SMU Data Science Review

Volume 1 | Number 3 Article 1

2018

Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis

Jethin Abraham
Southern Methodist University, jethina@smu.edu

Daniel Higdon
Southern Methodist University, dhigdon@smu.edu

John Nelson
Southern Methodist University, nelsonjohn@smu.edu

Juan Ibarra
Southern Methodist University, jibarralopez@smu.edu

Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis

Jethin Abraham, Daniel Higdon, Jack Nelson, Juan Ibarra

Master of Science in Data Science
Southern Methodist University
Dallas, Texas USA
{jethina, dhigdon, nelsonjohn, jibarralopez}@smu.edu

Abstract. In this paper, we present a method for predicting changes in Bitcoin and Ethereum prices utilizing Twitter data and Google Trends data. Bitcoin and Ethereum, the two largest cryptocurrencies in terms of market capitalization represent over $160 billion dollars in combined value. However, both Bitcoin and Ethereum have experienced significant price swings on both daily and long term valuations. Twitter is increasingly used as a news source influencing purchase decisions by informing users of the currency and its increasing popularity. As a result, quickly understanding the impact of tweets on price direction can provide a purchasing and selling advantage to a cryptocurrency user or a trader. By analyzing tweets, we found that tweet volume, rather than tweet sentiment (which is invariably overall positive regardless of price direction), is a predictor of price direction. By utilizing a linear model that takes as input tweets and Google Trends data, we were able to accurately predict the direction of price changes. By utilizing this model, a person is able to make better informed purchase and selling decisions related to Bitcoin and Ethereum.

1 Introduction

By May 2018, the two largest cryptocurrencies, measured in terms of market capitalization, had a combined market value of 160.9 billion dollars[1]. Bitcoin alone made up nearly $115 billion of this value. Given the significant value of these currencies, some people see value in them through use as actual currencies, while others view them as investment opportunities. The result has been large swings in value of both currencies over short periods of time. During 2017 the value of a single Bitcoin increased 2000% going from $863 on January 9, 2017 to a high of $17,550 on December 11, 2017. By eight weeks later, on February 5, 2018, the price of a single Bitcoin had been more than halved with a value of $7,9643. The promising technology behind cryptocurrencies, the blockchain, makes its likely that cryptocurrencies will continue to be used in some capacity, and that their use will grow.

[1] https://bitinfocharts.com/

The volatility in the value of cryptocurrencies means uncertainty for both investors, and people who wish to use them as a currency rather than an investment. Cryptocurrencies are a relatively new store of value (Bitcoin was created in 2009 [1]) relative to fiat currencies such as the U.S. dollar (USD) or the Japanese Yen. What causes price changes in this new store of value is an area of debate. Researcher Ladislav Kiroufek found that Bitcoin is a unique asset in that its price behaves in ways similar to both a standard financial asset and a speculative one [2]. Given that cryptocurrency prices do not behave like traditional currencies the prices are extremely difficult to predict.

In this paper, we present a solution to predicting cryptocurrency price changes.To accomplish this, methods utilizing sentiment analysis of tweets are reviewed. This is involved utilizing Twitter's API and a Python library called "Tweepy"[2] to collect and store tweets which mentioned Bitcoin or Ethereum. The tweets were then analyzed to create a sentiment score by day and compared to the price changes to that day to determine if a relationship between Twitter sentiment and cryptocurrency price changes could be determined.

Several researchers including Choi and Varian [3] and Ettredge et al. [4] found that web-based search data in the case of Ettredge et al., and Google Trends data specifically in the case of Choi and Varian, could be used to make predictions of several macroeconomic statistics including automobile sales and unemployment rates. Building off of this research we identify the correlations that exist between Bitcoin and Ethereum prices and Google Trends data. Finally, we include tweet volume about each cryptocurrency as inputs into a linear model and summarize the results.

The findings of our analyses show that sentiment analysis is less effective for cryptocurrency price changes in an environment in which prices are falling. This is because tweets about cryptocurrencies tend to be objective in nature (not having a clear sentiment) or positive regardless of price changes. While their value has exploded in recent years, they still pale in use compared to traditional fiat currencies. In addition, cryptocurrencies are actually a part of a larger technology (the blockchain). As such, Twitter activity about them can be driven by people with a special interest in the currency or the technology rather than just a store of value, as a traditional stock may be viewed. Google Trends data and tweet volume better reflects the overall interest in owning cryptocurrencies as they increase and decrease with prices.

Section 2 provides a brief overview of important topics the reader will need to understand to follow the research. These include cryptocurrency, Twitter, sentiment analysis, and Google Trends. Section 3 discusses previous work including the rich fields of research that this paper builds from. Section 4 describes how the data was collected and what the final data set looked like. Section 5 discusses the research methodology and how we chose the inputs for the model. Section 6 describes the model outputs. Section 7 discusses the ethical considerations when using publicly acquired data like tweets. Finally, Section 8 summarizes our conclusions and the opportunities for future work.

[2] http://docs.tweepy.org/en/v3.5.0/

2 Cryptocurrency, Twitter, and Sentiment Analysis Background

The analysis detailed later in this paper requires an understanding of where and why the data was collected, and how cryptocurrencies may vary from standard fait currencies or stocks in companies from traditional stock markets. In this section we will provide more background on these data sources and why they were chosen so that the final analysis is put in the proper context for the reader.

2.1 Blockchain Technology and Cryptocurrencies

In this paper we analyze data about the world's two largest cryptocurrencies in terms of market capitalization. The largest is Bitcoin followed by Ethereum. Bitcoin was the first cryptocurrency ever created. The creation of Bitcoin is mysterious as it was created by a person or group of people using the name "Satoshi Nakamoto" and released in 2009[3]. Along with the launch of Bitcoin "Satoshi Nakamoto" published a paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" which described a peer-to-peer payment system using electronic cash (cryptocurrencies) that could be sent directly from one party to another without the use of a third party to validate the transaction. This innovation is created by the use of the "blockchain" which is like a shared ledger on the peer-to-peer network where all transactions are verified by the network so they cannot be forged [1].

The applications of blockchain technology go beyond peer-to-peer payment systems. Blockchain technology provides security, privacy, and a distributed ledger which makes them applicable for internet-of-things applications, distributed storage systems, healthcare, and more [5]. The range of applications of the blockchain has led to many more blockchains and cryptocurrencies being created (1,658 cryptocurrencies are in existence[4]). Cryptocurrencies are tied to the blockchain because they provide the incentive for machines, and the electricity they consume, to run and validate the blockchain. As use of blockchains increases so too will the use of cryptocurrencies. This gives them an inherent value, but what that value is depends on many factors. Because this is a new type of currency, and store of value, improving the understanding of what can lead to price changes brings with it value.

2.2 Twitter

Twitter was launched in July of 2006 as an application in both the social media space (which includes other applications/websites such as Instagram, Facebook, LinkedIn and others) and microblogging. Microblogging is a medium that allows

[3] https://en.wikipedia.org/wiki/Bitcoin
[4] https://www.fool.com/investing/2018/03/16/how-many-cryptocurrencies-are-there.aspx

for smaller and more frequent updates than blogging[5]. Twitter allows users to post messages publicly (which are referred to as "tweets") with a maximum length of 140 characters[6]. In November of 2017 that limit was doubled to 280 characters. In addition, users can add "hashtags" to a tweet, which is the "#" symbol followed a consecutive string of characters. This is used to identify the topic or theme of a tweet and to make them searchable. This is used later when we collect the tweets in the data section.

Since its launch in 2006 Twitter has grown rapidly in popularity. One of the early examples of its reach and power was on January 15, 2009 when a US Airways flight crashed into the Hudson river. An image posted to Twitter broke the news before traditional media outlets did[7]. Twitter has 330 million monthly active users, 1.3 billion accounts have been created, 83% of the world's leaders have a Twitter account, approximately 23 million of Twitter's active users are actually bots rather than humans, and 500 millions tweets are sent each day[8]. The result of all of these impressive statistics is that Twitter can be a very rich source of data on how people feel about nearly any given topic. With the ability to see when a tweet was posted it is also possible to tell how those feelings change over time. This makes Twitter an excellent resource to collect text data on a topic like cryptocurrencies to explore the possible relationships between that and prices.

2.3 Sentiment Analysis

It is estimated that 90% of the data in the world has been created in the last two years[9]. Much of that data is in the form of unstructured text data whether it be in the form of tweets, articles posted to the internet, text messages, emails, or other forms. This vast amount of unstructured data has led to the creation of "natural language processing" (NLP) as an area of study or development. NLP is a collection of methods for computers to analyze and understand text[10].

In this paper we use a set of natural language processing tools commonly referred to as "sentiment analysis". Sentiment analysis is the act of extracting and measuring the subjective emotions or opinions that are expressed in text. There are multiple ways to do this. We chose the "VADER" (Valence Aware Dictionary and sEntiment Reasoner) [6] system in this analysis, which will be described in more detail in the methodology section. The end goal of this analysis is to apply sentiment analysis to collected tweets so that it can be determined if the tweets are generally positive or negative in their opinions of cryptocurrencies. In addition, we also want to use sentiment analysis to identify tweets which

[5] https://en.wikipedia.org/wiki/Microblogging
[6] https://en.wikipedia.org/wiki/Twitter
[7] https://www.brandwatch.com/blog/44-twitter-stats/
[8] https://www.brandwatch.com/blog/44-twitter-stats/
[9] http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/
[10] https://blog.algorithmia.com/introduction-natural-language-processing-nlp/

express an opinion (subjective tweets) versus those that just provide information without a positive or negative angle to them (objective tweets).

2.4 Google Trends

In many parts of the world nearly every aspect of day-to-day life now involves the internet. How the internet is navigated is through search engines and Google is far and away the world's most popular search engine as it accounts for 74.52% of all internet searches[11]. This means that Google search data can provide incredible insights into what the world is interested in, and how interested in any given topic it is.

Google makes this data available through "Google Trends". Google Trends data provides information on how popular given search terms are relative to other search terms at any given time. In addition, these search term popularity values can be compared over time. This provides a proxy metric for the general interest there is in cryptocurrencies at any given time, which could have a relationship with cryptocurrency prices over time as general interest increases and decreases.

3 Related Work

This paper builds on the ideas of a wide range of research and topics. Behavioral economists like Daniel Kahneman and Amos Tversky established that decisions, even ones involving financial consequences, are impacted by emotion and not just value alone [7]. R. J. Dolan's work in "Emotion, Cognition, and Behavior" further supports that decision making is highly impacted by emotions [8]. The insights from these researchers opens up the possibilities to find advantages through tools like sentiment analysis as it indicates that demand for a good, and therefore price, may be impacted by more than it's economic fundamentals.

Later researchers found specifically that purchase decisions people made were being impacted from information gathered online. Galen Thomas Panger found that Twitter sentiment correlated with people's general emotional state. Additionally, he found that social media like Twitter tended to have a calming affect on the end-user rather than amplifying their emotional state [9]. Chen et al. performed textual analysis on a social platform aimed at investors called "Seeking Alpha" and found that views expressed in articles posted on "Seeking Alpha" were associated with returns and could even predict earnings surprises [10]. In a similar vein Paul Tetlock found that high levels of media pessimism of the stock market impacted trading volumes [11]. Finally, Gartner found in a study that the majority of consumers relied on social media to influence purchase decisions [12].

Other researchers have specifically studied the efficacy of sentiment analysis of tweets. Kouloumpis et al. found that standard natural language processing techniques such as sentence level and document level sentiment scoring was ineffective due to the short nature of tweets and uniqueness of language

[11] https://www.reliablesoft.net/top-10-search-engines-in-the-world/

used [13]. Alexander Pak and Patrick Paroubek showed that separating tweets into positive, negative, or neutral categories could result in effective sentiment analyses [14]. O'Connor et al. showed that the sentiment found in tweets was reflective of public opinion on various topics in national polling [15]. Their research identified sentiment analysis as a cost saving option versus national polling, but the implication that sentiments from tweets do accurately reflect the larger population's feelings on topics suggests that it could also be used to predict demand, and therefore price changes of products.

Web data beyond Twitter and social media has been a rich area of research as well. To our knowledge one of the first papers to find that web search data could be used to predict macroeconomic indicators was by Ettredge et al. where they found that searches relating to employment was associated with unemployment rates [4]. Bordino et al. found that query volumes were correlated with trading volumes for stocks in the NASDAQ [16]. Specific research into Google Trends data has been done as well by Hyunyoung Choi and Hal Varian with the conclusion that simple seasonal auto-regressive models which included Google Trends data as inputs outperformed models that did not use Google Trends data by 5% to 20% [3]. Asur et al. found that tweet volume about recently released movies accurately predicted box-office receipts [17].

Having established that decisions are influenced by emotions, that social media can impact decisions, that sentiment analysis of social media can accurately reflect the larger population's opinions towards something, and that web search data can predict changes in macroeconomic statistics, much research built off of those findings to see if they applied to the stock market. Alan Dennis and Lingyao Yuan collected valence scores on tweets about the companies in the S&P 500 and found that they correlated with stock prices [18]. Pieter de Jong et al. analyzed minute-by-minute stock price and tweet data for 30 stocks in the DOW Jones Industrial Average and found that 87% of stock returns were influenced by the tweets. However, they also looked for the inverse happening, that stock prices were influencing tweets and found little evidence for it [19]. Bollen et al. used a self-organizing fuzzy neural network, with Twitter mood from sentiment as an input, to predict price changes in the DOW Jones Industrial average and achieved 86.7% accuracy [20].

With the introduction of cryptocurrencies similar work has been done to see if such methods effectively predict cryptocurrency price changes. In the paper "Predicting Bitcoin price fluctuation with Twitter sentiment analysis" by Evita Stenqvist and Jacob Lönnö, the authors describe their process in which they collected tweets related to Bitcoin, and Bitcoin prices from May 11 to June 11 in 2017. Tweets were cleaned of non-alphanumeric symbols (using "#" and "@" as examples of symbols removed). Then tweets which were not relevant or determined to be too influential were removed from the analysis. The authors then used VADER (Valence Aware Dictionary and sEntiment Reasoner) to analyze the sentiment of each tweet and classify it as negative, neutral, or positive. Only tweets that could be considered positive or negative were kept in the final analysis [21]. Connor Lamon et al. used sentiment of news headlines and tweets to

predict changes in Bitcoin, Litecoin (one of the many alternate cryptocurrencies now available on the market), and Ethereum. The study found that logistic regression performed best to classify these tweets and that they were able to correctly predict 43.9% of price increases correctly and 61.9% of price decreases [22]. Colianni et al. collected tweets from November 15, 2015 to December 3, 2015 and used Naive Bayes and Support Vector Machines to classify tweets and achieved a 255 accuracy increase [23]. Finally, Shah et al. successfully established a trading strategy using historical prices and Bayesian regression analysis [24].

Another branch of research in this area involves various applications of neural networks. Kimoto et al. used a modular neural network to create a buying and selling timing system for stocks on the Tokyo stock exchange and achieved profitability using their system with simulated stock purchases [25]. Guresen et al. compared various neural network performance in forecasting stock exchange rates and found that a multilayer perceptron (MLP) neural network performed best [26]. Xhu et al. used stock trading volumes as a neural network inputs and found that they modestly improved prediction performance over the medium and long terms [27].

The research presented in this paper builds off of everything above, but is unique in that we solve the problem of predicting cryptocurrency prices changes by combining web search data (in the form of Google Trends) and tweet volume as inputs into a linear model. In addition, we show why sentiment analysis is less useful in its predictive capabilities of cryptocurrencies despite its potential in other areas.

4 Data

To solve the problem of predicting cryptocurrency price changes several different data sources are considered as possible inputs to the model. The first input considered is sentiment analysis of collected tweets about Bitcoin or Ethereum. The second was Google Trends data, and the third was tweet volume. This section details how each of these data sources were gathered, cleaned, and adjusted when necessary.

4.1 Collecting Tweets from Twitter's API

The first step in collecting the desired tweets was to find the hashtag for the cryptocurrencies. For this we utilize Tweepy - an open-source Python library for accessing the Twitter API, to collect Twitter data. Tweepy allows for filtering based on hashtags or words. There are multiple ways in which the cryptocurrencies of interest may be referred to in tweets. The most direct way is by using a hastag ("#") followed by "bitcoin" or "ethereum". Other likely possibilities are using a hashtag and either currencies abbreviation ("#btc" for Bitcoin, and "#eth" or Ethereum). Early collections of tweets using only the "#bitcoin" and "#ethereum" hashtags quickly provided a large data set. As these hashtags had little ambiguity they were selected as the only ones we would use to collect tweets

through Tweepy. Additional data collected for each post included the user ID, a unique identifier which cannot be changed, the time stamp, and how many times the tweet was "retweeted" (someone posted the exact same tweet so that their followers could see it) and how many times it was "favorited".

The tweets are filtered for the language English for this analysis as tweets could be multilingual. The tweets were collected using a script to run every 15 minutes and collect 1,500 tweets for an instance. The script was scheduled to run automatically every 15 minutes to collect tweets for our analysis. This process was followed for 60 consecutive days and the volume of tweets vary depend on the range of active session at the time tweets were collected. In total the final tweet dataset was made up of 30,420,063 tweets.

4.2 Cleaning the Tweets for Analysis

With the tweets collected further processing is required. Tweets come in a format with characters which do not provide "information" for a sentiment analysis.

VADER (Valence Aware Dictionary for sEntiment Reasoning) sentiment analysis was used to analyze the collected tweets. VADER analysis provides several benefits including the fact that it not only classifies text as positive, negative, or neutral, but also measures the intensity, or polarity, of words used. For our purposes, we also benefit from the fact that the words and scores used in VADER are specifically tuned to microblog and social media contexts. To eliminate noise from the analysis we first clean the collected tweets.

Tweets contain a large amount of noise, such as hashtags, URLs, and emotions. These characters make Twitter sentiment analysis a challenging assignment. Preprocessing of the data is a very important step as it decides the efficiency of the other steps down in line for sentiment analysis. Preprocessing generally consists of removing capitalizing so that common words are represented as such, and stemming words so that tense is removed as they provide similar sentiment (run, ran, and running all represent similar information). For this we used pre-processing packages readily available and also the use of regular expressions. Broadly speaking, regular expressions are a collection of patterns that can be used to identify certain kinds of text and to clean text of erroneous patterns. Regular expressions was used to remove the # tags , quotes and question marks were also removed as it causes biased results for sentiment analysis. The https links were removed as well using regular expressions.

4.3 Collecting and Adjusting Google Trends Data

Google provides trends data, which is an unbiased sample of search data, as far back as 2004[12]. The data Google provides is not search volumes, but a search volume index (SVI). The search volume index is calculated by dividing each data point by the total searches within a geographic region and time range. The numbers are then scaled between 0 and 100 on a search term's proportion to all

[12] https://support.google.com/trends/answer/4355213?hl=en

searches on all topics[13]. When trends data is queried for a period of longer than 90 days the SVI returned are aggregated at a weekly level. In order to compare these SVIs across periods and adjustment has to be made. For our research we used the method detailed by Erik Johansson.

The method has four primary steps. First, collect all of the daily SVI data you need in 90 day increments and combine them into a single increment covering the entire time period of interest. Second, line up the data for the same entire time period, but aggregated at a weekly level to get the weekly SVI. Determine an adjustment factor which is done by dividing the weekly SVI with the daily SVI value where the dates overlap. Finally, multiple the daily SVI values by the adjustment factor[14]. In cases when the SVI was less than 1, the value was returned by the Google Trends query as < 1. To allow for an adjustment calculation we changed that value to 0.5. Google does not provide any more information on what the specific value was, so the halfway value of 0.5 was used as a substitute.

As we did when collecting tweets, we elected to collect Google Trends data using the least ambiguous terms as possible which were "bitcoin" and "ethereum". Each currency's abbreviation, "BTC" and "ETH" respectively were not used.

4.4 Collecting Tweet Volume

Twitter's API which allowed us to collect tweets through Tweepy also limits the number of tweets that can be collected to 1,500 tweets per instance. While this created a large data set of tweets that were a random sample of what was being tweeted while the program was running, it did not allow us to get a count of total tweets about the cryptocurrencies for any given day.

However, www.bitinfocharts.com freely provides the number of tweets by day about both of these cryptocurrencies dating back to April of 2014. This website was our source for total tweet volume of Bitcoin and Ethereum tweets.

5 Data Analysis

With the data collected, cleaned, and adjusted where needed the data was analyzed to determine if it would be a valuable input to the final model. In the case of tweets there are two remaining issues after cleaning them. First it must be determined how many of the tweets actually have a sentiment at all. If most of the tweets are not objective in nature, then a sentiment analysis of them adds little information to the model. Second, it has to be established that a relationship between the sentiment of tweets about cryptocurrencies and cryptocurrency price changes exists. Otherwise it will only introduce noise into the model. This is similarly true of Google Trends data and tweet volume. If a relationship between these metrics and price changes does not appear to be present, then they

[13] https://support.google.com/trends/answer/4365533?hl=en

[14] http://erikjohansson.blogspot.co.uk/2014/12/creating-daily-search-volume-data-from.html

will not be valuable model inputs. This section details the analyses to determine which data are suitable as model inputs.

5.1 Sentiment Analysis of Tweets

Not all tweets are posted by humans. A substantial number of users and tweets are actually from bots. Twitter has estimated that as many as 23 million of its active users are actually bots. If the bots were sending tweets which contained positive or negative sentiment about the cryptocurrencies then those tweets may still have an influence on people's demand to own cryptocurrencies, and therefore impact the prices. However, many of the tweets do not contain any sentiment and instead provide only facts or are serving the function of advertising. Beyond bots there is concern of the subject matter. Conversations about cryptocurrencies can be very neutral in nature. What the current USD price of a single bitcoin is a fact and does not carry any sentiment. Therefore, sentiment analysis of the tweets may provide limited information to the model. After pre-processing the collected tweets, algorithm results showed the information gained from the tweets through sentiment analysis is still of limited value. Overall tweets were generic, generated by bots, or advertisements. Bitcoin (Figure 1) and Ethereum (Figure 2) show only half of the tweets collected on any given day had any objective VADER score. All other tweets were strictly neutral.

Furthermore, tweets where an objective VADER score was extracted, positive or negative, found scores well below a 0.5 threshold. Gamma kernels for objective and neutral VADER sentiment scores on all tweets with a non-neutral score above 0.0 show little overlap in distributions. Bitcoin (Figure 3) and Ethereum (Figure 4) distribution plots indicate that even though positive or negative sentiment was gathered, VADER sentiment analysis determined tweets to be more neutral than objective overall.

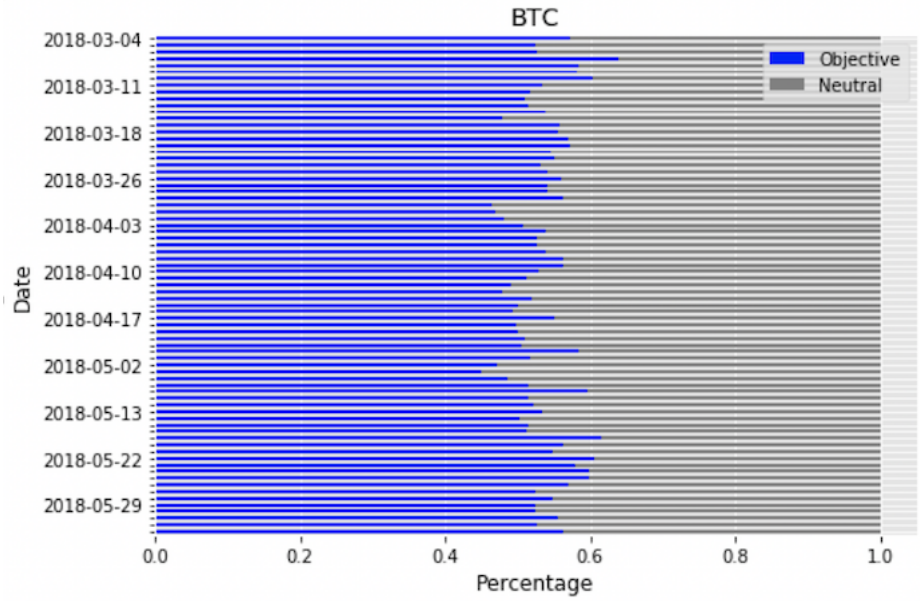

Fig. 1. Bitcoin percent of objective (having a positive or negative sentiment) versus neutral tweets *(blue bar represent objective tweets, gray bars represent neutral tweets)*. Charts created in Python.

Although only half of the tweets collected provide positive or negative sentiment, of which, is below a level to consider the tweet objective overall, it is still possible that the positive or negative sentiment gathered could provide valuable information to the model if a relationship between sentiment and price changes is present. Figures 5 and 6 below have data from March 4, 2018 to June 3, 2018. In both figures the blue line, relating to the left vertical axis, shows the day-over-day (Bitcoin with respect to figure 5, and Ethereum with respect to figure 6). The purple line, relating to the right vertical axis, shows the tweet sentiment by day. This period shows consistent day-over-day price variability with 11 of 19 days seeing a price decrease, and 8 of 19 days seeing a price increase. Conversely, the tweet sentiment remains consistent. Only one day, March 7, 2018, saw tweet sentiment drop below 0 in the case of Bitcoin. There was not a single day where tweet sentiment about Ethereum dropped below 0 despite the price fluctuations.

The analyses establishes that Twitter sentiment is not consistent price changes when prices are falling during the period of March 4, 2018 to March 24, 2018. Due to the lack of a clear relationship between the sentiment of tweets and prices changes the sentiment analysis will not be used as an input to the model.

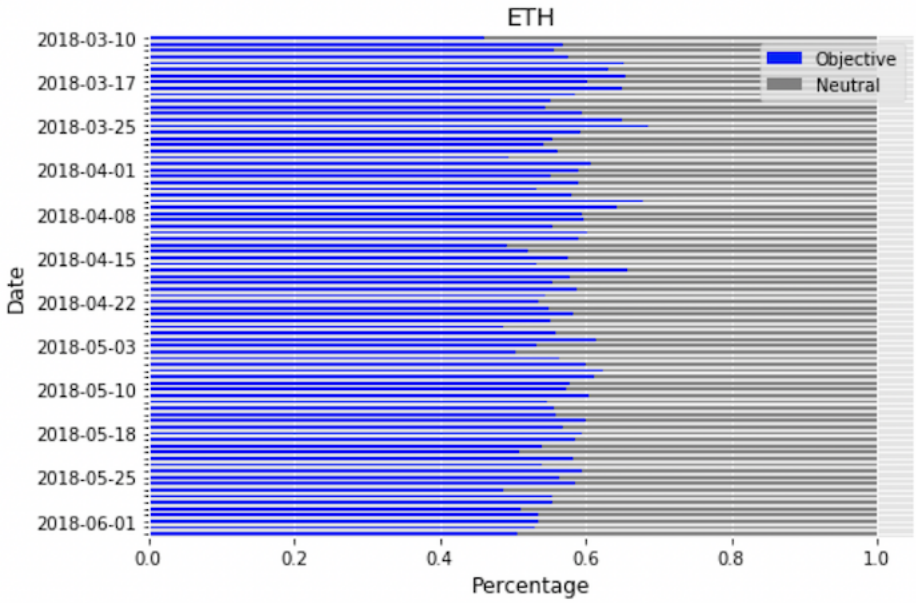

Fig. 2. Ethereum percent of objective (having a positive or negative sentiment) versus neutral tweets *(blue bar represent objective tweets, gray bars represent neutral tweets)*. Charts created in Python.

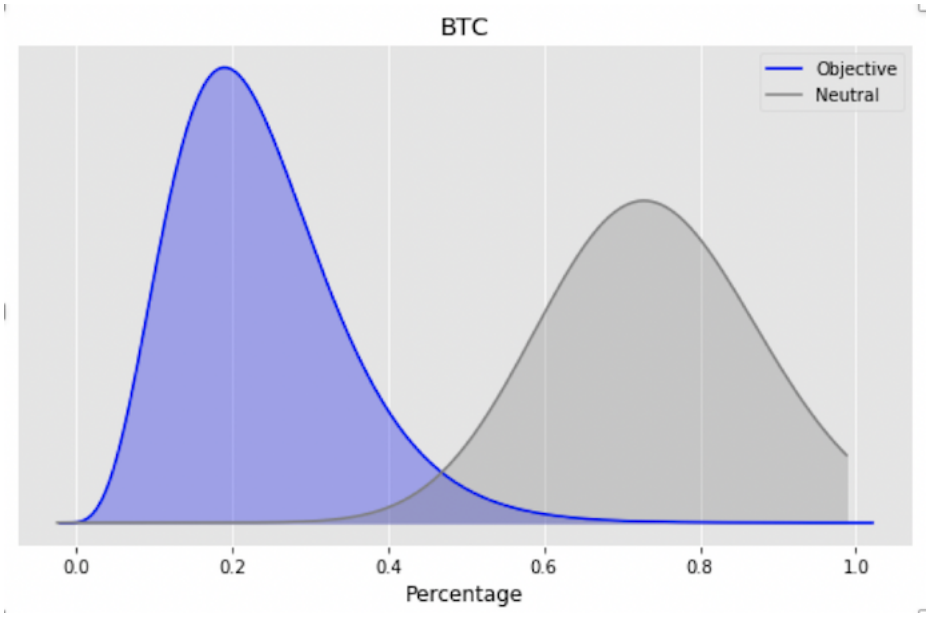

Fig. 3. *Bitcoin objectivity distribution (blue is objective, gray is neutral)*. Chart created in Python.

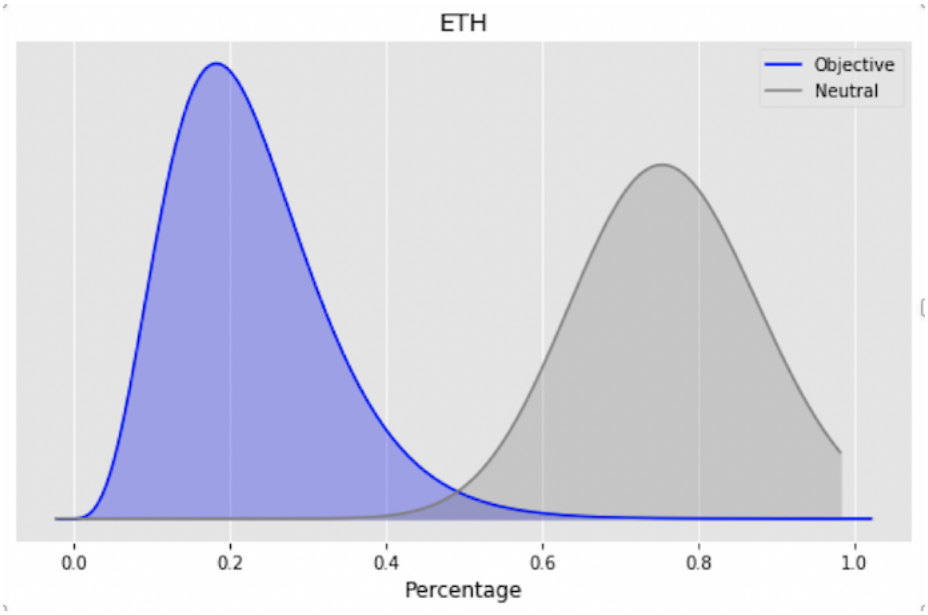

Fig. 4. *Ethereum objectivity distribution (blue is objective, gray is neutral).* Chart created in Python.

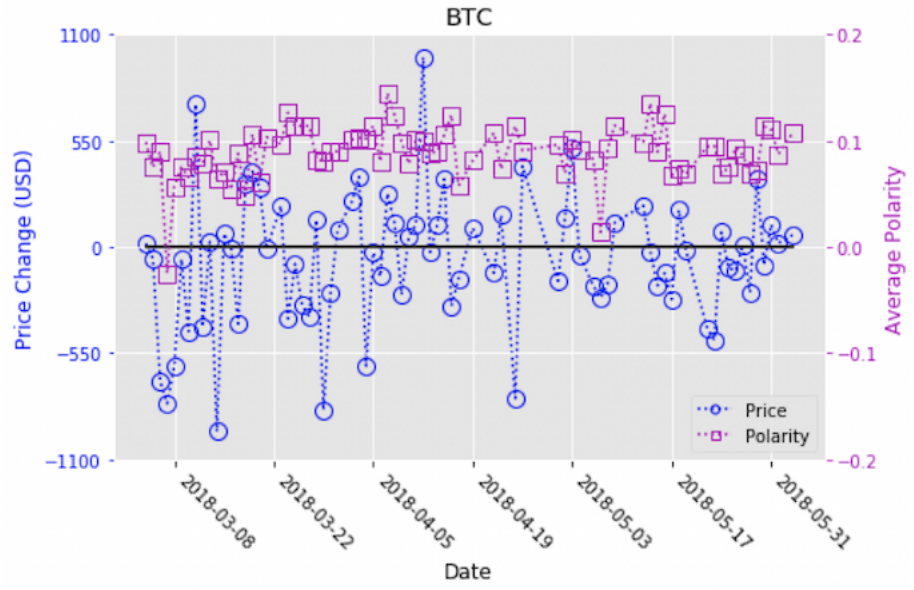

Fig. 5. *Bitcoin price change (blue line, left vertical axis)* and *Daily Average Tweet Polarity (purple line, right vertical axis)* by *date (horizontal axis).* Chart created in Python.

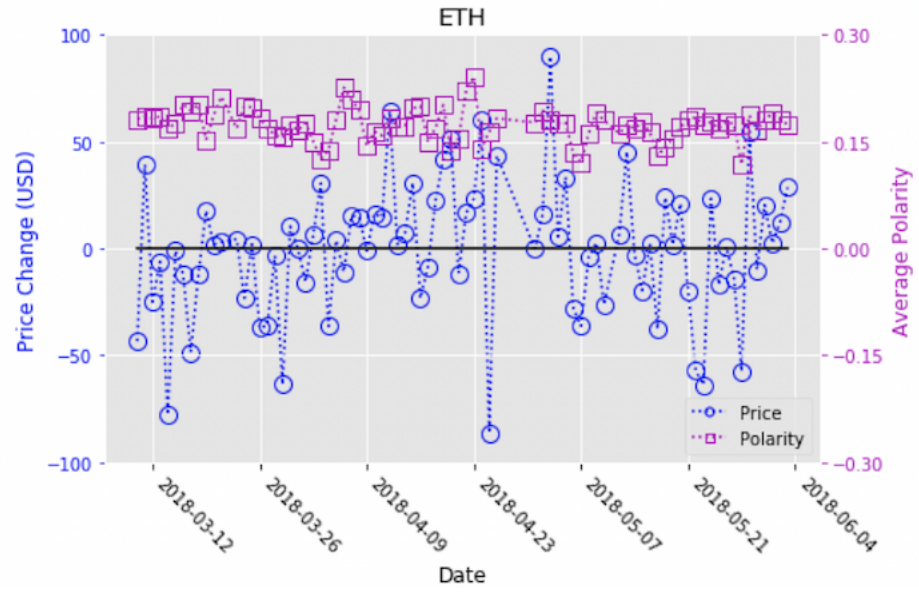

Fig. 6. *Ethereum price change (blue line, left vertical axis)* and *Daily Average Tweet Polarity (purple line, right vertical axis)* by *date (horizontal axis).* Chart created in Python.

5.2 Google Trends and Cryptocurrency Prices

To determine if a relationship is present between Google Trends search data and cryptocurrency price changes a correlation was ran for both currencies. The two primary metrics to determine this are the "Pearson R" and the p-value. The Pearson R is a measure of the strength of the correlation. Its value ranges from -1 to 1. A positive value means that the two variables are positively correlated, or that an increase in one variable is associated with an increase in the other variable (this is correlation, not causation, so we can't conclude that the change in one variable causes a change in the other, just that there is a relationship). Conversely, a negative value means that the two variables are negatively correlated, or an increase in one variables value is related to a decrease in the other variable. The p-value tells us how likely it is that these correlation measures would have been found by random chance. So the smaller the p-value, the more confident we can be that a relationship is in fact present, and not the result of random chance. Figure 7 below shows the correlation between Bitcoin Google trends data. The line chart shows that the price is highly correlated with Google Trends data. The pattern holds both in periods of increasing and decreasing prices. The Pearson R of the correlation is 0.817 with a p-value of 0.000.

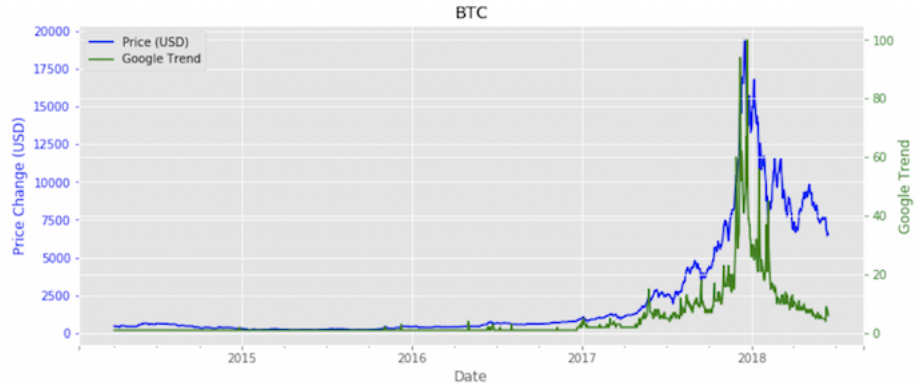

Fig. 7. Google Trends SVI values*(green line)* for Bitcoin after performing Erik Johansson's adjustment to compare daily data across time horizons longer than 90 days. Bitcoin price shown by the blue line.The plot was created in Python using matplotlib.

5.3 Tweet Volume and Cryptocurrency Prices

The final model input to review is tweet volume. Sentiment of tweets tended to stay positive regardless of price changes (when the tweets had any sentiment at all). This could be because when prices are falling people who still tweet about cryptocurrencies have an interest in cryptocurrencies for their other attributes besides value such as privacy. However, that is a fact of the technology and not something which ebbs and flows as price does. This would suggest that tweet volume is a better metric than sentiment as the amount of people talking about cryptocurrencies on Twitter may fluctuate with prices. Figure 8 below shows the Bitcoin price by day with tweet volumes. Similar to Google Trends the chart shows a strong correlation. It is also promising that the correlation holds when prices increase or when they are decreasing. The Pearson R was 0.841 with a p-value of 0.000.

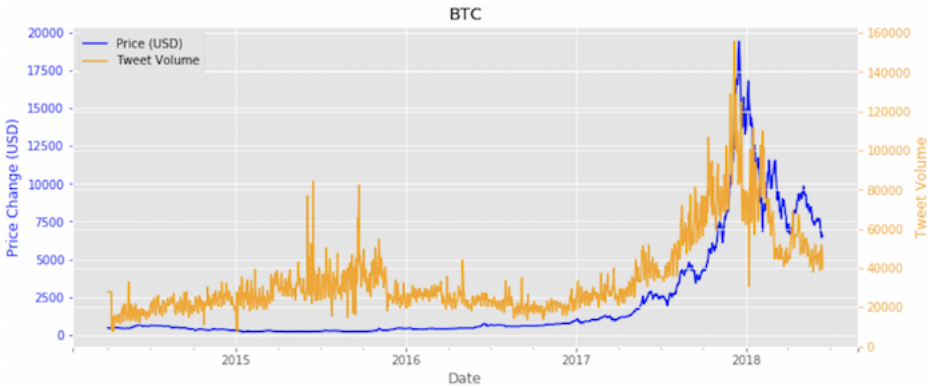

Fig. 8. Bitcoin tweet volumes*(yellow line)* by day. Bitcoin price shown by the blue line.The plot was created in Python using matplotlib.

6 Results

Three model inputs were considered. Sentiment of tweets was determined to not be a reliable indicator when cryptocurrency prices were falling so it is not included. Both Google Trends and tweet volume were highly correlated with price. In addition, the correlation held during periods of increasing and decreasing prices suggest that the relationship is robust to periods of high variance and non-linearity. A direct one to one comparison was made by using a linear model since input variables followed the same non-linear tendencies as the response. In addition, the inclusion of input variables from the past 15 days as separate inputs allowed for short term trends to be considered.

The model procedure followed fundamental machine learning protocol, the overall dataset was split into two sections, 80 percent for training the model and 20 percent for testing it. Cross validation was not included into the model, however, further exploring its use in future work can determine if a more generalized model is produced. Multiple linear regression was selected as the modeling algorithm of choice due to strong correlation metrics.

Figure 9 shows model residuals when using a 15-day window of Google Trends and tweet volume data in linear regression algorithm to predict Bitcoin closing daily price. Model price predictions are plotted on the y axis while actual prices are plotted on the x axis. A blue line shows where perfect prediction resides, and green and red markers indicate if the point resided in the training or test dataset respectively. Figure 10 takes the same training and testing predictions but plots them on a time series with actual closing daily price is on the y axis. The predictions climb up and down the non-linear portions of the time series as was observed in the initial input variable correlation plots.

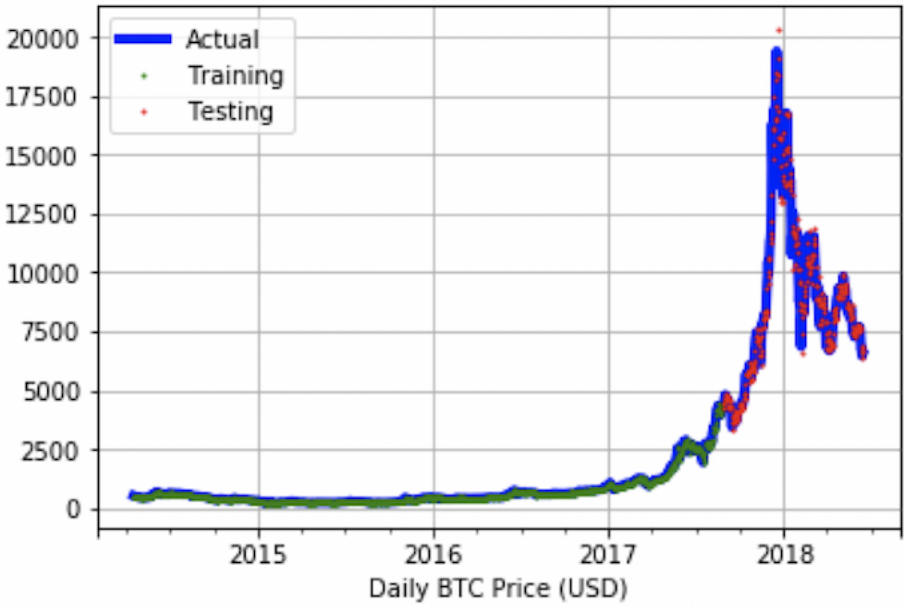

Fig. 9. Model fit shows as actual price*(blue line)* for Bitcoin, training data shown by green dots, and test results as red dots. The plot was created in Python using matplotlib.

7 Ethics

Working with Twitter data introduces ethical concerns around the privacy of the author of the tweets collected and the ethical responsibility to maintain the privacy of those entrusting their data to you. All of the tweets collected are made publicly available by the author through the acting of writing and sending the "tweet" through Twitter. Additionally, Twitter makes the tweets accessible through their API (application programming interface). This does not mean that the authors of the tweets realize the many different ways in which what they tweet can be accessed and used. Some may believe that only their friends or other people who choose to "follow" them on Twitter will ever see their tweets. As such, it is possible that collecting and working this data is making something more "public" than the author intended or realized it would be. With this in mind we chose to protect as much of the privacy of these public tweets as possible. Collected tweets were never stored in publicly viewable places such as GitHub. Additionally, when providing examples of tweets in this study usernames (often referred to as "Twitter handles") were removed as was any other personally identifiable information.

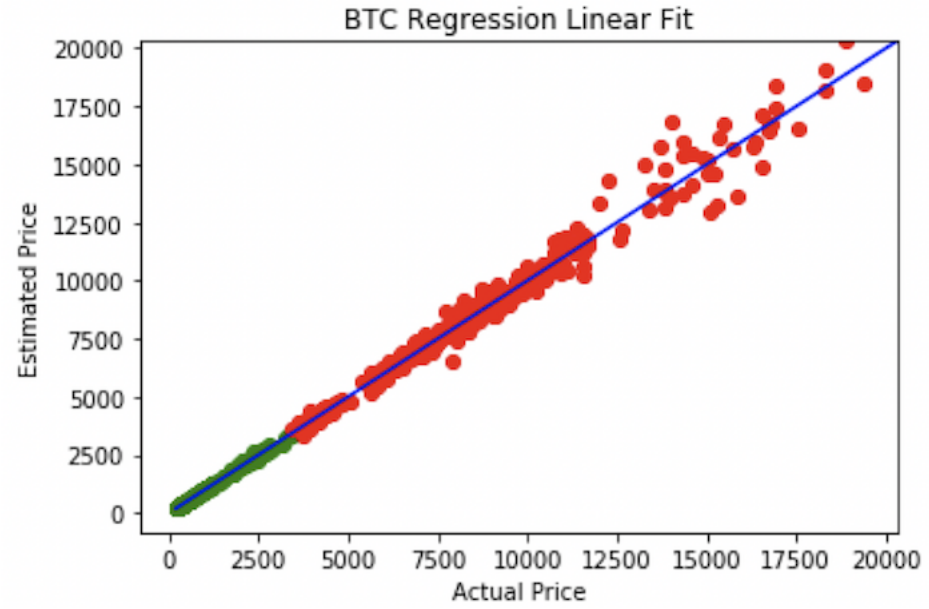

Fig. 10. Bitcoin regression fit shown as estimated price on the y-axis and actual price on the x-axis. Green dots are training data. Red dots are testing results. The plot was created in Python using matplotlib.

Additional ethical concerns surround the cryptocurrencies themselves. They are currently unregulated, which for many people is part of their appeal. Conversely, this allows for the possibility that the unregulated currency could empower people with harmful intents. One example is the case of the "Silk Road" on the "dark web" (basically internet sites that cannot be indexed and only accessed through special software)[15] where the anonymity of crypto-coins made it attractive as a currency of exchange for illegal goods or services[16]. However, another aspect of that is their value may be considered less predictable due to this lack of regulation. Providing a model to predict cryptocurrency price changes could lead to a false confidence in the efficacy of the model. This in turn could lead to people increasingly using cryptocurrencies as a store of value. If the market for the these cyrpto-coins was to experience a down turn, as it did in early 2018, and the model either did not predict or recognized it too late, people had stored their wealth in crypto-coins could stand to lose significant amounts of money. To address this efforts have been made to clearly explain what the model considers, and to honestly represent the results and the small sample of data they are based on. Extrapolating that to other time periods is an area of future study, or at the very least should be done knowing that it carries significant risks with it.

Finally, to the extent the model is useful in predicting future crypto-coin price changes, it also allows for manipulation of others. This could happen either by convincing others to invest as you have due to the model, or by disseminating false information to further increase the advantage provided by such a model. Any information given should be done so as honestly as possible. Additionally, to greatest extent possible and reasonable the caveats and risk of the information should be provided so that each person can make informed decisions.

The model should be viewed as a tool to help understand what can cause price changes in various cryptocurrencies. This allows people to make smarter decisions with how they store their wealth and to decrease their risk. However, not understanding the risks associated with it, or hiding these risks from others can result in detrimental results including a significant loss of wealth that someone was using to plan their future. This is not an area that should be taken lightly. Caveat emptor (let the buyer beware).

8 Conclusions and Future Work

Previous efforts to predict cryptocurrency fluctuations relied on Twitter sentiment analysis to serve as a proxy for future cryptocurrency demand which would result in increasing or decreasing prices. We have shown that these results were in part due to the study occurring at a time when cryptocurrency prices were always going up. Additionally, Twitter sentiment with respect to cryptocurrencies tend to be positive regardless of future price changes. People who tweet about

[15] https://www.iflscience.com/technology/what-dark-web/

[16] https://news.law.fordham.edu/jcfl/2018/02/21/silk-road-the-dark-side-of-cryptocurrency/

cryptocurrencies even when their prices drop have an interest in them beyond investment opportunity making the tweets biased towards positive. A more robust model would incorporate a measure of overall interest in terms of volume. This paper's recommendation is to use proxies for general interest such as Google Trends or tweet volumes. We have shown that the search volume index is highly correlated with cryptocurrency prices both when prices rise and when they fall, as are tweet volumes. With these inputs a multiple linear regression model, with the addition of lagged variables, accurately reflected future price changes.

Future work should determine if these results continue to hold in varying pricing environments. Additionally, more complex models, and not just linear ones like we used, could be fit using Google Trends and tweet volumes as inputs to see if results could be improved further.

References

1. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system. In: Cryptography Mailing list at https://metzdowd.com. (03 2009)
2. Kristoufek, L.: What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. PLOS ONE **10**(4) (04 2015) 1–15
3. HYUNYOUNG, C., HAL, V.: Predicting the present with google trends. Economic Record **88**(s1) 2–9
4. Ettredge, M., Gerdes, J., Karuga, G.: Using web-based search data to predict macroeconomic statistics
5. Miraz, M.H., Ali, M.: Applications of blockchain technology beyond cryptocurrency. CoRR **abs/1801.03528** (2018)
6. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: 8th International AAAI conference on weblogs and social media (IDWSM),. (2014)
7. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica **47**(2) (1979) 263–291
8. Dolan, R.J.: Emotion, cognition, and behavior. Science **298**(5596) (2002) 1191–1194
9. Panger, G.T.: Emotion in Social Media. PhD thesis, University of California, Berkeley (2017)
10. Chen, H., De, P., Hu, Y.J., Hwang, B.H.: Customers as advisors: The role of social media in financial markets (2011)
11. Tetlock, P.C.: Giving content to invsotry sentiment: The role of media in the stock market. The Journal of Finance (2007)
12. Gartner: Gartner says majority of consumers rely on social networks to guide pruchase decisions (2010)
13. Kouloumpis, E., Wilson, T., Moore, J. In: Twitter Sentiment Analysis: The Good the Bad and the OMG! AAAI Press (2011) 538–541
14. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC. (2010)
15. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Fourth International AAAI Conference on WEblogs and Social Media. (2010)
16. Bordino, I., Battiston, S., Caldarelli, G., Cristelli, M., Ukkonen, A., Weber, I.: Web search queries can predict stock market volumes. PLOS ONE **7**(7) (07 2012) 1–17

17. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Volume 1. (Aug 2010) 492–499
18. Sul, H., Dennis, A.R., Yuan, L.I.: Trading on twitter: The financial information content of emotion in social media. 2014 47th Hawaii International Conference on System Sciences (2014) 806–815
19. de Jong, P., Elfayoumy, S., Schnusenberg, O.: From returns to tweets and back: An investigation of the stocks in the dow jones industrial average. Journal of Behavioral Finance **18**(1) (2017) 54–64
20. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2011) 1 – 8
21. Stenqvist, E., Lönnö, J.: Predicting bitcoin price fluctuation with twitter sentiment analysis. KTH Royal Institute of Technology School of Computer Science and Communication (2017) 3–28
22. Lamon, C., Nielsen, E., Redondo, E.: Cryptocurrency price prediction using news and social media sentiment. Master's thesis, Standord (2015)
23. Colianni, S., Rosales, S.M., Signorotti, M.: Algorithmic trading of cryptocurrency based on twitter sentiment analysis (2015)
24. Shah, D., Zhang, K.: Bayesian regression and bitcoin. 2014 53nd Annual Allerton Conference on Communication, Control, and Computing (Allerton) (2014) 409–414
25. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural networks. In: 1990 IJCNN International Joint Conference on Neural Networks. (June 1990) 1–6
26. Guresen, E., Kayakutlu, G., Daim, T.U.: Using artificial neural network models in stock market index prediction. Expert Systems with Applications **38**(8) (2011) 10389 – 10397
27. Zhu, X., Wang, H., Xu, L., Li, H.: Predicting stock index increments by neural networks: The role of trading volume under different horizons. Expert Systems with Applications **34**(4) (2008) 3043 – 3054