DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2017

Predicting Bitcoin price fluctuation with Twitter sentiment analysis

EVITA STENQVIST

JACOB LÖNNÖ

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF COMPUTER SCIENCE AND COMMUNICATION

Predicting Bitcoin price fluctuation with Twitter sentiment analysis

EVITA STENQVIST

JACOB LÖNNÖ

Master in Computer Science
Date: June 14, 2017
Supervisor: Jeanette Hellgren Kotaleski
Examiner: Örjan Ekeberg
Swedish title: Förutspå Bitcoin prisändringar med hjälp av semantisk analys på Twitter data
School of Computer Science and Communication

Abstract

Programmatically deriving sentiment has been the topic of many a thesis: it's application in analyzing 140 character sentences, to that of 400-word Hemingway sentences; the methods ranging from naive rule based checks, to deeply layered neural networks. Unsurprisingly, sentiment analysis has been used to gain useful insight across industries, most notably in digital marketing and financial analysis.

An advancement seemingly more excitable to the mainstream, Bitcoin, has risen in number of Google searches by three-folds since the beginning of this year alone, not unlike it's exchange rate. The decentralized cryptocurrency, arguably, by design, a pure free market commodity – and as such, public perception bears the weight in Bitcoins monetary valuation.

This thesis looks toward these public perceptions, by analyzing 2.27 million Bitcoin-related tweets for sentiment fluctuations that could indicate a price change in the near future. This is done by a naive method of solely attributing rise or fall based on the severity of aggregated Twitter sentiment change over periods ranging between 5 minutes and 4 hours, and then shifting these predictions forward in time 1, 2, 3 or 4 time periods to indicate the corresponding BTC interval time.

The prediction model evaluation showed that aggregating tweet sentiments over a 30 min period with 4 shifts forward, and a sentiment change threshold of 2.2%, yielded a 79% accuracy.

Sammanfattning

Ämnet sentiment analysis, att programmatiskt härleda underliggande känslor i text, ligger som grund för många avhandlingar: hur det tillämpas bäst på 140 teckens meningar såväl som på 400-ords meningar a'la Hemingway, metoderna sträcker sig ifrån naiva, regelbaserade, till neurala nätverk. Givetvis sträcker sig intresset för sentiment analys utanför forskningsvärlden för att ta fram insikter i en rad branscher, men framförallt i digital marknadsföring och financiell analys.

Sedan början på året har den digitala valutan Bitcoin stigit trefaldigt i sökningar på Google, likt priset på valutan. Då Bitcoins decentraliserade design är helt transparant och oreglerad, verkar den under ideala marknadsekonomiska förutsättningar. På så vis regleras Bitcoins monetära värde av marknadens uppfattning av värdet.

Denna avhandling tittar på hur offentliga uppfattningar påverkar Bitcoins pris. Genom att analysera 2,27 miljoner Bitcoin-relaterade tweets för sentiment ändringar, föutspåddes ändringar i Bitcoins pris under begränsade förhållningar. Priset förespåddes att gå upp eller ner beroende på graden av sentiment ändring under en tidsperiod, de testade tidsperioderna låg emellan 5 minuter till 4 timmar. Om en förutspånning görs för en tidsperiod, prövas den emot 1, 2, 3 och 4 skiftningar framåt i tiden för att ange förutspådd Bitcoin pris interval.

Utvärderingen av förutspåningar visade att aggregerade tweet-sentiment över en 30-minutersperiod med 4 skift framåt och ett tröskelvärde för förändring av sentimentet på 2,2 % gav ett resultat med 79 % noggrannhet.

Contents

1 Introduction **1**
1.1 Related research 1
1.2 Problem statement 2
1.3 Scope 2
1.4 Purpose 3

2 Background **4**
2.1 Bitcoin 4
2.2 Opinion mining 5
2.3 Sentiment analysis 6
2.3.1 Polarity classification 7
2.3.2 Lexicon-based approach 7
2.3.3 VADER 7

3 Method **9**
3.1 Data collection 9
3.1.1 Bitcoin price data 9
3.1.2 Gathering tweets in real-time 10
3.2 Sentiment analysis process 11
3.2.1 Reducing noise in the twitter dataset 11
3.2.2 Individual tweet sentiment analysis 11
3.2.3 Aggregating sentiment scores 12
3.3 Deriving predictions from sentiment data 12
3.3.1 Frequency length and prediction shifts 12
3.3.2 Preliminary predictions 13
3.4 Model evaluation 13
3.4.1 Creating prediction vectors given threshold 13
3.4.2 Creating historical price fluctuation vector 13

3.4.3 Comparing predictions with historical price fluctuation . . . 13
3.4.4 Measurements . . . 14

4 Results **16**
4.1 Data collection . . . 16
4.1.1 USD/BTC Exchange rate data-set . . . 16
4.1.2 Twitter data-set . . . 17
4.2 Noise reduction . . . 18
4.3 Polarity classification . . . 19
4.4 Prediction performance . . . 20
4.4.1 The prediction model . . . 23

5 Discussion **24**
5.1 Prediction model . . . 24
5.1.1 Evaluating predictions . . . 24
5.1.2 Lackluster prediction opportunities . . . 24
5.2 Reconnecting with Problem statement . . . 25
5.3 Weaknesses . . . 26
5.3.1 Static threshold . . . 26
5.3.2 Domain specific lexicon . . . 26
5.3.3 No indication of success . . . 26
5.3.4 Lack of data . . . 27
5.4 Future research . . . 27
5.4.1 Picking frequency/shift based on historical accuracy . . . 27
5.4.2 Machine learning . . . 27

6 Conclusion **28**

Chapter 1

Introduction

It's 2017, the people of the world generate 2.5 million terabytes of information a day [1]. 500 million tweets, 1.8 billion pieces of shared information on Facebook, each and every day [2]. These snippets of information regard anything under the sun; from what the user had for lunch, to their disgust over a referee in a football match. Twitter specifically has become known as a location where news is quickly disseminated in a concise format.

When regarding a financial commodity, the public confidence in a particular commodity is a core base of its value. Social media has served as platform to express opinions since their inception, and as such tapping into the open APIs provided of the likes of Facebook and Twitter, these arguably biased pieces of information become available with a sea of meta-data.

Bitcoin (BTC), the decentralized cryptographic currency, is similar to most commonly known currencies in the sense that it is affected by socially constructed opinions; whether those opinions have basis in facts, or not. Since the Bitcoin was revealed to the world, in 2009 [3], it quickly gained interest as an alternative to regular currencies. As such, like most things, opinions and information about Bitcoin are prevalent throughout the Social Media sphere [4].

1.1 Related research

In the paper *Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns* by Sul et al. [5], 2.5 million tweets about S&P 500 firms were put through the authors own sentiment classifier and compared to the stock returns. The results showed that sentiment that disseminates through a social network quickly is anticipated to be reflected in a stock price on the same trading day, while slower spreading sentiment is more likely to be reflected on future trading days. Basing a trading strategy on these predictions are prospected to yield 11-15% annual gains.

The paper *Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis* by Colianni et al. [6], similarly analyzed how tweet sentiment could be used to impact investment decisions specifically on Bitcoin. The authors used supervised machine learning techniques that yielded a final accuracy of above 90% hour-by-hour and day-by-day. The authors point out that the 90% accuracy was mustered through robust error analysis on the input data, which on average yielded a 25% better accuracy. Colianni et al. together with Hutto and Gilbert both mentioned levels of noise in their dataset, and the former team got a significant reduction in error rates after cleaning their dataset for noise.

1.2 Problem statement

When analyzing the sentiment of opinions and snippets of information distributed on Twitter regarding Bitcoin and comparing with Bitcoin's price,

- Is there a correlation between Twitter sentiment and BTC price fluctuation?
- Can a naive prediction model based on sentiment changes yield better than random accuracy?

1.3 Scope

Sentiments are only collected from one micro-blogging source; Twitter. Due to Twitters establishment in the micro-blogging sphere, as well as its accessible programmatic interface for data collection.

Similarly, the decision to limit the cryptocurrency to Bitcoin came down to Bitcoin being the most established cryptocurrency both in age and cryptocurrency market share, reflecting its acceptance in the public's eye. Although, the presented prediction model can be tweaked to any other cryptocurrency by providing the underlying data collection mechanism with identifying keywords. The accuracy estimations would have to be recomputed and would likely vary vastly to the presented results in this paper.

The sentiments as well as the currencies price are analyzed on a short term basis, disregarding how micro-blogging sentiment correlates to macro trends in a cryptocurrency or attempting to identify if they exist. Short term in this paper is defined to be within the 24h mark, (based on the findings of Colianni et al. [6]).

The sentiments classification is limited to the most naive binary form of positive or negative, not attempting to capture sentiment on a more complex emotional level. On the BTC side, the key value will be limited to an increase and decrease in price over specific time intervals, disregarding volume and other key metrics. Further, BTC transactions are collected for the BTC/USD currency pair, and only collected from the Coindesk exchange due to difficulty in finding open-source aggregated exchange data.

1.4 Purpose

Micro-blogging sentiment value has been studied in relation to a variation of commodities, including S&P 500 firms [5] and even Bitcoin [6]. Although, in the context of Bitcoin, former researchers mused that accounting for negations in text may be a step in the direction of more accurate predictions. In this paper, by not only taking into account negation, but also valence, common slang and smileys [7], a more accurate sentiment analyser is hoped to yield more accurate predictions on the Bitcoin price. Furthermore, by comparing sentiment and Bitcoin price at different intervals of time, and optimizing a prediction model given these intervals, a short term analysis of correlation between sentiment and market change can be examined.

Chapter 2

Background

2.1 Bitcoin

Bitcoin is the most popular and established cryptographic digital currency. Unlike "normal" currencies the value of Bitcoin lay not in a physical commodity, but in computational complexity. In the most basic sense, Bitcoin is an open source software program, run on networked computers (nodes). Together these nodes share a distributed database, the blockchain, which serves as the single source of truth for all transactions in the network, and allows for Bitcoin to function according to its original design - touching upon the subjects of cryptography, software engineering and economy [4].

While the Bitcoin currency is the most commonly known application of the blockchain, the blockchain itself can be used for any system in which one would exchange value [4, 8] as it disallows for duplications of an asset.

Most currencies in the world are issued and regulated by a government, either directly or indirectly (i.e. through a central bank). In both cases, the goals and policies of a government are what guide and regulate it's currency [4]. In the case of a central bank, the above is still true, though the more direct control is left to the central bank - a semi-independent relationship between bank and government. The central banks' job is to achieve the goals set by its governing institute in areas including economic stability, economic growth and stability of the currency value [4].

The value of a currency depends on several factors, the more notable being; public confidence, acceptance, and social expectancy (of value) [4]. While Fiat, the de facto currencies governed by a commodity and centralized institute, might have started out with actual physical commodity value guarantees, this is rarely the case in the current financial climate [4, 9]. Since Fiat currencies are controlled, there are vulnerabilities in how the central agency decides to influence a currency. Irresponsible monetary policies can lead to an artificial long term deflation by using short methods (one of which is printing money, i.e. increasing monetary supply but decreasing value) to solve problems or crisis [4].

Bitcoin, on the other hand, has no central authority and no direct way to influence either Bitcoin value or supply of Bitcoins [8, 4]. By design, this removes the middle man that most monetary systems are created around, i.e. central bank and the banking system [4]. The only way to increase the supply of Bitcoins is to partake in transaction calculations, which leads to a predictable growth of Bitcoin supply [8, 4] and pays for the infrastructure. At the same time the monetary value of the currency are influenced by the same variables as a fiat currency [4].

The decentralized approach can also be seen in the architecture of the Bitcoin network. Bitcoin is intended to be a decentralized peer-to-peer network of nodes [4, 10], so any changes to the architecture or specific implementation parts of Bitcoin must be agreed upon by at least half of the peers [4]. Part of the decentralised design is the shared database, of which all nodes have a copy - commonly referenced as ledger and formerly mentioned as blockchain. This ledger contains all past transactions as well as all current Bitcoin owners [4, 3]. The database is created in blocks of chronological transactions. A new block is created by gathering current transactions and is then sealed cryptographically together with earlier blocks, creating a chain of blocks - a blockchain [3, 10]. This design makes it hard to censure or edit a preceding block in the chain, rendering it secure and transparent [4].

The Bitcoin design and theoretical work was first published in (2008) by Nakamoto [3, 10].

2.2 Opinion mining

Social Networks have grown rapidly since their inception early this millennium [11, 12, 6]. Global total users surpassed 2 billion in 2015, with continuous steady growth to come according to Statista.coms estimation [13]. Social networks provide users with a platform to express their thoughts, views, and opinions [12].

Twitter, the micro-blogging platform and public company, was launched in 2006. 10 years on, in 2016, the platform has over 300 million monthly active users [14]. As is characteristic for micro-blogging services, Twitter provides it's users with the possibility to publish short messages [15]. These messages are commonly called tweets and are limited in length to 140 characters. User can also include metadata inline in the text of a tweet [15, 12], by either # ('hashtag') or @ ('at'). The two operators have different intentions, with the former (#) used as a symbol to signal a specific context while the latter (@) references to another Twitter user [16]. Hashtag contextually link together tweets, creating a web of contextually similair datapoints. Twitter also provides facilities for both searching and consuming a real-time stream of tweets based on specific hashtags [12, 17].

Twitter is a centralised location to publish (as well as consume) internally and externally generated content [12]. For some companies, it has become an additional channel to communicate with the marketplace, and for others to use as a resource [11, 12, 6, 18]. Twitter has over the years become a platform for media: news, company announcements, government communication, to individual users personal opinions, world views, or daily life [15]. Together, Twitter users are generating millions of short messages, all public and some already labelled with contextual data [6, 12].

Due to the message length restriction and the classifying nature of tweets hashtags, Twitter has become a gold mine for opinionated data, in its semi-structured form [11]. Researchers and other entities are regularly mining Twitter for tweets, attempting to gain value, information or understanding in varying subjects and disciplines [11, 12, 6, 18]. As such Twitter is widely used as a source when looking for sentiment data sets [11, 6].

2.3 Sentiment analysis

In a nutshell, sentiment analysis is about finding the underlining opinions, sentiment, and subjectivity in texts [19, 20], which all are important factors in influencing behaviour [21]. The advent of machine learning tools, wider availability of digital data sets, and curiosity, has greatly reduced the cost of performing sentiment analysis – allowing for more research [21]. This type of data is highly sought after and has attracted the attention from researchers and companies alike [20, 21].

2.3.1 Polarity classification

Since the rise of social media, a large part of the current research has been focused on classifying natural language as either positive or negative sentiment [19, 20].

Polarity classification have been found to achieve high accuracy in predicting change or trends in public sentiment, for a myriad of domains (e.g. stock price prediction) [11, 12, 6].

2.3.2 Lexicon-based approach

A lexicon is a collection of features (e.g. words and their sentiment classification) [21, 19]. The lexicon-based approach is a common method used in sentiment analysis [19, 7] where a piece of text is compared to a lexicon and attributed sentiment classifications. Lexicons can be complex to create, but once created require little resources to use [20, 21]. Well designed lexicons can achieve a high accuracy and are held in high regard in the scientific community [7, 21, 20, 19].

2.3.3 VADER

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a combined lexicon and rule-based sentiment analytic software, developed by Hutto and Gilbert [7]. VADER is capable of both detecting the polarity (positive, neutral, negative) and the sentiment intensity in text [7]. The authors have published the lexicon and python specific module under an MIT License, thus it is considered open source and free to use [22]. VADER was developed as a solution [7] to the difficulty in analysing the language, symbols, and style used in texts in primarily the social media domain [7, 11].

Hutto and Gilbert [7] express the goals on which they based the creation of VADER as the following:

> *"...1) works well on social media style text, yet readily generalizes to multiple domains, 2) requires no training data, but is*

> *constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon 3) is fast enough to be used online with streaming data, and 4) does not severely suffer from a speed-performance tradeoff."* [7, section 3]

VADER was constructed by examining and selecting features from three previously constructed and validated lexicons as a candidate list [7]; Linguistic Inquiry and Word Count (LIWC) [23], Affective Norms for English Words (ANEW) [24], and General Inquirer (GI) [25]. The authors also added common social media abbreviations, slang, and emoticons. Each feature was allocated a valence value, with this additional information, 7500 features were selected to be included in the VADER lexicon. In addition to the word-bank, Hutto and Gilbert analysed the syntax and grammar aspects of 800 tweets and their perceived valence value. The aforementioned analysis resulted in five distinct behaviour that is used to influence a tweets intensity, which was formulated into rules. Together, these rules and lexicon constitute VADER [7].

For performance review, VADER was compared against eleven other semantic analysis tools and techniques for polarity classification (positive-, neutral- and negative sentiment), across four different and distinct domains. VADER consistently performs among the top in all test cases and outperformed the other techniques in the social media text domain [7].

Chapter 3

Method

This study began with a literature review. The purpose of the review was to explore the background of sentiment analysis and financial time series prediction methods. The chapter is structured in the chronological order of events: starting with data gathering, followed by dataset pre-processing, analyzing for sentiment, to finally describe the prediction model and it's evaluation.

3.1 Data collection

Two different data sources were collected during the study; the first consisting of historical BTC/USD exchange rate data and the other of tweets. The datasets were collected using a dedicated server, allowing for uninterrupted continues data gathering.

3.1.1 Bitcoin price data

Historical price points for Bitcoin were gathered daily from CoinDesk publicly available API [26]. Depending on the time interval length of the requested data, CoinDesk returns different levels of detail, such as transaction based or aggregated in the form of OHCL [1]. An interval length of a day returns price data for every minute. Listing 3.1 shows an API call for the pricing data for the 9th of May from 00:00 to 23:59.

Listing 3.1: CoinDesk API request for USD/BTC exchange price data

[1] *OHCL* - Open, High, Close, Low. Commonly aggregated values over financial time series data for different intervals of time.

```
\url{http://api.coindesk.com/charts/data?output=csv\&data
    ↪ =close\&index=USD&startdate=2017−05−09&enddate
    ↪ =2017−05−09\&exchanges=bpi&dev=1}
```

3.1.2 Gathering tweets in real-time

To collect data for the sentiment analysis Twitter's streaming API [17] was used in combination with Tweepy. Tweepy, an open source framework written in Python, facilitates tweet collection from Twitter's API [27]. Tweepy allows for filtering based on hashtags or words, and as such was considered as an efficient way of collecting relevant data. The filter keywords were chosen by selecting the most definitive Bitcoin-context words, for example "cryptocurrency" could include sentiments towards other cryptocurrencies, and so the scope must be tightened further to only include Bitcoin synonyms. These synonyms include: Bitcoin, BTC, XBT and satoshi.

Listing 3.2: Example function that gathers a stream of filtered tweets

breaklines

```
def btc_tweet_stream():
    api = twitterAPIConnection()
    listener = StdOutListener()
    stream = tweepy.Stream(api.auth, listener)
    stream.filter(track=
        ['btc', 'bitcoin', 'xbt', 'satoshi']
        , languages=['en'])
```

3.2 Sentiment analysis process

Sentiment analysis of the twitter dataset has three different phases: scrubbing bot generated content, sentiment analysis of individual tweets with VADER, and aggregation of individual tweet sentiment score into a combined score for each time series interval.

3.2.1 Reducing noise in the twitter dataset

As mentioned in section 2.2, automatically generated content and irrelevant or influential tweets are undesirable for the analysis. To avoid that the result to be too influenced by these undesirable tweets a filter was developed using the following strategy:

A subset from the greater twitter dataset of one hundred thousands tweets used as a basis for finding common attributes among duplicate or dot generated content. Those 100 000 tweets were scrubbed from any non-alphabetic symbols (excluding "#" & "@"). All non-unique tweets (message) text were then fed through a frequency analysis script identifying high prevalence hashtags, words, bigrams and trigrams. The most frequent in the previous mentioned groups were lastly put to manual scrutiny to identify suspicious patterns. Suspicious n-grams were deemed to be those that; coax users to do something, offer users free Bitcoin, are clearly bots announcing current exchange rates. Some of these n-grams intersected with many other n-grams on one token, either a hashtag or word. The identified tokens together constituted the basis for the construction of a filter. Table 4.3 displays the variables used for the filter. This filter combined with dropping duplicates was applied on the full tweet dataset and substantially reduced size of the set (see section 4.2). The filtering and dropping duplicate tweets constitutes the cleaning phase.

3.2.2 Individual tweet sentiment analysis

VADER (see section 2.3.3) is used to derive a sentiment score from each tweet. VADER provides a compound sentiment score between -1.0 and 1.0 for the text fed to it. Each tweets sentiment score is then compared to a (compound sentiment) threshold for classification as either positive or negative. Following the recommendation given by VADER's creators: *compound sentiment threshold*$= 0.5$ [22]. Any tweets that do not fall in to either categories is left unclassified, and is considered undesirable [7]. The result of this process is that each tweet row in the dataset is appended with it's individual sentiment score.

3.2.3 Aggregating sentiment scores

With the sentiments returned by VADER, the individual tweet sentiment scores are grouped into time-series (see table 3.1 for period duration). For each group the sentiment mean is taken on the underlying tweets to indicate the average sentiment. This is the last phase of the sentiment analysis, resulting in a dataset consisting of groups, ordered in time based on passed interval length with sentiment score.

3.3 Deriving predictions from sentiment data

Predictions for a time interval depend on a combination of frequency length, shift, fluctuation between periods, and the threshold the fluctuations are compared.

3.3.1 Frequency length and prediction shifts

In an attempt to substance the possible identification of correlation between Bitcoin price change and Twitter sentiment change, two temporal aspects are taken into account; frequency length and shift. Short term predictions are made on discrete time intervals, i.e. time series. Short term is regarded as a time series ranging from 5 minutes, to 4 hours in length. (see table 3.1).

Intervals	5 min	15 min	30 min	45 min	1 hour	2 hours	4 hours

Table 3.1: The chosen time intervals

Each time series is evaluated over four different shifts forward: 1, 2, 3, and 4. Where a shift indicates that an event predicts for that period in time in the future, e.g. if a positive event occurs in a test with freq. 30 mins at 14:30:00, and the shift is 2, the prediction is set for the BTC price change at 15:30:00.

3.3.2 Preliminary predictions

A periods sentiment score is used to measure rate of change in opinion in subsequent periods. This is done by calculating the difference between neighbouring periods sentiment score. If the sentiment change rate is positive (i.e. sentiment score has increased), this is deemed as increased positive sentiment shared about Bitcoin. Any such events are classified as a 1 - predicting an increase in price during a future period. Respectively, periods with a negative sentiment rate growth is classified as 0. This rate value is then available to be compared against a threshold to filter predictions.

3.4 Model evaluation

Given the baseline predictions, the prediction model creates binary classified vectors of predictions for a certain threshold to ultimately compare the predictions to actual historical price data.

3.4.1 Creating prediction vectors given threshold

At this point, the dataset contains the prediction vectors. Each vector includes predictions for an interval length and as mentioned in sector 3.3.2, predictions can be filtered based on sentiment change rate value, i.e. only include those prediction with a sentiment change higher or lower, than "threshold". The thresholds used ranged from 0% to 10% with a step of 0.05% in-between values.

3.4.2 Creating historical price fluctuation vector

The USD/BTC exchange price dataset contains minute-per-minute updates during the entire tweet collection period. The detailed price data is then aggregated into the frequencies mentioned in table 3.1. Lastly, each frequency is classified as 0 or 1, depending on the price change.

3.4.3 Comparing predictions with historical price fluctuation

In order to find out how well the various prediction vectors perform, each prediction vector is compared against the (corresponding) historical data. Each pair of elements that are compared between the two vectors is classified as one out of four classes (represented as a confusion matrix[2] in table 3.2). The four comparison classification classes from the binary vector comparison are:

1. *True Positive* - A correct positive prediction
2. *False Negative* - An incorrect negative prediction
3. *False Positive* - An incorrect positive prediction
4. *True Negative* - A correct negative prediction

		Predicted price	
		increase:	**decrease:**
Historical price	**increase:**	*True Positive (TP)*	*False Negative (FN)*
	decrease:	False Positive	True Negative

Table 3.2: Confusion matrix for comparison classifying predictions and real values

3.4.4 Measurements

From the definition and classes presented in section 3.4.3 the concepts of accuracy, recall[3], precision[4], and F1-score. The following numbered list displays definitions used in this paper the four measurements:

1. *Accuracy* (eq 3.1) - the fraction of correctly predicted prediction and all predictions.
2. *Recall* (eq 3.2) - the fraction of correctly identified (positive) predictions and all (positive) events.
3. *Precision* (eq 3.3) - the proportion between correctly predicted (positive) predictions in relation to all (positive) predictions made.
4. *F1-Score* (eq 3.4) - the harmonic mean between precision (3.3) and recall (3.2), where the weight between both ith both variables values as equals.

[2] Confusion matrix are also called contingency table, or error matrix.
[3] Recall is also called true positive rate (tpr), sensitivity, or probability of detection
[4] Precision is also called positive predictive value (ppv)

$$\textbf{Accuracy} = \frac{\sum(\text{True Positive}) + \sum(\text{True Negative})}{\sum(\text{Total population})} \tag{3.1}$$

$$\textbf{Recall} = \frac{\sum(\text{True Positive})}{\sum(\text{False Negative}) + \sum(\text{True Positive})} \tag{3.2}$$

$$\textbf{Precision} = \frac{\sum(\text{True Positive})}{\sum(\text{False Positive}) + \sum(\text{True Positive})} \tag{3.3}$$

$$\textbf{F1-score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3.4}$$

These four metrics are applied on all combinations of frequency and shifts over all thresholds, facilitating comparisons between predictions vector accuracy given the variation of variables.

Chapter 4

Results

4.1 Data collection

Data collection, for both sets, began on 11th of May and ended on the 11th of June, 2017. The datasets totaled 31 days of sequential Bitcoin-related tweets and USD/BTCexchange rate data.

4.1.1 USD/BTC Exchange rate data-set

Once daily, over the 31 days, CoinDesk's API was requested for the previous day's USD/BTC pricing data. The data arrived as a .csv-file, with pricing data in one minute intervals. Figure 4.1 shows Bitcoin's

Figure 4.1: USD/BTC price

price (in USD) over the entire span of USD/BTC exchange rate data-set. Table 4.1 showcase some sample data from API request described above.

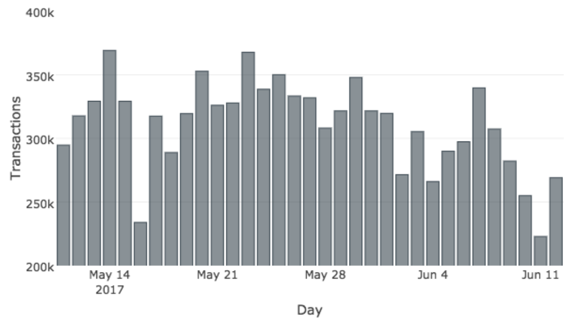

Figure 4.2: Daily amount of Bitcoin transactions

Figure 4.2 displays day-to-day total amount of trades for the collection period.

Date, "Close Price"	
2017-05-09 00:00:00, 1639.32	2017-05-09 00:01:00, 1639.71

Table 4.1: Examples from USD/BTC pricing data

4.1.2 Twitter data-set

During the month a total of 2 271 815 Bitcoin related tweets were gathered from the Twitter API. Figure 4.3 shows the distribution of the number of collected tweets per day, over the entire collection period. In figure 4.3: The loss of tweets on June 7th was due to a server crash.

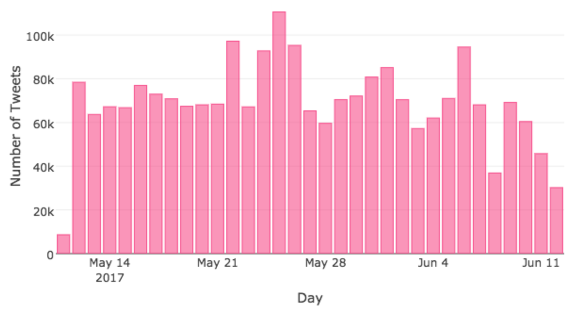

Figure 4.3: Daily amount of tweets collected

4.2 Noise reduction

The process of finding bots in the twitter-realm included analysing 100 000 tweets, collected between 22nd of April 08:30 and 24th of April 17:00, to identify suspicious looking n-grams manually. Table 4.3 displays the tokens identified. Table 4.4 has small selections of examples of discarded tweets. When the filter was run on the entire tweet data-set, the filtering and dropping of duplicates resulted in 44.8% reduction of data-set size, (see table 4.2).

	size all tweets	size after	%-reduction
100k tweets	100 000	58764	58.8%
All tweets	2 271 815	1 254 820	55.2%

Table 4.2: Reduction of tweet noise

Hashtags	#mpgvip, #freebitcoin,#livescore, #makeyourownlane, #footballcoin
Words	{entertaining, subscribe}
Bigrams	{free, bitcoin}, {current, price}, {bitcoin, price}, {earn, bitcoin}
Trigrams	{start, trading, bitcoin}

Table 4.3: Tokens identified as suspicious

RT @mikebelshe: I'm incredibly risk averse. That's why I have all my money in Bitcoin.
RT @EthBits: EthBits ICO status: `https://t.co/dLZk2Y5a88` #bitcoins #altcoins #blockchain #ethereum #bitcoin #cryptocurrency
Margin buying- profitable way of doing online trading
#tradingbitcoin on #margin. $ellBuy `https://t.co/aiYYyaCZhK` #Bitcoin
RT @coindesk: The latest Bitcoin Price Index is 1241.17 USD `https://t.co/lzUu2wyPQN` `https://t.co/CU1mmkP5mE`

Table 4.4: Examples of discarded tweets

4.3 Polarity classification

Table 4.5 contains examples of how VADER evaluated tweets as positive-, neutral-, and negative-sentiment on individual tweet level, these tweets was randomly selected from the twitter data-set after sentiment analysis with VADER.

Classification	Vader analysde tweet text examples
Positive	:D :D :D ....[Bitcoin performance assessment (+6.18%)] #bitcoin
	it's pretty cool BTC and Alts are being so bullish and fun eh? yeeeaa......just remember that winter exists. respect the cycles.
Neutral	RT LouiseMensch: According to #Steele the hacker network needed Micropayments too.
	I know somebody who is ALL ABOUT THE BITCOIN
	RT RandyHilarski: #Bitcoin News Blockchain Land Registry Tech Gets Test in Brazil https://t.co/MXTaSOGhaX
Negative	CYBER ATTACK FEARED AS MULTIPLE U.S. CITIES HIT WITH SIMULTANEOUS POWER GRID FAILURES OVER LAST 24 HOURS https://t.co/BzWfzlpZrc #Bitcoin
	I can't stand btc like that, that. E that fake shit role playing shit like btc just be yourself damn https://t.co/FSq222kTbl

Table 4.5: Example of tweets classified as positive-, neutral-, or negative-sentiment

4.4 Prediction performance

This section presents data on the predictions performance for all frequencies and shifts. Firstly, presenting the figure 4.4 showing how number of predictions decline, almost exponentially, given an increasing threshold.

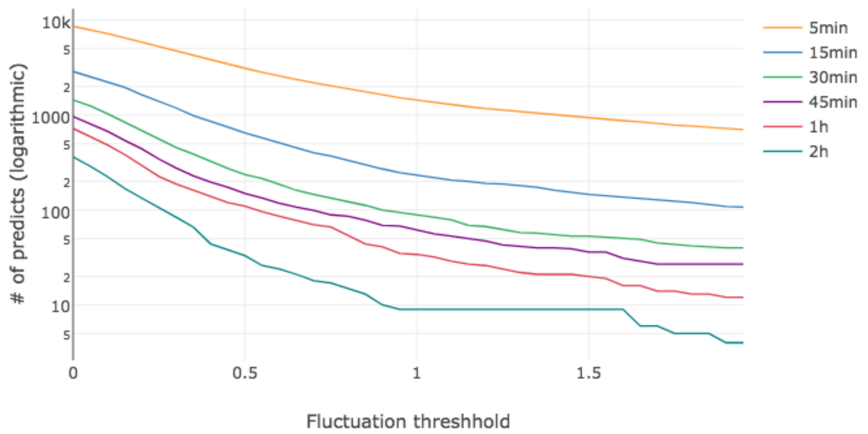

Figure 4.4: Predict count for given threshold (**Note: logarithmic y-axis**)

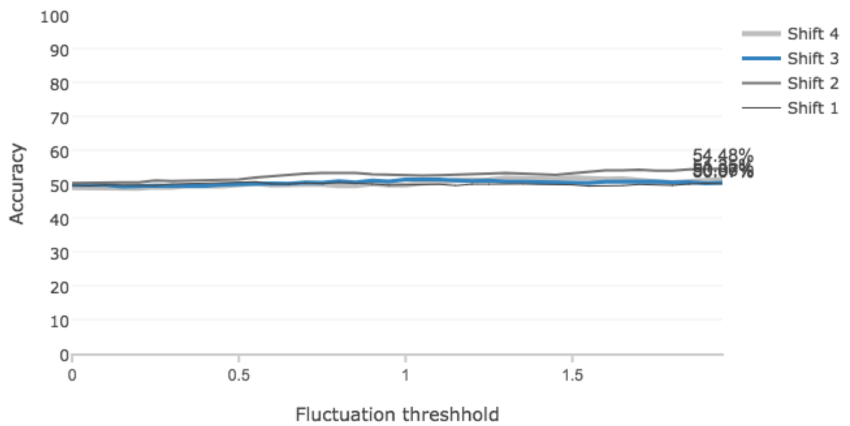

Figure 4.5: 5 minutes interval

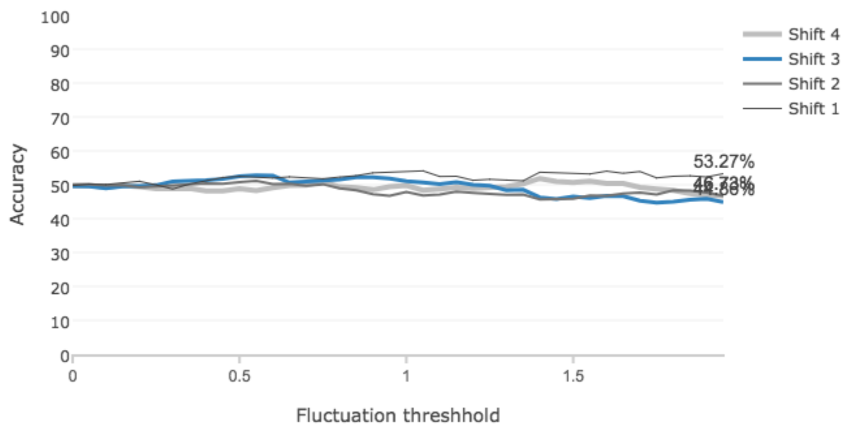

Figure 4.6: 15 minutes interval

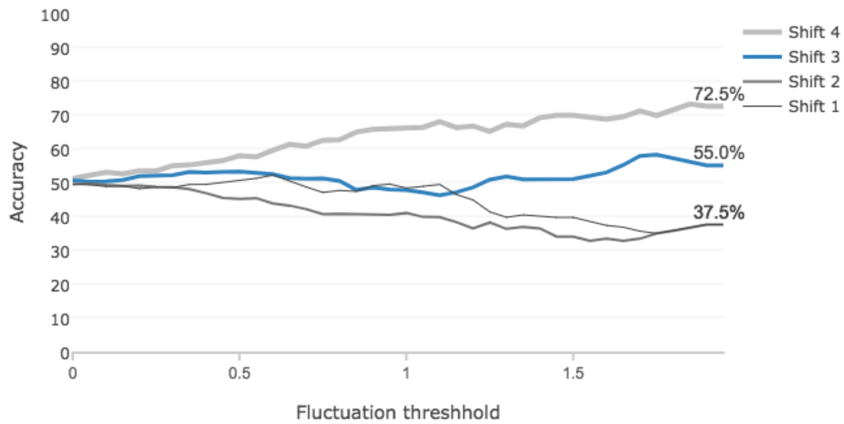

Figure 4.7: 30 minutes interval

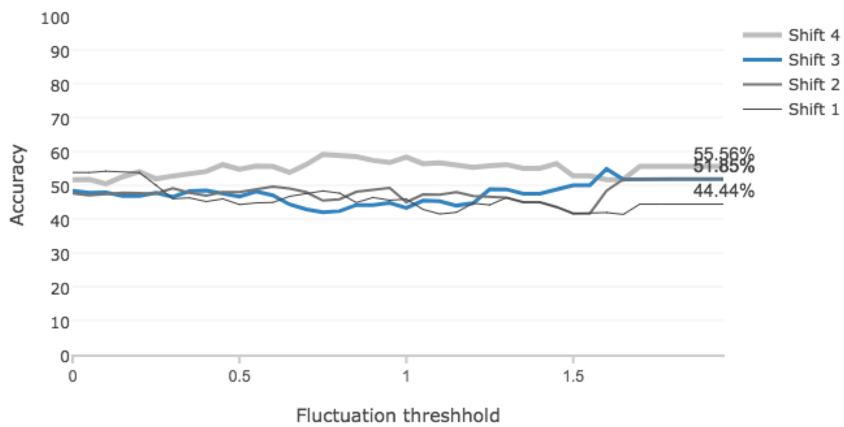

Figure 4.8: 45 minutes interval

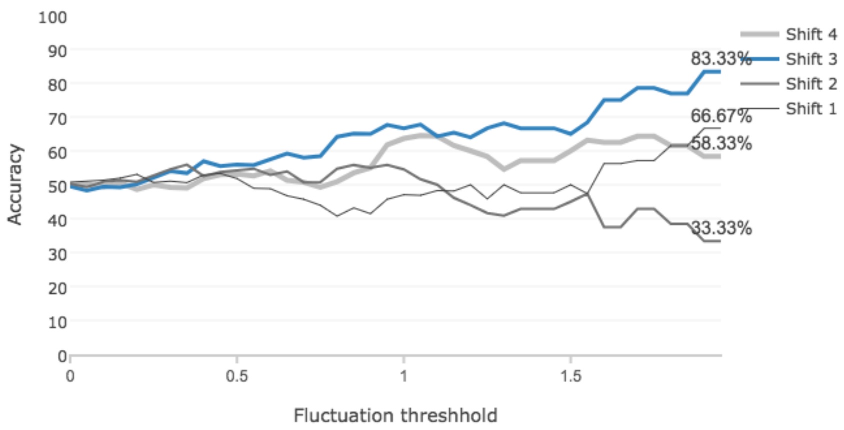

Figure 4.9: 1 hour interval

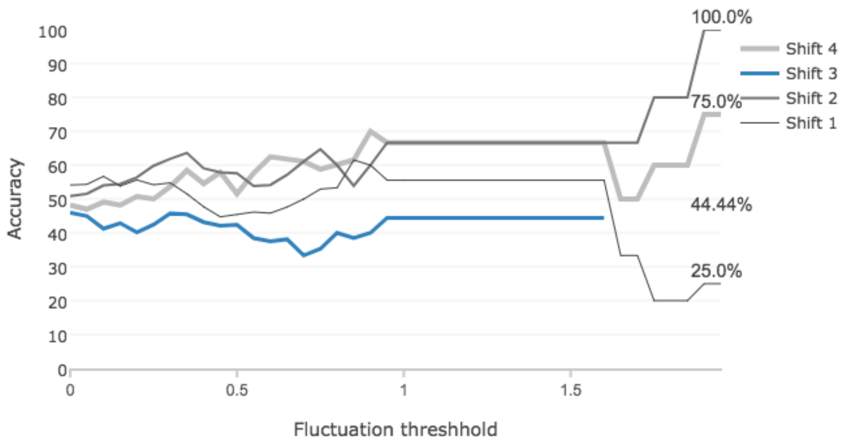

Figure 4.10: 2 hours interval

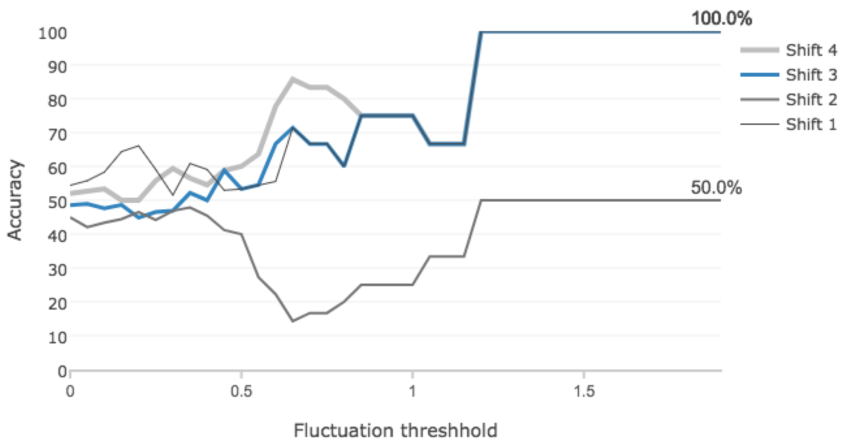

Figure 4.11: 4 hours interval

4.4.1 The prediction model

Table 4.6 contains the accuracy of the prediction model over the entire dataset, with prediction accuracy depending on threshold.

Freq-Shift	Accuracy	F1 score	Precision	Recall	Threshold
1h-3	0.833333	0.800000	1.000000	0.888889	1.90
30min-4	0.787879	0.866667	0.722222	0.787879	2.25
45min-3	0.705882	0.700000	0.777778	0.736842	3.15
4h-1	0.661017	0.658537	0.818182	0.729730	0.20
2h-2	0.647059	0.777778	0.636364	0.700000	0.75
5min-2	0.630137	0.658537	0.675000	0.666667	9.10
15min-4	0.586207	0.777778	0.636364	0.700000	7.40

Table 4.6: The shift and threshold evaluation for each frequency that maximizes accuracy. **Note that freq/shift/threshold combinations resulting in less than 10 predicts are discarded.**

Freq-Shift	|Predicts|	Chances Taken
1h-3	12.0	0.016129
30min-4	33.0	0.022177
45min-3	17.0	0.018280
4h-1	59.0	0.317204
2h-2	17.0	0.045699
5min-2	146.0	0.016353
15min-4	29.0	0.009745

Table 4.7: Number of predicts and ratio of intervals predicted for.

Chapter 5

Discussion

5.1 Prediction model

5.1.1 Evaluating predictions

The figures in section 4.4 represent how the predictions perform accuracy wise as threshold increases towards 2%. The two smaller frequencies barely change for any shift as the threshold increases, the 45 minute figure 4.8 leaves little to be desired too. The 30 minute interval's shift 4, on the other hand, increases in an almost linear fashion from 50% to 72.5%. Resulting in a prediction error decrease by 45%. The 1 hour shift 3 similarly seems promising with the slightly more crooked climb from ca. 50% to 83% accuracy, decreasing the prediction error by 66%.

The remaining frequencies and shifts behave erratically, leaving little room for drawing any reasonable conclusions on whether a specific shift or threshold has the magic touch overall.

Note that the seemingly very accurate shifts for 2 and 4 hours aren't presented in the table 4.6 as these predict under 10 times.

5.1.2 Lackluster prediction opportunities

The number of predictions decrease rapidly as the fluctuation threshold increases, see table 4.6), and intuitively so, as the selected intervals to predict for are smaller and smaller subsets of each other as the minimum percent fluctuation increases. Notably, only the 5 and 15 minute interval can produce more than 100 predictions at a sentiment change of 2% for a given interval. As the table 4.7 shows, the 30 minute interval only makes 2.2% of the "possible" (i.e. the number of half hours during a month) predictions when the sentiment threshold is set to 2.25%. The other frequencies show similar results, except for 4h-1. The 4 hour, 1 shift, combination predicted for 32% of possible intervals, comparing sentiment change to a threshold of 0.2%.

It seems notable that merely 30% of the time (for 4 hours), the sentiment fluctuated by 0.2% or more. The sentiment compound value itself, calculated by Vader, seems questionable when observing these values. Although, the observation could come down to any number of reasons; objective tweeters (unlikely), too coarse or fine spam filtering, a normal distribution of sentiment (resulting in neutral averages on sentiments over a time frame) or lack of domain specific lexicon (missing the most damning or proving statements about the commodity due to fintech lingo).

5.2 Reconnecting with Problem statement

Is there a correlation between Twitter sentiment and BTC price fluctuation?

Given the method used, discussing correlation in relation to sentiment change and price fluctuation must be confined within the binary notion of if price indeed went up or down depending on a prediction. Note that according to the data presented in table 4.6, when not providing any threshold to compare sentiment fluctuations to, i.e. taking every sentiment fluctuation into account, the accuracy of predictions for all freq/shift combinations hover around 50%; indicating that the binary fluctuation for sentiment and BTC price has neither a positive nor negative correlation value. Although, for certain freq/shift counts, accuracy increases as the threshold for predictive fluctuations increases; indicating that the subsets with more prominent fluctuations can indeed be identified to have a positive correlation value. This would indicate a partial correlation between binary sentiment and price change for small subsets of data, dependant on threshold.

Can a naive prediction model based on aggregated sentiment fluctuation yield better than random accuracy?

By taking into account that accuracy corresponds to no correlation at 50%, this would also mean that a random accuracy corresponds to 50%. As can be seen in the table 4.6, there are viable prediction options given certain frequency/shifts and thresholds. Most notably 1h-3 (figure 4.9), 30min-4 (figure 4.7), 45min-3 (figure 4.8) all yielding an accuracy above 70% for a subset of the data. As touched upon previously, the predictions made are notably scarce, see figure 4.4 or figure 4.7.

5.3 Weaknesses

5.3.1 Static threshold

Prior to selecting a static threshold for the prediction model, a variation of more dynamic methods were attempted, including; comparing the sentiment change to simple rolling averages, exponential rolling averages, quantiles based on these averages, and all with a variation of window lengths and weights. Although, merely checking severity of one interval change to the next proved for the most accurate results. As priory noted though, the static threshold seems a possible candidate for the low ratio of predictions made. One possible solution not attempted is checking for patterns during time periods of the day, i.e. looking to identify more appropriate thresholds for noon, evening, or likewise, and possibly dynamically setting them with moving historical data.

5.3.2 Domain specific lexicon

The lexicon provided by VADER **??** is an all-round lexicon, capturing sentiment for the most common social media expression. Although, financial and cryptocurrencies trade terms could arguably be the most indicative of sentiment towards Bitcoin. In this sentiment analysis, a term such as "short" wouldn't be classified as negative, neither would the "bullish" in the table 4.5 example be classified as positive.

5.3.3 No indication of success

The predictions state if the price will rise or fall, not by how much or with any level of differentiating certainty (e.g. higher fluctuations or larger tweet volume could indicate higher certainty).

5.3.4 Lack of data

Even though 2 million tweets at first seem a lot, conclusively stating that a 1h-3 frequency/shift is a fair predictive basis, for the Bitcoin price, for so-and-so threshold is naive. Only 12 predictions were made for the highest accuracy, and all-though this may still prove true in the future, the analysis would have to run for far longer than a month to surely state any basis. How the prediction model would evaluate a down-trending Bitcoin remains uncertain; as the 4.1 shows, the analyzed month saw a large Bitcoin price increase.

5.4 Future research

5.4.1 Picking frequency/shift based on historical accuracy

Before the presented prediction model in this thesis took form and was ultimately chosen for the final revision, an attempt was made at a dynamic prediction model which picked frequency/shift based on the success of historical accuracy for the combination. The initial findings looked promising when predicting based on comparing sentiment change to the upper and lower 25% quantile for the previous 12 hour sentiment change. Although, as the dataset grew, the implemented method became in-feasible due to the large number of calculations and further proved difficult to reason with evaluation wise.

5.4.2 Machine learning

When first exploring the data and calculating correlation between Bitcoin data and Tweet data, the highest correlation value came from number of Tweets and Bitcoin price. This path was not taken any further due to the lack of connection to sentiment. Applying machine learning to the problem would be the next step; taking into account variables such as the number of Tweets, Bitcoin volume, weighted sentiments depending on historical accuracy of users, etc.

Chapter 6

Conclusion

This thesis studied if sentiment analysis on Twitter data, relating to Bitcoin, can serve as a predictive basis to indicate if Bitcoin price will rise or fall.

A naive prediction model was presented, based on the intensity of sentiment fluctuations from one time interval to the next. The model showed that the most accurate aggregated time to make predictions over was 1 hour, indicating a Bitcoin price change 4 hours into the future. Further, a prediction was only made when sentiment mean was limited by a minimum 2.2% change.

The primary conclusion is that even though the presented prediction model yielded a 83% accuracy, the number of predictions were so few that venturing into prediction model conclusions would be unfounded.

Further improvements of the analysis would begin with the lexicon, as improving the classifier by adding a domain-specific lexicon would identify financial and cryptocurrency terms and yield a more representative sentiment, hopefully improving the prediction accuracy.

Bibliography

[1] IBM Corporation. IBM - what is big data? URL https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[2] Lisa Lowe. 125 amazing social media statistics you should know in 2016. URL https://socialpilot.co/blog/125-amazing-social-media-statistics-know-2016/.

[3] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

[4] Pedro Franco. *Understanding Bitcoin: Cryptography, engineering and economics*. John Wiley & Sons, 2014.

[5] Hong Kee Sul, Alan R Dennis, and Lingyao Ivy Yuan. Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 2016.

[6] Stuart Colianni, Stephanie Rosales, and Michael Signorotti. Algorithmic trading of cryptocurrency based on twitter sentiment analysis. 2015. URL http://cs229.stanford.edu/proj2015/029_report.pdf.

[7] C.J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[8] Douglas Wikström. Lecture 12. May 2017.

[9] Barry J Eichengreen. *Globalizing capital: a history of the international monetary system*. Princeton University Press, 1998.

[10] Collin Thompson. How does the blockchain work (for dummies) explained simply. URL https://medium.com/the-intrepid-review/how-does-the-blockchain-work-for-dummies-explained-simply-9f94d386e093.

[11] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[12] Leif W. Lundmark, Chong Oh, and J. Cameron Verhaal. A little birdie told me: Social media, organizational legitimacy, and underpricing in initial public offerings. *Information Systems Frontiers*, pages 1–16, 2016. ISSN 1572-9419. doi: 10.1007/s10796-016-9654-x. URL http://dx.doi.org/10.1007/s10796-016-9654-x.

[13] Inc Statista. Number of worldwide social network users 2010-2020. URL https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[14] Inc Twitter. Company | about, . URL https://about.twitter.com/company.

[15] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[16] Inc Twitter. The twitter glossary | twitter help center, . URL https://support.twitter.com/articles/166337.

[17] Inc Twitter. Api overview — twitter developers, . URL https://dev.twitter.com/overview/api.

[18] Pieter de Jong, Sherif Elfayoumy, and Oliver Schnusenberg. From returns to tweets and back: An investigation of the stocks in the dow jones industrial average. *Journal of Behavioral Finance*, 18(1):54–64, 2017. doi: 10.1080/15427560.2017.1276066. URL http://dx.doi.org/10.1080/15427560.2017.1276066.

[19] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[20] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[21] Bing Liu. Opinion mining and sentiment analysis. In *Web Data Mining*, pages 459–526. Springer, 2011.

[22] C.J. Hutto. cjhutto/vadersentiment: Vader sentiment analysis. vader (valence aware dictionary and sentiment reasoner). URL `https://github.com/cjhutto/vaderSentiment`.

[23] Inc Pennebaker Conglomerates. Liwc | linguistic inquiry and word count. URL `http://liwc.wpengine.com/`.

[24] University of Florida. Center for the study of emotion and attention. URL `http://csea.phhp.ufl.edu/media/anewmessage.html`.

[25] Roger Hurwitz. General inquirer home page. URL `http://wjh.harvard.edu/~inquirer/`.

[26] Bitcoin price index api - coindesk, . URL `http://www.coindesk.com/api/`.

[27] Tweepy, . URL `http://www.tweepy.org/`.