2016 IEEE 17th International Conference on Information Reuse and Integration

Rapid Prototyping of a Text Mining Application for Cryptocurrency Market Intelligence

Marek Laskowski
Department of Information and Computing Technology
Seneca College
Toronto, Ontario
marek.laskowski@senecacollege.ca

Henry M. Kim
Schulich School of Business
York University
Toronto, Ontario
hkim@yorku.ca

Abstract— **Blockchain represents a technology for establishing a shared, immutable version of the truth between a network of participants that do not trust one another, and therefore has the potential to disrupt any financial or other industries that rely on third-parties to establish trust. In order to better understand the current ecosystem of Blockchain applications, a scalable proof-of-concept pipeline for analysis of multiple streams of semi-structured data posted on social media is demonstrated, based on open source components. Deep Web as well as conventional social media are considered. Preliminary analysis suggests that data found in the Deep Web is complimentary to that available on the conventional web. Future work is described that will scale the system to cloud-based, real-time, analysis of multiple data streams, with Information Extraction (IE) (ex. sentiment analysis) and Machine Learning capability.**

Keywords- Natural Language Processing; Blockchain; Cryptocurrency; Open Data; Open Source; Rapid Prototyping; Deep Web

I. INTRODUCTION

A cryptocurrency is a system of token exchange between users underpinned and mathematically verifiable by virtue of the same cryptographic principles that underlie encryption on the Internet. Cryptocurrencies are typically implemented as distributed (Peer-to-Peer) systems based on the same Blockchain technologies that are widely argued to have the potential to revolutionize payment, financial, and monetary systems [1].

Recent trends in computing including: prevalence of Free and Open Source Software (FOSS); easy access to High Performance Computing (HPC i.e. 'The Cloud'); and increasingly advanced analytics capabilities including Natural Language Processing (NLP) and Machine Learning (ML) allow for rapidly prototyping applications for analysis of trends in the emergence of Blockchain technology.

This paper introduces a multidisciplinary approach, including Computer Systems Engineering, Software Engineering, Natural Language Processing (NLP) to perform analysis of social network data in order to better understand factors underlying the price and other trends in emerging cryptocurrency markets.

In this context, analysis of multiple data streams has been demonstrated [2] [3], however this work is unique in that it combines publicly available social networking website posts with data it scrapes from the "deep web" [4] or portion information on the Internet that cannot readily be accessed using search engines or indexes.

Although currently limited to Cryptocurrency markets, this project lays the groundwork for the real-time data fusion and analysis of multiple data streams relating to Blockchain-based ecosystems in future work.

II. BACKGROUND & MOTIVATION

Presently the only observable instances of Blockchain technology "in the wild" are Cryptocurrencies, and Cryptocurrency-like instruments, but this is expected to change [5]. Therefore, the study of emerging Cryptocurrency "ecosystems" could be a chance to understand how wider adoption of Blockchain-based technologies may unfold in the future.

A. Blockchain

The value of Blockchain technology is that it provides a mathematically verifiable means of settling exchanges between counterparties that do not trust one another. The problem of reaching consensus between computers in a trust-less network of computers is known as Byzantine Generals Problem [6] in the context of Distributed Systems [7]; and Blockchain represents the first practical solution to this problem [8][9].

Blockchain has the potential to make a huge splash in the Fintech (Financial Technology) landscape. Blockchain can be used to implement a distributed ledger that could be leveraged by financial institutions to settle transactions. Within the next decade, some argue an estimated $20 Billion USD [10] of overhead could be saved yearly, in bank settlements and securities exchanges by switching to a distributed ledger technology.

Blockchain can be used to control ownership of any asset, even real world assets, following arbitrarily complex rules i.e. "smart contracts" or "programmable money" [5][11]. In some cases, the role traditionally occupied by trusted third parties that charge a premium to assume counterparty risk could be challenged, or certainly made more efficient by employing smart-contracts implemented on Blockchain technology. However, the full extent of technological and organizational (legal/regulatory) challenges to implementing such systems in practice remains unclear. To better understand any barriers to adoption, contemporary deployments of Blockchain technology can be studied.

978-1-5090-3207-5/16 $31.00 © 2016 IEEE
DOI 10.1109/IRI.2016.66

B. Bitcoin

Bitcoin [9], and other cryptocurrencies leverage Blockchain technology to implement a distributed ledger, enabling a network of users to maintain and transfer ownership of Bitcoin tokens that are cryptographically verified and cannot be double-spent by virtue of the protocol.

What's different about Bitcoin compared to previous currency systems is that there is no central authority issuing Bitcoin tokens; they are issued according to a formula embedded within the protocol itself. Instead, the distributed ledger is maintained by an adversarial network of computers running the Bitcoin protocol, rather than a central authority. These networked computers form a Peer-to-Peer (P2P) network of "miners" in the case of Bitcoin, and they are rewarded with newly created Bitcoins as compensation for processing transactions and maintaining the integrity of the ledger.

When a bitcoin transaction occurs (e.g. when a user initiates a purchase by sending bitcoins) the sender broadcasts a request to the Bitcoin network. Miners receive and aggregate these requests into a block and "sign" the block by producing a hash, or cryptographic digest of the transactions in the block as well as the hash or signature of the previous block, forming a chain (hence blockchain).

By design there are many possible "correct" versions of the current block's hash so that miners can compete to be the first to compute a "winning" hash and broadcast it along with the miner's identifying address to its peers. Once the block is verified by the network of peers, the reward of 25 Bitcoins (as of May, 2016) is assigned to the address of that miner. This is the way that all Bitcoins have come into existence. The hashing function is relatively cheap to compute and verify; Miners continually try different combinations of padding at the end of the block (called the "nonce") in order to generate a winning hash. Therefore, the more computing power a miner has the more likely they are to generate the winning hash for that block, although it's effectively stochastic.

The difficulty or probability of generating the winning hash corresponds to the number of leading zeros in the hash, as a hash with four leading zeroes (ex. 0000123…) is more difficult, or improbable, to generate than one with three leading zeroes (ex. 000789…). The difficulty is adjusted by the protocol such that a block is created (a winning hash is generated) approximately every 10 minutes. For example, the winning hash of block #412717 mined on May 1, 2016 at 10:31:58 is: 00000000000000000025486306feab0dce320b922059256852f9812103c170720.

Bitcoin is one of many possible currencies and applications that can be built on top of Blockchain technology. Indeed, numerous of Cryptocurrencies exist, however the capitalization of the most prolific Cryptocurrency after Bitcoin is orders of magnitude smaller, making Bitcoin the largest Blockchain implementation available for study.

III. Materials and Methods

We employed a number of tools to gather and analyze data from multiple sources that we describe in this section.

A. Data Sources

Although the sources of social media data represent semi-structured data sources, each has their own peculiarities and irregularities that must be dealt with and understood when attempting to extract insights from them. Both data sources are potentially complicated by Internet connectivity issues with the usual causes (ex. service availability). Because of this, additional sources of data, paradoxically, are potentially additional sources of uncertainty, and care must be taken to avoid confusing the unavailability of data with a true drop in messaging volume in response to market conditions.

1) Twitter: The microblogging and social networking site[1], permits users to share short messages called Tweets that are accessible through Twitter's website, and also through a RESTful [12] Web API[2] . Hashtags, ex. #bitcoin, are used to identify topics or entities (users are identified with the @ symbol), and users can subscribe to or search hashtags and other users.

In order to handle disconnections gracefully, a Python script was written that requests from the Twitter RESTful web API all Tweets that have the hashtag #bitcoin or use the word Bitcoin in the topic or body of the Tweet (case is ignored). Data is returned and stored in JSON (BSON) format. If the connection is lost the script attempts to reconnect, but backs off for an increasingly long period of time if successive reconnection attempts fail. This is to avoid "hammering" the Twitter server with many connection attempts in rapid succession, and is generally considered polite in the sense of protocols.

The analysis of Twitter posts for understanding Bitcoin market conditions has been previously demonstrated [13], although not in combination with a Deep Web data source such as IRC, as described below. Twitter data was collected using the described approach for seven months between June 1, 2015 and December 31, 2015.

2) Internet Relay Chat(IRC): Internet Relay Chat (IRC) [14] is an archaic form of internet messaging, however it's analogous to modern instant messaging applications whereas users transiently join groups or channels organized by topic. Channels are named typically by topic and begin with a hash sign (e.g. #bitcoin) and are in this way similar to Twitter hashtags. IRC servers or networks of servers are accessible by the public using IRC clients adhering to the protocol, at a particular IP address or URI. Freenode (chat.freenode.net) is one such network purposed for discussion of Open Source projects and as such includes channels for Bitcoin, Open Source Blockchain development, and general discussion related to these topics.

In contrast to Twitter, IRC represents a Deep Web [4] data source, in the sense that it is information that is not readily

[1] http://www.twitter.com

[2] https://dev.twitter.com/rest/public

searchable or indexed by search engines. Most IRC clients by default log network and chat messages in similar but ultimately ad-hoc formats, and in this respect different than Tweets. The Konversation IRC client[3] was used, because it is conveniently packaged with many linux distributions.

In late May 2015, a list of public bitcoin related channels on freenode were compiled. In the end, the following channels were logged: #bitcoin-assets, #bitcoin-otc, #bitcoin-pricetalk, #bitcoin, and #dogecoin. The latter channel, used for the discussion of an alternate cryptocurrency, Dogecoin[4], was also included in order to sample the Cryptocurrency space outside of Bitcoin. IRC logs were filtered to remove network messages such as: Join, Topic, Quit, Mode, Created, Part, Nick, and Notice, as these were not of interest to this study.

IRC data was collected using the described approach for a period of six months and two-weeks between June 1, 2015 and December 12, 2015; ending after some of the studied IRC channels changed their policy to be invite-only (mode +i), preventing continued data collection efforts.

B. Natural Language Processing (NLP) Pipeline

As Engineers we espouse the principles of software re-use and modularization. GATE (General Architecture for Text Engineering) [15] is an Open Source framework for rapidly prototyping NLP applications. The GATE framework is designed to be highly modularized, permitting the interchangeability of many plugins and components in order to customize capabilities for the particular needs of the target application. Some features that are available out-of-the-box or easily by loading plugins include Information Extraction [16], Machine Learning [17], scalability to High Performance Computing (HPC) or Cloud environments [18], and tools to manage and organize large volumes of textual data into corpora.

We used the TwitIE (Twitter Information Extraction) [19] pre-packaged application pipeline in order to expedite development. Within the pipeline, the first step is to import the raw JSON into a GATE Document. The document is passed through a language identification module that identifies the language (e.g. English) used in each Tweet. Then the result is passed through a specialized tokenizer that can recognize and annotate parts of Tweets, in addition to common symbols such as "$" and "&" it handles URLs, hashtags(#), and user mentions (@). Then a gazetteer lookup is performed in order to identify any terms on which to perform named entity recognition (for proper names) at a later stage. A sentence splitting module then applies rules in order to determine where one sentence ends and the next begins, which may occur in Tweets, otherwise each Tweet is treated as a sentence. Next, a normalization module attempts to reduce linguistic noise such as abbreviations and other colloquialisms. Then, an adapted Stanford Part of Speech Tagger is applied, specialized to handle the nuances of Twitter-based communication. Named Entity Recognition is performed (in order to classify names and other entities found by the Gazetteer) before the annotated graph is outputted for post-processing as described below.

The GATE Developer tool is provided with GATE to enable rapid application development by permitting the user to reconfigure existing pipeline components using a graphical user interface (GUI), as well as, permitting the browsing of documents and their annotations interactively. A screen capture of GATE developer is shown in Figure 1, illustrating some of the features of this tool. The pipeline can be altered by adding, removing, or reconfiguring components by selecting them from the list of components on the left. On the right hand side are shown some example tweets that have been annotated using the pipeline.

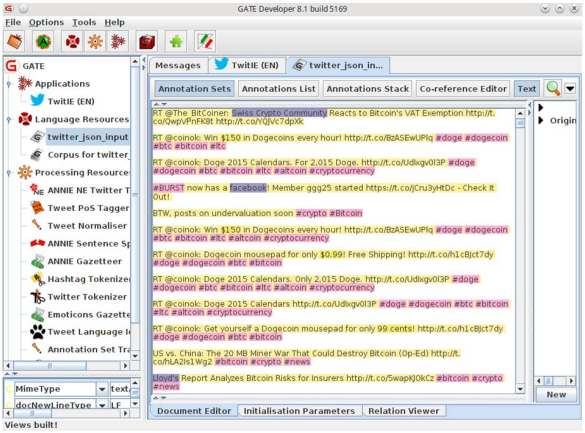

Figure 1. Gate Developer GUI displaying TwitIE pipeline components (left) and annotated tweets (right).

Applications developed using GATE Developer can be exported, and the resulting processing pipeline can re-used within any application written in Java (thanks to GATE's Java API) permitting the rapid development of post-processing resources. The API abstracts the annotated document as a graph-like or network structure, which can be traversed and the arbitrarily complex relationships between document (tagged) entities can be examined programmatically.

Before the collected could be processed using the TwitIE pipeline, we discovered we had to perform pre-cleaning and normalization. For example, escaped Unicode symbols such as "\u2026" (representing a horizontal ellipsis) commonly used on the Web had to be cleaned from the data (Ex. replaced by spaces to avoid corrupting the payload size of the tweet data) to prevent the TwitIE pipeline from crashing. A unix pipeline filter program was written for maximum flexibility and later reuse, that reads in an un-sanitized JSON file from standard input, and writes the cleaned file to standard output.

The modular architecture of GATE and the developer-centric API, have made it possible to rapidly prototype plugins and create processing pipelines for other semi-structured data formats such as IRC logs.

IV. RESULTS

Our results were ultimately bounded by available computing time. Working on a modest single 2GHz processor

[3] http://konversation.kde.org/

[4] http://dogecoin.com

core, 71 days of computing time were needed to process the 200GB of data that were collected during the 7 months. GATE used 1.95GB on average from the available 4GB of RAM. As GATE requires a considerable amount of heap memory a 64-Bit Operating System is recommended to permit increased limits on the Java Virtual Machine (JVM) virtual heap space. However, some Operating Systems enforce arbitrary limits to JVM heap space, making memory intensive Java processes such as GATE pipelines somewhat constrained by the operating environment.

TABLE I. SUMMARY OF FINDINGS; COUNT OF TOTAL OBSERVED MESSAGES, AND PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT BETWEEN MESSAGES-PER-DAY IN SEVERAL DATA STREAMS AND BITCOIN MARKET METRICS

Data Source	Total Messages	Bitcoin Volume Correlation	Bitcoin Price Correlation
Twitter	12105833	0.5239	-0.0191
#bitcoin-assets	189393	-0.1201	-0.2991
#bitcoin-otc	111499	-0.0568	-0.0675
#bitcoin-pricetalk	64712	0.7714	0.5715
#bitcoin	214283	0.0130	-0.1355
#dogecoin	1113243	-0.1682	-0.3333

The Pearson product-moment correlation coefficient (as implemented by the Microsoft Excel CORREL function) was computed between daily message counts in each social media stream, Bitcoin volume, and bitcoin price indicators, respectively. Table 1 compares total messages and correlations between daily message counts in each social media stream and Bitcoin market activity indicators. One thing that can be immediately noted from Table 1 is the average volume of Twitter messages is considerably higher than IRC messages. Figures 2, 3, and, 4 further illustrate some highlights from the results, and are discussed in the following section. Daily Bitcoin price and trading volume on USD exchanges was downloaded from http://blockchain.info.

Despite efforts to maintain a connection, unexplained significant gaps exist in collected Twitter data June 14, 2015, and August 29, 2015, and in some IRC data between November 19-21, 2016.

V. DISCUSSION

According to Table 1 and Figure 2 Twitter message volume is somewhat correlated with the volume of bitcoin transactions on USD-based exchanges as has been noted by previous work [13]. The count of messages per day in the IRC channel #bitcoin-pricetalk, interestingly is positively correlated with both Bitcoin price and trading volume according to Table 1 and Figures 3 and 4 respectively. As of December 2015 #bitcoin-pricetalk has changed its policy to allow only invited users to join.

Some other results bear brief mention; the number of messages per day related to another Cryptocurrency, Dogecoin, shows negative correlation with the price of Bitcoin. The number of messages tends to decrease as the price of Bitcoin increases but it remains unclear whether is it due to general trend in Dogecoin, or perhaps interest in Dogecoin wanes as Bitcoin price increases? We also note there's some negative correlation between bitcoin price and discussion messages per day in #bitcoin-assets. Again it remains unclear whether discussion there has waned because members are satisfied with their Bitcoin assets, or for some other reasons. Further analysis is required.

Overall, the results suggest that even with relatively simple measures such as message frequency, we can make

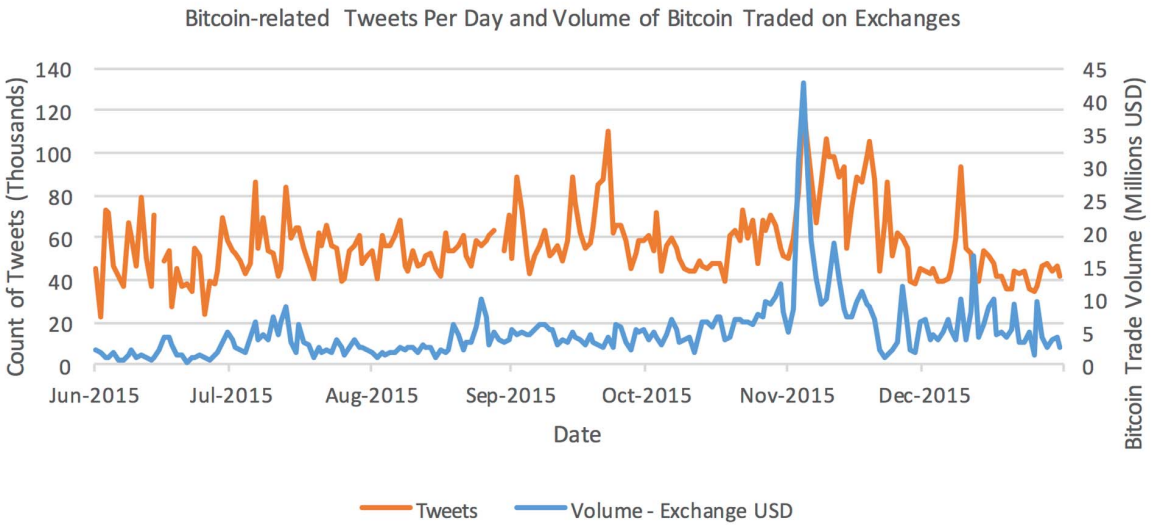

Figure 2. Count of tweets per day and Bitcoin trading volume as measured by USD exchanges (Source: Blockchain.info).

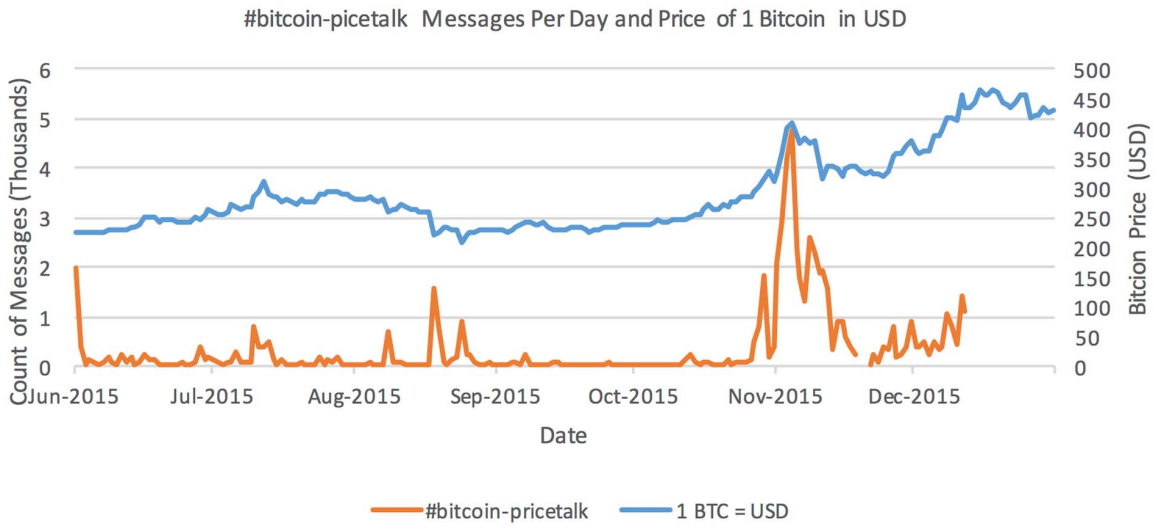

Figure 3. Count of messages per day in #bitcoin-pricetalk and price of Bitcoin in USD (Source: Blockchain.info).

some interesting observations from the combination of Deep and conventional Web data sources. More sophisticated analysis is expected to yield further insights. Counterintuitively, despite the volume of messages being lower, owing to the specificity and relevance of user discussion, and due to the low level of spam, IRC turned out to be quite valuable and a complementary counterpart to Twitter. However, the relentless heartbeat of spam messages on Twitter were useful for identifying network performance issues. Finally, a surprising amount of computing time was required, and the scalability concerns brought to light are to be addressed in Future Work.

VI. Conclusions & Future Work

We have demonstrated the rapid prototyping of a framework that's capable of fusing multiple semi-structured social media data streams into a coherent picture of a Cryptocurrency marketplace. Some preliminary results demonstrate the utility of combining data from conventional and Deep Web sources for getting a more complete picture of complex Blockchain ecosystems.

Although the groundwork for real-time analysis of multiple data streams has been demonstrated, the

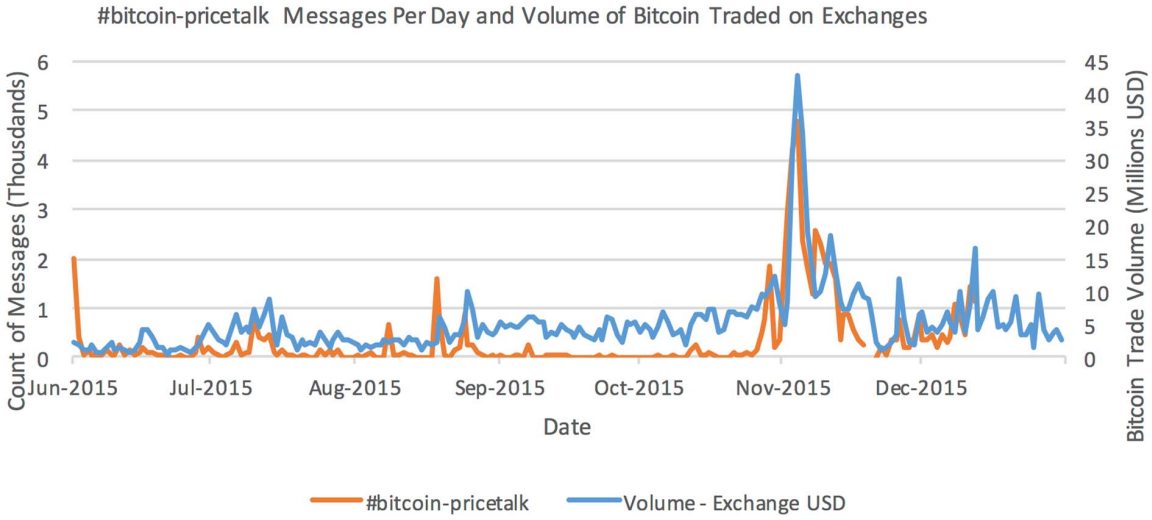

Figure 4. Count of messages per day in #bitcoin-pricetalk and Bitcoin trading volume as measured by USD exchanges (Source: Blockchain.info).

computational limits we encountered suggest that we will have to harness additional compute resources within HPC systems, or the Cloud. The GATE framework has been demonstrated as a viable tool for perusing these goals.

Our intention is to extend this work to cover other cryptocurrencies. Using the data we have collected, we will identify additional relevant keywords or hashtags to include as part of data collection. Additional data sources will also be considered such as search engine results and message board postings. Finally, to delve deeper into the semantic meaning of the data, we will perform Information Extraction [20] including but not limited to Sentiment Analysis [21] which may require the creation of domain specific ontologies [22] that would capture the semantic meaning and relationships between entities and concepts in Bitcoin and Blockchain ecosystems.

REFERENCES

[1] Swan, Melanie. Blockchain: Blueprint for a New Economy. " O'Reilly Media, Inc.", 2015.

[2] Kristoufek, Ladislav. "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era." Scientific reports 3 (2013).

[3] Garcia, David, Claudio J. Tessone, Pavlin Mavrodiev, and Nicolas Perony. "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy." Journal of the Royal Society Interface 11, no. 99 (2014): 20140623.

[4] Bergman, Michael K. "White paper: the deep web: surfacing hidden value." Journal of electronic publishing 7, no. 1 (2001).

[5] Swan, Melanie. "Connected car: quantified self becomes quantified car." Journal of Sensor and Actuator Networks 4, no. 1 (2015): 2-29.

[6] Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine generals problem." ACM Transactions on Programming Languages and Systems (TOPLAS) 4, no. 3 (1982): 382-401.

[7] Coulouris, George F., Jean Dollimore, and Tim Kindberg. Distributed systems: concepts and design. Pearson education, 2005.

[8] Andreessen, Marc. "Why Bitcoin Matters." New York Times 21 (2014).

[9] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." (2008).

[10] Rennick, Emmet, for Oliver Wyman. “The Fintech 2.0 Paper: rebooting financial services.” Oliver Wyman Anthemis Group, and Santander Innoventures, 2015. (Online: Accessed May 19, 2016).

[11] Buterin, Vitalik. "Ethereum white paper: a next generation smart contract & decentralized application platform." (2013).

[12] Richardson, Leonard, and Sam Ruby. “RESTful web services." O'Reilly Media, Inc., 2008.

[13] Kaminski, Jermain, and Peter Gloor. "Nowcasting the Bitcoin Market with Twitter Signals." arXiv preprint arXiv:1406.7577 (2014).

[14] C. Kalt, Internet Relay Chat: Client Protocol, RFC Editor, 2000.

[15] Cunningham, Hamish. "GATE, a general architecture for text engineering." Computers and the Humanities 36, no. 2 (2002): 223-254.

[16] Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. "A framework and graphical development environment for robust NLP tools and applications." In ACL, pp. 168-175. 2002.

[17] Y. Li, K. Bontcheva and H. Cunningham. Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, 15(02), 241-271, 2009.

[18] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. GATECloud.net: a Platform for Large-Scale, Open-Source Text Processing on the Cloud. Philosophical Transactions of the Royal Society A, 371(1983), 2013.

[19] K. Bontcheva, L. Derczynski, A. Funk, M.A. Greenwood, D. Maynard, N. Aswani. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013).

[20] Cowie, Jim, and Wendy Lehnert. "Information extraction." Communications of the ACM 39, no. 1 (1996): 80-91.

[21] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5, no. 1 (2012): 1-167.

[22] Staab, Steffen, Rudi Studer, Hans-Peter Schnurr, and York Sure. "Knowledge processes and ontologies." IEEE Intelligent systems 1 (2001): 26-34.