Visual Exploration of Blockchain News via Sentiment Index and Topic Models

Songye Han*
Zhejiang University
3160102019@zju.edu.cn

Shaojie Ye*
Zhejiang University
JaYE@zju.edu.cn

Hongxin Zhang
Zhejiang University
zhx@cad.zju.edu.cn

Abstract

Understanding, analyzing, and visualizing market sentiment has been one of the emerging issues in the blockchain-related study. Most of the previous work mainly focused on the exploration of blockchain trade and market patterns. Recently, there has been growing interests in exploring sentiment analysis of blockchain news. However, it is difficult for users to interpret such a huge amount of unstructured textual data, dig out the accurate meaning of semantics, and detect blockchain event patterns of sentiment polarity. In our work, we propose a novel visualization-driven blockchain text data semantic analysis model to cope with the major challenges mentioned above. We first collect blockchain news from open sources and encode sentences into its vectorized representation through the state-of-the-art technique BERT. Then we extract sentiment index employing the LSTM network to detect blockchain event patterns. Subsequently, we design and implement a highly-interactive visualization system to explore the blockchain news topic relationship and its development. We adopt the online LDA training and TSNE algorithm to visualize corpus topic semantics. In addition, the automatic detection of events enables users and domain experts to interactively explore the inner correlation among market semantic, topic distribution, and event patterns. Based on the framework, we develop a web-based interactive visualization system. We demonstrate the applicability as well as the effectiveness of our proposed system in case studies.

Index Terms— blockchain, semantic analysis, topic model, event detection

1. Introduction

With the increasing popularity of the digital currency market and blockchain technologies, investors and financial institutes are paying more attention to the blockchain market. The news and trade data is easier to reach because of its great attention, which lays the foundation of visualizing blockchain information visualization. However, blockchain technology is still at its early stage. Market sentiment is subject to external and internal disturbance, which makes news sentiment analysis difficult but meaningful.

Also, due to the information explosion, the need to interpret and extract core information of large amounts of data increases. Therefore, event detection, as well as other visualization tools are effective ways, for users to grasp the meaning of news data. Moreover, to grasp the general trend of the sentiment of blockchain news, pattern analysis is also needed to conduct further exploration.

In our work, we utilize machine learning models to extract latent semantics information from blockchain news. Previous work like Opinionflow [30] also focuses on extracting public opinions from social media and interpreting its development and patterns. Differently, due to the great fluctuation of the market emotion, extracting semantic information from social media is more important. Therefore, our work attempt to build a sentiment index to extract useful information from blockchain news, and we focus on blockchain event detection and pattern analysis afterward based on the sentiment index. Accordingly, we propose a novel interactive visualization system to help users like investors and domain experts better understand the sentiment of blockchain news. With these tools, they can explore event patterns of sentiment polarity and predict the trend better.

However, finding useful indicators are difficult and task-based in event detection. In our analysis, we use the blockchain news semantics as the most important indicator to detect an event. Specifically, we use the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) [6] for text encoding and Long Short Term Memory (LSTM) network [12] for downstream sentiment classification. And our case studies prove that our sentiment index is efficient in blockchain event detection.

The sentiment index is effective when reflecting general market semantics, and LSTM network is suitable to extract long-term sentiment due to its feature. However, interpreting and visualizing the mystery of deep learning models are still immature. We need to introduce topic models to help better capture the sentiment in the blockchain news. As Latent Dirichlet Allocation (LDA) [3] is the state-of-the-art model in many topic modeling cases, we use online learning for LDA to illustrate the multi-dimensional semantic topics behind the corpora.

In addition, we conduct several case studies using our real-time interactive visualization systems. We will specifically show several typical events in the blockchain domain. We discuss further improvement and sum up work at the end.

Our main contributions are as follows:

- First, our work adopts a novel method for event detection of sentiment polarity. We utilize text semantics which we get from BERT and LSTM to detect blockchain-related events.

[*] equal contribution

Moreover, our sentiment index demonstrates to be a strong indicator for event detection.

- Second, our visualization system explores blockchain news from multiple perspectives. We combine topic modeling visualization, pattern analysis, and sentiment line graph for helping users to understand text semantics comprehensively. Users are able to conduct their research with different granularity like drilling into a pipeline layer by layer.
- Third, we provide LDA topic modeling and TSNE cluster method to visualize corpus distribution pattern and its evolution pattern with different topics. We further combine our semantics and topic distribution to trace and fully unveil the inner sentiment of events.

2. Related work

This section reviews some important previous works that are related to our analysis in abstracting, visualizing, and exploring blockchain news.

2.1. Blockchain pattern analysis

Blockchain [22] allows all the network participants to reach an agreement, commonly known as consensus. All the data stored on a blockchain is recorded digitally and has a common history that is available for all the participants. This way, the chances of any fraudulent activity is eliminated without the need of a third-party. With the idea, the blockchain industry is rapidly growing nowadays. Researches towards blockchain cover a wide range of area including security, resources, usability, privacy, smart contracts, cryptocurrencies, botnets, P2P broadcasting, and trustworthiness [33].

The Skyrocketing and tubulating market of Bitcoin has caught colossal attention. Thus, most of the blockchain applications focus on studying the trade pattern of digital currency, which concentrate on the price and transactions. BitexTract [35] focuses on delving into transaction patterns between exchange-exchange and exchange-client. Also, it uses a mass sequence view to illustrate the evolution pattern of multiple exchanges, providing users with a comparison between exchanges. BitCoin View [7] seeks a reliable way to detect illegal activities in blockchain by tracking bitcoin flows and abnormal transaction logs. [25] made use of transaction data to construct a directed hyper-graph, from which transaction patterns of Bitcoin exchanges can be revealed and leveraged in fraudulent pattern mining. Other perspectives are also discussed, such as identifying the performance of the bitcoin transaction system and interpreting dynamic system behavior [19]. Our work differs from all those works that the analysis we have applied is not linked with the direct market data, but based on only open news regarding blockchain.

2.2. Text sentiment analysis

The rapid development of the Internet has brought numerous websites. Researchers can obtain overall opinions from a large amount of text without many human resources, thanks to the development of automatic opinion mining. Delving into the sentiment of text data is one of the main focuses of our work. There have been efforts to visualize text sentiment. The data cube is a way to organize data in multi-dimension and provide interactive and intuitive queries and exploration by slicing, dicing, and drilling through cube cells. [17] utilizes text cube to analyze and visualize social media semantic. It visualizes a heat map and the hotspot on the map to extract Point of Interest (POI). Other works also focus on embedding unstructured text data into structured vectors. Zhu *et al*. introduced a vector embedding technique of urban location data based on situation awareness [37]. They emphasize the continuity of trajectory data and regard trajectories as sequential data just as a word within a sentence. Besides, topic models are widely used in extracting semantics behind unstructured data such as urban trajectories [34] and news text [9].

The visualization of opinions (sentiments in our work) extracted from the text can be classified into three categories: document-level, feature-level visualization, and these two combined. Feature level visualization concentrates on details. [16] proposed a method to extract customer opinions, and augmented traditional bar charts to facilitate visual comparison of extracted feature-level data. Oelke et al. [23] introduced visual summary reports, cluster analysis, and circular correlation maps to facilitate visual analysis of customer feedback data at the feature level. Document-level visualization focuses on visualizing opinion data at the document level. Morinaga et al. [20] suggested a 2D scatterplot called a positioning map to show the group of positive or negative sentences. Chen et al. [5] presented a visual analysis system with multiple coordinated views to help users understand the nature and dynamics of conflicting opinions. OpinionSeer [31] combined those two perspectives, working on analyzing relationships among multiple data dimensions and comparing opinions of different groups.

2.3. Social event detection

Visual analysis of event detection on social media has been the subject of increasing attention from the industry or academia. Our analysis is focusing on blockchain news event detection. Before our work, most event detection-related study focuses on social event detection. StreamExplorer [29] specialized in tracking streaming social data interactively. They use current tweets volume as the only feature to detect subevent, which proves to be effective. Other works adopt a more complex method to detect events. [26] utilizes Bayesian location inference to unveil event, and [11] proposes a novel method that uses attention and LSTM network to detect abnormal event *etc*. Visual Backchannel uses a similar approach for representing dynamic tweets keywords varying in time, evolving topics reflected in social media text [8]. [36] developed the FluxFlow system for detecting and visualizing anomalous information propagation processes on Twitter. Opinionflow [30] conducted a time-oriented visual analysis tracking the procedure of the diffusion of opinions among social media users. Our work follows the idea of Opinionflow but focus on different aspect of the subject.

3. Data modeling and processing

This Section mainly introduces the whole process of our analysis and some key models we use in our system.

3.1. Data collection and feature interpretation

To avoid bias, we crawl blockchain news text from different websites instead of only one source. Also, to illustrate the general applicability of our model, we also crawl Chinese blockchain news websites such as [1] and [2] besides English websites. Next, we clean our data by filtering useless texts with concise length. We also sort out data chronologically. Here are the main features we get, which will be used in further analysis.

- Semantic Value: The semantic value of our passage which will be thoroughly introduced next.
- Time: Time when the passage published.
- Reading numbers: The page view of passage. Moreover, it well be used in event detection just as a number of tweets in a time window used in social stream event detection [29].
- Labels: The labels of the passage (might be used in further work like modifying our topic model as labeled LDA [24].)
- Author Followers: Number of fans of the author. Indicator as the power of the author.

3.2. Semantic quantization of news corpus

This section mainly introduces our process of getting text semantics. And this can be mainly divided into two steps. First, we use BERT to embed our sentences. Second, we represent each paragraph as 2-dimensional tensor and use deep learning methods to achieve text sentiment classification.

3.2.1 Tensor embedding

In upstream natural language processing tasks, one of the biggest challenges is the lack of annotated data. So a variety of techniques are developed for training general-purpose language representation models using an unlimited amount of unlabeled text on the web. Also, the upstream pre-trained models can be fine-tuned for numerous downstream NLP tasks such as question answering and text semantic classification. Among those models, Some are independent from context such as [13]. Others rely on context such as BERT [6]. BERT demonstrates the state-of-the-art performance in 11 classical NLP tasks. So it is reasonable to choose it as our model for the accuracy of our downstream analysis.

In our experiment, we do not use specific task-based variants of BERT. We only use BERT-as-Service to get static representations of sentences since fine-tuning with BERT's parameters with our task is time-consuming.

We use the BERT-Base-Chinese model, which has 12 layers, 768-hidden, and 12-head and BERT-Base-uncased for our English corpora. We divide each corpus into sentences and send each sentence to BERT-as-Service for encoding. BERT-as-Service produces a matrix of L*768D, where L is the sentence length. Then the model reduces the matrix into a 768D vector. We get a 768-Dimensional vector for each sentence. To unify the representation of each corpus, we set a sentence number N for each passage. news which have smaller number of sentences will be padded, whereas bigger will be truncated.

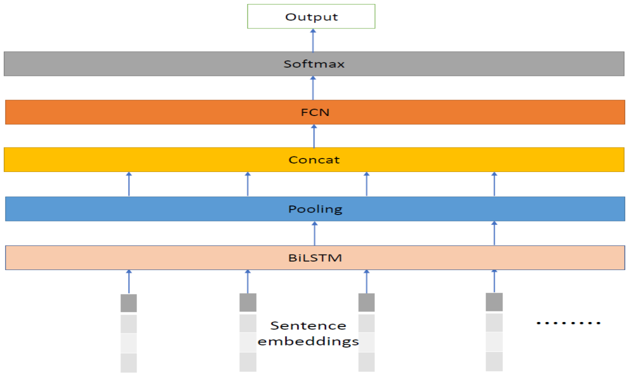


Figure 1. Network Overview.

Model	Accuracy	F1
CNN	0.85	0.84
SVM	0.82	0.86
Random Forest	0.85	0.90
Adaboost	0.80	0.83
LSTM	0.91	0.93

Table 1. Model Evaluation.

[a] We divide our dataset to the proportion 7:2:1 for train, validation, test respectively to optimize the hyperparameter.

[b] After optimizing hyperparameters in each mode, we randomly split our dataset to 7:3 proportion to train and test set for 50 times, and the accuracy and F1 shown above is the result of averaging.

3.2.2 Sentiment index

Next, we finish our downstream text semantic analysis using the tensor we mentioned above. Since BERT encodes our documents based on sentence-level, our downstream models should focus on extracting the relationship between sentences in a corpus. Since no available open related data is for us, we labeled 2052 passages to positive sentiment or negative sentiment.

Before trying with LSTM [12], we also experiment with other models such as Adaboost, Support Vector Machine (SVM), Random Forest (RF) and Convolutional Neural Network (CNN). In building our CNN model, we utilize 2-D convolution and multiple kernel size to extract local sequence information. Different kernel size can be seen as different time step length. So we get a list of models and outputs from different time steps. We concatenate the outputs to feed to the full connection layer. The method we use is learned from [32]. As to model evaluation, we split our data into three parts: Train set, validation set, and test set. Using a grid search to optimize each model's hyperparameters, including learning rate, batch size, epoch, max number of leaves, and regularization penalty coefficients. Especially for the CNN model, We also set our hidden size by conducting a pre-experiment on the dataset. The experiment shows that the network with a hidden size of 2 far outperforms the CNN network with the other number of layers. According to this, we set the hidden size to 2. We use accuracy and F1-score to evaluate the performance of our model, The result is shown in Table 1:

$$F1 = \frac{2 * precision * recall}{precision + recall}. \tag{1}$$

However, CNN is more frequently used in computer vision, such as 2-D image classification. LSTM might be more suitable in most cases when dealing with sequence data. Also, CNN has to pad each passage with the same dimension. Thus many short passages will send useless messages to the network, which will likely affect the performance of CNN. Therefore, we choose the most frequently used deep learning model in NLP: Long Short Term Memory Network. It is better at learning long-sequence dependence and can use pack techniques to deal with different sentence lengths in batch training.

We utilize two-layer stacked bidirectional LSTM (BiLSTM) [21]. Bidirectional allows us to learn both leftward and rightward contextual information and then concatenate the leftward and rightward hidden state and then combine the corresponding hidden state as h$_{output}$ = [h$_L$,h$_R$] . Besides, we use 2-layer BiLSTM based on an experiment. Table 1 is the comparison of our downstream models.

Assuming our LSTM network's hidden layer's dimension is H, the current paragraph's sentence number is L. After BERT's sentence encoding, we get an L*768 tensor. Then we use the pad and pack method to uniform the input data dimension for batch training. After our 2-layer stacked bidirectional LSTM, we get a $L*2H$ tensor for each passage. We then put a pooling layer and full connection layer afterward to get our sentiment. Figure 1 shows our downstream network in a more intuitive sense. After the output of the softmax layer, we also get each passage's sentiment label. We use label*O$_{softmax}$ as our raw sentiment. (Here label is -1 for negative 1 for positive, since the output of softmax Layer is always the bigger positive real numbers. Modifying the negative label to -1 can more accurately reflect the sentiment index. While in training, we use 0 for negative 1 for positive. And O$_{softmax}$ is the max element of softmax layer.) Finally, we use z-score normalization to smooth the distribution of our raw sentiment:

$$S = \frac{S_{raw} - \mu}{\sigma}. \tag{2}$$

where S_raw denotes the raw sentiment, and S denotes the processed sentiment.

3.3. Analyzing topic models

For the need for text sentiment visualization and multiple semantics interpretation, topic modeling is included. In our topic analysis, we use traditional LDA [3] to extract topics. It is a bag-of-words model, so we have to participle first. We use jieba for our Chinese corpora and NLTK for our English corpora. We filter out stopwords and also add blockchain domain terminology to the dictionary, such as Bitcoin, Dapp *etc*. And in English corpora, we also stem the uncased word.

We utilize online LDA traning method [14] which is less computationally expensive compared to the traditional. In [14], an online varational Bayes algorithm is developed instead of variational Bayes algorithm used in traditional method. It uses a stochastic adaptive strategy to update topic posterior parameter:

$$\lambda \leftarrow (1 - \rho_t) * \lambda + \rho_t * \tilde{\lambda}. \tag{3}$$

Where $\tilde{\lambda}$ is calculated by mini-batch to reduce noise and ρ_t is similar to Simulated annealing algorithm's weight [28]. In particular:

$$\rho_t \triangleq (\tau + t)^{-\kappa}. \tag{4}$$

However, choosing LDA models have been quite difficult so far and no decisive method is acknowledged. In online LDA model training, the best metrics to evaluate how well the model fits might be log perplexity:

$$P(n_i^{test}, \lambda, \alpha) \triangleq \exp\{\sum_i \log p(n_i^{test} \mid (\alpha, \beta))/\sum_{i,w} n_{iw}^{test}\}. \tag{5}$$

Where n_i^{test} denotes the vector of word counts for the ith document. α and β is the posterior parameters of the LDA model.

However, it is still hard to find the optimal model for our analysis. Log perplexity mainly measures the model's likelihood function. Since our data contains 70 thousand corpora, so there should be no overfitting. Thus models with more topic numbers would likely fit better. The results seem meaningless for our analysis because too many topics may cause memory overload for our users. Instead, other metrics such as AIC and BIC can help balance the simplicity of a model and its performance. However, we use visualization methodology: LDAvis [27] to explore the feasibility of our LDA model, Since our original goal is to let our end-users to decide which model to choose intuitively. Also, topic model selection is not a clear-cut issue. A higher loss does not necessarily mean the model is suitable for the event analysis process. A visualization of the model itself, including each topic's most related terms, each topic's relative relationship between each other, would be more acceptable for non-tech users. And it can customize the process of topic model selection. This visualization technique will be further discussed in Section 4.1. Also, we develop a user interface to let users decide which model they want.

Another issue is whether or not using TF-IDF to measure the word frequency. In most case, it is not necessary to use TF-IDF matrix to measure the frequency of words because LDA is developed to lessen the shortcomings of the TF-IDF method. Also our LDAvis shows that models with TF-IDF produce some strange words and topics when the topic number is relatively high(usually greater than 10). So we do not use TF-IDF method in our model training.

3.4. Event detection

The last part of our analysis is event detection, which is also the endpoint of our system. Since the definition of an event is often vague, and there is no open data of labeled blockchain events for us to conduct supervised learning, we provide our event detection for exploratory purposes. Since the main focus of our analysis is to use the sentiment index to extract and detect abnormal points, we do not elaborate on our method by adding additional neural networks or sophisticated algorithms. We use Bollinger band [4], which is frequently used in quantitative finance to seek trading opportunities. Later in our work, it will be shown that it is excellent at detecting events when dealing with smoother data.

Also, since some users may be uncertain about our sentiment index, as social event detection often uses tweets volume in a window, we deploy reading numbers R as additional features to satisfy personal preference. We provide a parameter β for users to adjust and get a synthetic feature called power to detect event:

$$power = \beta * \mid S \mid + (1 - \beta) * R. \tag{6}$$

Where S, R denote the sentiment index, reading number respectively.

The reading number and semantics are all scaled using z-score normalization. And we use feature power to build our Bollinger band:

$$\begin{aligned} BOLU &= MA(Power, n) + b * \sigma(Power, n). \\ BOLD &= MA(Power, n) - b * \sigma(Power, n). \end{aligned} \tag{7}$$

Where $BOLU, BOLD$ denote the upper and lower boundary of a Bollinger Band, $MA(Power, n)$, $\sigma(Power, n)$ denotes the moving average and standard deviation of the n Power indexes within the chosen time window. b is the bandwidth, which users can set by themselves.

We give a window length $\frac{n}{2}$ corresponding to time granularity for users to choose from. Since reading numbers and the absolute value of our semantics both have Positive correlation to the probability of event occurrence, we only use the upper bound of the Bollinger band as our standard for choosing an event. More specific discussions are presented at Section 4.2.1 's event detection part.

4. Visualization

This Section mainly introduces the structure and function of our visualization system and its user interface. Also, we will discuss several tasks and focuses on our visual analysis using the system.

4.1. Visual analysis

To design the visualization system, we ask some advice from professors in computer science and blockchain experts. Eventually, the following tasks and design goals are proposed after several discussions.

T1. Multi-phase event exploration

T1.1 News events often incorporate multiple subevents. Therefore our system should allow users to keep drilling into the interior of an event layer by layer and keep tracking the sentiment of an event in different time granularities to get closer to the truth.

T1.2 To make our exploration more interactive, our system should highlight critical events as suggestions and provide timeline structure to unveil the inner order of subevents inside an event. And auxiliary tools to represent text meaning is also required.

T2. Event Timeline analysis

T2.1 A Timeline analysis is also needed because of the rapid acceleration of blockchain technology and the unstable market semantic. Therefore, our system should enable end-users to understand the development of an event and reflect the entry point, polarities as well as the endpoint of an event. Thus we include a sentiment line graph to visualize the development of an event.

T2.2 Moreover, during the exploration, change of topic composition inside an event should be presented, which can provide an understanding of the events from a development perspective.

T3. Pattern analysis

- Events usually have patterns. Our design goal is to visualize the underlying patterns of the subevents inside an event during the development of an event. We should visualize the relationships between each subevents within an event. And with T1 and T2, our users can focus on the polarities, entry point, endpoint or other special points' patterns to get a more comprehensive understanding of an event.

4.2. System design

Our visualization system consists of five parts:

- Timeline visualization of the sentiment index.
- Visual tree graph.
- LDA model visualiation.
- Wordcloud.
- Topic document map.

Our visualization system is built using D3.js, which is interactive, fast and functional. All of our sub-systems provide our users with multi-stage exploration. Since BERT encoding is computationally expensive, we store the pre-trained semantics in our database and implement detached timeline visualization.

4.2.1 Timeline visualization of sentiment index

As suggested by experts and our end-users, a sentiment line should be displayed to show the trend of our sentiment index. It is shown in Figure 2.

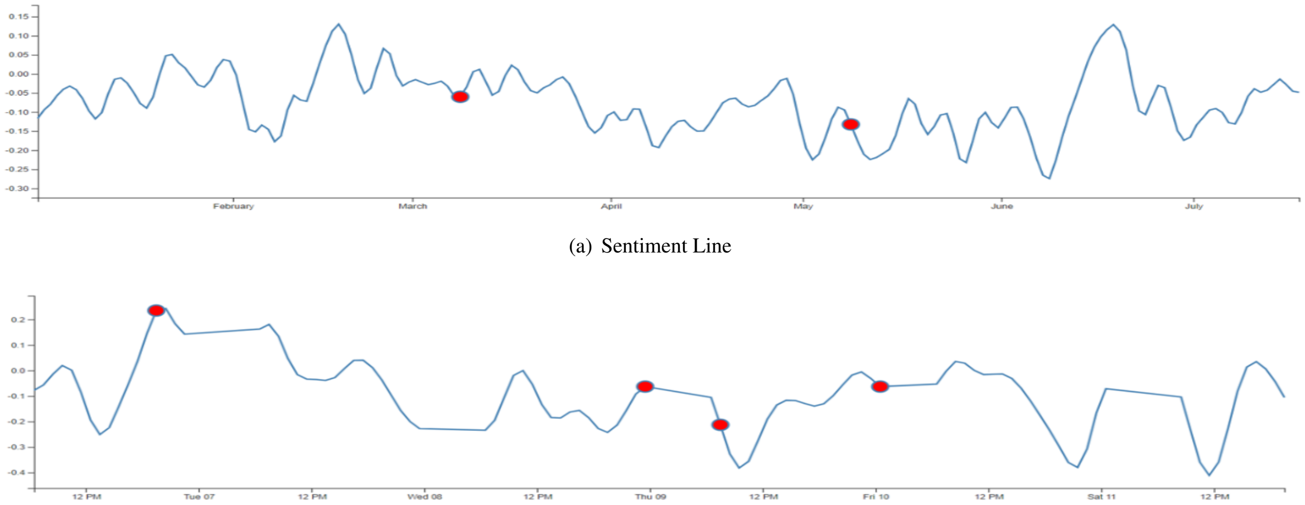
Aggregation and smoothing We utilize our scaled semantic from text semantic analysis proposed in Section 3.2.2. Also, since our data contains about $70,000$ corpora, aggregation, and smoothing methods are required to visualize our sentiment line more intuitively.

We first aggregate our data by time granularity. We use the average value as the results of aggregation. And granularity is chosen by our end-users. In the multi-phase event exploration process, our system initializes a suitable granularity for visualization, and granularity will be reduced by one level automatically when our users drill into the inner subevents of an event.

Next, we smooth our aggregated data. Currently, there are many smoothing methods widely used, such as linear smoothing [15], local polynomial smoothing and spline methods [10]. Learning from binomial smoothing [18], we use simple combination number as our smoothing weights:

$$w_k = \frac{C_h^k}{\sum_{i=1}^{h} C_h^i}, \quad k = 1, 2, ...h. \tag{7}$$

Where h is the window size, which is an odd number to make our smoothing weights symmetric. So $\frac{h+1}{2}$ is the center of the window, and the point smoothed. So every point's semantic value is calculated by:



(a) Sentiment Line

(b) Sentiment Line after click

Figure 2. Sentiment Line

$$semantic_k = \sum_{i=k-(h-1)/2}^{k+(h-1)/2} w_i semantic_i. \tag{8}$$

Where w_i is the smoothing weights defined in Equation 7, and the window size is chosen by end-users, though our system will initialize a generally feasible value.

Event detection Notice the red dots on Figure 2(a), our system automatically detect events in the chosen time range. And give red dots on the timeline visualization as a hint to users. Moreover, users can click the red dots to further explore the text, sentiment as well as topic semantics in the range of current event point. The next phase will be based on the point in the window of the current event point (**T1.1,T1.2**).

For instance, when users click the red dots in Figure 2(a), since the current time window is 5, and granularity is day. When users click one of the red dots in the line graph, next-stage exploration will be triggered. And all the points within the time window of the center point will be selected. The current time granularity of Figure 2(a) is day. So five days within the window will be selected and all the data will be aggregated by hour (granularity downgrades 1) as presented in Figure 2(b). Also, new red dots will be detected in Figure 2(b). And our end-users could click the event point dot further exploration until there is no event detected or the time granularity is down to the most granular granularity (minute).

4.2.2 Visual tree graph and wordcloud

In this part, we introduce two auxiliary tools, which help users to explore their interests more efficiently.

Visual tree graph Visual Tree graph aid users in keeping track of the current timeline and drilling into the subevents of the current chosen event. Figure 5 shows our tree graph visualization for the event node. Our tree visualization will be displayed if end-users click the red dots suggested by our system in the sentiment line. And all the time within the time window will be selected as child nodes. Also, users could click further to explore the interior feature of a parent node. And the inner node will be divided according to the next-level granularity. And when granularity is down to minute, accurate report time will be displayed since minute is the most subtle granularity (**T2.1**).

Besides, each node is composed of a donut chart characterized by three features: Semantics, reading numbers and author followers. All of the features are scaled and range from -1 to 1. Red and green color represents positive and negative value respectively. And a parent node feature is computed by the average of its child nodes. **Wordcloud** Also, our system provide users with wordcloud to visualize word frequencies in the corpora within the chosen time range. It has been proved to be a useful tool for text semantic analysis and extraction. It is shown in Figure 5.

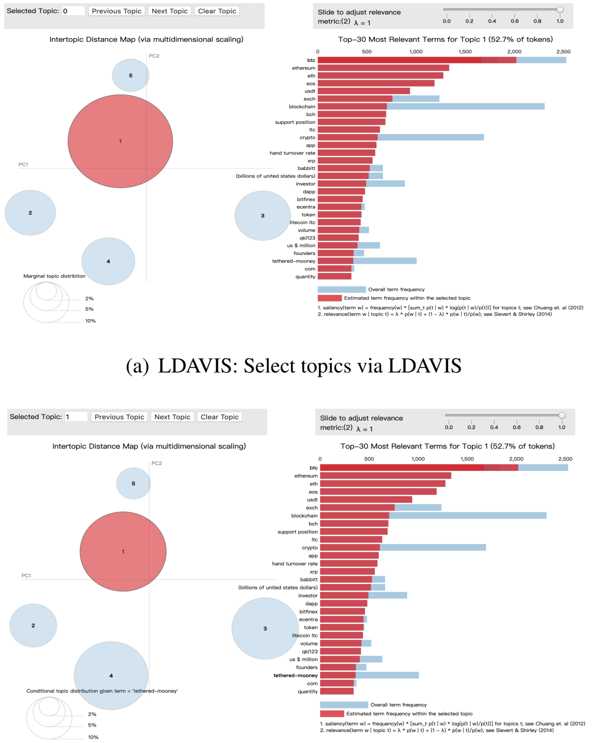
4.2.3 Topic model visualization

Also, we deploy a visualization of the chosen LDA model for our end-users by LDAvis [27], a web-based visualization for estimated chosen LDA topics. In LDAvis, the relationship between topics and terms are evaluated by relevance:

$$r(w, k \mid \lambda) = \lambda \log(\phi_{kw}) + (1-\lambda) \log(\frac{\phi_{kw}}{p_w}). \tag{9}$$

Where ϕ_{kw} denotes the probability of term $w \in \{1, 2, ..., V\}$ for topic $k \in \{1, 2, ..., K\}$. V is the total number of terms, and K is the topic number. p_w denotes the marginal probability of term w in the corpus. λ is the free weight for users to adjust. Also, we can choose a topic on the left panel. In Figure 3(a), the top-30 relevant term for the chosen topic is shown in the right panel. The width of the red bar of a term denotes the term frequency within the chosen topic, whereas grey denotes the overall frequency of the term. For simplicity of topic interpretation, we set $\lambda = 1$ first, which makes relevance totally denotes the term frequency within the topic.

On the left panel, two main features of topics are quantified. First, the size of a topic is proportional to its prevalence in all



Selected Topic: 0 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 − λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(a) LDAVIS: Select topics via LDAVIS

Selected Topic: 1 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:(2) λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 − λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Intertopic Distance Map (via multidimensional scaling)
Top−30 Most Relevant Terms for Topic 1 (52.7% of tokens)
Conditional topic distribution given term = 'tethered−mooney'
2%
5%
10%

(b) LDAVIS: Select term via LDAVIS

Figure 3. Topic model visualization

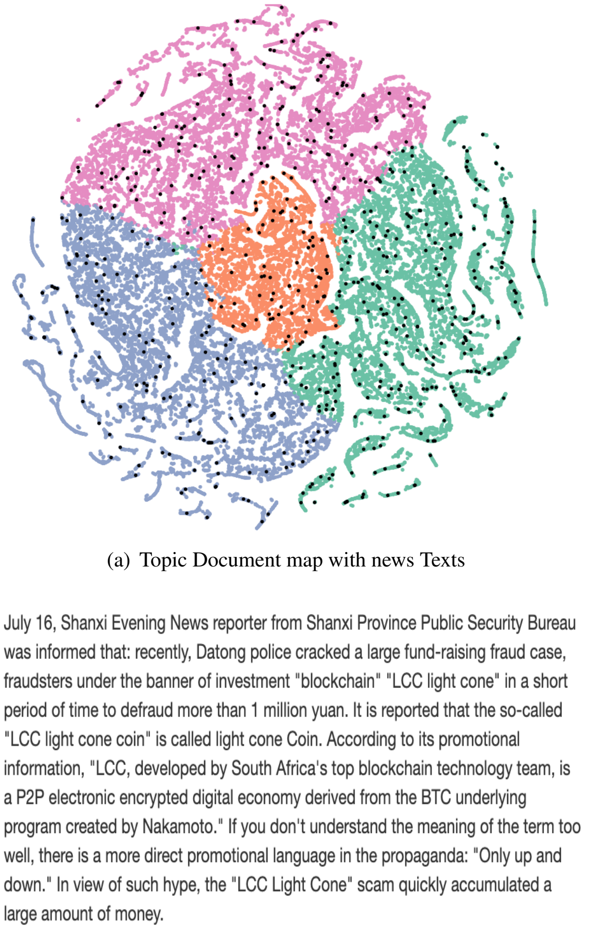
texts. Also, the relative position among all topics is computed by the Jensen-Shannon divergence. And the principal component analysis algorithm is also used to scale the inter-topic distances.

Users can also choose a term within a topic to reveal its conditional distribution among all topics. For example, in Figure 3(b), the term "tethered-mooney" is chosen, and the corresponding size of each topic circle changes, representing the actual weight of that term. Also, when topic numbers are comparatively large, clicking by term can indicate topic clusters, though it is not our primary focus in the study.

In a word, web-based LDAvis is incorporated into our system for LDA model interpretation. Users can judge the interpretability of their chosen model by looking at each topic's most related term, and the relative position and distance between each other. In our analysis, we mainly evaluate a topic model's feasibility by looking into the overlap among all topics and every topic's most relevant term.

4.2.4 Topic document map

For further exploration of pattern analysis, event detection, and topic composition visualization, we develop a document visualization map. We first utilize the document-topic weights in our topic model to vectorize each document as a K-dimensional vector. Afterward, a dimension reduction method is required to visualize each document. We use the TSNE algorithm in our analysis. It uses T-distribution to model probability matrix in low dimensional space. And it is a nonlinear reduction method, which is suitable for nonlinear data or pattern. However, it is computationally expensive. Therefore it is challenging to use the TSNE algorithm for interactive visualization. In our analysis, we use an offline method. We first store the LDA model frequently (topic number below 10) asked by our users, and train the TSNE model to get and store coordinates for different topic number(using corpus-topic weight).



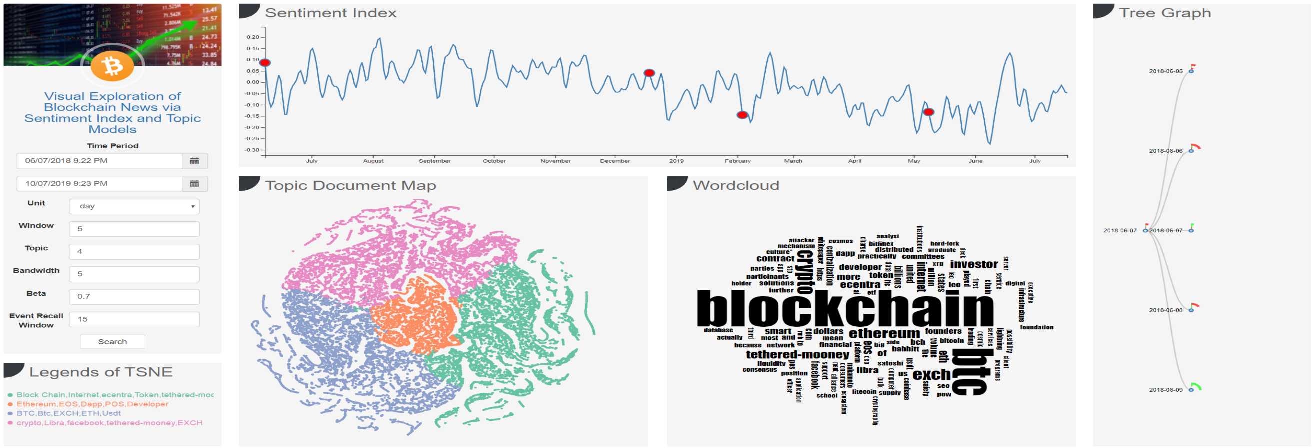
(a) Topic Document map with news Texts

July 16, Shanxi Evening News reporter from Shanxi Province Public Security Bureau was informed that: recently, Datong police cracked a large fund-raising fraud case, fraudsters under the banner of investment "blockchain" "LCC light cone" in a short period of time to defraud more than 1 million yuan. It is reported that the so-called "LCC light cone coin" is called light cone Coin. According to its promotional information, "LCC, developed by South Africa's top blockchain technology team, is a P2P electronic encrypted digital economy derived from the BTC underlying program created by Nakamoto." If you don't understand the meaning of the term too well, there is a more direct promotional language in the propaganda: "Only up and down." In view of such hype, the "LCC Light Cone" scam quickly accumulated a large amount of money.

(b) Topic Document map with event indication

Figure 4. Topic Document Map: (a) shows the map label event as black points in the view when users hover on the red dots on the sentiment line graph mentioned above. (b) shows that our topic document map can show texts when users click on the dots in the scatter plot.

Also, our topic document map includes text data source, as is shown in 4(b), which helps users to track text directly and facilitates their research. Moreover, when the mouse hovers on the red dots in the sentiment line graph, the corresponding event point will be displayed as black points in the topic document map, as shown in Figure 4(a). Therefore, our end-users can track events by the data source and explore topic distribution as well as event patterns by looking into the suggested event points in the topic document map. (**T3**)



Visual Exploration of Blockchain News via Sentiment Index and Topic Models

Time Period

06/07/2018 9:22 PM

10/07/2019 9:23 PM

Unit: day

Window: 5

Topic: 4

Bandwidth: 5

Beta: 0.7

Event Recall Window: 15

Legends of TSNE

- Block Chain,Internet,ecentra,Token,tethered-moc
- Ethereum,EOS,Dapp,POS,Developer
- BTC,Btc,EXCH,ETH,Usdt
- crypto,Libra,facebook,tethered-mooney,EXCH

Figure 5. System Overview: Our system is mainly composed of five parts: A sentiment line graph, a Tree visualizaion graph, a wordcloud graph, a topic document map, and a LDAvis on the ”LDA” channel.

4.3. User interface

As is shown in Figure 5, our system incorporates five parts. And each part co-work interactively. First, users give their interested time range to our system. And our system automatically chooses granularity depending on the number of points within the time range given by users. Then, the entrance of our system, which is the red dots auto-detected by our event detection method. Meanwhile, the topic document map visualizes all the corpora in 2-D space, and wordcloud visualize the frequency of hot words in those same corpora. Our end-users can hang their mouse on the red dots, correspondingly all the documents within the window of the hovered red dots are labeled as black points in the topic document map below.

Worth paying attention, our system also provides users several parameters to choose from. Users can choose a topic number by the LDAvis comparing different models' composition and interpretation. After choosing the time-related argument as mentioned above, users should try to adjust their optimal event-related argument. Bandwidth is the width of the Bollinger band introduced in section 3.4, $beta$ is the free weight reflecting personal preferences between reading numbers and semantics, event recall window is the traceback window used in Bollinger band.

For further microscopic analysis, users could point the red dots to enter the next phase of analysis. Our system can automatically give the time range and granularity. Though users could adjust those time-related features after the first stage, we do not suggest doing so. After red dots are clicked, wordcloud correspondingly update visualization of the new corpora, and tree graph initializes and show the time and feature of clicked points as a parent node and the corresponding time divided by granularity within the window as child nodes. Besides, the topic document map also changes its visualization with the newly chosen corpora. (**T2.2**) And events are detected as red dots in the line graph again as the entrance of the next-stage analysis. And when the time unit is down to the most subtle(minute), or no events are detected, the analysis reaches its endpoint. (**T1.2**)

5. Case study

This Section presents several case studies according to the goals and tasks of analysis proposed in section 4.1. It is worth noting that our selected cases are the results of our end-users' discovery using our visualization system. And we select several typical and persuasive cases to demonstrate the effectiveness of our system.

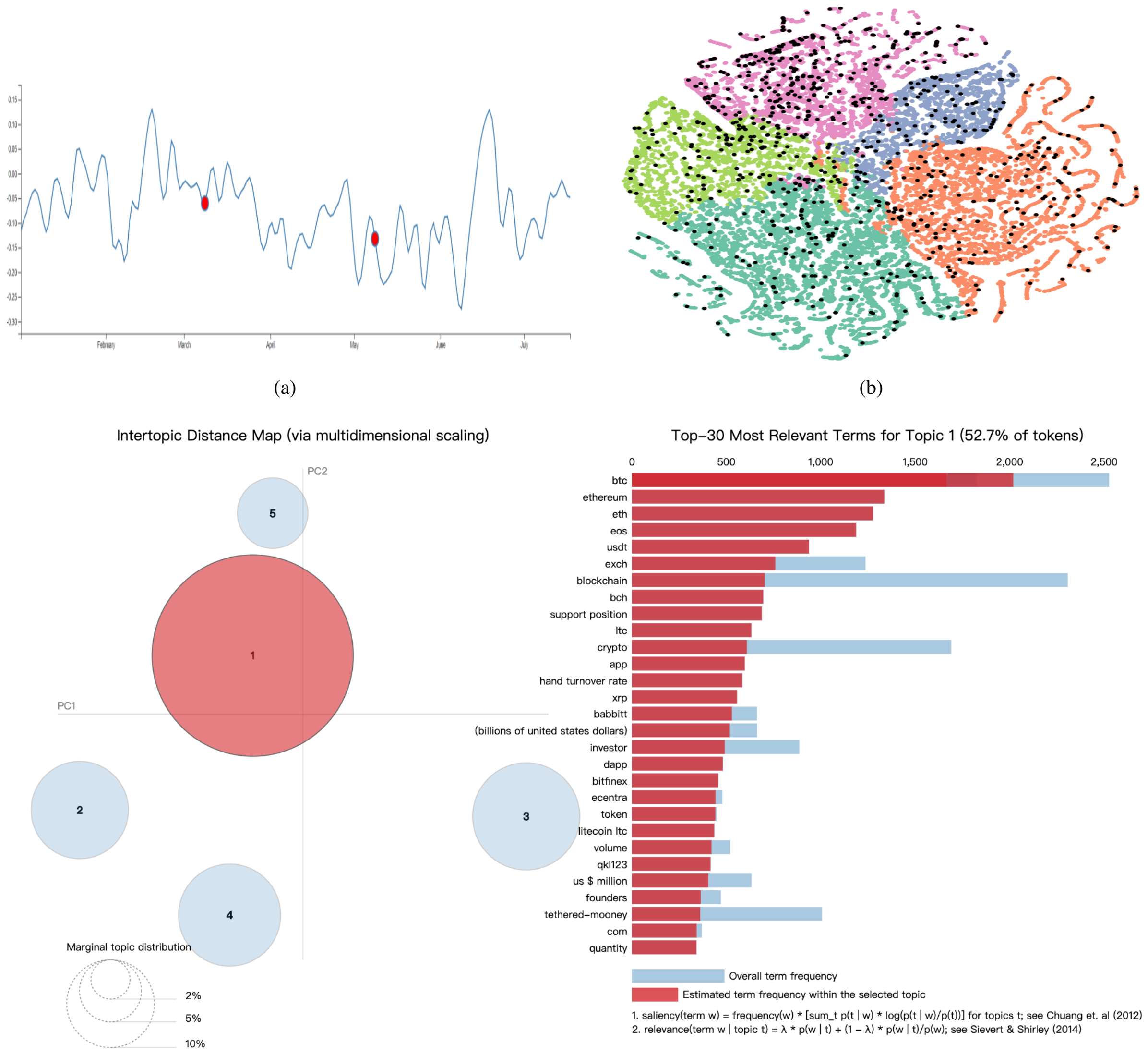
5.1. Hack attacking Ethereum

We first present a case study about hack attacking. One of our test users is interested in the market this year. Therefore he chooses the time from January 2019 to the present. Since the longevity of the chosen time range is about hundreds of days, Our systems automatically set the time granularity as day. The initialized sentiment line is shown in Figure 6(a). Besides, the bandwidth is set large to prevent too many events detected for peace of mind. We notice from Figure 6(b) has a biased topic distribution toward the pink-colored topic, which mainly represents different types of cryptocurrencies and exchanges according to the LDAvis 6(c). Compared to another suggested event point's topic distribution, the right dot is preferred because of its lower entropy.

And the user steps into the next stage of exploration. By clicking the red dots on the sentiment line, the user enters into the interior of the chosen events. Figure 7 is the interior sentiment development of the chosen event. We can see from the newly detected subevents also labels the turning point or polarity of the whole events.

The market semantic first increased to the highest point labeled as a subevent, and then drastically went down, fluctuating afterward.

Hovering on the red dots comparatively, we can see from Figure 8 that the distribution concentrates on the pink-colored topic first, scatters to other topics afterward. It is worth noting that the pink-colored topic has high correlations with different types of digital currencies including ETH. We click by each red dot to debunk what happened chronologically. First, the subevent which locates in the highest point in 7, is around 8 Am May 6, 2019, and



(a)

(b)

(c)

Figure 6. The Hacking Case: This figure presents the topic distribution difference between the detected two events. And c represents the topic interpretation.

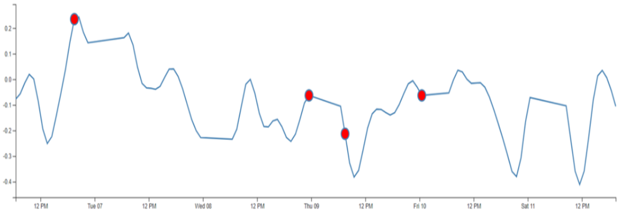


Figure 7. The Hacking case: Phase2 Sentiment Line

the topic is mainly pink-colored. Further reading on the texts, the news was excited about the prediction of a coming bull market. Therefore the semantics around the time window was abnormally high. Entering the next red dots, after experiencing the drastic decline, we can see the texts inside the point. Though the topics are scattered, a large portion of the news was reporting that ETH was attacked by some hackers in the early morning on May 8, 2019. Different reports provide different views about ETH being stolen which may explain the diversity of themes: technologies such as hard fork and Pos mechanism, the effect on Bitcoin market and the measures taken to stop the significant loss. Moving on further, we can see the focus shift to other issues like bitfinex and tether's league dispute, the warm of the Bitcoin market, and the latent danger of the Dex system. After May 10, little news was still focusing on ETH being stolen, and topics are more disorderly distributed, which symbolizes the endpoint of the event.

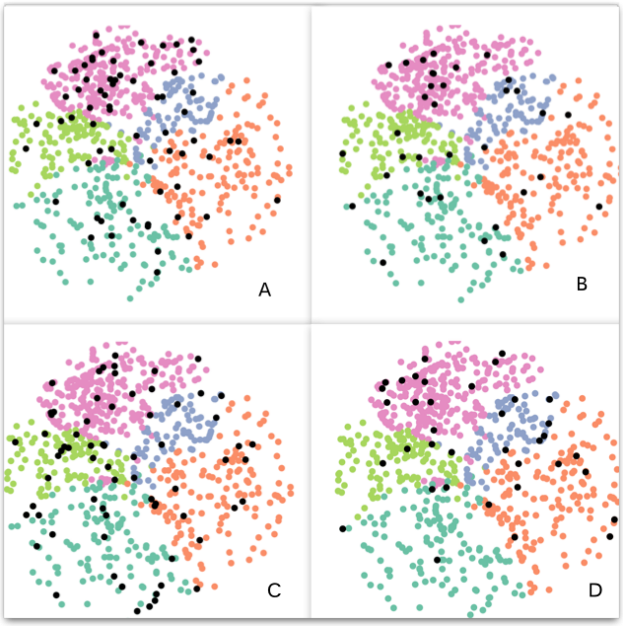


Figure 8. The Hacking case: Topic distribution changes inside the event: The distribution shifts in alphabetical order.

5.2. Facebook releasing Libra

This case shows the usefulness of our system in detecting new events, multi-phase extraction of an event and trend analysis of an event.

Besides, one of our users is interested in the publishing of Libra by Facebook. So a relatively short period is chosen near this event. It is From June 2019 to August 2019.

Figure 9(a) shows that during the chosen period, the cyan-colored topic is more dominant than before. According to LDAvis, cyan-labeled topic is highly related to Libra-related terms. Also, the word ”Libra” and ”Facebook” are more significant in the Wordcloud view. According to the sentiment line graph, the semantics went high around June 17, which may indicate the entry point of the whole event. Also an event is detected by our system around the peak. Hence the left red dot is more critical for exploring the event.

Entering further into the event, we can see from Figure 9(b) that the semantic went to the highest point around June 17. And event-related words such as Libra, Facebook, Ethereum become dominated in the wordcloud view in Figure 9(b). Further exploring the topic document map, still most of the events are distributed in the cyan-colored topic, which mainly consists of Libra-related terms according to LDAvis. Unveiling the texts around the peak, which is shown in 9(c), most of the news was claiming that the releasing of Libra was an unprecedented milestone of cryptocurrency technology and was about to transform the whole industry chain. Also the cyan-related topics become more dominent in the document topic map, so as related terms in the wordcloud. The semantic began fluctuating after the highest point. Clicking the red dots and viewing the news texts, it seems that the market became more uncertain about the releasing of Libra. Some still held the previous positive attitude, but other negative mouths became more prevalent than before: the government’s antitrust policy against Libra, the disputes about the releasing wallet and whitepaper from Facebook, and the underlying fraud behavior it may bring about.

After around June 22, though there are still quite a lot of derived dispute such as the impact on other digital currencies’ market and Paypal’s cooperation with Libra. But cyan-colored topic and Libra-related terms become less significant, which may indicates the end of the event.

6. Future work

The results of several case study proves that our blockchain news visualization system is applicable for users and experts to find blockchain news event and explore semantic behind news texts interactively and efficiently. However, with the drawback from our test users and experts, we realize that our work has some downsides.

First, since we can get labeled news text, Labeled LDA model might be able to better interpret the underlying topic semantics inside corpora. Furthermore, labeled LDA model can be employed to categorize each corpus by the type of currency, which facilitates single currency analysis.

Besides, the performance of our event detection algorithm has no guarantees. Since there are little blockchain news data labeled with events. We can not implement supervised learning to implement model selection and parameter optimization. Also, the event itself can not be strictly defined, which makes the evaluation of our work difficult. Moreover, some events might be simply classified as positive or negative. Our event detection algorithm might not be able to detect event with lots of controversies since we use a one-dimensional sentiment index. Therefore, we shall try to incorporate our topic weights as indicators for event detection in our future work.

7. Conclusion

This study adopts a novel framework for news event detection, text sentiment analysis, and topic modeling visualization. Besides, a multi-stage interactive visualization system is designed for users to drill into the pattern of events. In the future, we will optimize our system to conduct higher performance. Also, we will improve our topic models and enable multiple choice of cryptocurrency type for users to specify in their interested domains.

8. Declaration

8.1. Availability of data and materials

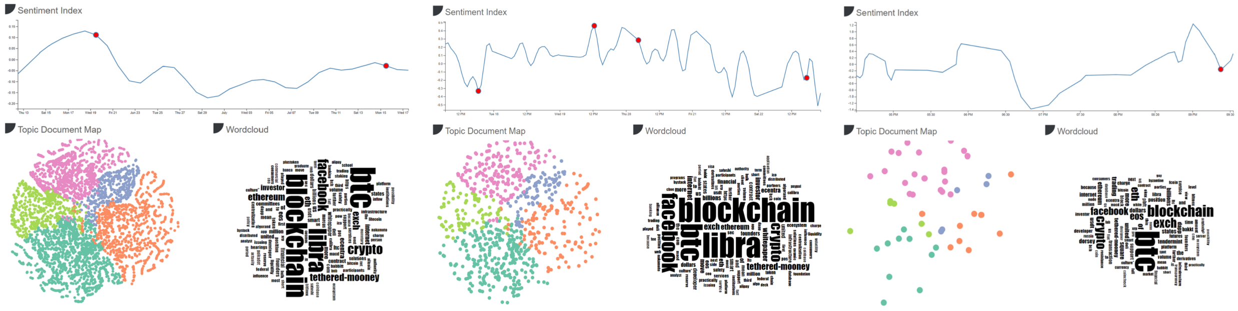
We have included our discussion on the availability of data and materials in section 3.1.

8.2. Competing interests

The authors declare no conflict of interest.

8.3. Funding

This work was supported by the National Key Research and Development Project of China (No. 2017YFC0804401) and the National Natural Science Foundation of China (No. U1909204).



(a) P 1

(b) P 2

(c) P 3

Figure 9. The Libra Case: a,b,c includes sentiment index, topic document map and wordcloud in the first phase, the second phase and the third phase respectively.

8.4. Authors' contributions

Conceptualization, S.H., S.Y., and H.Z.; methodology, S.H.; software, S.Y. and S.H.; formal analysis, S.H.; data visualization, S.H and S.Y.; model research, S.H.; writing---original draft preparation, S.H. and S.Y.; writing--- review and editing, S.H., S.Y. and H.Z.; supervision, H.Z.; funding acquisitions, H.Z. All authors have read and agreed to the published version of the manuscript.

8.5. Acknowledgements

The work was supported by Wei Chen, who provided suggestions on how to build an interactive visualization system structure, and Liyan Liu, who provided us valuable ideas on how to conduct standardized tests for our system.

References

[1] https://www.8btc.com/. 3

[2] http://www.bitcoin86.com/. 3

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 1, 4

[4] J. Bollinger. Using bollinger bands. *Stocks & Commodities*, 10(2):47–51, 1992. 4

[5] C. Chen, F. Ibekwe-SanJuan, E. SanJuan, and C. Weaver. Visual analysis of conflicting opinions. In *2006 IEEE Symposium On Visual Analytics Science And Technology*, pages 59–66. IEEE, 2006. 2

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 3

[7] G. Di Battista, V. Di Donato, M. Patrignani, M. Pizzonia, V. Roselli, and R. Tamassia. Bitconeview: visualization of flows in the bitcoin transaction graph. In *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)*, pages 1–8. IEEE, 2015. 2

[8] M. Dörk, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. *IEEE transactions on visualization and computer graphics*, 16(6):1129–1138, 2010. 2

[9] S. Doumit and A. Minai. Online news media bias analysis using an lda-nlp approach. In *International Conference on Complex Systems*, 2011. 2

[10] R. L. Eubank. *Nonparametric regression and spline smoothing*. CRC press, 1999. 5

[11] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural networks*, 108:466–478, 2018. 2

[12] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999. 1, 3

[13] Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014. 3

[14] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010. 4

[15] T. Kailath and P. Frost. An innovations approach to least-squares estimation–part ii: Linear smoothing in additive white noise. *IEEE Transactions on Automatic Control*, 13(6):655–660, 1968. 5

[16] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, 2005. 2

[17] X. Liu, K. Tang, J. Hancock, J. Han, M. Song, R. Xu, and B. Pokorny. A text cube approach to human, social and cultural behavior in the twitter stream. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 321–330. Springer, 2013. 2

[18] P. Marchand and L. Marmet. Binomial smoothing filter: A way to avoid some pitfalls of least-squares polynomial smoothing. *Review of scientific instruments*, 54(8):1034–1041, 1983. 5

[19] D. McGinn, D. McIlwraith, and Y. Guo. Towards open data blockchain analytics: a bitcoin perspective. *Royal Society open science*, 5(8):180298, 2018. 2

[20] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 341–349, 2002. 2

[21] A. Mousa and B. Schuller. Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis. In *Proceedings of the 15th Conference of the European Chapter*

of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1023–1032, 2017. 4

[22] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Technical report, Manubot, 2019. 2

[23] D. Oelke, M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 187–194. IEEE, 2009. 2

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009. 3

[25] S. Ranshous, C. A. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters. Exchange pattern mining in the bitcoin transaction directed hypergraph. In *International Conference on Financial Cryptography and Data Security*, pages 248–263. Springer, 2017. 2

[26] T. Reuter, S. Papadopoulos, G. Petkos, V. Mezaris, Y. Kompatsiaris, P. Cimiano, C. de Vries, and S. Geva. Social event detection at mediaeval 2013: Challenges, datasets, and evaluation. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013*, 2013. 2

[27] C. Sievert and K. Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014. 4, 6

[28] P. J. Van Laarhoven and E. H. Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987. 4

[29] Y. Wu, Z. Chen, G. Sun, X. Xie, N. Cao, S. Liu, and W. Cui. Streamexplorer: A multi-stage system for visually exploring events in social streams. *IEEE transactions on visualization and computer graphics*, 24(10):2758–2772, 2018. 2, 3

[30] Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1763–1772, 2014. 1, 2

[31] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–1118, 2010. 2

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. 3

[33] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander. Where is current research on blockchain technology?—a systematic review. *PloS one*, 11(10):e0163477, 2016. 2

[34] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 27(3):712–725, 2014. 2

[35] X. Yue, X. Shu, X. Zhu, X. Du, Z. Yu, D. Papadopoulos, and S. Liu. Bitextract: Interactive visualization for extracting bitcoin exchange intelligence. *IEEE transactions on visualization and computer graphics*, 25(1):162–171, 2018. 2

[36] J. Zhao, N. Cao, Z. Wen, Y. Song, Y.-R. Lin, and C. Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE transactions on visualization and computer graphics*, 20(12):1773–1782, 2014. 2

[37] M. Zhu, W. Chen, J. Xia, Y. Ma, Y. Zhang, Y. Luo, Z. Huang, and L. Liu. Location2vec: a situation-aware representation for visual exploration of urban locations. *IEEE Transactions on Intelligent Transportation Systems*, 2019. 2