicted theme weight values using ARIMA for Privacy

Privacy				
Themes	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
Group Related	0.000360	(0,1,1)	0.000258	72%
Financial	0.000861	(6,0,0)	0.000576	67%
Technical	0.000470	(0,0,2)	0.000304	65%
Emotional	0.000619	(4,1,0)	0.000363	59%
Marketing	0.000874	(8,0,0)	0.000147	17%
Mixed	0.001655	(2,0,1)	0.001114	67%
Financial/Emotional	0.001070	(1,1,1)	0.000680	64%
Financial/Technical	0.000273	(1,0,0)	0.000220	81%
Technical/Marketing	0.000311	(0,0,1)	0.000194	63%

Table 8: MSE comparison between Naive and ARIMA predictions for Privacy

It is possible to affirm that Privacy themes had the best predictive ability of all groups due to the overall lower MSE ratio. Marketing, in particular, has the lowest one of the entire set of themes analyzed in this thesis. Its predicted curve is very close to the shape of the observed time series in the test set for the same point in time. In contrast, Financial/Technical is the closest prediction to Naive in terms of performance. Predicted curves for the rest of the themes follow the test set but are significantly smoother than the observed data in general.

Finally, real and predicted theme weight values were put together in Figures 27, 28 and 29 for the three groups. Both Faster transactions and Smart Contracts show predicted data very similar to the actual data for the last days of the timeframe, with the exception of Emotional in Faster transactions, having a more linear predicted weight curve. Each theme stayed in its respective positions, which may be explained by the low predictive ability of the themes for the two groups. On the other hand, forecast data for Privacy keeps the overall dynamic and it is slightly less noisy than the actual data. There are fewer intersections amidst themes and it is possible to see competition between pairs, namely Financial and Financial/Emotional, Technical and Group Related, along with Marketing and Technical/Marketing.

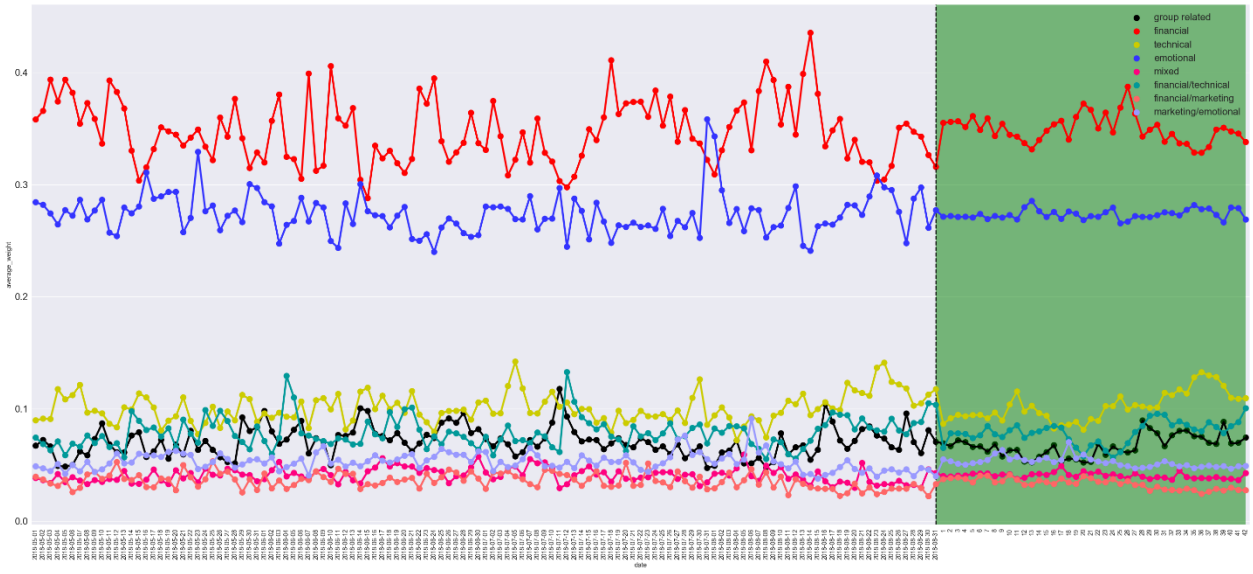


Figure 27: Average of theme weights per theme with forecast (ARIMA) shown in the green-colored area for Faster transactions

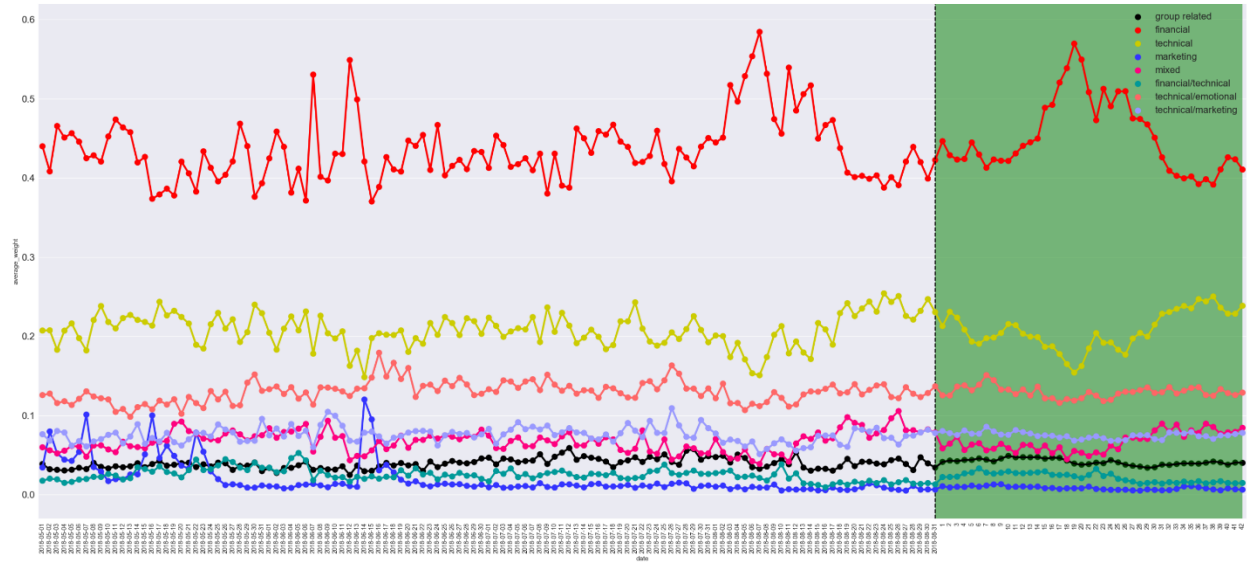


Figure 28: Average of theme weights per theme with forecast (ARIMA) shown in the green-colored area for Smart Contracts

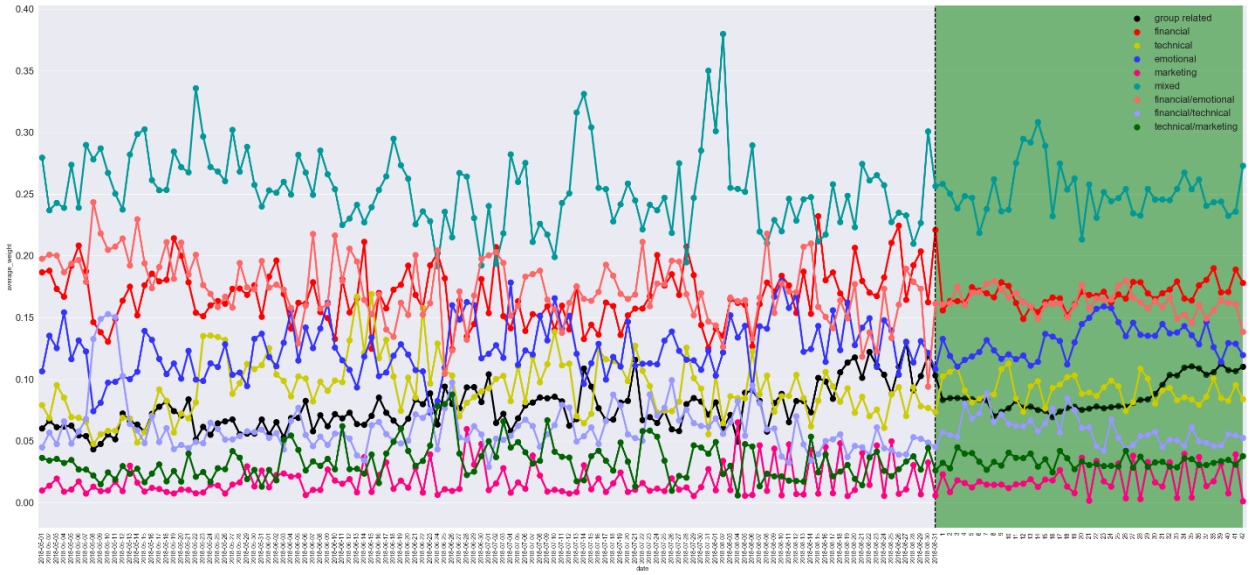


Figure 29: Average of theme weights per theme with forecast (ARIMA) shown in the green-colored area for Privacy

Experiments

Plotting an hourly view of the average topic weights did not prove itself useful. The number of data points per hour was heavily increased compared to daily amounts, creating noisy curves instead of well-defined lines, making it very difficult to see single values and which timestamp it belongs to. Besides, of the 18 topics, only 3 to 5 were visible depending on the cryptocurrency group, while the others are hidden behind those. The same situation holds after the topics were aggregated into themes. Due to the difficulty of analyzing such noisy data, forecasting did not proceed for the hourly view.

There were more interesting results regarding the modeling with 8 topics instead of the previously used amount, 18 topics. With 8 topics, most of the top 7 words for each topic were the same from the previous modeling, with some positions shuffled, both within a topic and with others in the same group. New additions were quite a few. The most noticeable for Faster transactions was the absence of the topic representing Group Related, again exhibiting the problematic handling of the categorization for this group. Aside from this, the themes represented by the new topics were identical to the old ones although the equivalency was not perfect, with the Mixed and Financial/Marketing themes missing. Contrarily, Group Related was present for the other groups. Both Smart Contracts and Privacy had all the themes from the previous modeling and the same

number of absences. Technical, Marketing and Technical/Emotional were not present for Smart Contracts while Marketing, Financial/Technical and Technical/Marketing were missing for Privacy. The lack of some themes evidences the risk of getting very broad topics using a low *k* for the LDA modeling. Even so, the major themes are still shown, making these results a good representation of the general discussion.

Faster transactions	
topic words	*Theme*
buy, good, sell, time, market, long, price	Financial
price, analysis, market, hour, last, chart, usd	Financial
usd, price, change, live, min, increase, analysis	Financial
use, good, people, say, make, think, know	Emotional
blockchain, new, news, security, mining, miner, post	Technical
blockchain, new, exchange, payment, use, add, bank	Financial/Technical
free, blockchain, bounty, earn, start, trade, mining	Financial/Technical
effect, broad, thank, new, love, good, video	Marketing/Emotional

Table 9: 8-topic representation of Faster transactions

Smart Contracts	
topic words	*Theme*
blockchain, platform, contract, smart, base, use, build	group related
buy, good, think, time, make, market, sell	Financial
price, analysis, market, usd, last, chart, hour	Financial
bounty, blockchain, free, start, earn, follow, day	Financial
blockchain, project, crowdsale, good, great, team, ico	Mixed
blockchain, exchange, classic, mining, new, news, trading	Mixed
network, fund, test, faucet, request, jpy, less	Financial/Technical
blockchain, new, security, classic, founder, community, news	Technical/Marketing

Table 10: 8-topic representation of Smart Contracts

Privacy	
topic words	*Theme*
privacy, blockchain, use, good, project, private, transaction	group related
price, live, buy, hour, last, target, market	Financial
mining, mine, miner, use, new, asic, malware	Technical
wallet, new, fork, add, support, update, soon	Technical
buy, good, make, time, worth, think, know	Emotional
wave, top, big, masternode, hot, day, list	Mixed
news, exchange, add, fintech, switzerland, render, blockchain	Mixed
trade, start, reward, friend, free, amazing, referral	Financial/Emotional

Table 11: 8-topic representation of Privacy

Results for average topic weights also match the previous findings for Faster transactions and Smart Contracts: the best-positioned topics represented practically the same themes as the best-positioned topics of the previous modeling, with a minor discrepancy for the latter (Technical/Emotional before, now as Technical/Marketing) due to a slightly different word arrangement. Privacy shows more notable differences, now having a clear topic in the first position and distant from the remaining ones, while there were two topics competing for dominance before. The best-positioned one somehow resembles one of these two topics, containing three of its top words and representing the Emotional theme (the previous topic that it resembles was Financial/Emotional).

Considering that the objective was not aggregation and simply to compare the new topics directly with the themes, the prediction task proceeded topic-wise. Once again, no considerable differences from previous results were seen: predictive ability measured by comparing MSE scores from ARIMA and Naive were similar, again with the best prediction performance within the Privacy group, and overall topic positions remained unaltered for the most part.

In the end, the 8-topic results mostly support the claims inferred from the 18-topic modeling. All figures from the latest experiments are presented in the Appendix section.

6. CONCLUSION AND FUTURE WORK

This thesis performed analysis on text data from Twitter posts that referred to a total of 25 cryptocurrencies for 4 months. After separating these cryptocoins into three groups (Faster transactions, Smart Contracts, and Privacy), the main objective was to verify whether the group categories were adequate, which were the significant themes discussed and forecast which one of these themes were most likely to be the subject of future discussions.

According to the collected topics, the selection of the cryptocoins into each one of the groups seems accurate. There is little or no interference of features associated with a particular group into another one, as well as the lack of any other major aspect that could determine a fourth group. However, Faster transactions is the only case that requires more attention, due to having its categorization defined by just one coin, XRP. It is vague how much the remaining eight coins have contributed to the overall meaning of the group or even if they are indeed a good fit.

None of the Group Related themes were the most important ones for their groups, suggesting that the common aspects that designated each group placed a secondary role in the discussion. The expected clash between technical and financial features resulted in a clear dominance of the second one. Financial was the most important theme for both Faster transactions and Smart Contracts, especially for the latter, being isolated from the other themes. Likewise, the two groups had a second strongest theme (Emotional for Faster transactions and Technical for Smart Contracts) while all the others remained in lower positions. Their importance is then diminished regardless of the differences in the final scores. Privacy exhibited slightly different behavior: although Mixed had the highest weight, followed by Financial and Financial/Emotional, weight differences were smaller and there was more competition among the themes.

Forecasting theme weights did not prove itself meaningful for Faster transactions and Smart Contracts. The final configuration of both groups along with the low predictive ability of its themes resulted in future weights that merely repeated a pattern seen in the final days of the research, also not being able to detect events that were the cause of spikes. Privacy once again sets apart from the first two groups with a modestly improved forecast, still displaying predicted curves with a similar pattern from the actual data but smoothing it and pointing competition for dominance between pairs of themes.

The research done for this thesis can serve as a foundation to further studies in cryptocurrency. For example, using the same methods but expanding the time for data collection to a full year and beyond. Hence, it would be possible to answer some questions that could not be addressed in a short period of four months: does the data present some sort of seasonal behavior? How much do the themes differ for the same month in different years? Furthermore, similar research could be conducted just by using another source, such as Reddit or other dedicated online message boards. Despite narrowing the audience in comparison to Twitter, this would greatly minimize the effect of bots while at the same time it would offer the perspective of only real enthusiasts of the subject. Comparing the results of both platforms for the same timeframe would be of substantial relevance.

This application of topic modeling could be associated to other aspects in cryptocurrency similarly to what Linton et al. conducted in their study, analyzing comments on message boards and comparing the results with indicators to fraudulent schemes; for instance, investigating variations in topic weights related to financial features with fluctuations of prices or overall volume of transactions for a group of cryptocoins, or even understanding the correlation between topics that have a technical meaning and the announcements of new technologies.

In the end, future work could also shift the focus to individual cryptocoins instead of groups. By associating each one with a set of topics, their separate roles would be evident, avoiding the situation aforementioned for Faster transactions, where the contributions done by every other cryptocoin excluding XRP were left unclear. Besides, sentiment analysis could be used to further investigate topics expressing emotional opinions and detect which coin generates more positive or negative feelings.

REFERENCES

[1] CoinMarketCap. [Online]. Available: https://coinmarketcap.com/currencies/bitcoin/. Accessed: February 10, 2019.

[2] IEEE. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

[3] Lee, H. J., Choi, M. S., & Rhee, C. S. (2003, October). Traceability of double spending in secure electronic cash system. In *2003 International Conference on Computer Networks and Mobile Computing, 2003. ICCNMC 2003.* (pp. 330-333). IEEE.

[4] CoinMarketCap. [Online]. Available: https://coinmarketcap.com/all/views/all/. Accessed: March 29, 2018.

[5] Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday*, *2*(9).

[6] "What exactly is Turing Completeness?". [Online]. Available: https://medium.com/@evinsellin/what-exactly-is-turing-completeness-a08cc36b26e2. Accessed: July 29, 2018.

[7] "The Ultimate Guide to Understanding Smart Contracts ". [Online]. Available: https://www.blockchaintechnologies.com/smart-contracts/. Accessed: May 07, 2019.

[8] "How does Monero's privacy work?". [Online]. Available: https://www.monero.how/how-does-monero-privacy-work. Accessed: July 29, 2018.

[9] "Privacy Considerations for Official Zcash Software & Third-Party Wallets". [Online]. Available: https://z.cash/support/security/privacy-security-recommendations.html. Accessed: July 30, 2018.

[10] "How cryptocurrency Verge (XVG) pushes privacy and anonymity to the next level". [Online]. Available: https://medium.com/@MartinRosulek/how-cryptocurrency-verge-pushes-privacy-and-anonymity-to-the-next-level-4b367e16a8a4. Accessed: July 30, 2018

[11] Zhao, Y. (2017). Bitcoin Zero.

[12] "10 things you need to know about Ripple". [Online]. Available: https://www.coindesk.com/10-things-you-need-to-know-about-ripple. Accessed: August 01, 2018

[13] Stellar. [Online]. Available: https://www.stellar.org. Accessed: August 01, 2018.

[14] "What is the Difference Between Litecoin and Bitcoin?". [Online]. Available: https://www.coindesk.com/information/comparing-litecoin-bitcoin. Accessed: August 01, 2018.

[15] “Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2018 (in millions)”. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ Accessed: September 13, 2018.

[16] Lebanon, G., Mao, Y., & Dillon, J. (2007). The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, *8*(Oct), 2405-2441.

[17] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[18] Darling, W. M. (2011, December). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*(pp. 642-647).

[19] Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

[20] Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012, July). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952-961). Association for Computational Linguistics.

[21] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262-272). Association for Computational Linguistics.

[22] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408). ACM.

[23] Syed, S., & Spruit, M. (2017, October). Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 165-174). IEEE.

[24] Brockwell, P. J., Davis, R. A., & Calder, M. V. (2002). *Introduction to time series and forecasting* (Vol. 2). New York: springer.

[25] Gujarati, D. N. (2009). Basic econometrics. Tata McGraw-Hill Education.

[26] Greene, W. H. (2003). Econometric analysis. Pearson Education India.

[27] Blau, B. M. (2017). Price dynamics and speculative trading in bitcoin. *Research in International Business and Finance*, *41*, 493-499.

[28] Jang, H., & Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access*, *6*, 5427-5437.

[29] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003, November). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE international conference on data mining* (pp. 427-434). IEEE.

[30] Matta, M., Lunesu, I., & Marchesi, M. (2015, June). Bitcoin Spread Prediction Using Social and Web Search Media. In *UMAP Workshops* (pp. 1-10).

[31] Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis.

[32] Laskowski, M., & Kim, H. M. (2016, July). Rapid Prototyping of a Text Mining Application for Cryptocurrency Market Intelligence. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)* (pp. 448-453). IEEE.

[33] Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, *11*(8), e0161197.

[34] Linton, M., Teo, E. G. S., Bommes, E., Chen, C. Y., & Härdle, W. K. (2017). Dynamic topic modelling for cryptocurrency community forums. In *Applied Quantitative Finance* (pp. 355-372). Springer, Berlin, Heidelberg.

[35] "19 Bitcoin Accounts You Should Follow on Twitter". [Online]. Available: http://fortune.com/2017/12/27/bitcoin-twitter/. Accessed: August 19, 2018.

[36] "The 100 Most Influential People in Crypto - 2018 Edition -". [Online]. Available: https://cryptoweekly.co/100/. Accessed: August 19, 2018.

[37] "Top 50 Blockchain and Crypto Twitter Accounts to Follow". [Online]. Available: https://medium.com/@Uttoken.io/top-50-blockchain-and-crypto-twitter-accounts-to-follow-3a343948d176. Accessed: August 19, 2018.

[38] "Bitcoin and cryptocurrency on Twitter: The most important people to follow". [Online]. Available: https://www.marketwatch.com/story/bitcoin-and-cryptocurrency-on-twitter-the-most-important-people-to-follow-2017-12-04. Accessed: August 19, 2018.

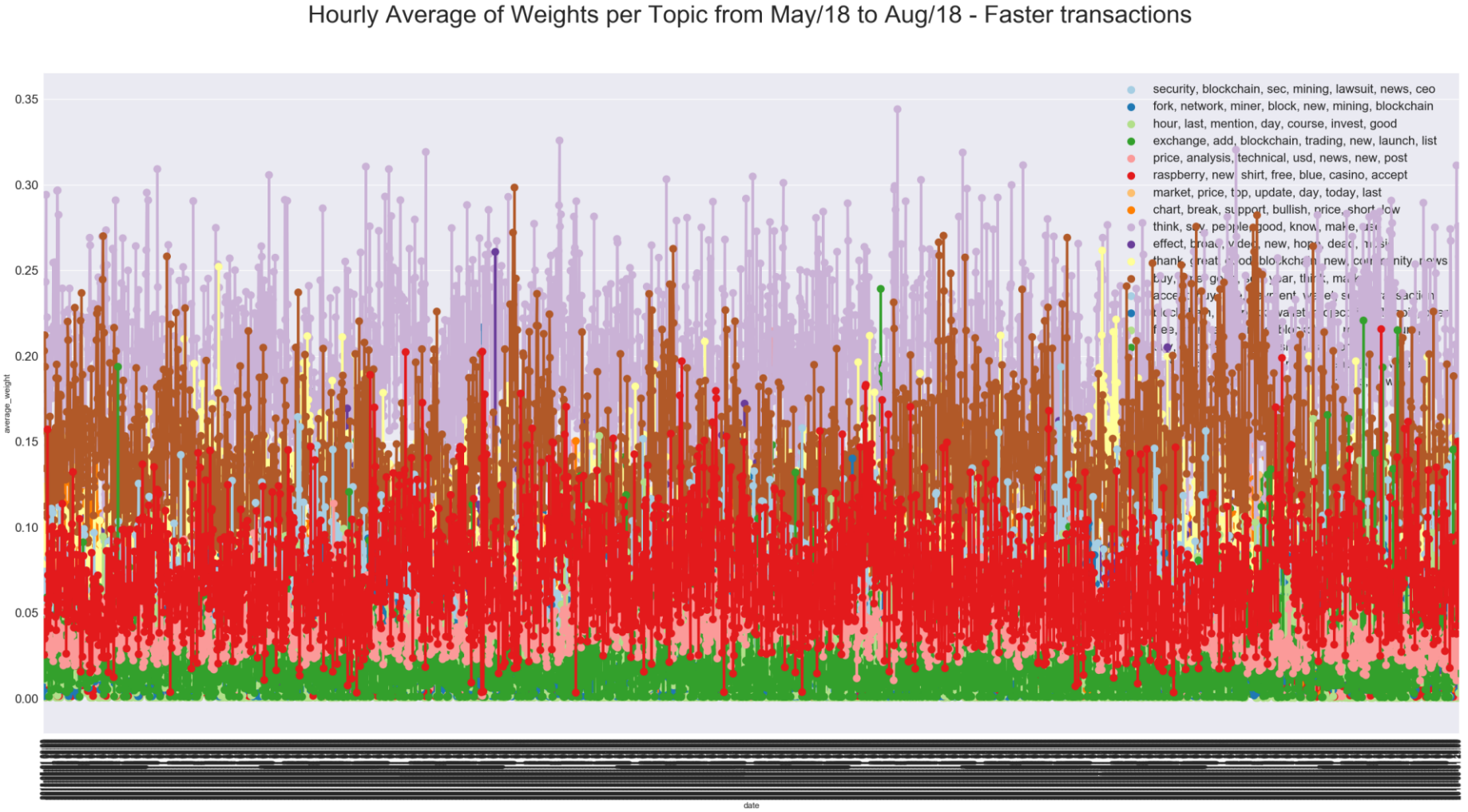
[39] "MALLET - MAchine Learning for LanguagE Toolkit". [Online]. Available: http://mallet.cs.umass.edu/. Accessed: November 23, 2018.

[40] "Ripple: Source Liquidity – xRapid". [Online]. Available: https://ripple.com/ripplenet/source-liquidity. Accessed: January 17, 2019.

[41] "ERC-20 and What Does it Mean for Ethereum?". [Online]. Available: https://www.investopedia.com/news/what-erc20-and-what-does-it-mean-ethereum/. Accessed: January 17, 2019.

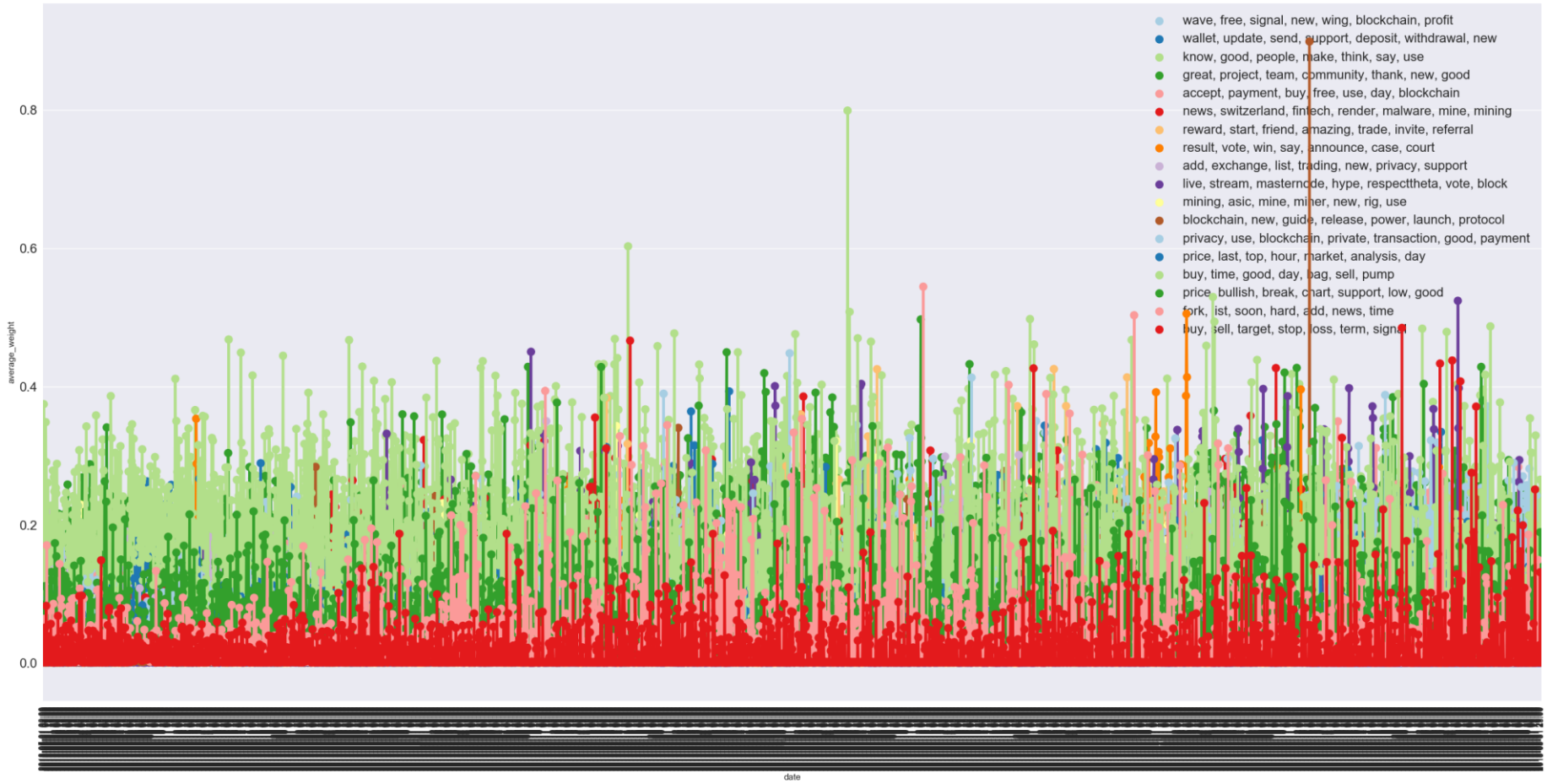
APPENDIX

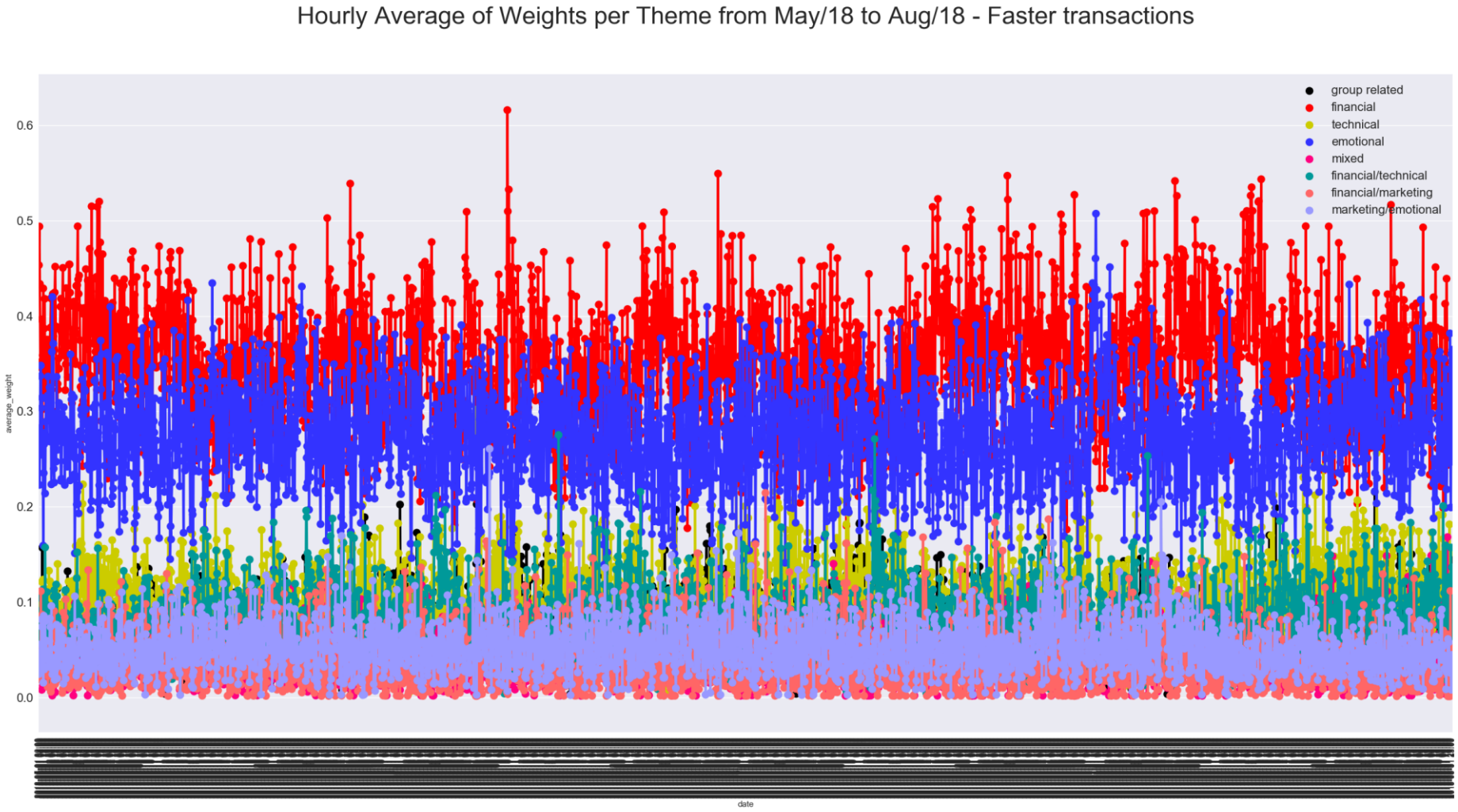
This section contains the results from the additional tests for the three groups. Pages 67 to 72 show the charts for topic weights and theme weights from the hourly modeling. Pages 73 to 87 concern the results from the modeling for 8 topics. They are presented in a logical way, according to their counterparts described in Chapter 5: obtained topics, daily topic weights, average of daily topic weights and forecasting results.

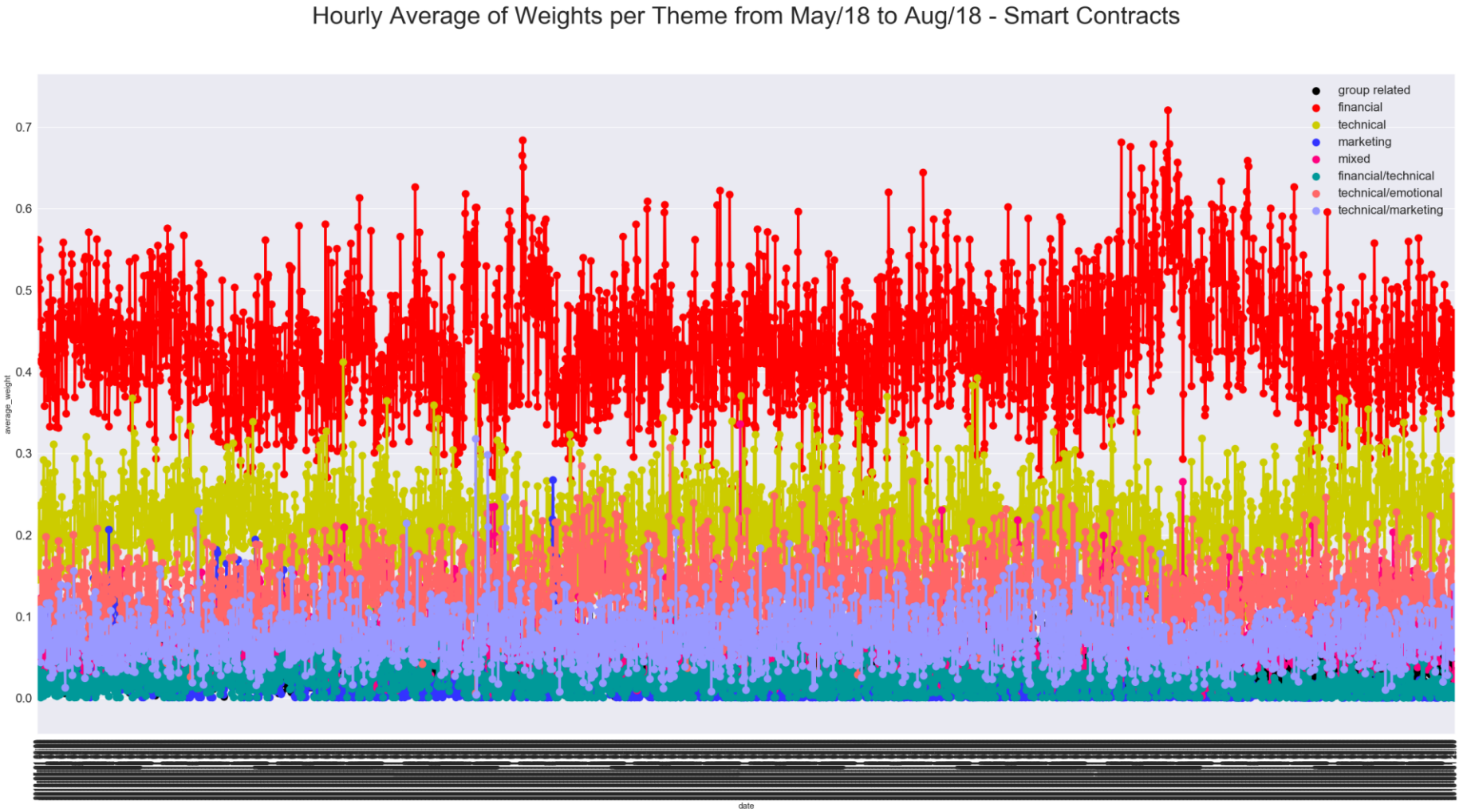

Hourly Average of Weights per Topic from May/18 to Aug/18 - Faster transactions
security, blockchain, sec, mining, lawsuit, news, ceo
fork, network, miner, block, new, mining, blockchain
hour, last, mention, day, course, invest, good
exchange, add, blockchain, trading, new, launch, list
price, analysis, technical, usd, news, new, post
raspberry, new, shirt, free, blue, casino, accept
market, price, top, update, day, today, last
chart, break, support, bullish, price, short, low
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
average_weight
date

Hourly Average of Weights per Topic from May/18 to Aug/18 - Smart Contracts

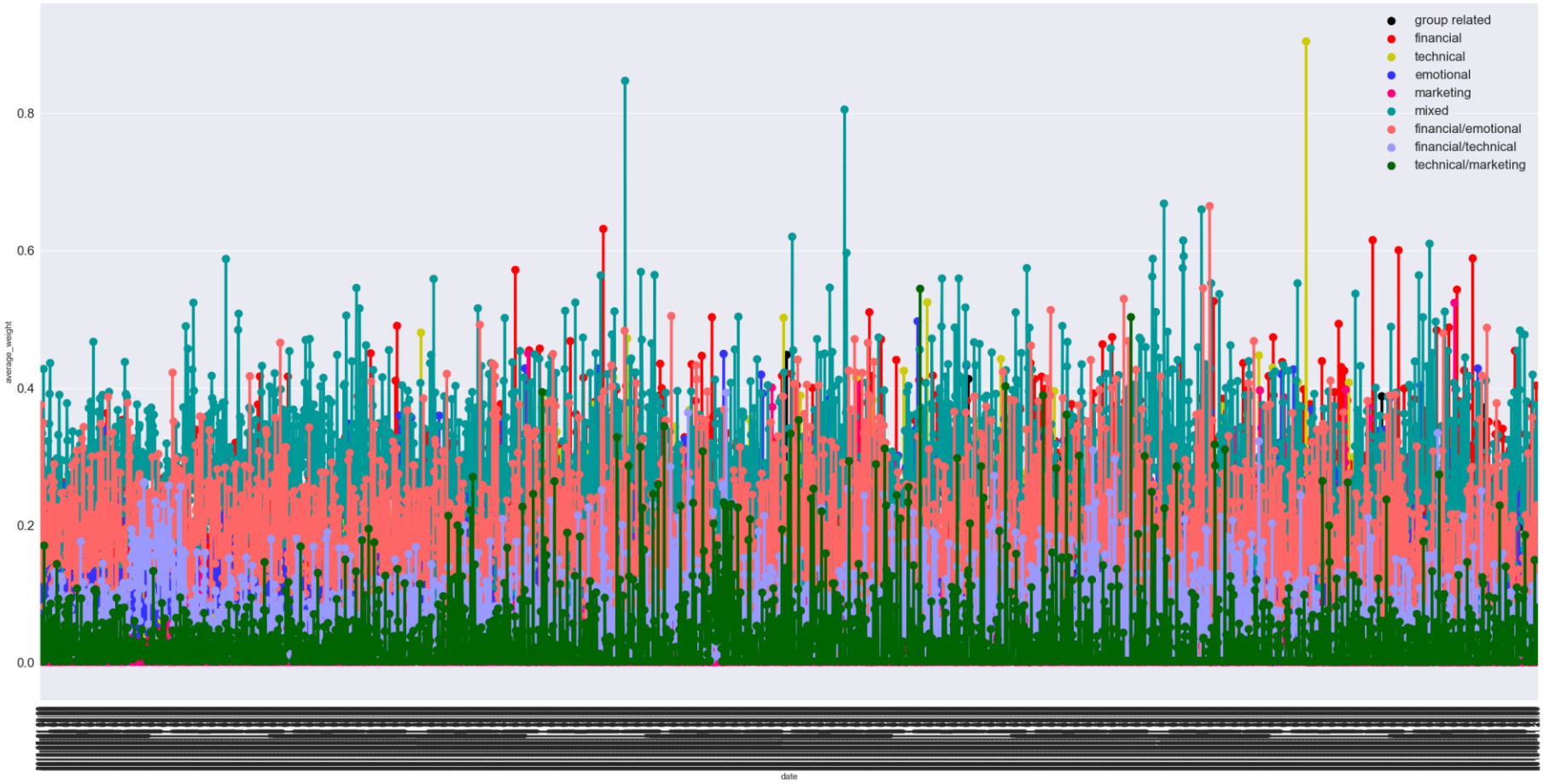
Hourly Average of Weights per Topic from May/18 to Aug/18 - Privacy



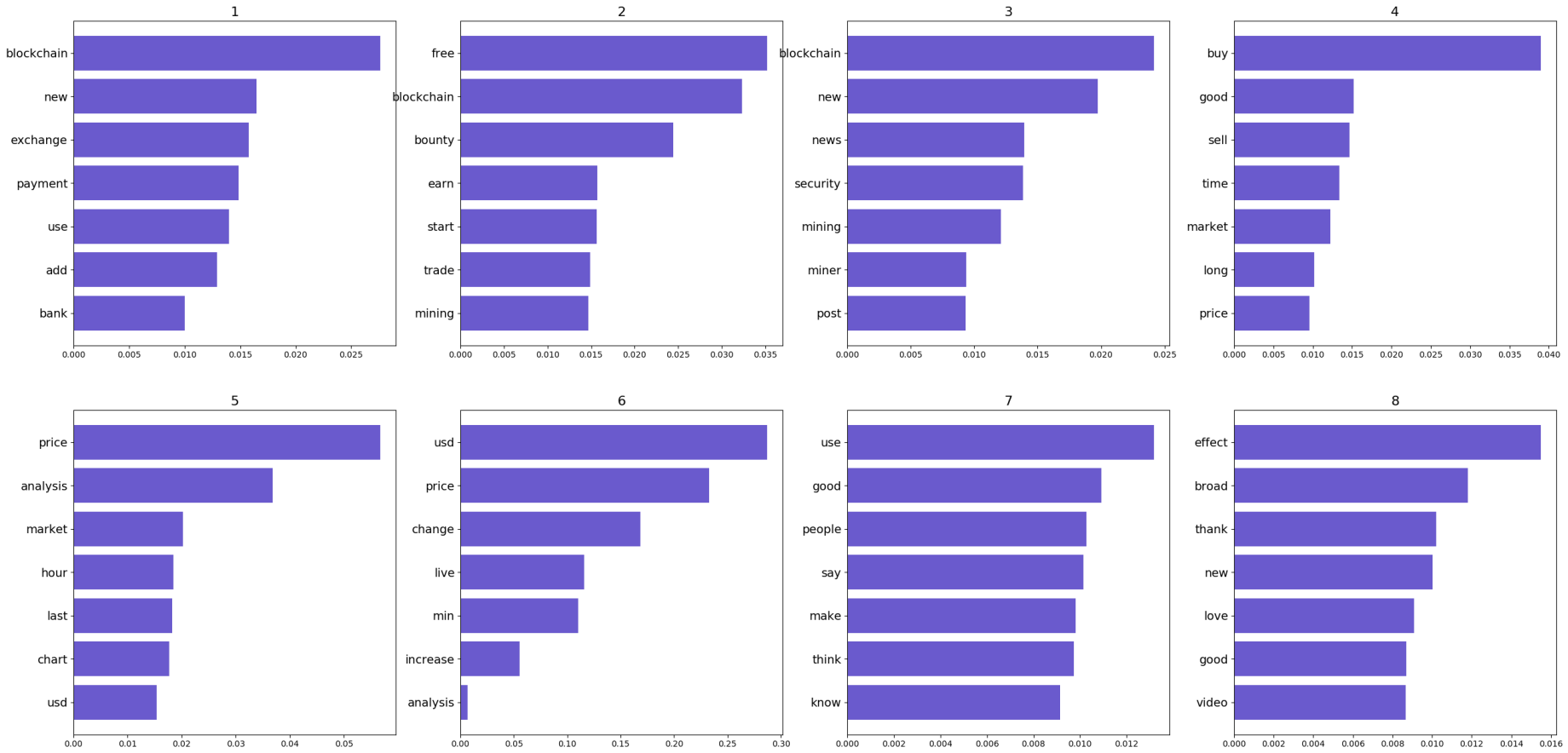

Hourly Average of Weights per Theme from May/18 to Aug/18 - Faster transactions
group related
financial
technical
emotional
mixed
financial/technical
financial/marketing
marketing/emotional
0.6
0.5
0.4
0.3
0.2
0.1
0.0
average_weight
date


Hourly Average of Weights per Theme from May/18 to Aug/18 - Smart Contracts
group related
financial
technical
marketing
mixed
financial/technical
technical/emotional
technical/marketing
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
average_weight
date

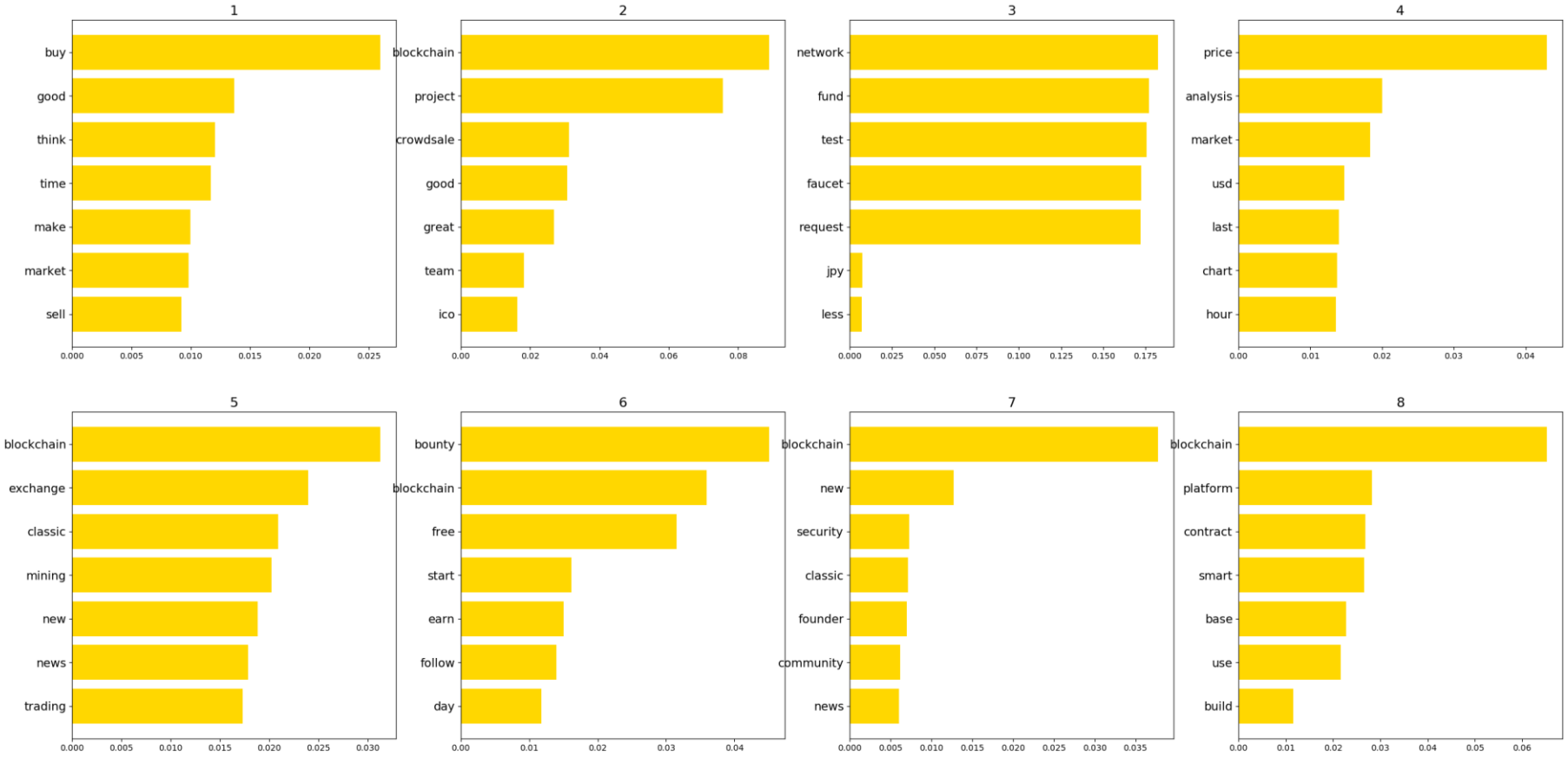
Hourly Average of Weights per Theme from May/18 to Aug/18 - Privacy



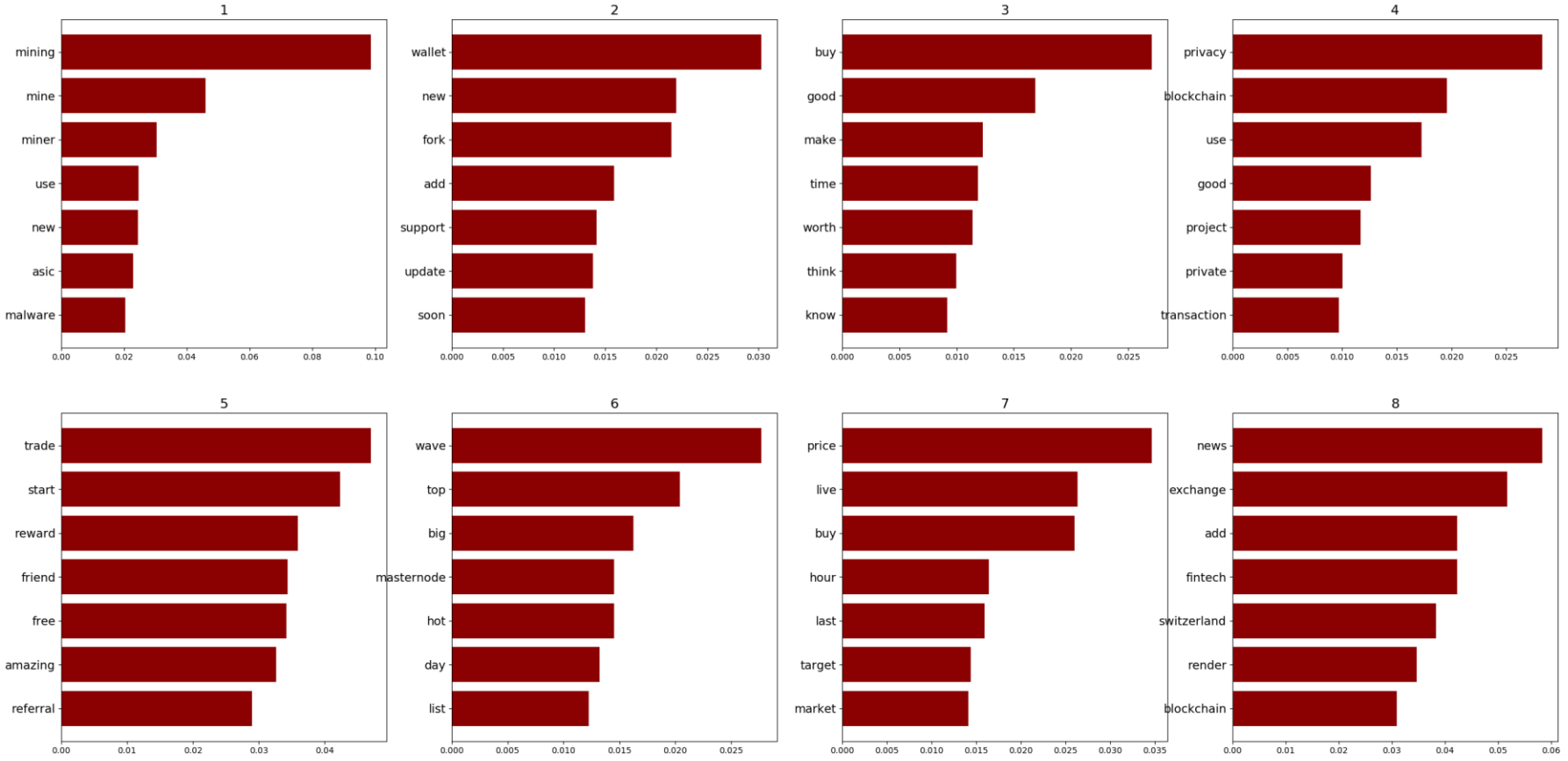
8 Topics for Faster transactions



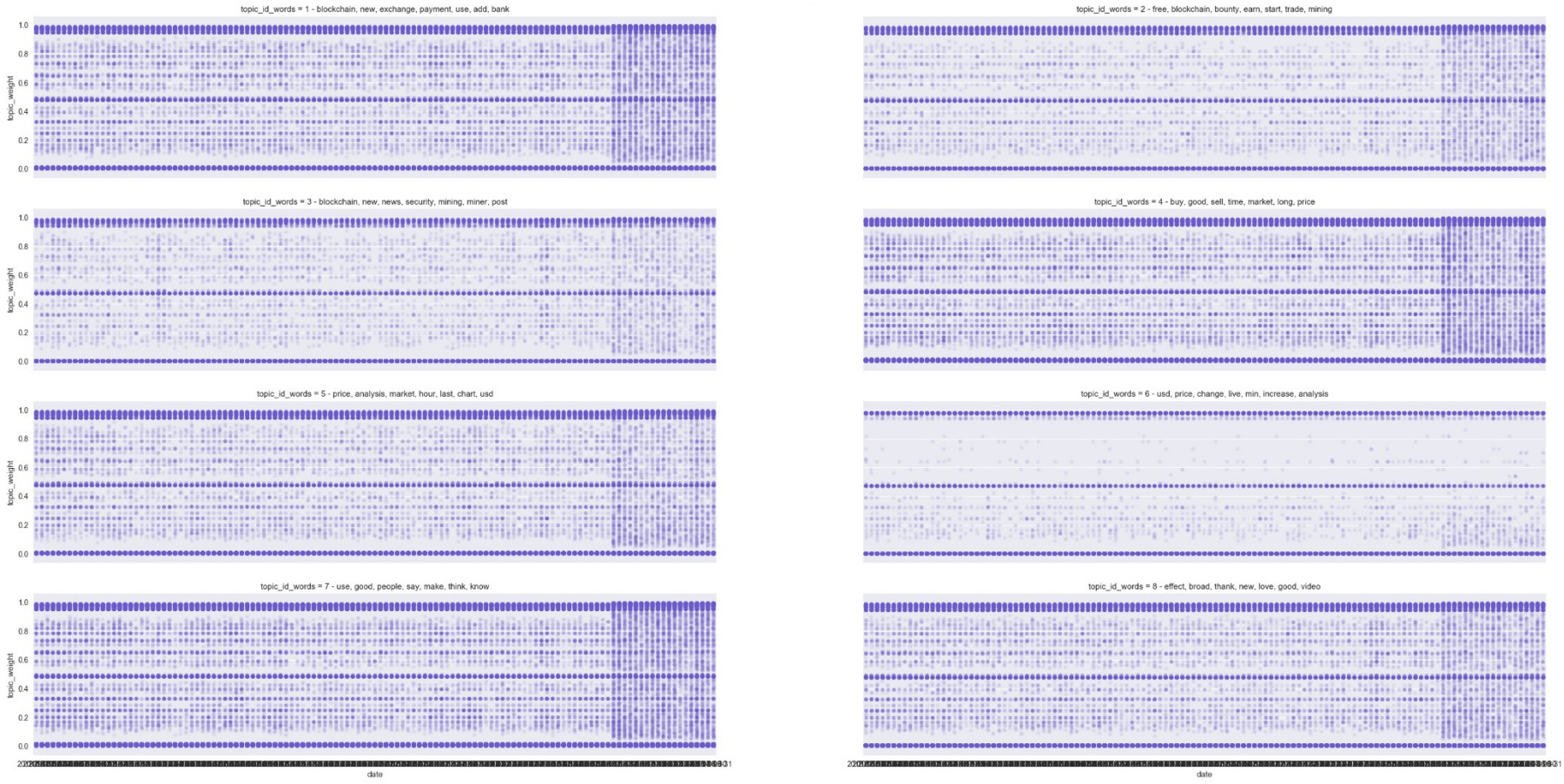
8 Topics for Smart Contracts



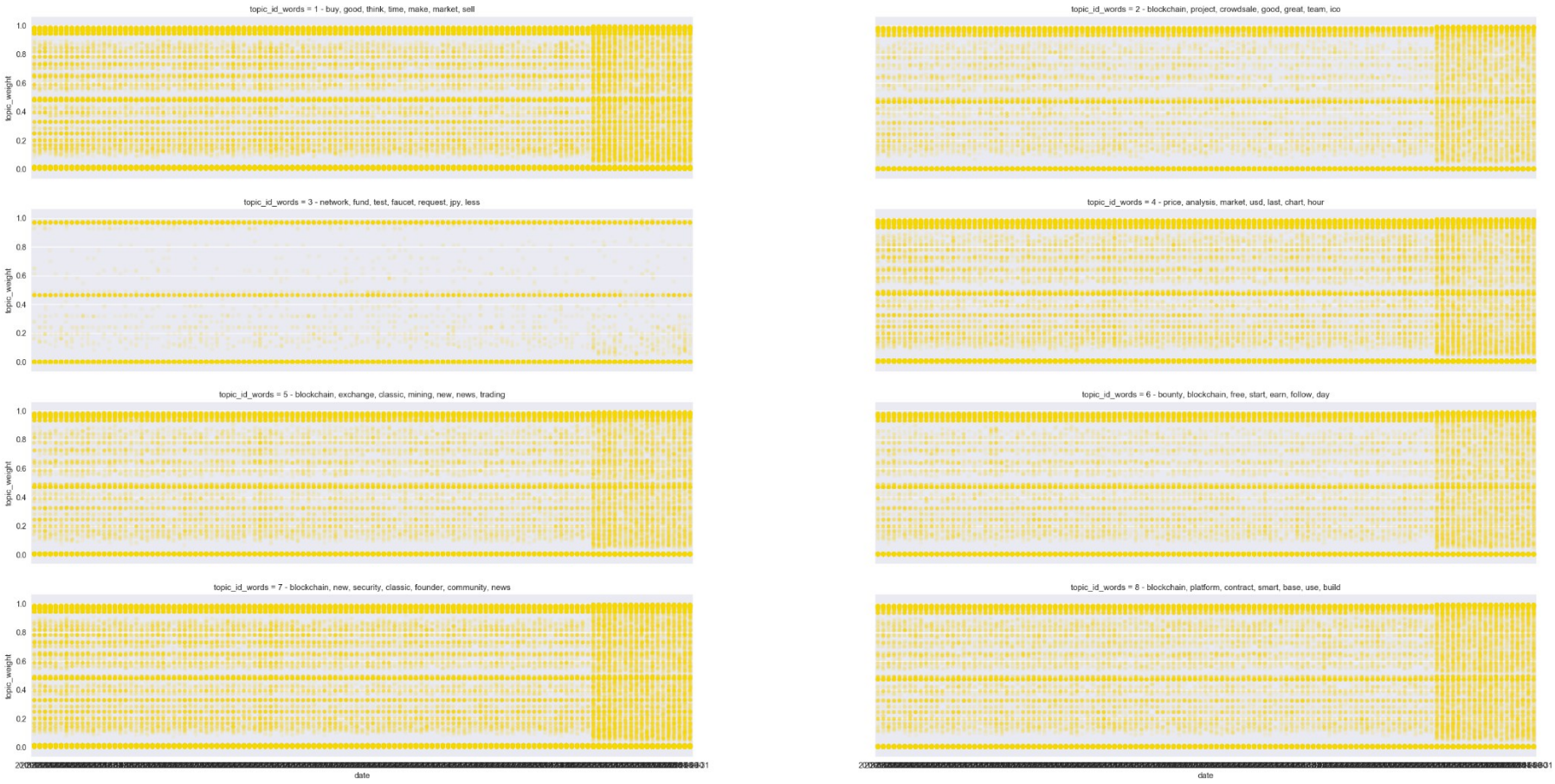
8 Topics for Privacy



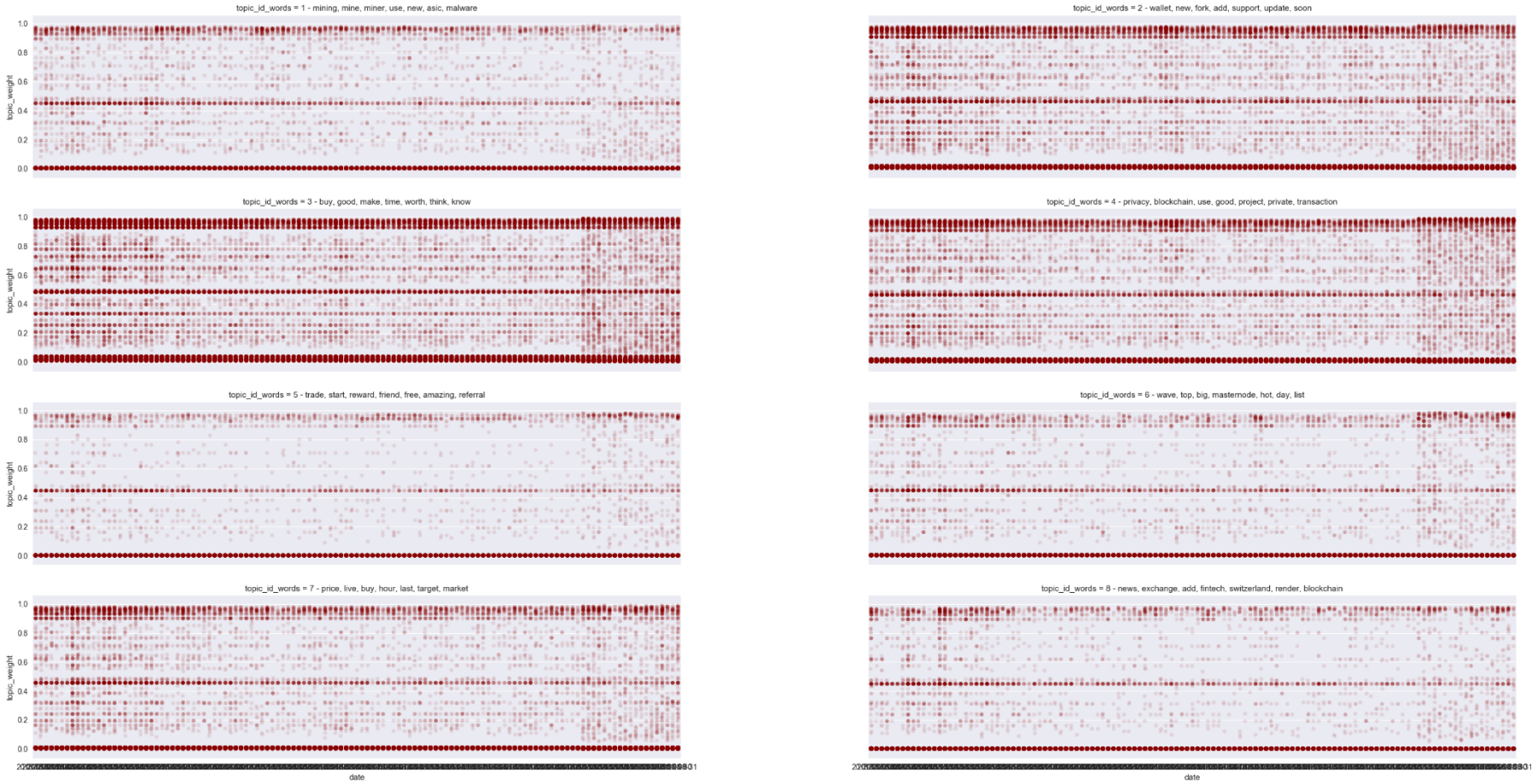
Scatterplot of Topic Weights by Topic, for 8 Topics, from May/18 to Aug/18 – Faster transactions



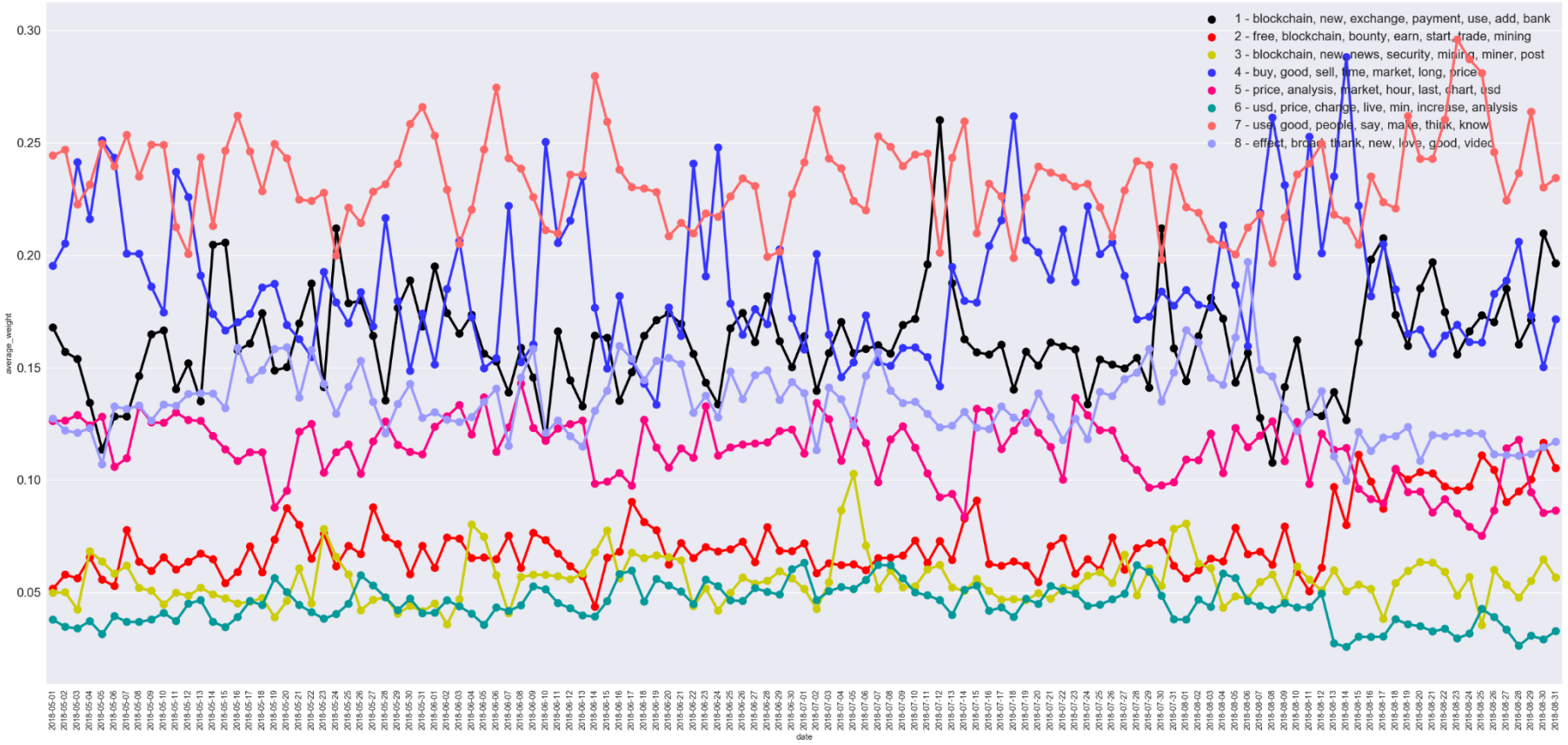
Scatterplot of Topic Weights by Topic, for 8 Topics, from May/18 to Aug/18 – Smart Contracts



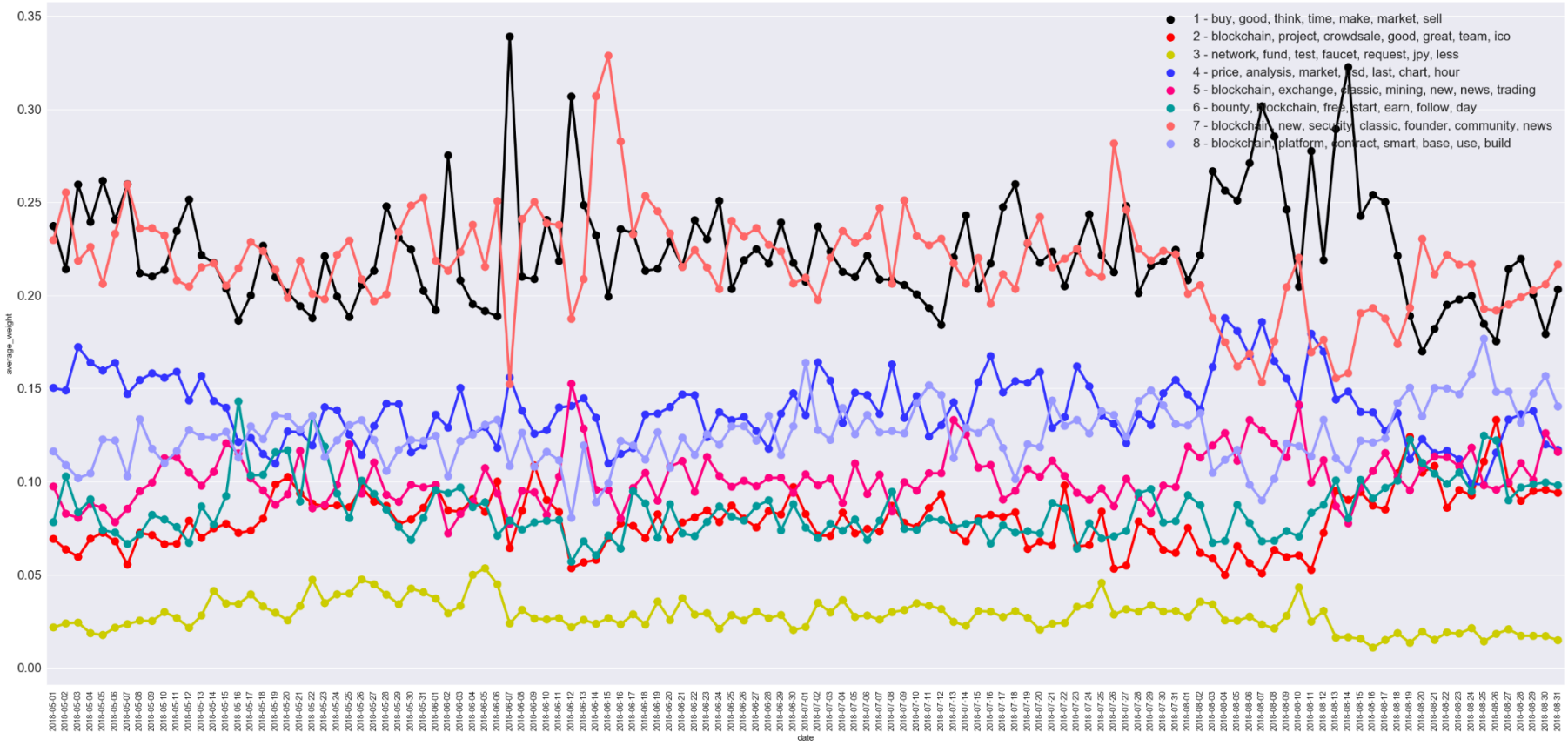
Scatterplot of Topic Weights by Topic, for 8 Topics, from May/18 to Aug/18 – Privacy



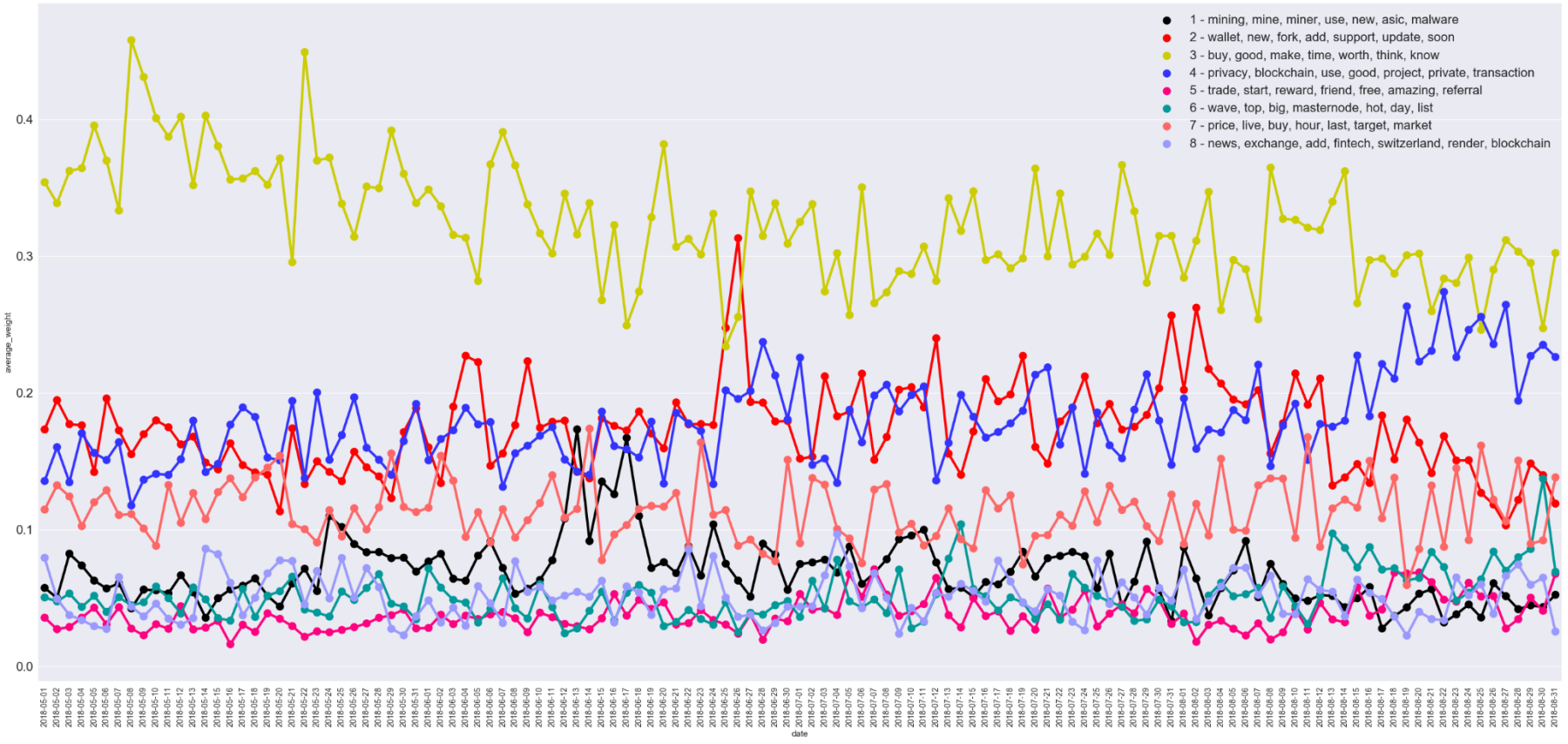
Daily Average of Weights per Topic, for 8 Topics, from May/18 to Aug/18 – Faster transactions



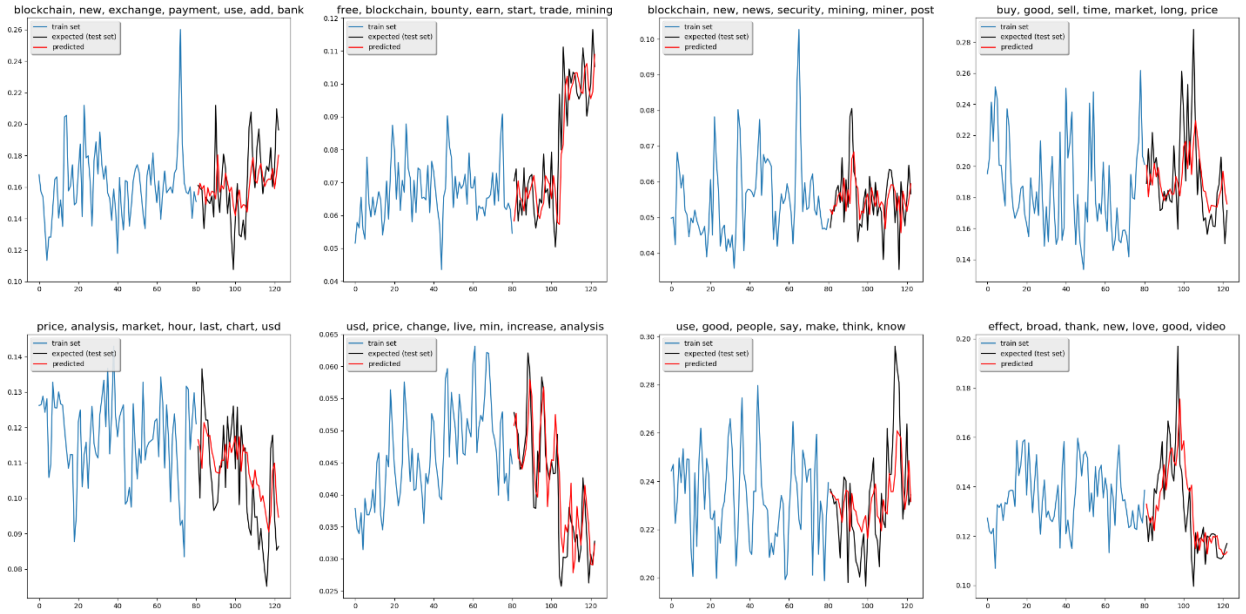
Daily Average of Weights per Topic, for 8 Topics, from May/18 to Aug/18 – Smart Contracts



Daily Average of Weights per Topic, for 8 Topics, from May/18 to Aug/18 – Privacy

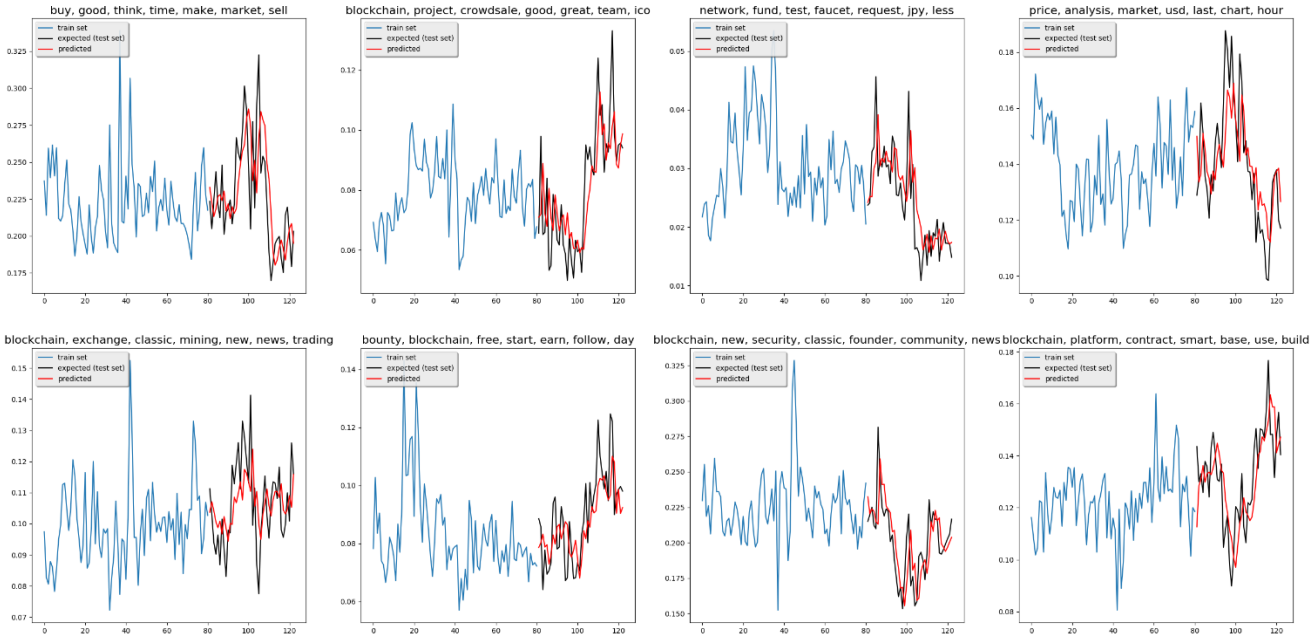


Predicted Weight Values for 8 Topics using ARIMA – Faster transactions



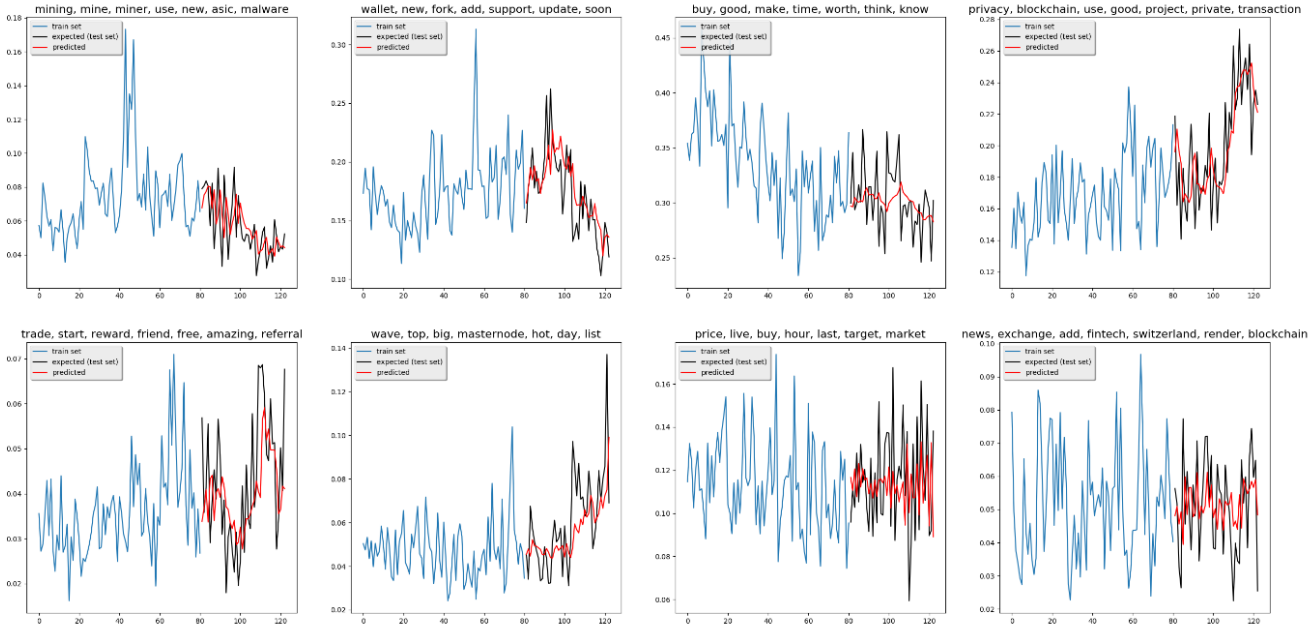
Faster transactions				
Topics	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
blockchain, new, exchange, payment, use, add, bank	0.000564	(2,0,2)	0.000429	76%
free, blockchain, bounty, earn, start, trade, mining	0.000158	(2,1,0)	0.000135	85%
blockchain, new, news, security, mining, miner, post	0.000114	(1,0,0)	0.000076	67%
buy, good, sell, time, market, long, price	0.000840	(4,0,0)	0.000658	78%
price, analysis, market, hour, last, chart, usd	0.000176	(1,1,1)	0.000152	86%
usd, price, change, live, min, increase, analysis	0.000044	(10,0,0)	0.000028	64%
use, good, people, say, make, think, know	0.000382	(1,0,0)	0.000336	88%
effect, broad, thank, new, love, good, video	0.000208	(1,1,2)	0.000192	93%

Predicted Weight Values for 8 Topics using ARIMA – Smart Contracts



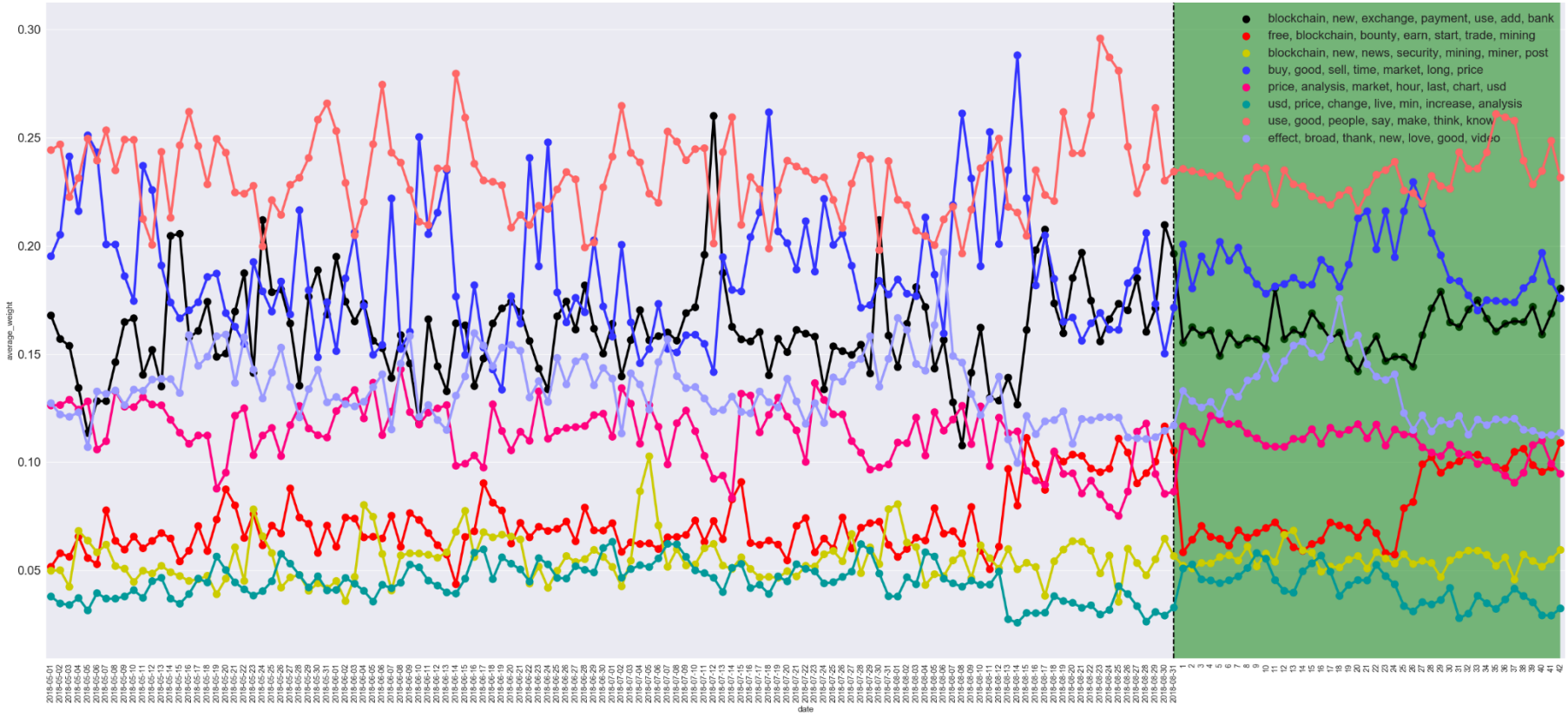
Smart Contracts				
Topics	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
buy, good, think, time, make, market, sell	0.000951	(2,1,0)	0.000837	88%
blockchain, project, crowdsale, good, great, team, ico	0.000237	(10,0,0)	0.000189	80%
network, fund, test, faucet, request, jpy, less	0.000044	(1,1,1)	0.000035	80%
price, analysis, market, usd, last, chart, hour	0.000236	(1,0,0)	0.000223	95%
blockchain, exchange, classic, mining, new, news, trading	0.000195	(1,1,1)	0.000152	78%
bounty, blockchain, free, start, earn, follow, day	0.000170	(8,0,1)	0.000126	74%
blockchain, new, security, classic, founder, community, news	0.000432	(2,1,0)	0.000407	94%
blockchain, platform, contract, smart, base, use, build	0.000190	(2,1,0)	0.000178	94%

Predicted Weight Values for 8 Topics using ARIMA – Privacy

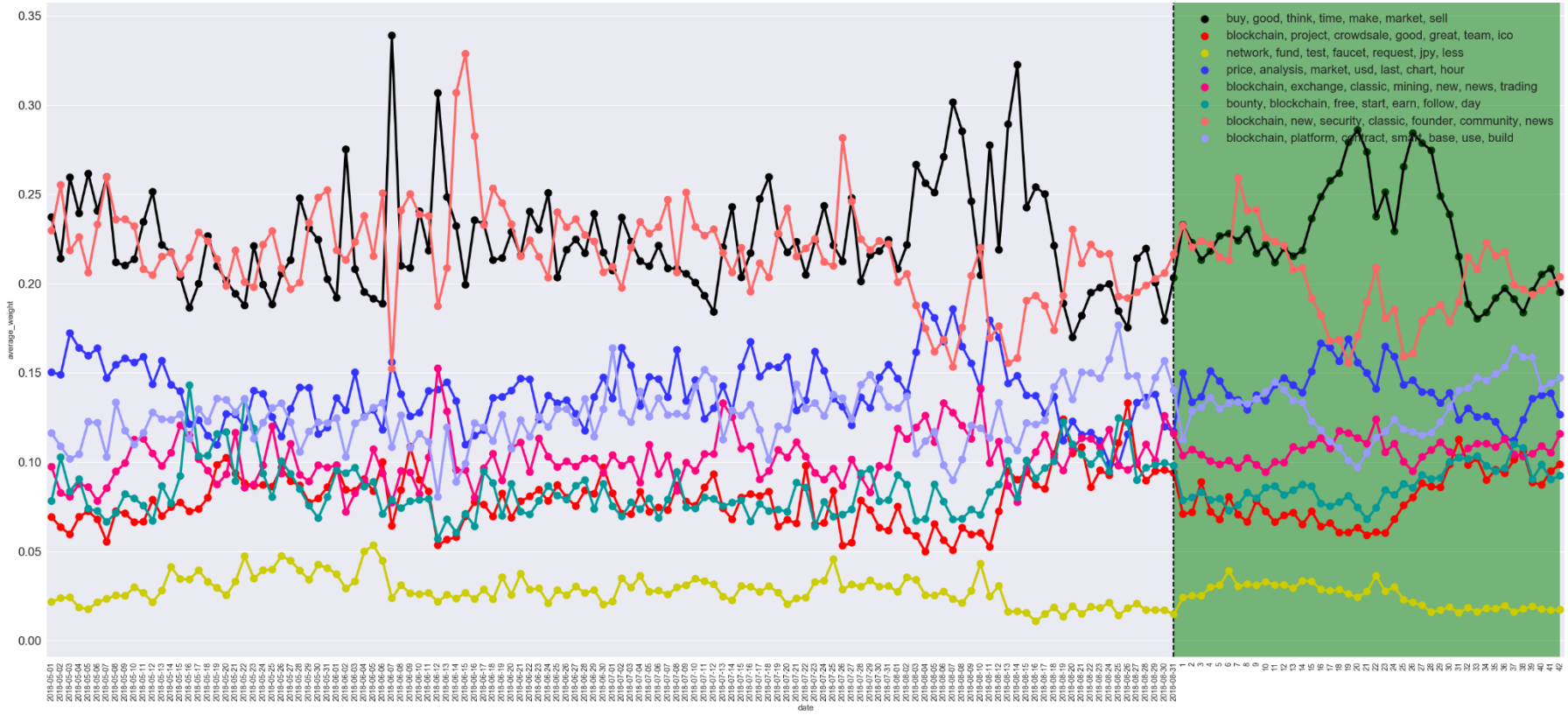


Privacy				
Topics	MSE (Naive)	ARIMA parameters	MSE (ARIMA)	Ratio MSE (ARIMA/Naive)
mining, mine, miner, use, new, asic, malware	0.000396	(0,1,2)	0.000259	65%
wallet, new, fork, add, support, update, soon	0.000846	(6,1,0)	0.000652	77%
buy, good, make, time, worth, think, know	0.001679	(0,1,1)	0.000882	52%
privacy, blockchain, use, good, project, private, transaction	0.001198	(2,1,0)	0.000693	58%
trade, start, reward, friend, free, amazing, referral	0.000218	(8,0,0)	0.000156	72%
wave, top, big, masternode, hot, day, list	0.000396	(0,1,2)	0.000322	81%
price, live, buy, hour, last, target, market	0.001515	(8,0,0)	0.000498	33%
news, exchange, add, fintech, switzerland, render, blockchain	0.000377	(8,0,0)	0.000189	50%

Average of Weights per Topics, for 8 Topics, with forecast (ARIMA) shown in the green-colored area – Faster transactions



Average of Weights per Topics, for 8 Topics, with forecast (ARIMA) shown in the green-colored area – Smart Contracts



Average of Weights per Topics, for 8 Topics, with forecast (ARIMA) shown in the green-colored area – Privacy