Ni, Xinwen; Härdle, Wolfgang Karl; Xie, Taojun

Working Paper

A Machine Learning Based Regulatory Risk Index for Cryptocurrencies

IRTG 1792 Discussion Paper, No. 2020-013

Provided in Cooperation with:
Humboldt University Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series"

Suggested Citation: Ni, Xinwen; Härdle, Wolfgang Karl; Xie, Taojun (2020) : A Machine Learning Based Regulatory Risk Index for Cryptocurrencies, IRTG 1792 Discussion Paper, No. 2020-013, Humboldt-Universität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", Berlin

This Version is available at:
https://hdl.handle.net/10419/230819

IRTG 1792 Discussion Paper 2020-013

A Machine Learning Based Regulatory Risk Index for Cryptocurrencies

Xinwen Ni [*]
Wolfgang Karl Härdle [* *2 *3 *4 *5]
Taojun Xie [*6]

[*] Humboldt-Universität zu Berlin, Germany
[*2] Singapore Management University, Singapore
[*3] Xiamen University, China
[*4] Charles University, Czech Republic
[*5] National JiaoTong University, Taiwan
[*6] National University of Singapore, Singapore

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

A Machine Learning Based Regulatory Risk Index for Cryptocurrencies

Xinwen Ni * [a], Wolfgang Karl Härdle[a,b,c,d,f], and Taojun Xie[g]

[a]School of Business and Economics, Humboldt-Universität zu Berlin, Berlin, Germany
[b]Sim Kee Boon Institute for Financial Economics, Singapore Management University, Singapore
[c]W.I.S.E. - Wang Yanan Institute for Studies in Economics, Xiamen University, Fujian, China
[d]Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[f]Department of Information Management and Finance, National JiaoTong University, Taiwan
[g]Asia Competitiveness Institute, Lee Kuan Yew School of Public Policy
National University of Singapore, Singapore

August 9, 2020

Abstract

Cryptocurrencies' values often respond aggressively to major policy changes, but none of the existing indices informs on the market risks associated with regulatory changes. In this paper, we quantify the risks originating from new regulations on FinTech and cryptocurrencies (CCs), and analyse their impact on market dynamics. Specifically, a **C**ryptocurrency **R**egulatory **R**isk Inde**X** (CRRIX) is constructed based on policy-related news coverage frequency. The unlabeled news data are collected from the top online CC news platforms and further classified using a Latent Dirichlet Allocation model and Hellinger distance. Our results show that the machine-learning-based CRRIX successfully captures major policy-changing moments. The movements for both the VCRIX, a market volatility index, and the CRRIX are synchronous, meaning that the CRRIX could be helpful for all participants in the cryptocurrency market. The algorithms and Python code are available for research purposes on www.quantlet.de.

Keywords: Cryptocurrency, Regulatory Risk, Index, LDA, News Classification
JEL classification: C45, G11, G18

*The authors gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft through the International Research Training Group IRTG 1792 "High Dimensional Non Stationary Time Series", as well as the Czech Science Foundation under grant no. 19-28231X

1 Introduction

Today, there are nearly 2,500 cryptocurrencies worth more than $252.5 trillion trading in the market (Dolan, 2020). The original boom of cryptocurrencies occurred in an unregulated environment. Even as news outlets and investors paid closer attention to the market, regulators and international actors remained largely distant from the action, and prices continued to soar unabated. However, the situation changed since 2014.

Regulations are designed to protect the investors, to put a stop on money laundering, or to prevent the fiat currency from being crowded out. Despite these good wills, speculation and implementation of regulations have resulted in volatile price movements in the cryptocurrency markets. Recent incidents, including China's ban on cryptocurrency exchanges and the rumours of Korea doing the same, have caused major sell-offs and losses among investors. It is therefore important to identify the extent to which new regulations and speculations on them have affected the cryptocurrency markets. Ignoring this source of risk, regulators could end up with self-destroying outcomes and create thus systemic risk bias.

In this paper, we aim to quantify the risks originating from introducing regulations on the Cryptocurrency (CC) markets and identify their impact on the cryptocurrency investments. In order to measure the regulatory risk, particularly the effects of regulations, some researchers considered event-study methods (Binder, 1985; Buckland and Fraser, 2001; Binder, 1985; Binder, 1985; Schwert, 1981 and etc.). However, cryptocurrency market is young and so different from other financial market that the previous regulatory event may not appear again, such as a certain country ban the market. Therefore, we need a measurement tool, which is able to represent the risk level, be comparative and track the changes over time. An index matches all those requirement.

Indices have been applied to track the Cryptocurrency markets already. The CC index developed in Trimborn and Härdle (2018), known as CRIX [1] is a benchmark and tracks the price movements in the CC markets on a daily basis. A volatility measure, VCRIX (Kim et al., 2019), similar to VIX, is also presented. there to reflect the market's volatility How-

[1] seen in crix.berlin, or thecrix.de

ever, even though the VCRIX shows jumps that are to some extent attributable to political decisions, neither of these indices and other indices, like e.g. CCI30 [2] directly address regulatory risks although it plays an important role for the future of CCs and in addition might be created by text mining techniques from a sufficiently rich corpus.

The here proposed Cryptocurrency Regulatory Risk Index (CRRIX) will be constructed by evaluating regulation-related news articles. The indices introduced in Baker et al. (2016) reveal economic policy uncertainty. Similar to their indices, our index is also based on the policy-related news coverage frequency. Unlike their algorithm, which involved a meticulous manual process to label a pool of 12,000 articles, we pursue a Machine Learning (ML) technique to classify policy-related news in our data. We use the support vector machine (SVM), a widely used text classifier, as benchmark and apply Latent Dirichlet Allocation (LDA) method to capture the distances between articles and further classify the unannotated news according to their similarity to policy-related news.

The Regulatory Risk Index (CC) will subsequently be used to analyse the association between regulatory risk and market activities. The algorithms have been programmed in Python. All numerical calculations are available for research purposes on www.quantlet.de and also on Github in the organization QuantLet , see Borke and Härdle (2018).

In the next Section 2, we discuss the background and research questions. The Section 3 presents the data in detail and shows the basic statistics. The Section 4 enters into the methodology and the Section 5 presents results from LDA model and the time series of our regulatory risk index. Finally, Section 6 concludes.

2 Background and Research Questions

In this section, we first generally review the development of cryptocurrency. Based on that, we display the picture of trends of regulatory dynamics in the CC market. The research questions touch the identification of the regulatory risk, the construction of the regulatory

[2] The CCi30 is a rules-based index designed to objectively measure the overall growth, daily and long-term movement of the blockchain sector. It does so by tracking the 30 largest cryptocurrencies by market capitalization, seen in http://cci30.com

risk index, and the interaction of the regulatory risk index with market variables.

2.1 Development of Cryptocurrency

The history of digital or programmable monies can be traced back to as far as four decades ago, although the word "cryptocurrency" became popular only recently. In the 1980s, the concept of E-cash was introduced by David Chaum in a paper entitled "Blind Signatures for Untraceable Payments" (Fiorillo, 2018). Subsequently, DigiCash, B-Money, and Bit Gold were proposed by Chaum, Wei Dai, and Nick Szabo in the 1990s, respectively. Most of these digital monies did not survive as they failed to address the practical issues of "double spending" and "third-party trust".

The milestone development in this field came in after the global financial crisis. Nakamoto, in the seminal white paper, 2008, proposed the bitcoin. This is a peer-to-peer electronic cash system implemented via the blockchain technology, with the participation of a network of computer owners known as miners. The blockchain technology, later known as the distributed ledger technology, or DLT, ensures that transaction records are easy to be updated but costly to be changed, avoiding the "double spending" issue. Miners, after solving complex mathematical puzzles, are rewarded with a predetermined amount of bitcoins. The amount of the reward can only be amended with the agreement from majority of the miners. Such a mechanism avoids the "thrid-party trust" problem in fiat monies, whose issuance depends on the central banks' sole discretion. Since its birth, the BTC model has defined the meaning of "cryptocurrency", which now typically refers to a decentralized digital network that facilitates secured transactions using cryptographic methods.

Bitcoin has sparked a series of events since 2009. In Figure 1, we show a timeline of a few major events in the last decade. Some events pertained to innovations that aimed at improving the technology behind cryptocurrencies (highlighted in blue). For example, competing CCs, also known as altcoins, began to emerge in 2011. The events highlighted in yellow are the ones involving CCs being used in legitimate real-life applications. A well-known example was the two pizzas in 2010 that cost 10,000 bitcoins. Through this series of

Figure 1: Cryptocurrency timeline

events, the general public became aware of the strengths and weaknesses of CCs. As good and bad news took turns to be reported, the price of bitcoin, highlighted in gray, and that of the other CCs also experienced volatile movements. Notably, events relating to losses of cryptocurrency exchanges have been associated with the largest price movements. For instance, as Mt. Gox went bankruptcy in 2014 after losing over 850,000 bitcoins, the price of bitcoin fell from over $1,000 to the around $400. This event has been foreseeable through textual analysis of BTC blog and other solid media channels. Linton et al. (2017) and the chapter 3 of Härdle et al. (2017) employ a technique similar to ours to evaluate discussions in social media. A recent price movement was an upswing of 1300% in year 2017, followed by a fall of more than half in May 2018.

2.2 Regulations of the Cryptocurrency Market

Among the features of cryptocurrencies, anonymity has been the most controversial one. Users of cryptocurrencies like this feature for it is difficult to trace one's spending history, but regulators dislike it for the exact same reason. We thus see an interesting interaction here. Users have proposed numerous improvements to enhance anonymity. Zcash and Monero were designed to facilitate anonymous transactions. At the same time, incidents such as the Silk Road going live and terrorists using cryptocurrencies for remittance got the regulators on the toes. Regulators, on the one hand, insisted on know-your-customer (KYC) measures to trace any illegitimate transactions, but on the other hand, prepared to launch their own cryptocurrencies (Barrdear and Kumhof, 2016; George et al., 2020). Fighting against illegitimate transactions became one of the first tasks for the cryptocurrency regulators.

Trading activities at the exchanges was the next issue that regulators reacted to. The anonymity feature and a lack of regulation at the cryptocurrency exchanges cultivated illegal and unethical behaviors, such as money laundering, pump-and-dump activities and scams. Ignorant users of the cryptocurrency exchanges faced high risks while trading. Responding to these, starting in 2011, the US Treasury Department's Financial Crimes Enforcement Network (FinCEN) began oversight of cryptocurrency exchanges, transmitters, and admin-

istrators under the Bank Secrecy Act related to anti-money laundering and combating the financing of terrorism (AML/CFT) (Lee and Deng, 2018). In the same year, the United States Department of Homeland Security (DHS) initiated first investigation relating to cryptocurrency. The number of cases raised to over 200 in 2017 (seen in Figure 1).

However, there has been a dilemma in regulating the cryptocurrency space. While it was important to protect retail investors and to prevent unlawful transactions, new technology driving the development of cryptocurrencies needed to be incubated until the ecosystem was matured. It was then imperative for the regulators to move towards systematic governance. In Figure 2, we show a timeline of countries' publications of guidance on the cryptocurrency space.

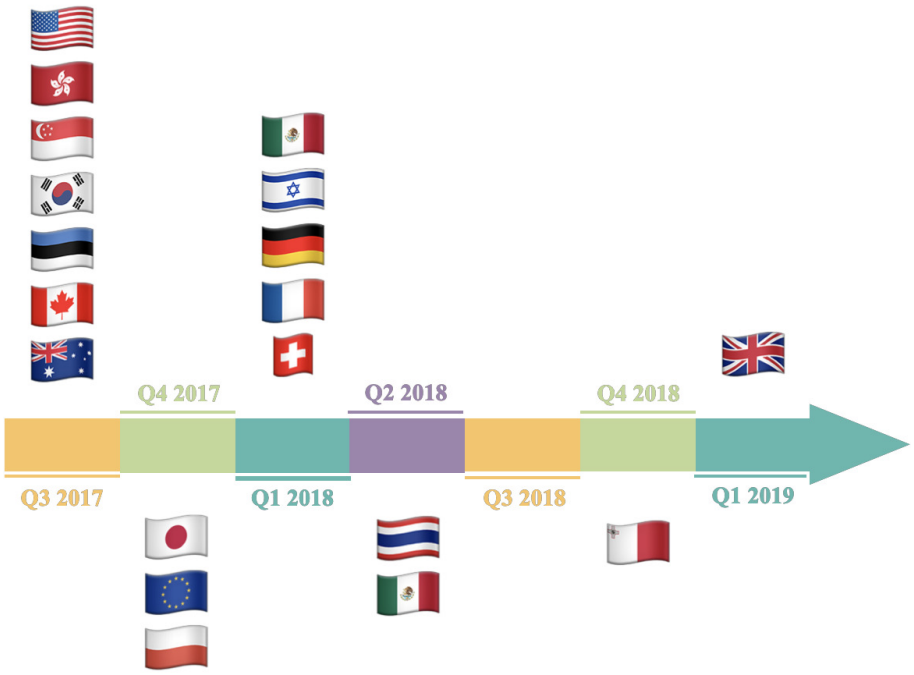

Figure 2: Timeline for Cryptocurrency Guidance

In order to apply the existing law and regulations, the very first step for the governors is to identify the cryptocurrency nature as means of holding, transferring or investing "money". Since the birth of cryptocurrency, the debate, whether BTC, or generally cryptocurrencies, are currency or asset, eolved with the growth of the market (Glaser et al., 2014, Baur et

al., 2018). Despite using "currency" as the name name, most countries that have permitted cryptocurrencies view these monies as assets. In 2014, the US Internal Revenue Service (IRS) stipulated that virtual currency is considered property for the purpose of federal tax". Purchases using cryptocurrencies as the media of exchange are considered barter trades, or exchanges between properties and services (Blandin et al., 2019). The Swiss Financial Market Supervisory Authority (FINMA), in 2018, published guidelines for initial coin offerings (ICOs). In those guidelines, tokens were categorised into three types based on their economic functions: payment tokens, utility tokens and asset tokens (Caytas, 2018).

Regulatory responses to cryptocurrencies vary from strict bans to government adoption. Most of these regulations were designed to protect retail investors or to ensure that cryptocurrencies are not used for illegal activities. The Financial Action Task Force (FATF), recommended that regulations be implemented to prevent the use of cryptocurrencies in money laundering and terrorist finance (Gold and McBride, 2019). At the G20 Summit in 2018, the FATF urged on all countries to take necessary preventive measures towards the misuse of cryptocurrencies. In the European Union, by 2020, all member states will implement AML/CFT[3] rules to cryptocurrency exchanges and wallet operators (Houben and Snyers, 2018). On the contrary, some countries began to change their attitude towards the adoption of DLT. In China, the use of BTC as currency for financial institutions was prohibited in 2013 (Glaser et al., 2014), but in Octobor 2019, the president of China announced that the country will encourage enterprises to seize the opportunity in the up-and-coming technology. Subsequently, the People's Bank of China announced that it would launch digital Yuan, commonly known as the central bank digital currency (Zhong, 2019). Such support for DLT lead to a new round of debate globally.

This disparity in regulatory approaches creates interesting dynamics in the cryptocurrency markets. As poited out earlier, good and bad news is likely to induce different movements in the price of cryptocurrencies. As more regulations are on their way, and because the cryptocurrency market is globally unified, policy changes in one country or even the rumor about the attitude adjustment of one government, would be widely discussed in the

[3] AML/CFT: Anti-money laundering / combating the financing of terrorism.

news platform and bring turbulence to the whole market. Under these circumstances, and in consistent with Auer and Claessens (2020), the construction of a regulatory risk measurement for the whole market is necessary and the news data, which includes information from different sources and countries, would be helpful.

2.3 Research Questions

For the most cases, the regulation changes do not come out from nowhere. A good example is the Fed's rate cut (or hike). Before the Fed's announcement of changes, there are discussions and predictions about the Fed's move in newspapers. Numerous studies have attempted to examine whether the text mining technology would contribute to the forecasting for financial markets (Nassirtoussi et al., 2014), e.g. Ghiassi et al. (2013), Geva and Zahavi (2014), Tu and Härdle (2018). But not many use textual data to analyze financial regulatory risk. Gulen and Ion (2016), Baker et al. (2016) and Kang and Ratti (2013) argue that news from newspapers could be a good indicator for macro policy uncertainty.

As discussed before, indices have been introduced to trace the movement of CC market, but none of them addresses regulatory risks which plays an important role for the future of CCs. We try to, in this paper, quantify the risks brought by introducing policies on the CC market and further discuss its impact on the CC investment. A regulatory risk index for the CC market, which could serve as a tool for passive investors, for the fund manager, and even for policy-makers.

There are mainly three kinds of indices as to the data sources which were applied to construct the index. First, and most commonly, some indices use real data, e.g. VIX, S&P 500, and DAX, which employed real market price or volume data. Second, some are based on a regular survey, e.g. IFO business Climate Index and Purchasing Managers' Index (PMI), which is relied on a monthly survey of supply chain managers. Recently, the third source, news data, or generally speaking the text data, becomes popular, e.g. Thomson Reuters MarketPsych Indices [4], sentiment Indices, which are standard input to trading desks and are

[4] seen in https://www.refinitiv.com/en/products/world-news-data/

provided by a variety of news data channels. The majority of these indices are based on text mining tools that establish a dictionary and then calculate in a bag of words approach the frequency of key words with identifiable sentiment, like good, bad etc. Such input for asset trading and risk management is a refined data source that in many cases contains useful directional information. Baker et al. (2016) introduced an index to reveal economic policy uncertainty based on newspaper coverage frequency. They calculated the number of articles which contained "economic" or "economy"; "uncertain" or "uncertainty"; and one or more of "congress", "deficit", "Federal Reserve", "legislation", "regulation" or "White House" from 10 large newspapers. Their index successfully represented movements in policy-related economic uncertainty and they also found that policy uncertainty rise stock price volatility.

Unlike Baker et al.'s algorithm, which involved a meticulous manual process to label a pool of 12,000 articles, we propose a machine learning based approach to classify policy-related news in our data. With a sufficiently rich corpus, it might be able to conduct the regulatory risk index by some NLP techniques, such as topic modelling methods. We are also curious about whether the policy-related uncertainty index could be helpful for the market participants. Therefore, the research questions discussed in this paper are as follows:

Research Question 1 *How to indentify the regulatory risk for Cryptocurrencies?*

Research Question 2 *How to construct an index of regulatory risk for Cryptocurrency market based on news data?*

Research Question 3 *What is the impact of regulatory risk to the market?*

3 Data

3.1 News from Top Cryptocurrency Online Platforms

Since CCs are frequently mentioned in newspapers only for the very recent years, we choose to use news data from the top online cryptocurrency news platform (Guides, 2018). The

representative news platform Coindesk [5] and Bitcoin Magazine [6] were considered in this paper, because both of them are not only pioneers and leaders in the market but also they offer news data which could trace back to the beginning of BTC's boost.

Coindesk is a news website with a particular focus on Blockchain, Bitcoin and Cryptocurrencies as a whole. The site launched in April 2013 and released close to 25000 articles. The articles have been classified already in categories: markets, technology, business, policy & regulation and people. The aim of this paper is to introduce a policy uncertainty index by calculating the frequency of policy-related news. The pre-classified news data from Coindesk perfectly matches our demand and will be further applied as training data for the ML models. The textual data from the source was collected via a dynamic web scraper.

After checking for duplicates, eventually we keep 16,528 articles from 01 April 2013 to 18 July 2019. The data covers 76 months, 329 weeks and 2300 days. Out of the total over 16,000 articles, 2,468 are marked as policy-related news. The data is available for further research at the Blockchain Research Center (BRC) [7] and on Quantlet.de.

Figure 3 represents the average number of news per week related to policy and the average number of all news. The number of daily articles increased in 2014 and 2018 both in total and in regulation-related term. In those years, the price of Bitcoin underwent volatile movements, declining by more than 70 % in 2014 in 2018. Before these massive corrections, the end of 2013 and 2017 marked periods of price discovery and all-time highs in USD valuation being broken every other day. During the same periods, number of blockchain-related news articles increased, indicating growing interests in blockchain and distributed ledger technology. There is no doubt that, simultaneously, the market attracted policy-makers' strong attention as well.

Bitcoin Magazine is another leading and pioneer platform supplying information on the new market. It was founded in Feb 2012, one year earlier than Coindesk. There are fewer

[5] https://www.coindesk.com
[6] https://bitcoinmagazine.com
[7] https://hu.berlin/BRC

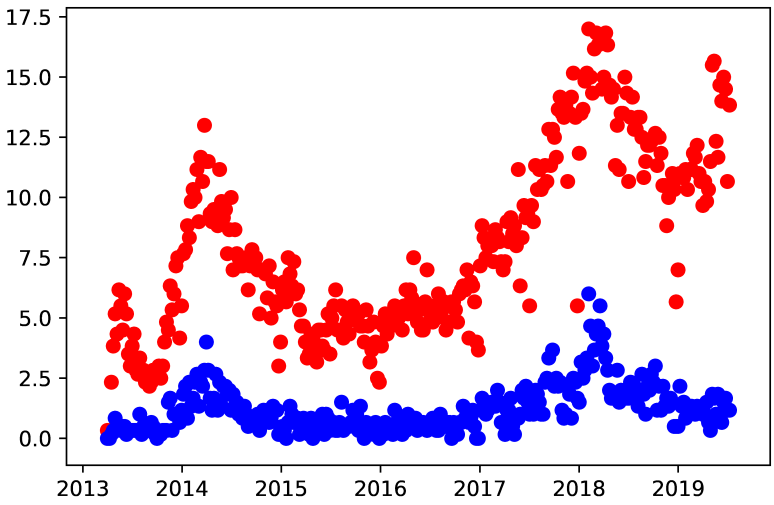

Figure 3: Weekly Average Number of total News (in red) and of Policy-Related News (in blue)

news collected from Bitcoin Magazine, only 4,586, but it covers longer period from 28 Feb 2012 to 18 July 2019. Similar to Coindesk, the platform creates channels with different topics like "Policy and Laws", "Payment", "Blockchain" and etc. But not all news have been classified. We found that 2,976 articles didn't belong to any category. Classification is needed and machine learning methods will be employed. However, our final goal is to apply our methods to multiple data sources, especially the other financial news platforms, such as the NASDAQ news platform in which 887, 018 articles are available since Jan 2013. In order to test the performance of applying our methods to other platforms, we manually marked each article from Bitcoin Magazine with "policy-related" and "non-policy-related" to compare with our ML classification results. 582 of 4,586 articles are related to cryptocurrency market regulations.

3.2 CRIX and VCRIX

The CRIX and VCRIX is chosen to represent the value and the volatility of the entire cryptocurrency market for the later analysis. The CRIX (CRyptocurrency IndeX), created by

Trimborn and Härdle (2018), closely tracks the entire cryptocurrency market performance. Its construction is robust in the sense it takes into account the dynamics of market structure, thus ensuring the representativity and the tracking performance of the index. It follows that constituents of CRIX change over time, depending on market conditions and the relative dominance of CCs. The CRIX series begins from July 2014, and is available through thecrix.de. Reallocation of the CRIX happens on a monthly and quarterly basis. It adopts a liquidity rule when incorporating a certain cryptocurrency into CRIX, and hence guarantees the trading of CRIX, which is good for ETFs and traders. CRIX has been widely investigated in the pioneering research on cryptocurrencies, including Hafner (2020), Klein et al. (2018), Trimborn et al. (2018), and da Gama Silva et al. (2019).

Like VIX or VDAX, which provide a measure for implied volatility, VCRIX, created by Kim et al. (2019), is a volatility index, able to grasp the risk induced by the cryptocurrency market. This index accurately addresses the market dynamics on the basis of CRIX and thus proved to be a proper basis for option pricing. Similar to CRIX, the data of VCRIX is also be able to be download from thecrix.de.

4 Methodology

Based on the rich text corpus, on can now enter the machine learning text mining step to identify policy-related news from others. In this paper, the classification problem is simply binary: policy-related or not. In the literature, SVM is widely applied to solve such kind of binary problem. However, this method didn't performance well with imbalanced cases, whilst our target is to classify the regulatory news (a very small subgroup) from all. In our training data, the ratio is $1:7$. Indeed there are multiple ways to solve the unbalanced data problem, such as oversampling or class-weighted SVM by assigning higher misclassification penalties. But the pre-process of oversampling or undersampling will change the distribution of labels and further change the distribution of test data. Our index is constructed based on regulatory news frequency and it is therefore sensitive to the distribution of classes.

On the other hand, we could assume that when policy-related topics are discussed, sim-

ilar topics with their key words are used and their distributions are close. Based on that assumption, the Latent Dirichlet Allocation (LDA) method can be employed to analyze the topic distribution and words distribution for the corpus and further identify the policy-related articles according to the similarity calculation.

4.1 LDA Topic Modeling

The Latent Dirichlet Allocation (LDA) technique proposed by Blei et al. (2003), is an unsupervised machine learning algorithm that learns the unobserved topics of a corpus (individual news articles in this paper). This technique is widely applied to establish the thematic structure of text and other discrete data in the linguistic, information retrieval, biologic and even engineering literatures (see Blei, 2012 for a review of topic modelling and its application to various text collections).

The LDA technique is based on a generative statistical method to identify the distribution of words that contribute to a topic, while simultaneously constructing documents with different probabilities of topics, meaning that each topic z is annotated with a collections of the most probable words w, and each document d is annotated with a collections of the most probable topics z. It is an unsupervised algorithm which requires no labeled texts and learns these two latent (unobserved) distributions $p\left(w|z\right)$ and $p\left(z|d\right)$ by acquiring model parameters that maximize the probability of each word appearing in each document with the number of topics K as given.

Then, with Bayes theorem, the probability of observed word w_n appearing in a document d_m is given by:

$$p\left(d_m, w_n\right) = p\left(d_m\right) p\left(w_n|d_m\right) \tag{1}$$

$$= p\left(d_m\right) \sum_{k=1}^{K} p\left(w_n|z_k\right) p\left(z_k|d_m\right) \tag{2}$$

where z_k is a latent variable indicating the kth topic from which the words were drawn (Z

in Figure 4), $p\left(w_n|z_k\right)$ is a distribution for each topic over the vocabulary (ϕ in Figure 4), and $p\left(z_k|d_m\right)$ denotes the topic proportions for the mth document (article in this paper) (θ in Figure 4). Intuitively, ϕ indicates which words weight more to a topic, while θ states the importance of those topics to a document.

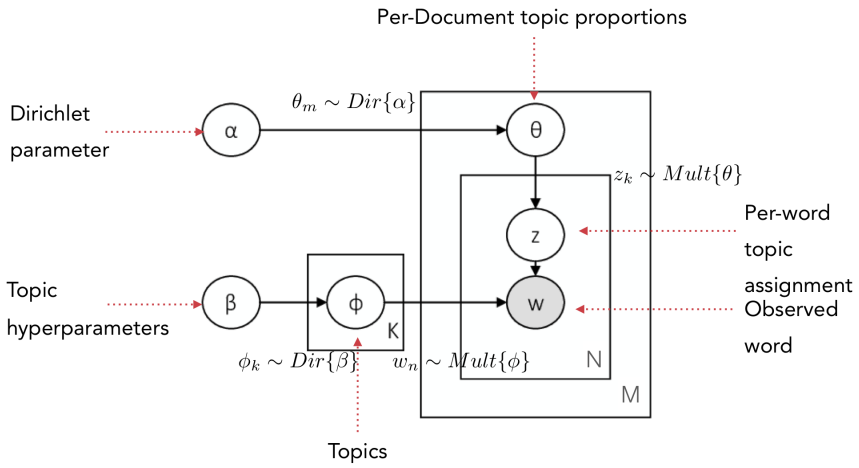

Figure 4: Graphic LDA Model

Both θ and ϕ follow the Dirichlet distribution with hyper-parameter α and β respectively. With higher α, the topic distribution per article turns to be more specific, while similarly, higher β leads to a more specific word distribution per topic. In general, α links to the similarity of documents, meaning that a higher alpha value implies that documents are embodied by more similar weights of each topic. The same holds for β but meaning that a higher beta value indicates that topics contents more similar weights of each word. In the Python package gensim, the symmetric or asymmetric hyper-parameters are learned from data. The generative process of LDA is based on the following joint distribution of the observed variables w and the unobserved variables z, θ, ϕ, α and β,

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k; \beta) \prod_{d=1}^{M} p(\theta_d; \alpha) \prod_{n=1}^{N} p(z_{d,k} \mid \theta_d) p(w_{d,n} \mid \phi_{z_{d,k}}), \tag{3}$$

4.2 Number of Topics for LDA

The standard LDA model proposed by Blei et al. (2003) has a significant weakness that it requires pre-determination of the number of topics, meaning that users should set the number of unobserved topics manually before applying the method. The quality of LDA model is heavily depended on the choice of topic number. Therefore, many believe that choosing the best value for the topic numbers is more art than science (Azqueta-Gavaldón, 2017).

A common solution is to plug in a set of values and pick the optimal topic number either based on some intrinsic criterion, such as the coherence of the topics, or based on some extrinsic criterion, such as accuracy on a specific task, e.g. paraphrase identification. There also exist other methods to help with choosing the number of topics. For example, nonparametric Bayesian models, e.g. Hierarchical Dirichlet process are employed to automatically generate the number of topics (Teh et al., 2004). However, it is computationally inefficient to apply such nonparametric models to LDA (Wallach et al., 2009).

Coherence measures which are based on word co-occurrence are widely applied to quantify the quality of topic models. The poor quality topics with the type of "chained", "intruded" and "random" could be detected using detected with coherence measures (Mimno et al., 2011). Newman et al. (2010) proposed a coherence measure which is comparable to the human rating of topics. Their coherence measure (C_{UCI}) takes the set of the top J words ($w_1, ..., w_J$) for a given topic and sum a confirmation measure over all word pairs. The function is given as follows:

$$C_{\text{UCI}} = \frac{2}{J \cdot (J-1)} \sum_{i=1}^{J-1} \sum_{j=i+1}^{J} \log \frac{P\left(w_i, w_j\right) + \epsilon}{P\left(w_i\right) \cdot P\left(w_j\right)} \tag{4}$$

where the probabilities are estimated on Wikipedia outperform which is used as external reference corpus. Mimno et al. (2011) employ an asymmetrical confirmation measure between top word pairs in the calculation of coherence C_{UMass}:

$$C_{\text{UMass}} = \frac{2}{J \cdot (J-1)} \sum_{i=2}^{J} \sum_{j=1}^{i-1} \log \frac{P\left(w_i, w_j\right) + \epsilon}{P\left(w_j\right)} \tag{5}$$

Unlike C_{UCI}, the probabilities in function 5 are estimated based on the original corpus with applied to trained topic models. Röder et al. (2015) build a coherence framework and report a measure (C_v) with the best performance. Different from C_{UMass} and C_{UCI}, C_v defines the confirmation using normalized point-wise mutual information (NPMI) for the $j-th$ element of the context vector $\overrightarrow{v_i}$ of word w_i:

$$v_{ij} = \text{NPMI}\left(w_i, w_j\right)^{\gamma} = \left(\frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log\left(P\left(w_i, w_j\right)+\epsilon\right)}\right)^{\gamma} \tag{6}$$

where γ denotes the weight for NPMI. In this paper, we use coherence value C_v as the criteria to select model.

4.3 LDA-Based Similarity Measurement and Classification

Semantic similarity problems can be classified according to different levels of granularity, specifically ranging from word-to-word to sentence-to-sentence to document-to-document similarities (Niraula et al., 2013). In this paper, our task is to analyze document-to-document similarity, particularly as a binary decision problem in which an article is policy related or not. We rely on one probabilistic method, LDA, which regards documents as distribution over topics and topics as distribution over words. So, we assume that policy-related articles have similar topic distributions.

The Hellinger distance can be applied to compute the distance between two distributions. For document p and document q, the distributions of topics are $z_p = (z_{p,1}, \ldots, z_{p,k}, \ldots, z_{p,K})$ and $z_q = (z_{q,1}, \ldots, z_{q,k}, \ldots, z_{q,K})$ respectively. Hellinger Distance d_H between those two news with K topics is given as follows:

$$d_H\left(z_p, z_q, X\right) = \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^{K}(\sqrt{z_{p,k}} - \sqrt{z_{q,k}})^2} \tag{7}$$

There are reasons that we choose Hellinger distance rather than other distances to calculate news similarity. First, if we denote $\hat{f}(x)$ as a kernel density estimator, the asymptotic

distribution of $\sqrt{nh}(\hat{f}(x) - f(x))$ depends on $f(x)$, the true distribution, however, after taking square root, the asymptotic distribution of $\sqrt{nh}(\sqrt{\hat{f}(x)} - \sqrt{f(x)})$ eliminates its dependency with $f(x)$ (see the proof in Appendix). Second, when applying Hellinger distance in a binary decision criterion, it is not sensitive to the class skew (Cieslak and Chawla, 2008), implying that it performs well with imbalance data. Besides, the results from Hellinger distance is bounded by $[0, 1]$ for all values of $z_{p,k}$ and $z_{q,k}$. It is easy to read and compare. The highest value, 1, indicates the maximized distance and therefore means that the compared two distributions differ from each other significantly, whereas the value 0 implies the highest similarity and shortest distances. Meanwhile, d_H is symmetric, meaning $d_H(z_p, z_q) = d_H(z_1, z_p)$.

As the next step, we calculate the average distance between the article i and all policy-related news d_i:

$$\bar{d}_{l,i} = \frac{1}{N_r} \sum_{j=1}^{N_r} d_H(z_{l,i}, z_{r,j}) \tag{8}$$

where N_r is the number of regulatory news, $z_{u,i}$ denotes the topic distribution of article i and $l = \{r, \ non \ or \ u\}$ meaning that the article i is regulatory news "r", non-regulatory news "non" or unclassified news "u". $z_{r,j}$ represents topic distribution of regulatory new j ($j = 1, ..., N_r$).

Since we have the assumption that regulatory news have smaller distances between each other than the other news, the average distance $\bar{d}_r = \{\bar{d}_{r,1}, ..., \bar{d}_{r,N_r}\}$ for all policy-related news should be relatively smaller than $\bar{d}_{non} = \{\bar{d}_{n,1}, ..., \bar{d}_{n,N_{non}}\}$ for all non-policy-related news. Then if $\bar{d}_{u,i}$ of the unclassified article i is small and close to $\bar{d}_r$, we mark that news as policy-related. In this paper, we set a threshold $\underline{d}$ equals to τ^{th} quantile of $\bar{d}_r$ and $\tau = 0.95$ in this paper.

4.4 Construction of CRRIX

As mentioned before, the construction of CRRIX is simply the coverage frequency of policy-related news, as followed:

$$CRRIX_t^s = \frac{N_{t,reg}^s}{N_{t,all}^s} \tag{9}$$

where s is the periodicity and $s = \{daily,\ weekly\ or\ monthly\}$. $N_{t,reg}$ and $N_{t,all}$ are the number of regulatory news and all news at time t.

5 Empirical Results

First we did some pre-processing of the data (words) : stopwords are eliminated (words that do not informatively or semantically contribute to an article, e.g. “at”, “or”, “and”); all words have been converted to lower case. We calculate the coherence value of LDA models with different topic numbers from 2 to 25 given an automatic generated hyperparameter α (α=0.01) and $\beta = 0.1$. The Figure 5 indicates that the best performanced model with optimal number of topics for the corpus in this paper is the model with $K = 14$. When $K = 14$, the model has the highest coherence value and when the number of topics increases after the optimal choice, the coherence value turns to relatively stable. We also test the robustness with different α and β from 0.01 to 0.3. The above mentioned combination performances best but the coherence value doesn’t change much for given topic number.

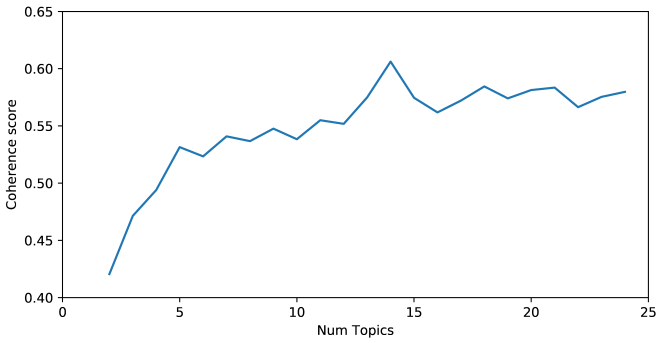

Figure 5: Coherence value for different numbers of topics K

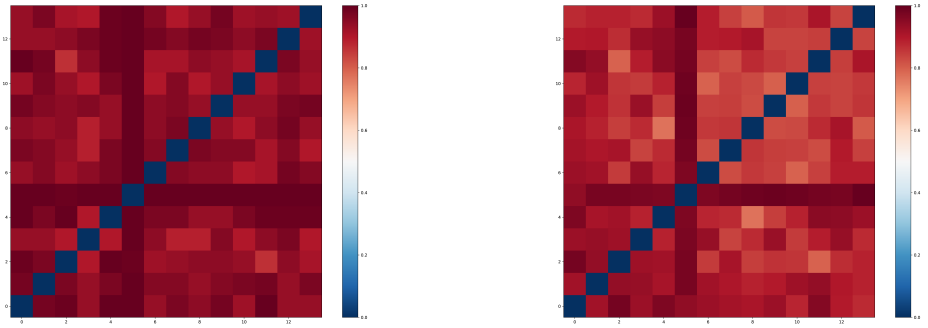

Figure 6: The distances between topics using Jaccard distance (left) and Hellinger distance (right)

Another criterion for determining the quality of the LDA model is to concern the diversification of topics. Ideally, we would like to see topics as different as possible. The heating maps in the Figure 6 exhibit the distances between topics ($K = 14$). The red cell represents strongly uncorrelated topics, while the blue cell indicates high correlation. The left diagram was generated using Jaccard distance and for the right one, we apply Hellinger distance to calculate the differences between topics. All elements except those in the diagonal in both diagrams are red or reddish, which means the 14 topics are relatively different from each other and our LDA model performance well in this aspect. However, compared with Hellinger distance, even though Jaccard distance is robust and wildly used in ML methodologies, it is less sensitive than Hellinger distance. Therefore, in the later discussion, we only apply Hellinger's method in the distance calculation.

In order to further show the performance of the trained LDA model, we try to compare the topics with other sources. In the leading Cryptocurrency news platform, Coindesk, the news are labelled as those categories: "Opinions", "Tech", "Business", "Policy & Regulations", "Market" and "Feature". These categories appear in Table 1 (column 1) together with their equivalent topic (column 2) which is generated by our LDA model and the list of representative words for each topic (column 3). From the table we can see that for the major

Coindesk Subcategory	LDA Topic	Top Keywords
Opinions	Opinions	Bitcoin, say, people, make, go get, take, would, could, way
Tech	Technology	System, blockchain, use, transaction, chain Technology, security, work, datum, network
Business	Business	Company, say, business, new, service base, startup, firm, founder, CEO
Policy & Regulation	Regulation	Currency, business, virtual, law, state Regulation, money, digital, exchange, tax
Market	Investment	Bitcoin, market, currency, price, exchange value, investor, Litecoin, trade, investment
	Trading and Exchange	Exchange, BTC, account, customer, trading User, deposit, page, trade, fund
Feature	Mining	mine, power, asic, block, hash chip, network, unit, hardware, pool
	Coins	Coin, project, Dogecoin, game, Altcoin developer, community, donate, crowdfunder, token

Table 1: Categories (Coindesk.com) matched by LDA topics

True\Pred	NB		SVM_{cw}		LDA		Total
	1	0	1	0	1	0	
1	0	582	0	582	361	221	582
0	0	4004	0	4004	188	3816	4004
Total	0	4586	0	4586	549	4037	4586
Accuracy	0.873		0.873		0.907		

Table 2: Confusion matrix of classification for three methods (Naive Bayes, Class-weighted SVM and LDA)

categories in the popular platforms, we could find the corresponding topics in our model. In the case of "Market", topics go beyond the categories proposed by the platform. We must admit that parts of the category "Business" overlaps with the category "Market". Even though we select the topic "Trading and Exchange" to match the category "Market", it could also be put under a bigger concept of "Business". In this sense, the machine learning LDA technique performs better and clearly identifies topics which keep distances with each other.

We use the trained LDA model to calculate the Hellinger distances. We find that the distributions of average distances between each regulatory news and all other regulatory news $\bar{d}_r$ and of average distances between each non-regulatory news and all regulatory news $\bar{d}_{non}$ are significantly different. Policy-related news are similar with smaller distances, whereas the most non-policy-related news are further away. Then we calculate the average Hellinger distance $\bar{d}_{u,i}$ for each unclassified article i. Those, which are smaller than 0.392, the 0.95 quantile of $\bar{d}_r$, will be classified to the policy-related group.

The classification results of our method based on LDA was compared with those generated by Naive Bayes and SVM, two broadly used supervised ML classification methods. The confusion matrices of classification results and the manually classified data can be found in Table 2. "True" value is given by human involved annotation, and "Pred" value is predicted by the here applied ML technique. Number 1 means that the given news is labelled as policy-related and number those marked by 0 is non-policy-related.

The accuracies of all three methods are relatively high, over 0.87 (seen in Table 2). How-

ever, for supervised ML methods Naive Bayes and Class-weighted SVM, they simply label all articles non-policy-related. Even with the high accuracy, those methods can not help with our research question. The purpose of classification in this paper is to find the ratio of policy-related news over all. With the increase of news taken in to the calculation, but 0 policy-related news was identified, the index will go towards destruction.

Meanwhile, the accuracy of our methods is higher, 0.91. Although from the Table 2 we can read that the type I error for LDA classification are also high, almost 40 percent, it still could contribute to build the index and in this sense performances much better than NB and SVM.

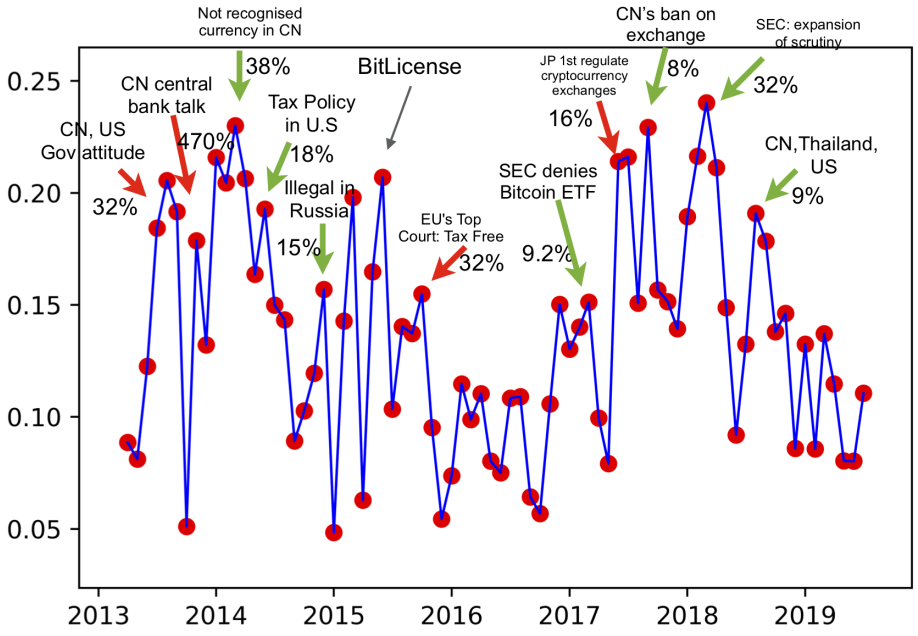

Figure 7: CRRIX (Monthly) with news and price highlights

Multiple reasons could contribute to the misclassification. One could be that the LDA based criteria are relatively strict. News, which have the distance to all policy-related news

as close as that of 95% the pre-identified regulatory news, would be counted for regulatory news. Those with a bit larger Hellinger distance are all excluded. Another reason could come from the data itself. Cryptocurrency market is young and the core policies, which were discussed by the public and announced by the governments, were time-varying. Using all time slot from the year 2013 to 2019 might be biased. A dynamic LDA with rolling window will solve this problem but it requires sufficient data points. In this paper, we didn't consider this method, since the data is limited.

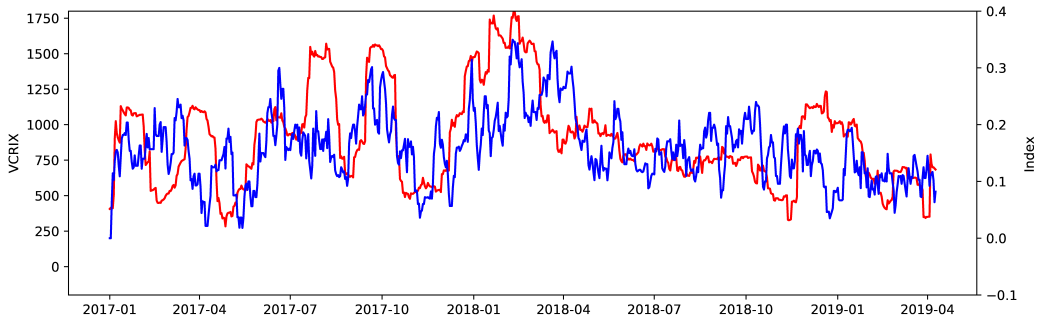

Figure 8: VCRIX (in red) and Regulatory Risk Index of Cryptocurrency Market (in blue)

QCRRIX

After each article was labeled according to its distance with other pre-identified policy-related news, we simply count the number of articles under the class "regulation" for a give time slot and divide it by the number of all articles for the same time. Since the daily time series is too noisy, especially at the early stage of the market, we only consider weekly time steps. Figure 7 indicates that the peaks and jumps of the index are mainly led by big policy changes. The red arrow means positive policy which brought an increase to the price of Bitcoin, whereas the green arrow is vice versa. The number next to arrows is the weekly return rate (positive with red arrow and negative with green arrow).

Figure 7 reveals that the changes of policies are accompanied with drastic price fluctuations which bring high risk to the market. Out index successfully captures those big changing moments.

Figure 8 shows that our regulatory risk index is closely related to VCRIX, a volatility

number of lags (no zero) 1				
ssr based F test:	F=23.1736	p=0.0000	df_denom=825	df_num=1
ssr based chi2 test:	chi2=23.2579	p=0.0000	df=1	
likelihood ratio test:	chi2=22.9372	p=0.0000	df=1	
parameter F test:	F=23.1736	p=0.0000	df_denom=825	df_num=1

Table 3: Granger causality test results for lag 1

index for CCs market. Especially for the period from Sep 2017 to March 2018, the extremely high volatility is driven by the policy uncertainty. The movements for both VCRIX and the regulatory risk index are synchronous. The correlation between these two indices is 0.44712. Our regulatory risk index could contribute to forecast the market movement.

We further test the causality between CRRIX and VCRIX. First we do Dicky Fuller test to confirm stationary of both time series. The results reject the non-stationary hypothesis ($p - value$ equals to 0.000895 for VCRIX and 0.004617 for CRRIX). Here we only show the Granger causality test results for lag 1 in the Table 3. The null hypothesis for Granger causality test is that the time series CRRIX, does NOT Granger cause the time series VCRIX. Here, for lag 1, we reject the null hypothesis. This means that the past (lag 1) values of CRRIX (lag 1) have a statistically significant effect on the current value of VCRIX. The results hold for lag 1 to 7.

6 Conclusion

In this paper, via the machine learning tool LDA, we quantify the risks originating from introducing regulations on the cryptocurrency market and identify their impact on the cryptocurrency investments. Indices have been constructed to track the Cryptocurrency markets, however, none of these indices directly address regulatory risks. The indices introduced in Baker et al. (2016) focus on economic policy uncertainty in general. Similar to that, we construct a regulatory risk index for cryptocurrencies that is based on the policy-related news coverage frequency. Unlike the classical annotation approach, which involves a meticulous

manual process, we employ the ML-LDA rather costless and efficient method to classify policy-related news.

We first generally reviewed the development of cryptocurrencies and the trend of regulatory dynamics. Based on that, we discussed the research questions: What exactly is the regulatory risk for Cryptocurrencies? How to construct an index of regulatory risk for Cryptocurrency market based on news data? What is the impact of regulatory risk to the market?

In order to address the answer to those questions, we first collected news data from the top online cryptocurrency news platform (Guides, 2018), Coindesk and Bitcoin Magazine, via a dynamic web scraper. In addition, the CRIX and VCRIX is chosen to represent the value and the volatility of the entire cryptocurrency market for the later analysis.

To calculate the coverage frequency, we tried to solve the problem of semantic similarity as a binary decision problem in which an article is policy related or not, using Latent Dirichlet Allocation (LDA), which models the underlining topics for a corpus of documents, where each topic is a mixture over words and each document is a mixture over topics.

The topics given by LDA were comparable with the leading Cryptocurrency news platform. For the major categories in the popular platforms, we could find the corresponding topics in our model and the clearly identified topics kept distances with each other. According to our model, the top words for regulation topic are: *currency, business, virtual, law, state, regulation, money, digital, exchange* and *tax*.

We use the trained LDA model to calculate the Hellinger distances. Those with small average distance will be classified to the group of policy-related news. The results were compared with that of Naive Bayes and Class-weighted SVM methods. Since our data is very imbalanced, the performance of those to supervised ML methods were not helpful. But our LDA based distance classification turned to a high accuracy 0.91 and could contribute to construct the index.

The final results of the regulatory risk index are shown in Figure 7 and in Figure 8. Our index successfully captures those big policy changing moments. The movements for both VCRIX and the regulatory risk index are synchronous, and the Granger test proved the causality of CRRIX to the market volatility.

References

Auer, Raphael and Stijn Claessens, "Cryptocurrency market reactions to regulatory news," Discussion Paper DP14602, CEPR April 2020.

Azqueta-Gavaldón, Andrés, "Developing news-based economic policy uncertainty index with unsupervised machine learning," *Economics Letters*, 2017, *158*, 47–50.

Bai, Yiqi and Jie Wang, "News classifications with labeled LDA," in "2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)," Vol. 1 IEEE 2015, pp. 75–83.

Baker, Scott R, Nicholas Bloom, and Steven J Davis, "Has economic policy uncertainty hampered the recovery?," *Government policies and the delayed economic recovery*, 2012, *70*.

_ , _ , and _ , "Measuring economic policy uncertainty," *The quarterly journal of economics*, 2016, *131* (4), 1593–1636.

Barrdear, John and Michael Kumhof, "The macroeconomics of central bank issued digital currencies," Staff Working Paper 605, Bank of England July 2016.

Baur, Dirk G, Kihoon Hong, and Adrian D Lee, "Bitcoin: Medium of exchange or speculative assets?," *Journal of International Financial Markets, Institutions and Money*, 2018, *54*, 177–189.

Binder, John J, "Measuring the effects of regulation with stock price data," *The RAND Journal of Economics*, 1985, pp. 167–183.

Blandin, Apolline, Ann Sofie Cloots, Hatim Hussain, Michel Rauchs, Rasheed Saleuddin, Jason G Allen, Katherine Cloud, and Bryan Zheng Zhang, "Global cryptoasset regulatory landscape study," *Available at SSRN*, 2019.

Blei, David M, "Probabilistic topic models," *Communications of the ACM*, 2012, *55* (4), 77–84.

_ , **Andrew Y Ng, and Michael I Jordan**, "Latent dirichlet allocation," *Journal of machine Learning research*, 2003, *3* (Jan), 993–1022.

Borke, Lukas and Wolfgang K Härdle, "Q3-D3-LSA: D3. js and Generalized Vector Space Models for Statistical Computing," in "Handbook of Big Data Analytics," Springer, 2018, pp. 377–424.

Buckland, Roger and Patricia Fraser, "Political and regulatory risk in water utilities: Beta sensitivity in the United Kingdom," *Journal of Business Finance & Accounting*, 2001, *28* (7-8), 877–904.

Caytas, Joanna Diane, "Regulation of Cryptocurrencies and Initial Coin Offerings in Switzerland: Declared Vision of a 'Crypto Nation'," *Includes Chapter News*, 2018, *31* (1), 53.

Chen, Jingnian, Houkuan Huang, Shengfeng Tian, and Youli Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, 2009, *36* (3), 5432–5435.

Chen, Xingyuan, Yunqing Xia, Peng Jin, and John Carroll, "Dataless text classification with descriptive LDA," in "Twenty-Ninth AAAI Conference on Artificial Intelligence" 2015.

Cieslak, David A and Nitesh V Chawla, "Learning decision trees for unbalanced data," in "Joint European Conference on Machine Learning and Knowledge Discovery in Databases" Springer 2008, pp. 241–256.

da Gama Silva, Paulo Vitor Jordão, Marcelo Cabus Klotzle, Antonio Carlos Figueiredo Pinto, and Leonardo Lima Gomes, "Herding behavior and contagion in the cryptocurrency market," *Journal of Behavioral and Experimental Finance*, 2019, *22*, 41–50.

Dnes, Antony W and Jonathan S Seaton, "The regulation of British Telecom: An event study," *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 1999, pp. 610–616.

Dolan, Shelagh, "How the laws & regulations affecting blockchain technology and cryptocurrencies, like Bitcoin, can impact its adoption," https://www.businessinsider.com/blockchain-cryptocurrency-regulations-us-global?r=DE&IR=T 2020. Accessed March 8 2020.

Fiorillo, Steve, "Bitcoin History: Timeline, Origins and Founder," *The Street*, 2018.

Flamary, R'emi and Nicolas Courty, "POT Python Optimal Transport library," 2017.

George, Ammu, Taojun Xie, and Joseph Alba, "Central Bank Digital Currency with Adjustable Interest Rate in Small Open Economies," Technical Working Paper, Asia Competitiveness Institute June 2020.

Geva, Tomer and Jacob Zahavi, "Empirical evaluation of an automated intraday stock recommendation system incorporating both market data and textual news," *Decision support systems*, 2014, *57*, 212–223.

Ghiassi, Manoochehr, James Skinner, and David Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," *Expert Systems with applications*, 2013, *40* (16), 6266–6282.

Glaser, Florian, Kai Zimmermann, Martin Haferkorn, Moritz Christian Weber, and Michael Siering, "Bitcoin-asset or currency? revealing users' hidden intentions," *Revealing Users' Hidden Intentions (April 15, 2014). ECIS*, 2014.

Gold, Zack and Megan McBride, "Cryptocurrency: A Primer for Policy-Makers," 2019.

Griffiths, Thomas L and Mark Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, 2004, *101* (suppl 1), 5228–5235.

Guides, Trading Strategy, "Top 10 Cryptocurrency Blogs (You Should Follow in 2019)," 2018.

Gulen, Huseyin and Mihai Ion, "Policy uncertainty and corporate investment," *The Review of Financial Studies*, 2016, *29* (3), 523–564.

Hafner, Christian, "Testing for bubbles in cryptocurrencies with time-varying volatility," *Available at SSRN 3105251*, 2018.

Hafner, Christian M, "Testing for bubbles in cryptocurrencies with time-varying volatility," *Journal of Financial Econometrics*, 2020, *18* (2), 233–249.

Härdle, Wolfgang Karl, Cathy Yi-Hsuan Chen, and Ludger Overbeck, *Applied quantitative finance*, Springer, 2017.

Houben, Robby and Alexander Snyers, *Cryptocurrencies and blockchain: Legal context and implications for financial crime, money laundering and tax evasion* 2018.

Kang, Wensheng and Ronald A Ratti, "Oil shocks, policy uncertainty and stock market return," *Journal of International Financial Markets, Institutions and Money*, 2013, *26*, 305–318.

Kim, Alisa, Simon Trimborn, and Wolfgang K Härdle, "VCRIX-A Volatility Index for Crypto-Currencies," *Available at SSRN 3480348*, 2019.

Klein, Tony, Hien Pham Thu, and Thomas Walther, "Bitcoin is not the New Gold–A comparison of volatility, correlation, and portfolio performance," *International Review of Financial Analysis*, 2018, *59*, 105–116.

Lee, David and Robert H Deng, "Handbook of blockchain, digital finance, and inclusion: Cryptocurrency, FinTech, InsurTech, and regulation," 2018.

Lee, Seonggyu, Jinho Kim, and Sung-Hyon Myaeng, "An extension of topic models for text classification: A term weighting approach," in "2015 International Conference on Big Data and Smart Computing (BIGCOMP)" IEEE 2015, pp. 217–224.

Lin, Yung-Shen, Jung-Yi Jiang, and Shie-Jue Lee, "A similarity measure for text classification and clustering," *IEEE transactions on knowledge and data engineering*, 2013, *26* (7), 1575–1590.

Linton, Marco, Ernie Gin Swee Teo, Elisabeth Bommes, CY Chen, and Wolfgang Karl Härdle, "Dynamic topic modelling for cryptocurrency community forums," in "Applied Quantitative Finance," Springer, 2017, pp. 355–372.

Mimno, David, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, "Optimizing semantic coherence in topic models," in "Proceedings of the conference on empirical methods in natural language processing" Association for Computational Linguistics 2011, pp. 262–272.

Nakamoto, Satoshi, "Bitcoin: A peer-to-peer electronic cash system," Technical Report, Manubot 2019.

_ et al., "Bitcoin: A peer-to-peer electronic cash system.(2008)," 2008.

Nassirtoussi, Arman Khadjeh, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo, "Text mining for market prediction: A systematic review," *Expert Systems with Applications*, 2014, *41* (16), 7653–7670.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin, "Automatic evaluation of topic coherence," in "Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics" Association for Computational Linguistics 2010, pp. 100–108.

Niraula, Nobal, Rajendra Banjade, Dan Ştefănescu, and Vasile Rus, "Experiments with semantic similarity measures based on lda and lsa," in "International conference on statistical language and speech processing" Springer 2013, pp. 188–199.

Prager, Robin A, "The effects of deregulating cable television: evidence from the financial markets," *Journal of Regulatory Economics*, 1992, *4* (4), 347–363.

Redondo, Connor Lamon Eric Nielsen Eric, "Cryptocurrency Price Change Prediction Using News."

Röder, Michael, Andreas Both, and Alexander Hinneburg, "Exploring the space of topic coherence measures," in "Proceedings of the eighth ACM international conference on Web search and data mining" 2015, pp. 399–408.

Schwert, G William, "Measuring the effects of regulation: evidence from the capital markets," *Journal of Law and Economics*, 1981, *24* (1), 121–58.

Sebastiani, Fabrizio, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, 2002, *34* (1), 1–47.

Teh, Yee Whye, Michael I Jordan, Matthew J Beal, and David M Blei, "Sharing clusters among related groups: Hierarchical dirichlet processes," in "Proceedings of the 17th International Conference on Neural Information Processing Systems" MIT Press 2004, pp. 1385–1392.

Tong, Simon and Daphne Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, 2001, *2* (Nov), 45–66.

Trimborn, Simon and Wolfgang Karl Härdle, "CRIX an Index for cryptocurrencies," *Journal of Empirical Finance*, 2018, *49*, 107–122.

_ , Mingyang Li, and Wolfgang K Härdle, "Investing with cryptocurrencies-A liquidity constrained investment approach," *Published version: Trimborn, S., Li, M. and WK Härdle (2019)" Investing with Cryptocurrencies—a Liquidity Constrained Investment Approach" Journal of Financial Econometrics, doi. org/10.1093/jjfinec/nbz016*, 2018.

Tu, Jun and Wolfgang Karl Härdle, "Information Arrival, News Sentiment, Volatilities and Jumps of Intraday Returns Ya Qian," 2018.

Wallach, Hanna M, David M Mimno, and Andrew McCallum, "Rethinking LDA: Why priors matter," in "Advances in neural information processing systems" 2009, pp. 1973–1981.

Zhong, Raymond, "China's Cryptocurrency Plan Has a Powerful Partner: Big Brother," https://www.nytimes.com/2019/10/18/technology/china-cryptocurrency-facebook-libra.html?_ga=2.218041619.426277731.1583491045-877433028.1583491045 2019. Accessed March 6, 2020.

Appendix

A Proofs for Hellinger Distance

A.1 Asymptotic distribution of $\sqrt{nh}[\hat{f}(x) - f(x)]$ depends on $f(x)$.

Proof Denote $\hat{f}(x)$ is a kernel density estimator,

$$\sqrt{nh}[\hat{f}(x) - f(x)] = \sqrt{nh}\{\hat{f}(x) - \mathrm{E}[\hat{f}(x)]\} + \sqrt{nh}\{\mathrm{E}[\hat{f}(x)] - f(x)\} \tag{10}$$

For the second part,

$$\begin{aligned}
\mathrm{Bias}\left\{\hat{f}_h(x)\right\} &= \mathrm{E}\left\{\hat{f}_h(x)\right\} - f(x) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{K_h\left(x - X_i\right)\right\} - f(x) \\
&= \mathrm{E}\left\{K_h(x - X)\right\} - f(x) \\
&= \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)du - f(x)
\end{aligned} \tag{11}$$

The transformation $s = \frac{u-x}{h}$, i.e. $u = hs + x$, $\left|\frac{ds}{du}\right| = \frac{1}{h}$. A second-order Taylor expansion of $f(u)$ around x is given by

$$f(x + hs) = f(x) + f(x)'hs + \frac{1}{2}f''(x)h^2s^2 + o\left(h^2\right) \tag{12}$$

Then,

$$\begin{aligned}
\text{Bias}\left\{\hat{f}_h(x)\right\} &= \int \frac{1}{h}K(-s)f(x+hs)hds - f(x) \\
&= \int K(s)[f(x) + f(x)'hs + \frac{1}{2}f''(x)h^2s^2 + o\left(h^2\right)]ds - f(x) \\
&= f(x)\int K(s)ds + f(x)'h\int sK(s)ds + \frac{1}{2}f''(x)h^2\int s^2K(s)ds - f(x) + o\left(h^2\right) \\
&= \frac{h^2}{2}f''(x)\mu_2(K) + o\left(h^2\right), \quad \text{as } h \to 0
\end{aligned} \tag{13}$$

where $\int K(s)ds = 1$, $\int sK(s)ds = 0$ and $\int s^2K(s)ds = \mu_2(K)$.

For the first part,

$$\begin{aligned}
\text{Var}\left\{\hat{f}_h(x)\right\} &= \text{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} K_h\left(x - X_i\right)\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left\{K_h\left(x - X_i\right)\right\} \\
&= \frac{1}{n}\text{Var}\left\{K_h(x - X)\right\} \\
&= \frac{1}{n}\left\{\text{E}\left[K_h^2(x-X)\right] - \left\{\text{E}\left[K_h(x-X)\right]\right\}^2\right\} \\
&= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f(t)dt - \frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{x-t}{h}\right)f(t)dt\right)^2 \\
&= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f(t)dt - \frac{1}{n}(f(x) + \text{Bias}(\hat{f}(x)))^2
\end{aligned} \tag{14}$$

Substituting $s = \frac{u-x}{h}$,

$$\text{Var}\left\{\hat{f}_h(x)\right\} = \frac{1}{nh}\int K(s)^2 f(x+hs)ds - \frac{1}{n}\left(f(x) + o\left(h\right)\right)^2 \tag{15}$$

Applying a Taylor approximation yields

$$\begin{aligned}\operatorname{Var}\left\{\hat{f}_h(x)\right\} &= \frac{1}{nh}\int K(z)^2\left(f(x)+hsf'(x)+\frac{1}{2}f''(x)h^2s^2+o(h^2)\right)ds-\frac{1}{n}\left(f(x)+o\left(h\right)\right)^2 \\ &= \frac{1}{nh}\|K\|_2^2 f(x)+o\left(\frac{1}{nh}\right), \quad \text{as } nh\to\infty\end{aligned} \tag{16}$$

where $\int K^2(s)ds = \|K\|_2^2$. With $\operatorname{Var}(\hat{f}(x)) \to 0$ as $nh \to \infty$

$$\frac{\hat{f}(x)-\mathrm{E}[\hat{f}(x)]}{\sqrt{\operatorname{Var}(\hat{f}(x))}} \xrightarrow{d} \mathcal{N}(0,1) \tag{17}$$

Substituting the expression for $\operatorname{Var}(\hat{f}(x))$

$$\sqrt{nh}\{\hat{f}(x)-\mathrm{E}[\hat{f}(x)]\} \xrightarrow{d} \mathcal{N}\left(0, f(x)\|K\|_2^2\right) \tag{18}$$

If the bandwidth tends to zero faster than the optimal rate, then

$$\sqrt{nh}\{\mathrm{E}[\hat{f}(x)]-f(x)\} \to 0 \tag{19}$$

and the bias term vanishes from the asymptotic distribution,

$$\sqrt{nh}[\hat{f}(x)-f(x)] \xrightarrow{d} \mathcal{N}\left(0, f(x)\|K\|_2^2\right) \tag{20}$$

A.2 Asymptotic distribution of $\sqrt{nh}[\sqrt{\hat{f}(x)}-\sqrt{f(x)}]$ does not depend on $f(x)$

Proof From transform theorems, we know that if If $\sqrt{n}(t-\mu) \xrightarrow{\mathcal{L}} N_p(0,\Sigma)$,

$$\sqrt{n}[f(t)-f(\mu)] \xrightarrow{\mathcal{L}} N_q\left(0, \mathcal{D}^\top\Sigma\mathcal{D}\right) \quad \text{for } n\longrightarrow\infty \tag{21}$$

Denote $g(x) = x^{1/2}$, then $\frac{dg}{dx} = \frac{1}{2}x^{-1/2}$. With $\sqrt{nh}(\hat{f}(x) - f(x)) \xrightarrow{d} \mathcal{N}(0, f(x)\|K\|_2^2)$, then

$$\sqrt{nh}[\sqrt{\hat{f}(x)} - \sqrt{f(x)}] \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{4}\|K\|_2^2\right) \tag{22}$$

The asymptotic distribution of $\sqrt{nh}[\sqrt{\hat{f}(x)} - \sqrt{f(x)}]$ does not depend on $f(x)$

IRTG 1792 Discussion Paper Series 2020

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Estimation and Determinants of Chinese Banks' Total Factor Efficiency: A New Vision Based on Unbalanced Development of Chinese Banks and Their Overall Risk" by Shiyi Chen, Wolfgang K. Härdle, Li Wang, January 2020.

002 "Service Data Analytics and Business Intelligence" by Desheng Dang Wu, Wolfgang Karl Härdle, January 2020.

003 "Structured climate financing: valuation of CDOs on inhomogeneous asset pools" by Natalie Packham, February 2020.

004 "Factorisable Multitask Quantile Regression" by Shih-Kang Chao, Wolfgang K. Härdle, Ming Yuan, February 2020.

005 "Targeting Cutsomers Under Response-Dependent Costs" by Johannes Haupt, Stefan Lessmann, March 2020.

006 "Forex exchange rate forecasting using deep recurrent neural networks" by Alexander Jakob Dautel, Wolfgang Karl Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2020.

007 "Deep Learning application for fraud detection in financial statements" by Patricia Craja, Alisa Kim, Stefan Lessmann, May 2020.

008 "Simultaneous Inference of the Partially Linear Model with a Multivariate Unknown Function" by Kun Ho Kim, Shih-Kang Chao, Wolfgang K. Härdle, May 2020.

009 "CRIX an Index for cryptocurrencies" by Simon Trimborn, Wolfgang Karl Härdle, May 2020.

010 "Kernel Estimation: the Equivalent Spline Smoothing Method" by Wolfgang K. Härdle, Michael Nussbaum, May 2020.

011 "The Effect of Control Measures on COVID-19 Transmission and Work Resumption: International Evidence" by Lina Meng, Yinggang Zhou, Ruige Zhang, Zhen Ye, Senmao Xia, Giovanni Cerulli, Carter Casady, Wolfgang K. Härdle, May 2020.

012 "On Cointegration and Cryptocurrency Dynamics" by Georg Keilbar, Yanfen Zhang, May 2020.

013 "A Machine Learning Based Regulatory Risk Index for Cryptocurrencies" by Xinwen Ni, Wolfgang Karl Härdle, Taojun Xie, August 2020.

IRTG 1792, Spandauer Strasse 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.