Multi-channel online discourse as an indicator for Bitcoin price and volume

Marvin Aron Kennis

[1] *Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*

Abstract. This research aims to identify how Bitcoin-related news publications and online discourse are expressed in Bitcoin exchange movements of price and volume. Being inherently digital, all Bitcoin-related fundamental data (from exchanges, as well as transactional data directly from the blockchain) is available online, something that is not true for traditional businesses or currencies traded on exchanges. This makes Bitcoin an interesting subject for such research, as it enables the mapping of sentiment to fundamental events that might otherwise be inaccessible. Furthermore, Bitcoin discussion largely takes place on online forums and chat channels. In stock trading, the value of sentiment data in trading decisions has been demonstrated numerous times [1] [2] [3], and this research aims to determine whether there is value in such data for Bitcoin trading models. To achieve this, data over the year 2015 has been collected from Bitcointalk.org, (the biggest Bitcoin forum in post volume), established news sources such as Bloomberg and the Wall Street Journal, the complete /r/btc and /r/Bitcoin subreddits, and the bitcoin-otc and bitcoin-dev IRC channels.

By analyzing this data on sentiment and volume, we find weak to moderate correlations between forum, news, and Reddit sentiment and movements in price and volume from 1 to 5 days after the sentiment was expressed. A Granger causality test confirms the predictive causality of the sentiment on the daily percentage price and volume movements, and at the same time underscores the predictive causality of market movements on sentiment expressions in online communities.

1 Introduction

After the internet forever changed the way investors trade, analyze, and acquire information [4], it is now giving rise to a new era of financial innovation through the introduction of cryptocurrencies. The expanding market of cryptocurrencies involves capital exceeding USD10 billion as of November 2016 [5], providing an unusual opportunity to study the emergence of value from currency created solely in the digital realm. Similar to traditional currencies, the value of cryptocurrencies is largely based on supply and demand, driven by the community's belief in the merit of these currencies. Although hundreds of different cryptocurrencies are now available, Bitcoin is considered to be the the virtual currency that set off the cryptocurrency revolution and now enjoys the largest market capitalization [6].

Research on the influence of online discourse on the Bitcoin exchange price has been scarce, and mostly focused on data from a single channel. [7] demonstrates that for the prediction of Bitcoin price movements online discourse on Twitter seems to lag behind events occurring on the exchange, rather mirroring what happens on the metaphorical trading floor than predicting what is going to happen in the future. Contradicting results were achieved by [8], where the increasing polarity of social signals preceded corresponding movements of the Bitcoin exchange price. [8] also shows that incorporating these signals into algorithmic trading models can lead to profitable trading strategies.

[9] looked into the less rational factors for Bitcoin price formation, and found sentiment to play an important role. Findings by [10] suggest that macro-economical financial trends do not significantly influence the Bitcoin price.

Studies regarding the influence of online discourse on traditional stock and foreign exchanges on the other hand have been plentiful, and provide the starting point for this paper. [11] investigates the role of media in the stock market, and found that using daily content from the Wall Street Journal, high media pessimism predicted downward pressure on market prices followed by a reversion to the fundamental value of the security. Unusually high or low pessimism can predict high market trading volume. In [12] the predictive value of news reports on stock prices in the short term is asserted. Besides news, social media and other online sources may be used to extract early indicators for investor sentiment, as demonstrated by [1], where Twitter data was shown to be a leading indicator for the closing prices of the Dow Jones Industrial Average index (DJIA). Google Trends data has previously been used in [13] to show that this information can provide meaningful indicators for upticks in market volume one or more days after increases in search volume, indicating that online behavior away from the exchanges carries anticipatory value for events occurring on exchanges. [14] also shows that in easily manipulated market environments, such as those of thinly traded stocks, online discussion boards can be effectively used to manipulate the price, even without the presence of fundamental news.

The added value of sentiment data in stock forecasting algorithms is the main motivator behind the research presented in this paper. Should there be a causal connection between Bitcoin-related sentiment and exchange movements, it can be integrated into trading models that trade on the Bitcoin markets and potentially lead to increased profits. These predictions may also aid in reducing risk, as increases in volume tend to be paired with higher volatility. Volatility in turn is an important metric in managing risk in (algorithmic) trading strategies [15].

1.1 Contributions of this paper

The research presented in this paper investigates the predictive value of online discourse originating from multiple channels on Bitcoin price and volume, by analyzing publication volume and sentiment data from these channels concurrently. It extends the current research in this domain that has demonstrated co-movement relations between social signals and market metrics such as price, volatility and volume on various types of financial exchanges. The hypothesis leading this paper will be that this online sentiment data has significant predictive causality in relation to inter-day Bitcoin market movements.

1.2 Outline

This paper will first discuss the workings of Bitcoin in section 2.1 before referring to financial theory to determine parallels between Bitcoin and traditional financial instruments in section 2.2. Drawing these parallels allows us to select established financial theories as a starting point, given the relative novelty of cryptocurrencies compared to traditional financial instruments. This will be followed by an overview of sentiment analysis and its value in financial markets, as well as a detailed description of relevant sentiment analysis algorithms in section 2.3. The complete data collection to classification approach is explained in section 3, evaluating each classifier against the collected data sources (3.5), and finally using the best performing classifier for each respective channel in the sentiment classification of all collected data per channel. Building on this, the process of establishing a correlation and causal relationship by means of a Granger causality test between the recorded sentiment and market data is described in section 4.

2 Background and related work

2.1 How Bitcoin works

Bitcoin is an electronic peer to peer payment system first introduced by Satoshi Nakamoto in 2008 and became functionally implemented by 2009 [16]. Bitcoin transactions are stored on a publicly accessible ledger called the blockchain, allowing everyone to verify, validate and execute transactions that are recorded in this ledger. Instead of having a single authority rule over a currency - such as is the case with fiat currencies and preceding digital currencies, it is now regulated by a decentralized network. Bitcoin's innovation lies in it being the first digital currency to eliminate the need for a central trusted party. It does so by relying on the HashCash proof of work function for transaction processing and verification [17]. This proof of work requires participating computers in the network (called Bitcoin miners) to conduct increasingly difficult computations. Specifically, miners are required to produce an SHA-256 hash with an arbitrary amount of leading zeros from the transaction block. Upon solving those computations they are awarded with newly minted Bitcoin (providing an incentive to solve the computations), and the transactions in the block for which the computation was solved become verified. While these proof of works are complex to compute, they are easy to verify [17].

With a market cap that surpassed USD10 billion in November 2016 according to Coindesk [5], Bitcoin can now be considered more than an idealistic hobby project rooted in ideas put forward by the cypherpunk community [17]. It has attracted attention from investors, consumers, criminals, and governments worldwide [18]. However still not adopted by the masses as a day-to-day currency, adoption of the cryptocurrency has moved past underground communities and mere hobbyists. The cryptocurrency is now accepted even by larger corporations such as Dell [19] and Overstock [20], raising the question whether Bitcoin can be considered a real currency as consumers can increasingly use it for their traditional purchases.

2.2 Investor reactions to news

Although primarily intended to be used as a currency by its inventor, Bitcoin should still be seen as a speculative investment according [21] and [10]. A proper currency acts as a medium of exchange, a store of value, and a unit of account. Bitcoin largely fails to satisfy these criteria. Firstly, it is still not widely accepted compared to established payment methods and as such can not be considered an effective medium of exchange. Next to that, the Bitcoin exchange price is highly volatile, making it a poor store of value. In the absence of large institutional investors, speculative investments in Bitcoin are more likely driven by retail or individual investors called noise traders. According to behavioral finance research by [22] and [23], noise traders are prone to exhibit less rational market behavior. The sentiment as it may be expressed online could therefore serve as an interesting indicator for exchange movements, as these noise traders might be influenced by them. In the context of Bitcoin, online messages can disclose new or previously private information that fundamentally alters Bitcoin valuations, such as when new merchants accept Bitcoin or forthcoming regulations limit its use, but may also disclose attempts of market manipulation in the absence of fundamental news.

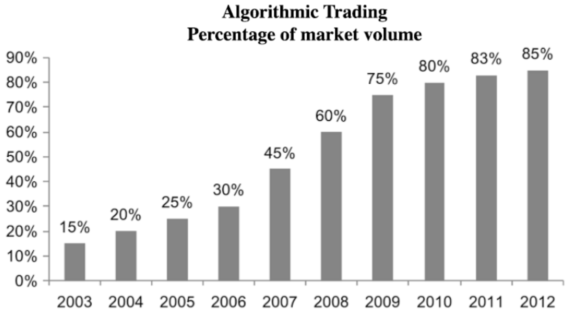


Figure 1. Increase of algorithmic trading as a percentage of market volume from 2003 to 2012 [28]

.

Research concerning the influence of news reports on stock market returns has been abundant. [24] has shown that elements of surprise in news reports are a strong indicator for stock price movements shortly after the surprising news has been published. This is in line with the (widely disputed) efficient market hypothesis postulated by Fama in 1970, which states that all available information regarding a security is already reflected in the stock price [25]. If the efficient market hypothesis were to be true, it would prevent anyone from exploiting securities that are mis-priced because large movements are only caused by unforeseen events, and not by misinterpretation of the true value of the security by other market participants (overvaluing and undervaluing participants will cancel each other out on average). This implies that only if a news report contains unexpected or previously unknown information, it will move the exchange price. In the meantime, the price will follow a random-walk. However, the sole presence of high-frequency sentiment analysis on financial exchanges discredits the efficient market hypothesis on short-term intervals [26]. This is where the value of computerized sentiment analysis becomes apparent; market participants reacting rapidly to new insights (whether this insight stems from opinion mining or elsewhere) will have a higher likelihood of being able to profit from it before the information reaches other parties. The rapid rise of high-frequency trading (HFT) firms over the past decade (trading in nanosecond time frames), precipitated by the introduction of large-scale algorithmic trading on stock exchanges attest to this [27]. Figure 1 shows the increase in algorithmically executed trade orders between 2003 and 2012.

[29] estimates that up to 40% of equity and 15% of foreign exchange trades were initiated by high-frequency traders in 2016. News-based trading is an established practice in the HFT industry. An illustrative example of the convergence of sentiment analysis and algorithmic high-frequency trading is the ‘Hash Crash’ that took place on April 23, 2013, in which a hacked Twitter account of the Associated Press spread false rumors about an attack on the White House, subsequently causing a drop in the Dow Jones Industrial Average of 143 points in minutes [30]. Besides the need for speed, unstructured data is being generated in such vast amounts over such diverse channels, that no single human will be able to find, let alone process, all of this. It then follows that different market participants will act upon the relevant information in varying speeds.

According to principles of behavioral economics, opinions which are not necessarily driven by the release of topic-related news are also thought to influence movements in financial markets, as investors are susceptible to cognitive bias and other predictable human errors [23]. It therefore seems valuable to examine the overall sentiment in Bitcoin-related communities and Bitcoin-related news in parallel. Bollen et al. investigate the use of Twitter mood as a forecasting mechanism for the Dow-Jones Industrial Average index (DJIA) in [1], without filtering on communities and find an accuracy of 87.7% in predicting daily up and down movements in the closing values of the index. Due to Bitcoin’s low penetration in the consumer market and comparatively small ecosystem, it is unreasonable to study the mood of a representative sample of the general population as they likely do not interact with Bitcoin on a daily basis and might consequently not represent the opinions of the individuals that do actively interact with Bitcoin. The analysis to be carried out later in this paper will hence be limited to Bitcoin communities as they can be found on online forums and chat channels.

Further, financial news may concern a single company, but news not directly written about a single company or index might also influence its price. Take for instance an announcement about a cut in oil production, which might influence stocks of industries and companies dependent on this resource. The same holds true for Bitcoin, where changes in international monetary policy might, for example, make Bitcoin a more attractive refuge, increasing attempts to regulate the currency will likely have a negative effect on the demand for the cryptocurrency. This type of secondary news will not be investigated in this paper.

2.3 Sentiment Analysis

Opinions play an important role in human decision making processes, and individuals are increasingly turning to the (social) web to aid them in making those decisions [31]. Relevant data is fragmented across multiple sources or buried in lengthy forum discussions and blog posts. As the amount of unstructured data on the web keeps amassing in volume, both identifying sources and processing information become increasingly challenging tasks [32]. Sentiment analysis concerns itself with (computationally) extracting subjective information regarding entities such as products, locations, or people and encapsulates the analysis of opinions, sentiments, evaluations, appraisals and attitudes [32]. From rule-based to statistical approaches and recent developments such as deep learning, research in this space is flourishing, driven by commercial incentives and the far-reaching domains in which it can be applied. Due to the varying sentiment analysis methods between domains, a single cross-domain state-of-the-art accuracy to which to aspire does not exist. Furthermore, the subjective nature of opinions as it surfaces in inter-annotator agreements should be considered. If we for example consider an 80% annotator agreement, it implies that humans will disagree with *any* assigned label about 20% of the time. Finally, it should be noted that the goal of this research is not to build the perfect sentiment classifier for the Bitcoin domain, but rather select classifiers of satisfactory accuracy for further analysis of sentiment in relation to market movements.

Early work on machine learning based sentiment analysis methods was carried out by Pang and Lee in [33], wherein it was shown that machine learning techniques such as Naive Bayes, Logistic Regression, and Support Vector Machines on unigram and bigram features could rival human-generated baselines on movie reviews (82% accuracy). These techniques are still relevant today as they achieve decent accuracy and are relatively easy to implement with available toolkits. By analyzing the data with these algorithms first, we can discover whether there is a correlation and causal relation at all. If there is, future research could investigate whether improving the sentiment classification can strengthen this correlation, such that misclassification does not propagate to trading models.

2.4 Data Labeling

Labeling representative data is an integral part to natural language processing and machine learning processes which try to learn from a correctly labeled dataset. In supervised machine learning, models are first trained on data containing the correct labels so that the models can learn a classification function from the features. As we try to determine whether news articles or social posts regarding Bitcoin are positive or negative, we first show the classifiers what features belong to positive and negative classes through training sets. Based on the labels and corresponding features the model has seen during the training process, it tries to estimate a label for a feature set belonging to an unlabeled item. Sometimes labels can be inferred from context or meta-data. In sentiment analysis of product reviews, the star rating might be used as a proxy label for the review text. Unfortunately the collected data is not augmented with labels or relevant meta-data and needs to be annotated.

Linguistic annotation tends to be carried out by experts and is expensive and time consuming. It then seems tempting to resort to cheaper and faster methods. However an evaluation of these methods should be taken into consideration. [34] compares expert annotations to annotations provided by Amazon Mechanical Turk (http://mturk.com) workers for affective text analysis on headlines of news articles. Amazon Mechanical Turk (hereafter MTurk) is an online platform where requesters can employ workers on human intelligence tasks (HITs), ranging from classification and categorization to collecting feedback and moderating content [35]. The platform includes a dedicated service for sentiment annotation. Requesters can upload a dataset they want to annotate and specify rating scales and instructions. [34] found that experts agree with each other more than non-experts agree with experts. Individual experts were found to provide better annotations than non-expert annotators. However, on average 4 non-expert annotators can rival the annotation accuracy of a single expert on affective text analysis tasks. Similarly, in [36] it is reported that for sentiment analysis, three MTurk annotations per item provide similar or better performance compared to a single expert annotation. These insights were used as a starting point for the MTurk process in this paper in that more than 4 non-expert annotators were assigned to each document. However, increasing the amount of annotators introduces disagreement, leaving us with the challenge of inferring the true label, given a set of assigned labels by various annotators. In doing so, annotator bias as well as ambiguity in the task have to be taken into account. Furthermore, there does not have be a single true label for each item, as some are more open to interpretation than others.

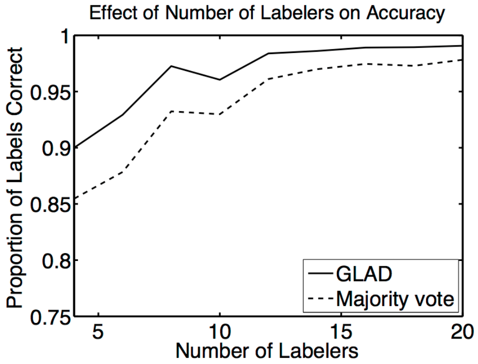


Figure 2. Majority vote versus GLAD approach as proposed by [38].

Leaving out low performing workers should increase the overall precision of assigned labels, especially when these low-performing workers have provided a large amount of annotations. The motive for low quality submissions should be clear; workers are paid based on the amount of tasks completed, and with no gold standard to which their submissions can be compared from the start, workers have to make a trade-off between speed and quality to optimize their earning. A simple frequentist approach may suggest the use of a majority vote, but this does not take individual noise into account. [37] discusses the inference of a true label by means of expectation maximization. In [38] this approach is applied specifically to Amazon Mechanical Turk by modeling the ability of individual workers.

As figure 2 demonstrates, this approach will achieve a maximum improvement of 5%. While this could be significant when the labels used in this research were to be applied to practical trading algorithms, for now it was decided to only determine whether there is a correlation between sentiment and exchange first, and then look at possible methods of strengthening this correlation if worthwhile.

3 Approach

The following sections will describe the data analysis process from data collection to classification in detail. First, the collection of data from online channels is described. Once data has been collected, a randomly selected subset of the data for each channel will get labeled by gathering crowd-sourced labels for each item in the selected subset. By collecting multiple labels for each item, we will be able to increase the likelihood of the true label being inferred. This labeled set will be preprocessed before transforming it into numerical feature vectors. Likewise, transformations are applied to the unlabeled dataset. Consecutively, the classifiers used to learn from the labeled data will be described. The performance of each individual classifier is then compared. As some classifiers are expected to perform better on data from a particular source compared to others, the classifier with the highest cross-validation per source will be selected to classify the unlabeled data for the respective channel. The output of this classification will be matched on daily timestamps to market data to determine whether there is a correlation between the amount of positive and negative online discourse and the upward and downward exchange movements on n days after the recorded sentiment occurred. Figure 3 illustrates the approach in detail.

3.1 Collected Data

3.1.1 News Reports

News articles were collected from Bloomberg, Reuters, Coindesk, news.bitcoin.com, Wallstreet Journal, and CNBC, yielding a total of 7,730 articles, of which 1,534 were published in 2015. Besides the article body, the author and the date on which the article was published have also been collected. To select only Bitcoin-related articles, the custom-built web scraper would either use the corresponding website's search function and traverse through all pages, or filter on article tags where available. The scraper did not discriminate on publication date and collected articles that were tagged with Bitcoin or showed up when searching for Bitcoin using the website's search function. Scraping articles from a wider date range (2012-2016) will allow us to train the classifier on a wider breadth of vocabulary; it could be possible that Bitcoin was plagued by a lot of highly similar negative events in a specific period, and that the vocabulary in describing those events largely centers around a single topic with the same sentiment. Training a classifier on such sets would limit the possible application of this research in trading models as it generalizes poorly to the wider domain and will therefore not be reliable in assessing unfamiliar events.

3.1.2 Forum and Reddit posts

With more than 550,000 topics and 1,500,000 posts since its inception in 2009, Bitcointalk.org is by far the biggest Bitcoin-related forum. Irrelevant subforums where off-topic discussions take place were filtered out by excluding URLs from those subforums from being scraped, before mining posts from the following subforums; 'Speculation', 'Economics', 'Trading discussion'. These three subforums contain discussions that are directly related to trading in Bitcoin, and have active market participants posting topics and replies. For forum topics, the topic title and body, the timestamp as well as the author name and total number of replies were indexed. The same data was collected from Reddit, but also included the score of the Reddit post (a function of 'upvotes' and 'downvotes' by users in the community).

3.1.3 IRC Chat

The two biggest IRC (Internet Relay Chat) channels were scraped for text messages through the BitcoinStats.com website over 2015. For each message, the author, content and timestamp was recorded. Due to the informal writing style that is prevalent in online chat rooms, the data collected from the Bitcointalk forum and Reddit, as well as IRC will likely be significantly noisier than that collected from news articles.

3.1.4 Market Data

Market data was collected through Blockchain.info through the CSV download option [39]. Blockchain.info is the sole source upon which this paper relies for data relevant to the fundamentals of Bitcoin being traded on exchanges, and contains the daily data from the Bitfinex, Bitstamp and BTC-e exchanges.

The downloaded CSV contains the following data, per day over 2015:

- Date (daily timestamp)
- Average (average price on daily timestamp)
- Ask (average ask price, sell offer, on timestamp day)
- Bid (average bid price, sell offer, on timestamp day)
- Last (Last price recorded on day of timestamp)

3.2 Crowdsourced data labeling

In order to quickly label the news articles, the Amazon Mechanical Turk sentiment analysis service was employed to create a labeled corpus of 1,000 randomly selected news articles. Using the Amazon Mechanical Turk service, English-speaking workers were asked to rate provided sentiment. For our task, a 5-point scale ranging from 'very negative' through 'neutral' to 'very positive' was used. By including a task description and title, workers could opt out of this task for any reason before starting.

In general, sentiment analysis aims to determine the attitude or polarity of the content with regard to some topic. The topic for the presented content will be Bitcoin. Due to the massive amount of different workers, there is a high likelihood that the vast majority of them do not have any expert knowledge of Bitcoin or trading.

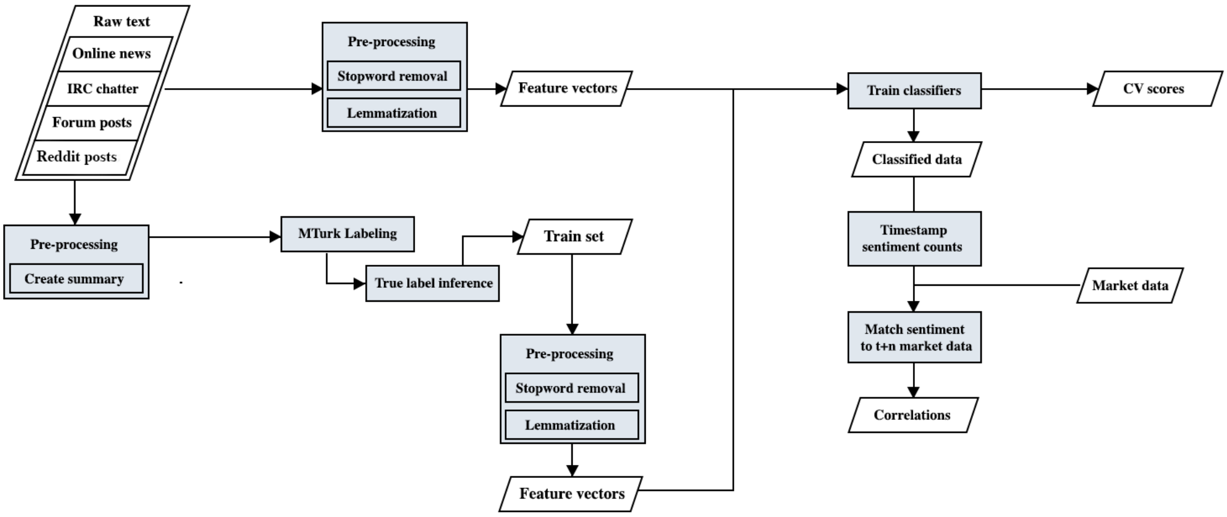


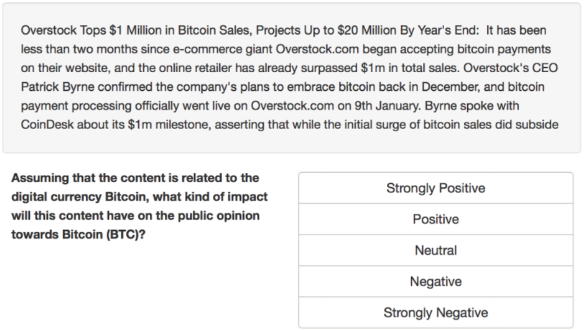
Figure 3. Flow diagram of the NLP process

Instead of directly asking workers what the impact of the content would be on the exchange price, they were asked to rate the perceived effect on the public opinion towards Bitcoin. By doing so, we aim to discover the true intended emotional communication of the content. Negative news about Bitcoin can inherently not have a positive effect on the public opinion towards it. To aid in rating the sentiment for the presented content, workers were provided examples for each item on the rating scale. A detailed view of this task setup can be found in the appendix. The direct task prompt states that the user should assume that the content is Bitcoin-related. Besides that, the task was kept as succinct as possible.

‘‘Assuming that the content is related to the digital currency Bitcoin, rate what kind of impact will this content have on the public opinion towards Bitcoin (BTC)‘‘

The direct prompt that was posed to the human intelligence workers is not a true sentiment analysis prompt, in that it assumes that the content is related to Bitcoin and that it asks for the impact on the public opinion towards Bitcoin. As such, the trained classifiers will not classify sentiment of the text, but rather try to classify the perceived intended effect of the article on the Bitcoin public opinion. This is valuable when applying this information in trading strategies, as traders will try to judge what decision other players in the market will make upon the release of the relevant information. Articles in which a Bitcoin competitor is mentioned negatively would have a negative classification in traditional sentiment analysis tasks, but might be positively classified by the human intelligence workers responding to the prompt.

Content for the HIT tasks was created by taking the article headline or topic title and the first 500 characters of the article or topic body where available. Titles of financial articles tend to give a summary of the general expressed sentiment in the article. Some forum topics consist solely of an expressive title, and in such cases only the title was used. The first 500 characters of the article or post body are included to provide context on the title. This 500-character limit was imposed to not overload MTurk workers with massive walls of text; analyzing this text would require more time as confounding information might be given in the full article, consequently increasing the costs of the labeling task. Figure 4 shows an example task as presented to HIT workers.



Overstock Tops $1 Million in Bitcoin Sales, Projects Up to $20 Million By Year's End: It has been less than two months since e-commerce giant Overstock.com began accepting bitcoin payments on their website, and the online retailer has already surpassed $1m in total sales. Overstock's CEO Patrick Byrne confirmed the company's plans to embrace bitcoin back in December, and bitcoin payment processing officially went live on Overstock.com on 9th January. Byrne spoke with CoinDesk about its $1m milestone, asserting that while the initial surge of bitcoin sales did subside

Assuming that the content is related to the digital currency Bitcoin, what kind of impact will this content have on the public opinion towards Bitcoin (BTC)?

Figure 4. Screen capture of an example task as it is presented to MTurk workers.

As the syntactical and semantic structure of generally well-composed news articles differs greatly from the structures used on online forums and chat channels, this annotation process was performed on 2,000 forum topics, 2,000 Reddit posts, and 2,500 IRC messages. Sentiment classifiers on each source will be trained separately.

The task setup was the same for news articles, forum topics, and Reddit posts. The approach for IRC chatter was different, as messages are typically very short and lack context compared to the former. To adjust for this, workers were simply asked to rate whether the general content of the message had an expressed sentiment ranging from very

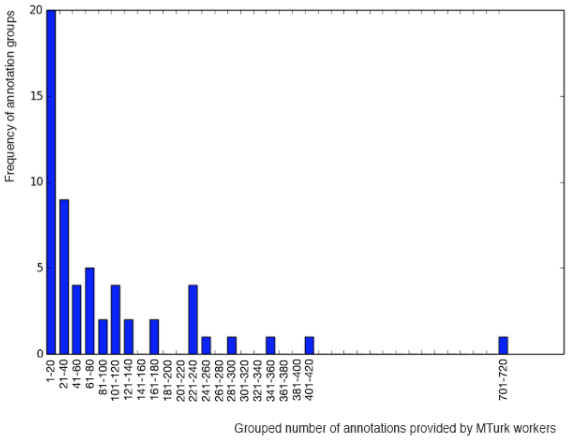


Figure 5. Worker activity distribution on news articles. Annotations per worker are grouped into bins of 20. Bins without any annotations in the 421 - 701 range are unlabeled.

negative through neutral to very positive, disregarding the Bitcoin context.

3.3 Mechanical Turk result analysis

The labeling of each dataset was completed within 48 hours of task submission. Workers were compensated with $0.02 for each provided label. An analysis of the labeled news article data shows that 57 workers provided a total of 5000 annotations (each item was labeled 5 times), for an average of 87.7 annotations per worker. The standard deviation of this population is 121.67, indicating a high variance in the amount of provided annotations between workers. Due to this high variance, it is important to control for individual worker bias. The graph of this variance is shown in Figure 5.

Table 1 shows a distribution of the labels provided by workers, aggregated over all 5,000 annotations provided. Very negative, negative, neutral, positive, and very positive annotations are indicated by the column headers - - , -, 0, +, + +, respectively. With 37.62% of instances being assigned 'neutral', and a slight bias towards positive news indicating a balanced news landscape. While all news outlets report more positive than negative news, WSJ, Bloomberg, news.bitcoin.com and Bitcoin report about three or more times as many positive than negative reports. This could potentially indicate a bias in their reporting, but can also be caused by the sampled content; the value of Bitcoin has steadily increased over the years the sampled content was published in. This bias is less pronounced in the forum, and IRC annotations. It can further be noted that the largest amount of labels for each source were assigned a neutral label.

Applying a simple majority vote with the aim of inferring the true labels for a given news article leads to a more pronounced bias towards positive news, outweighing negative news by a factor of 3 to 1, and increasing the relative size of the neutral class. The results can be seen in Table 5.

Table 1. Annotation distribution across news channels.

	- -	-	0	+	+ +
Coindesk	90	278	750	852	55
Reuters	168	494	868	589	31
NewsBitcoin	7	43	139	147	9
Bloomberg	12	45	127	99	2
WSJ	4	15	47	56	3
CNBC	20	83	134	148	25
Total	276 (5.5%)	847 (16.9%)	1,881 (37.7%)	1,876 (37.5%)	120 (2.4%)

Table 2. Annotation distribution across Bitcointalk subforums

	- -	-	0	+	+ +
Speculation	147	1,087	2,635	1,017	124
Economics	75	707	2,925	1,163	120
Trading	51	417	3,260	1,250	21
Total	273 (1.8%)	2,211 (14.7%)	8,838 (58.9%)	3,430 (22.8%)	265 (1.8%)

Table 3. Annotation distribution across Reddit pages

	- -	-	0	+	+ +
bitcoin	66	535	3,206	1,074	96
btc	101	600	3,051	1,150	111
Total	167 (1.7%)	1,135 (11.4%)	6,257 (62.6%)	2,224 (22.2%)	217 (2.2%)

Table 4. Annotation distribution across IRC channels

	- -	-	0	+	+ +
otc	133	1,590	8,944	1,634	54
dev	71	1,877	8,163	2,260	59
Total	204 (0.8%)	3,467 (14%)	17,107 (69%)	3,894 (15.7%)	113 (0.5%)

The majority vote on news data causes 47.9% of samples to fall into the neutral and furthers the imbalance between positive and negative classes. We can infer that such a majority vote on the other datasets will have a similar if not more pronounced effect due to the higher presence of neutral labels. The true difference will be established in section 3.5, where the errors of each classifier are analyzed. As the predictions we aim to make are binary, it is expected that leaving out the neutral label will lead to a clearer decision boundary between nearby labels in binary prediction. Before continuing, 'very positive' and 'positive' labels were merged into 'positive'. The same was done for negative sentiment, ensuring that each item fits in one of two classes (positive/negative).

3.4 Feature Extraction

What follows is an overview explaining the applied preprocessing and feature extraction techniques.

Table 5. Majority vote label distribution

- -	-	0	+	+ +
16 (1.6%)	122 (12.2%)	479 (47.9%)	379 (37.9%)	4 (0.4%)

3.4.1 Pre-processing Data

3.4.2 Lemmatization

Lemmatization is a normalization process by which morphological variation in data can be reduced. Word forms such as ‘diving’ and ‘dove’ will be mapped to their dictionary form ‘dive’. Compared to the more crude process of stemming (which is faster and less complex) the produced lemmas are linguistically valid [40].

To create accurate lemmas, each token has to be assigned a part-of-speech tag (POS tag) before determining the lemma. In POS-tagging lexical categories such as ‘Noun’ or ‘Adverb’ are added to each word. These lexical categories are valuable in handling disambiguation. The spaCy Python package [41] has been used to tokenize, parse, and lemmatize the data from each source. The Spacy parsing accuracy is within 2% of the state-of-the-art parser by introduced by [42] (92.8% for Spacy, 94.3% for Andor et al.) according to a comparative study carried out by Choi et al. [43]. Settings were left on default.

Texts may also contain various types of noise such as stop words and punctuation that may not have any influence on the polarity of the text. Before processing the data with classifiers, the data was cleaned. In the process of filtering stop words, negating terms were removed from the NLTK stop word corpus [44]. Punctuation was removed using a regular expression, regardless of the source channel. However, [45] postulates that removal of stop words from tweets has a negative influence on sentiment prediction accuracy due to the noisy nature of short text messages (abbreviations and irregular forms). In the collected datasets for this research, there was no demonstrated improvement upon stop word removal, and they were left in place.

The collected text documents have to be converted to numerical feature representations so that they can be used by the statistical classifiers. Text is converted to word n-grams (features), and then represented using a bag-of-n-grams approach.

Using the Scikit learn [46] CountVectorizer feature with n-gram range specifications, a vectorizer for unigrams, bigrams and trigrams was created. The corpus is then transformed to a sparse matrix counting the occurrences of each word n-gram per document, using the respective vectorizer for each of the n-grams.

The theoretical benefit of bigrams and trigrams over unigrams, is that they maintain word order. In the context of sentiment analysis the unigram features of the sentence 'I do not like Bitcoin' would be ['i','do', 'not', 'like', 'bitcoin']. A trigram would contain the feature [..., ['not','like','bitcoin'], ...], and would include the connection that it is the Bitcoin that is being disliked. A bigram containing [‘not’, ‘like’] allows for negation. This context is lost on unigrams and BOW approaches.

3.4.3 TF-IDF Transformation

To counter the overly present words in this specific domain, the counts in the feature vectors are weighted and selected using a TF-IDF transform [47]. This is different from stop word removal, as it is adaptive to domain contexts. If the resulting matrix from vectorization would be directly used in classification, common words would outweigh the less common but more relevant words. The main idea of TF-IDF is that a word that is very common in a specific article, but does rarely appear in other articles, will provide a strong indication that this word will offer good categorical differentiability.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad (1)$$

This process will return a weight $w_{i,j}$ multiplying the term frequency tf, the amount of times a term i occurs in document j, by the inverse document frequency (the total number of documents (N), divided by the total number of documents containing term i (df).

The performance and added value of the TF-IDF transform has been evaluated against the collected content, and has not shown an increased cross-validation performance for any discourse channel combined with any of the proposed classifiers. A possible explanation for this, is the relatively short texts which are being classified (recall that they were limited to the title plus the first 500 characters). TF-IDF performs well on longer documents, but short text is known to cause noisy TF-IDF values [48].

3.5 Sentiment classification

The application of each of the three classifiers (AlchemyAPI, logistic regression, naive Bayes) will be explained with the news data as an illustrative example. Although it might seem odd to leave out support vector machines in this comparison of text classifiers, Summarized results for all sources are presented at the end of this section, as the approach is the same for each source. Full results for each source can be consulted in the appendix.

3.5.1 AlchemyAPI

We start by establishing a baseline for our sentiment classification using the AlchemyAPI (part of the IBM Watson cloud offering), before employing classifiers from the scikit library. AlchemyAPI is a natural language machine learning service [49]. The Alchemy Language API provides out-of-the-box solutions for sentiment analysis, but can not be trained to suit a particular domain and might suffer from the domain-transfer problem [50].

The raw text data was sent to be processed by the AlchemyAPI the same way it was sent to the MTurk annotation service, as the AlchemyAPI has its own text pre-processing in place. This means that the pre-processing as applied in the previous section is irrelevant to this classifier. It should be noted that the way in which the data was labeled might affect the evaluation, as AlchemyAPI is a pure sentiment analysis API, whereas annotators were asked to label items according to the perceived effect on the public opinion towards Bitcoin.

Table 6. AlchemyAPI Confusion Matrix

Pred. / Real	**P**	**N**
P	422	127
N	87	237

3.5.2 AlchemyAPI Error Analysis

From the confusion matrix in Table 6, we can determine that the AlchemyAPI achieves a recall of 74.2%, and a precision of 81.3% on summarized news articles, using the average of MTurk annotations as gold standard. Recall is defined to be the ratio of correct classifications divided by the total amount of correct classifications.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

In the context of trading applications and a hypothetical integration into trading models, it is absolutely critical to maximize precision, as uninformed trades could lead to significant financial losses. It does not matter that some profitable trades are being missed (recall), we just want to ensure that the trades that are made are indeed correct. This is not to say that recall does not matter at all, you would still want to make enough profitable trades to outperform the benchmark of simply buying and holding Bitcoin.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

When analyzing misclassification (17% of true positive instances, 34.89% of true negative instances), the classification of negative instances performs significantly less than that of positive instances. In accounting for this, it is important to keep the bias of individual news sources in mind. A further possible explanation for this may be the wide ranging interpretations that can be given to a Bitcoin-related text. For example, a document could receive the labels [-2, -1, 0, 0, 1] from five individual annotators and subsequently receive a neutral label due to the majority vote. However, just as many annotators gave it a neutral label as a negative label, the negative annotators simply did not agree on the degree of negativity. The negative sentiment assignments are then completely disregarded in the final label inference. This negative sentiment is however still contained in an average of the five labels.

3.5.3 Logistic Regression

Logistic regression (also called maximum entropy or MaxEnt) has proven effective in various text classification tasks [51] and is often seen as the go-to method for binary classification problems. It is easy to implement and does not require any tuning. For now, we will again resort to the implementation as it is provided in the scikit-learn Python library. At its core, logistic regression is based on the logistic function which takes any input x and maps it to a value between the limits 0 and 1.

$$\frac{1}{1+e^{-z}} \tag{4}$$

As the predictions are boolean, the logistic regression model for determining the probability of feature set X containing individual features x_i belonging to class *1*, $P(1|X)$, can be given as

$$P(1|X) = \frac{1}{1+e^{w_0+\sum_{i=1}^{n} w_i X_i}} \tag{5}$$

with x_i representing each feature in the feature space. The weights w_i of each feature x_i are determined using a Maximum Likelihood Estimator.

As the probabilities must sum to 1, *P(0|X)* is given by

$$P(0|X) = \frac{e^{w_0+\sum_{i=1}^{n} w_i X_i}}{1+e^{w_0+\sum_{i=1}^{n} w_i X_i}} \tag{6}$$

We can now assign label *0* if the below condition holds. Otherwise label *1* is assigned.

$$1 < \frac{P(0|X)}{P(1|X)} \tag{7}$$

A first pass with logistic regression on word unigram features shows a 78% and 73% and 71% accuracy rating for news articles, forum topics and Reddit posts respectively. Unigrams achieving the highest accuracy is in line with the findings of Pang and Lee [33]. The underperformance of n-grams compared to unigrams can be attributed to the limited availability of labeled data (which will then be further split into train and test sets in cross-validation). Unique bi- and tri-grams are anticipated to occur less often across a sparse set of documents. This causes the classifier to come across more 'unseen' n-grams than unigrams.

3.5.4 Logistic Regression Error Analysis

Inspecting the classification errors shows that the majority (66.67%) of misclassified news documents contain 'neutral' as the majority vote assigned by MTurk workers. When comparing majority votes and assigned average scores and removing neutral labeled documents from the corpus, we find that an average of all five labels yields higher cross-validation scores, because this effectively leaves out the neutral label in a large amount of instances. A mere 16 instances received all neutral annotations, meaning that the rest would, on average, have either a positive or negative sentiment. The majority vote greatly increases the amount of neutral labels, up to a third of the total dataset. The classifier output is binary for positive or negative, and the third neutral class will never be predicted. Multi-class classification generally also suffers from higher error rates and in [52] it is also argued that maintaining a third neutral label would blur the decision boundary between positive and negative and decrease performance. For the above reasons, the average of MTurk annotations is used as the correct label in the classifier training process. This effectively eliminates neutral instances (very few documents received all-negative labels by all five annotators). Fully neutral documents were removed from the training set.

3.5.5 Naive Bayes

Naive Bayes (hereafter NB) is often considered the baseline for many text classification models as it is robust, accurate and fast to implement. Traditional NB models assume that all attributes of the sample are independent to each other. Although often false (relations between words offer context, which is lost in a simple bag-of-words approach), NB performs well in various real-world tasks. Domingos and Pazzani explain this apparent paradox in [53] by asserting that classification estimation is only a function of the sign (1, 0) of the function estimation. NB relies on Bayes' theorem to compute the posterior probability of a class, given the distribution of features in the input vector.

$$P(C_i|F_j) = \frac{P(F_j|C_i)P(C_i)}{P(F_j)} \tag{8}$$

In equation 8, $P(C_i)$ is the prior probability of class i existing, independent from any other factors. $P(C_i|F_j)$ is the prior probability that a given feature set j is classified as $C_i.P(F_j)$ is the prior probability that a given feature set occurs, again independent from any other factors.

Two common implementations of the NB model for text classification are the Bernoulli and Multinomial models. They differ in the method in which the features they use are represented. In the Bernoulli model, each document is represented by a binary vector over the word space, where dimensions match words from the vocabulary, indicating only whether or not a word is present in the document. Multinomial models on the other hand will also consider how often each word appears. The feature vectors built in the previous step do contain word counts and a multinomial model therefore seems better suited. Multinomial naive Bayes computes the probability of a document d belonging to class c as follows [40]:

$$(c|d) \propto logP(c) + \prod_{1\leq k\leq n_d} logP(t_k|c) \tag{9}$$

$P(t_k|c)$ is the conditional probability of feature t_k being present in c, and signifies the amount of evidence feature t_k contributes in determining whether this document belongs to class c. $P(c)$ is the prior probability of any document at all belonging to class c. The features in the vector of document d are represented by t_k, where n_d is the total amount of features in d. The multinomial NB classifier aims to maximize P based on the data that was used to train the model, effectively aiming to select the most likely class from the set of classes, given a certain set of features for the respective document. According to [40] multiplying probabilities can lead to floating point underflow. For this reason, the log probabilities are added instead of multiplied.

Table 7. Comparison of multinomial NB and Bernoulli NB classifiers on news article summaries

	Multinomial NB	Bernoulli NB
CV-score	0.82	0.78

$$argmax_{c\in\mathbb{C}}\hat{P}(c|d) = argmax_{c\in\mathbb{C}}log\hat{P}(c) + \prod_{1\leq k\leq n_d} log\hat{P}(t_k|c) \tag{10}$$

In equation 10 $\hat{P}$ is used in the above notation as the probabilities are not truly known, they are based on observations made from the training set.

A comparison confirms the assumption of Multinomial naive Bayes outperforming Bernoulli naive Bayes. Cross-validation scores showed a 4% increase Bernoulli to Multinomial using unigram features on the created news article summaries in table 7. This corresponds with the findings in [54].

Although both logistic regression and naive Bayes are used for classification, NB is a generative model, whereas logistic regression is a discriminative model. Generative models try to which model the underlying probability distribution whereas discriminative models aim to learn the boundaries between classes [55]. Logistic regression splits feature space linearly, and performs well even if some of the features are correlated (NB assumes independence). [56] states that it is preferable to select a discriminative model where possible, as one should aim to solve the classification problem directly, rather than addressing a more general problem (modeling the underlying distribution) as an intermediate step in the classification.

3.5.6 Multinomial Naive Bayes Error Analysis

To maintain a consistent comparison, the average of MTurk annotations is again used as the target in classifier training. Assessing the performance of the multinomial naive Bayes classifier from the confusion matrices, we find that this classifier is better at predicting negative news than positive news, which seems counter-intuitive upon inspecting table 8, given that there are more true positives in the training set. The same holds for all other sentiment channels (table 20, 21, 22). Table 8 shows the distribution of target labels for each channel. Class imbalance has decreased compared to the majority vote in table 5. This inconsistency indicates that the that the negative class has stronger predictive features compared to the positive class; given the same document length (title and first 500 characters), the features present in a negative document vector provide more evidence of the document belonging to the negative class than the features of a positive document.

Manual inspection of the errors does not show a clear pattern in document topic of misclassified instances.

3.5.7 Classifier Comparison

Table 9 compares precision, recall, F-measure, and cross-validation score (accuracy) between the applied classifiers

Table 8. True class distributions for each sentiment channel

	News	Forum	Reddit	IRC
Positive	515 (62.80%)	1.257 (64.42%)	979 (66.68%)	1.956 (56.66%
Negative	305 (37.20%)	694 (35.58%)	489 (33.32%)	1.496 (43.34%)

Table 9. News classifier comparison.

News	Precision	Recall	F-Measure	CV / Accuracy
AlchemyAPI Raw text	0.8290	0.7686	0.7976	0.76
NB MultiNom. unigram	0.782	0.728	0.754	0.82
LogRegression unigram	0.775	0.626	0.685	0.78

Table 10. Forum classifier comparison.

Forum	Precision	Recall	F-Measure	CV / Accuracy
AlchemyAPI Raw text	0.8708	0.5685	0.7976	0.69
NB MultiNom. unigram	0.585	0.584	0.584	0.71
LogRegression unigram	0.652	0.563	0.604	0.74

Table 11. Reddit classifier comparison.

Reddit	Precision	Recall	F-Measure	CV / Accuracy
AlchemyAPI Raw text	0.8298	0.6004	0.6967	0.71
NB MultiNom. unigram	0.603	0.546	0.573	0.73
LogRegression unigram	0.65	0.425	0.573	0.73

on news data with unigram feature vectors. The same is repeated for forum posts (table 10), Reddit posts (table 11), and IRC chatter (table 12 and 13). The process was further split by separating the IRC channels with the assumption that vocabulary might differ between the development (#dev) and over-the-counter (#otc) channel. The number of cross-validation folds (10) was chosen iteratively for the logistic regression and multinomial naive Bayes classifiers, optimizing for accuracy. A confusion matrix of each source and classifier can be found in the appendix.

Although the AlchemyAPI is easily scalable, does not require any preprocessing, and shows a decent accuracy score on IRC chatter and forum posts compared to the other models, the low precision and recall make it a poor choice for our application.

Without a true cost attached to mistaken decisions, a trade-off between precision and recall can not directly be made. The harmonic mean of precision and recall (the F-Measure, introduced by van Rijsbergen [57]) and accuracy will serve as the main selection criteria for now. The harmonic mean of the precision and recall is calculated as follows:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{11}$$

Table 12. IRC OTC classifier comparison.

IRC #OTC	Precision	Recall	F-Measure	CV / Accuracy
AlchemyAPI Raw text	0.57677	0.6720	0.6207	0.77
NB MultiNom. unigram	0.623	0.57	0.595	0.65
LogRegression unigram	0.7	0.538	0.609	0.69

Table 13. IRC DEV classifier comparison.

IRC #DEV	Precision	Recall	F-Measure	CV / Accuracy
AlchemyAPI Raw text	0.3099	0.8549	0.45489	0.658
NB MultiNom. unigram	0.641	0.59	0.616	0.69
LogRegression unigram	0.623	0.631	0.628	0.71

4 Correlation and Causation

Recall that sentiment was classified in a binary matter. This means that a single negative post cancels out a single positive post when counting sentiment cumulatively, without any regard for the degree of polarity. In order to establish the correlations of cumulative, negative and positive sentiments individually, correlations have been calculated separately.

Initially it is assumed that sentiment can correlate with market movements for more than one day after the sentiment is expressed. As we are mostly interested in predicting upward or downwards movement in exchange price and trade volume, we calculate the percentage and absolute changes in volume and average daily price between t (the date at which the sentiment was expressed) and $t + n$ (the date on which we calculate the correlation).

Table 23, 24, and 25 illustrate correlations of the positive, negative, and cumulative sentiment at t with the price and volume changes from t to $t + n$. For each channel the best respective classifier from the previous section was chosen to classify the dataset over 2015. The Pearson correlation coefficients (r) and corresponding p-values are calculated using the *pearsonr* module from the Python SciPy library [58].

For all channels except Reddit, positive sentiment seems to correlate negatively with changes in trading volume from *t+2* onwards at significance levels of at least 0.05 or below. Interestingly, positive Reddit sentiment is the only sentiment source which positively correlates with upwards price movements. No other positive sentiment channels exhibits such a correlation - all other positive sentiment correlates negatively. The differences in correlations of the same sentiment type between channels could offer interesting insight into how Bitcoin market participants who reside in specific online communities respond to news (which may or may not be exclusive to that community).

Overall, correlations are stronger for negative expressions of sentiment. In the context of the hypothesis, this could indicate that Bitcoin traders are more sensitive to negative news. This matches with the findings of [11]. In making that statement, it is important to take the classification errors into account. Incorrect classifications of positive sentiment spill over into the negative class. This means that the classifiers will recognize more negative news than there truly is. The same is true in the other direction, which on average compensates this effect to some degree. Future research improving the classification of positive sentiment will likely decrease the correlations of negative sentiment slightly.

The strongest correlations can be found in forum sentiment. A possible explanation for this is the subforum filtering that preceded data collection. Other sources discuss anything Bitcoin related, whereas the scraped subforums center around Bitcoin trading, speculation, and economics, and as such could have a larger population of active Bitcoin traders.

Although the correlations found thus far are weak to moderate, a perfect correlation was never expected; there are many other factors at play in Bitcoin price formation.

In a first effort of strengthening the correlations, we only consider sentiment at time *t* and market movements at *t+n* if the positive sentiment at *t* is above the 90th percentile of daily positive sentiment for 2015 on that channel and negative sentiment is below the 10th percentile of negative sentiment for 2015. This improved correlations, but decreased the statistical significance ($p > 0.05$) by large amounts. Due to the small remaining sample (10th percentiles of 365 daily observations over 2015), we can observe correlations even if there is no true correlation between variables at all. The increase of p-values reflects this uncertainty.

4.1 Granger Causality Test

To determine whether sentiment has any predictive value in forecasting Bitcoin exchange movements in price and volume, we can perform a Granger causality test [59]. The Granger causality test determines whether information in one time series (x) precedes the other (y) and whether including historical data from x improves the prediction accuracy of y over using only historical data from y. It is then said that x Granger-causes y. The Granger causality test accomplishes this by using the information in one time series to model the changes in the other time series. This test does not aim to prove true or philosophical causality. The test only aims to establish predictive causality [60].

As the tests are bi-directional (we want to determine the direction of causality), the null hypotheses H_0 for the performed Granger causality test are the following:

- $H_{0.1}$:*Series x does not Granger cause series y*
- $H_{0.2}$:*Series y does not Granger cause series x*

Series x will be the type of sentiment, and series y the market metric. Null hypotheses are rejected at significance levels beyond 0.05.

The Granger causality test requires that the analyzed time series are both covariance stationary. To determine this, an augmented Dickey-Fuller (ADF) [61] test is performed on each of the series that will be tested in H_0. Stationary series for each variable are then used in the Granger causality test. The ADF test confirms stationarity of percentage daily changes in price and volume, and all collected sentiment time series.

Table 14. Summarized results of the performed Granger causality tests, showing the direction of Granger-causality between series x and y at significance levels of 0.05.

x	y	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5
Negative News	% Price change	→	→	→	→	→
Positive News	% Price change					
Negative Reddit	% Price change	←	→←	←	←	←
Positive Reddit	% Price change	→←	→←	←	←	
Negative forum	% Price change		→	→	→	→
Positive forum	% Price change					
Negative news	% Volume change		→			
Positive news	% Volume change	→←	→←			
Negative Reddit	% Volume change		←	←	←	←
PositiveReddit	% Volume change				←	←
Negative forum	% Volume change	←	←	←	←	←
Positive forum	% Volume change	→←	←	←	←	←

The EViews 9 software [62] was used to perform the causality tests with various lag settings. A lag-length of for example 2 makes the assumption that data from one of the selected time series x does not help predict series y if it is more than two time steps removed from y.

Table 14 shows the summarized findings of the Granger Causality tests as performed on news articles, Reddit posts and forum topics, with percentage changes in volume and average daily price. For news and forum channels, we most notably find a Granger-causality from negative sentiment to the price change, indicating that negative news sentiment has value in predicting these movements. This observation is reversed in negative Reddit sentiment, where sentiment is Granger-caused by percentage changes in price. This surfaces an interesting pattern in the behavior of online Bitcoin communities. News and forum are seemingly used to collect trading intelligence, whereas Reddit seems contain discussion as to what happened on the markets. We also find that volume changes are leading indicators for negative and positive forum sentiment and negative Reddit sentiment. None of the analyzed channels show a predictive causality from negative or positive sentiment to changes in trading volume.

Full results of the Granger tests can be found in table 26, 27 and 28 in the Appendix.

5 Conclusion and further research

This research provides evidence of the value of including online discourse sentiment originating from various sources in predicting daily Bitcoin exchange movements in price and volume. In parallel, it confirms earlier findings by [7] that suggest social chatter mirrors preceding price and volume changes that occur on the exchanges. As we now find, this predictive causal relationship is not unidirectional for all channels and should be further dissected by channel, market metric, and sentiment. In line with the findings by [7], exchange movements do indeed seem to serve as a predictive indicator for the expressed sentiment on the examined social channels (forum and Reddit) for both volume and price. On the other hand, we do find predictive causality for both negative forum and news sentiment in relation to the average daily exchange price. This becomes apparent from the performed Granger causality test, in which we most notably find that negative news and forum sentiment leads price movements, but price and volume movements lead sentiment on forum and Reddit channels at significance levels of 0.05 or below.

This relationship can likely be strengthened by improving the classifier accuracies beyond the simple logistic regression and multinomial naive Bayes models used to reach these results. Feature engineering might also help, as there did not seem to be a clear set of separating features for positive sentiment documents. The true test is the application of these findings into algorithmic Bitcoin trading strategies to determine whether these findings hold up in practice.

Although the sentiment time series are not included into trading algorithms, the performed tests are sufficient to confirm the hypothesis. As the Granger causality test inherently measures if there is any added value of another time series y (in this case sentiment) in predicting series x, we can conclude that if this sentiment data were indeed included in a trading model, it has the potential to improve the prediction accuracy.

The main limitations of the conducted study consist of the relatively small and sparse data set and time frame over which the study was carried out. Collecting sentiment from a wider set of sources will make it possible to get a better sense of the overall sentiment amongst the Bitcoin community and see how sentiment changes over time. Single-year Bitcoin exchange prices are generally not stationary and follow a trend. Language and geographical bias also play a role. Bitcoin is for example heavily used in China. It then follows that the Chinese Bitcoin market participants will likely also be influenced by online discussions that were not included in this research.

Secondary news and discussions that do not mention Bitcoin directly were also not investigated in this research, although such news can carry important information concerning Bitcoin. News regarding foreign monetary policy for example could very well have a positive effect on Bitcoin, in that more stringent monetary controls can be circumvented by the use of Bitcoin, driving further demand for the cryptocurrency. For the conducted research, the assumption was made that if this information is truly relevant to Bitcoin, it will spill over into Bitcoin communities. This direct analysis and synthesis of secondary news and discussion in relation to Bitcoin would however bring a speed advantage if applied in trading algorithms.

References

[1] J. Bollen, H. Mao, X. Zeng, Journal of Computational Science **2**, 1 (2011), `1010.3003`

[2] C. Oh, O. Sheng (2011)

[3] J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (Springer, 2013), pp. 77–88

[4] Y. Zhang, P.E. Swanson, Journal of Economics and Finance **34**, 96 (2010)

[5] y...u..h.u... Coindesk LLC, title = Bitcoin Market Capitalization

[6] CoinMarketCap, *Cryptocurrency market capitalizations* (2017), `http://coinmarketcap.com`

[7] J. Kaminski, pp. 1–8 (2014), `1406.7577`

[8] D. Garcia, F. Schweitzer, Royal Society open science **2**, 150288 (2015)

[9] J. Bukovina, M. Marticek et al., Tech. rep., Mendel University in Brno, Faculty of Business and Economics (2016)

[10] P. Ciaian, M. Rajcaniova, d. Kancs, Applied Economics **48**, 1799 (2016)

[11] P.C. Tetlock, The Journal of Finance **62**, 1139 (2007)

[12] G. Gidofalvi, C. Elkan, Department of Computer Science and Engineering, University of California, San Diego (2001)

[13] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, I. Weber, PloS one **7**, e40014 (2012)

[14] S. Sabherwal, S.K. Sarkar, Y. Zhang, Journal of Business Finance & Accounting **38**, 1209 (2011)

[15] C.M. Jones, G. Kaul, M.L. Lipson, Review of financial studies **7**, 631 (1994)

[16] S. Nakamoto, Www.Bitcoin.Org p. 9 (2008), `43543534534v343453`

[17] A. Back et al., *Hashcash-a denial of service countermeasure* (2002)

[18] N. Popper, *Digital gold: Bitcoin and the inside story of the misfits and millionaires trying to reinvent money* (Harper, 2015)

[19] Dell Inc., *Dell now accepts bitcoin* (2014), `http://www.dell.com/bitcoin/`

[20] Overstock.com, Inc., *Bitcoin on overstock.com* (2014), `https://www.overstock.com/bitcoin`

[21] D. Yermack, NBER Working Paper Series **53**, 1689 (2013), `arXiv:1011.1669v3`

[22] A. Kumar, C. Lee, The Journal of Finance **61**, 2451 (2006)

[23] B.M. Barber, T. Odean (2011)

[24] D.K. Pearce, V.V. Roley, *Stock prices and economic news* (1984)

[25] E.F. Fama, The journal of Finance **25**, 383 (1970)

[26] J. Kleinnijenhuis, F. Schultz, D. Oegema, W. Van Atteveldt, Journalism **14**, 271 (2013)

[27] A.P. Chaboud, B. Chiquoine, E. Hjalmarsson, C. Vega, The Journal of Finance **69**, 2045 (2014)

[28] M. Glantz, R. Kissell, *Multi-asset risk modeling: techniques for a global economy in an electronic and algorithmic trading era* (Academic Press, 2013)

[29] Aldridge, Krawciw, *Real-Time Risk: What Investors Should Know About Fintech, High-Frequency Trading and Flash Crashes* (Wiley and Sons, Inc., 2017)

[30] M. O'Hara, D. Easley, *Financial markets are at risk of a 'big data' crash* (2013), `https://www.ft.com/content/48a278b2-c13a-11e2-9767-00144feab7de`

[31] B. Pang, L. Lee et al., Foundations and Trends® in Information Retrieval **2**, 1 (2008)

[32] B. Liu, Synthesis lectures on human language technologies **5**, 1 (2012)

[33] B. Pang, L. Lee, S. Vaithyanathan, *Thumbs up?: sentiment classification using machine learning techniques*, in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (Association for Computational Linguistics, 2002), pp. 79–86

[34] R. Snow, B.O. Connor, D. Jurafsky, A.Y. Ng, D. Labs, C. St, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 254–263 (2008)

[35] I. Amazon.com, *Amazon mechanical turk faq* (2005), `https://www.mturk.com/mturk/help?helpPage=overview`

[36] B. Mellebeek, F. Benavent, J. Grivolla, J. Codina, M.R. Costa-Jussa, R. Banchs, *Opinion mining of spanish customer comments with non-expert annotations on mechanical turk*, in *Proceedings of the NAACL HLT 2010 workshop on Creating speech and language data with Amazon's mechanical turk* (Association for Computational Linguistics, 2010), pp. 114–121

[37] A.P. Dawid, A.M. Skene, Applied statistics pp. 20–28 (1979)

[38] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, J. Movellan, Advances in Neural Information Processing Systems **22**, 1 (2009)

[39] Blockchain.info, *Average usd market price across major bitcoin exchanges* (2015), `https://blockchain.info/charts/market-price`

[40] C.D. Manning, P. Raghavan, H. Schütze et al., *Introduction to information retrieval*, Vol. 1 (Cambridge university press Cambridge, 2008)

[41] spaCy, *spacy, industrial-strength natural language processing* (2016), `https://spacy.io`

[42] D. Andor, C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, M. Collins, arXiv preprint arXiv:1603.06042 (2016)

[43] J.D. Choi, J.R. Tetreault, A. Stent, *It Depends: Dependency Parser Comparison Using A Web-based Evaluation Tool.*, in *ACL (1)* (2015), pp. 387–396

[44] nltk.org, *Nltk corpora* (2017), `http://www.nltk.org/nltk_data/`

[45] H. Saif, M. Fernández, Y. He, H. Alani (2014)

[46] scikit learn, *scikit-learn, machine learning in python* (2007), `http://http://scikit-learn.org`

[47] G. Salton, C. Buckley, Information processing & management **24**, 513 (1988)

[48] V. Jain, J. Mahadeokar, *Short-text representation using diffusion wavelets*, in *Proceedings of the 23rd International Conference on World Wide Web* (ACM, 2014), pp. 301–302

[49] IBM, *IBM Watson Developer Cloud: AlchemyLanguage* (2015), `https://www.ibm.com/watson/developercloud/alchemy-language.html`

[50] S. Tan, G. Wu, H. Tang, X. Cheng, *A novel scheme for domain-transfer problem in the context of sentiment analysis*, in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (ACM, 2007), pp. 979–982

[51] A.L. Berger, V.J.D. Pietra, S.A.D. Pietra, Computational linguistics **22**, 39 (1996)

[52] X. Zhang, Y. LeCun, arXiv preprint arXiv:1502.01710 (2015)

[53] P. Domingos, M. Pazzani, Machine learning **29**, 103 (1997)

[54] S. Wang, C.D. Manning, *Baselines and bigrams: Simple, good sentiment and topic classification*, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (Association for Computational Linguistics, 2012), pp. 90–94

[55] A.Y. Ng, M.I. Jordan, Advances in neural information processing systems **2**, 841 (2002)

[56] V.N. Vapnik, V. Vapnik, *Statistical learning theory*, Vol. 1 (Wiley New York, 1998)

[57] C.J. Van Rijsbergen (1979)

[58] SciPy.org, *Scipy.org — scipy.org* (2017), `https://www.scipy.org`

[59] C.W. Granger, Econometrica: Journal of the Econometric Society pp. 424–438 (1969)

[60] F.X. Diebold, *Elements of forecasting* (Citeseer, 1998)

[61] R. Mushtaq (2011)

[62] I.G. Inc., *Eviews 9sv* (2015), `http://www.eviews.com/EViews9/EViews9SV/evstud9.html`

A Appendix

Table 15. Confusion matrix for logistic regression on news articles

Pred. / Real	**P**	**N**
P	191	114
N	62	453

Table 16. Confusion matrix for logistic regression on Reddit posts

Pred. / Real	**P**	**N**
P	208	281
N	112	867

Table 17. Confusion matrix for logistic regression on forum topics

Pred. / Real	**P**	**N**
P	391	303
N	209	1.048

Table 18. Confusion matrix for logistic regression on IRC Chatter

Pred. / Real	**P**	**N**
P	371	318
N	159	692

Table 19. Confusion matrix for multinomial NB on news articles

Pred. / Real	**P**	**N**
P	222	83
N	66	449

Table 20. Confusion matrix for multinomial NB on Reddit posts

Pred. / Real	**P**	**N**
P	267	222
N	176	803

Table 21. Confusion matrix for multinomial NB on forum topics

Pred. / Real	**P**	**N**
P	405	289
N	287	970

Table 22. Confusion matrix for multinomial NB on IRC Chatter

Pred. / Real	**P**	**N**
P	393	296
N	238	613

Table 23. Correlation table of positive sentiment from source at time *t* with metric at *t+n*. Cells show Pearson correlation coefficient (r) and p-value (p)

Source	Metric	t + 1	t + 2	t + 3	t+4	t+5
News	% price change	r = -0.0608 p = 0.2486	r = -0.0268 p = 0.6118	r = -0.0279 p = 0.5970	r = -0.0383 p = 0.4672	r = -0.0570 p = 0.2797
News	% volume change	r = -0.1029 p = 0.0505	r = -0.2019 p = 0.001>	r = -0.2511 p = 0.001>	r = -0.2374 p = 0.001>	r = -0.18107 p = 0.001>
News	abs price change	r = -0.0879 p = 0.0949	r = -0.0710 p = 0.1776	r = -0.0768 p = 0.1449	r = -0.0814 p = 0.1221	r = -0.0880 p = 0.0944
News	abs volume change	r = -0.0837 p = 0.1121	r = -0.1843 p = 0.001>	r = -0.2306 p = 0.001>	r = -0.2165 p = 0.001>	r = -0.1700 p = 0.001
Forum	% price change	r = 0.0089 p = 0.8666	r = -0.0090 p = 0.8641	r = -0.0344 p = 0.5136	r = -0.0301 p = 0.5680	r = -0.0224 p = 0.6707
Forum	% volume change	r = -0.1731 p = 0.0009	r = -0.1992 p = 0.0001	r = -0.1892 p = 0.0003	r = -0.2347 p = 0.001>	r = -0.2496 p = 0.001>
Forum	abs price change	r = -0.0721 p = 0.1713	r = -0.1154 p = 0.0281	r = -0.1458 p = 0.0054	r = -0.1392 p = 0.0080	r = -0.1259 p = 0.0165
Forum	abs volume change	r = -0.2452 p = 0.001>	r = -0.3558 p = 0.001>	r = -0.3642 p = 0.001>	r = -0.4140 p = 0.001>	r = -0.4139 p = 0.001>
Reddit	% orice change	r = 0.1169 p = 0.0260	r = 0.1909 p = 0.001>	r = 0.2114 p = 0.001>	r = 0.1778 p = 0.001>	r = 0.1449 p = 0.0057
Reddit	% volume change	r = 0.0081 p = 0.8780	r = 0.0249 p = 0.6368	r = -0.0043 p = 0.9340	r = -0.0263 p = 0.6186	r = -0.0495 p = 0.3474
Reddit	abs price change	r = 0.1203 p = 0.0221	r = 0.1598 p = 0.0023	r = 0.1736 p = 0.001>	r = 0.1524 p = 0.0037	r = 0.1324 p = 0.0117
Reddit	abs volume change	r = 0.0359 p = 0.4958	r = 0.0084 p = 0.8739	r = -0.0687 p = 0.1925	r = -0.0756 p = 0.1509	r = -0.0854 p = 0.1046
IRC	% price change	r = -0.05217 p = 0.3222	r = -0.0794 p = 0.1317	r = -0.0814 p = 0.1223	r = -0.0944 p = 0.0728	r = -0.0858 p = 0.1032
IRC	% volume change	r = -0.0575 p = 0.2750	r = -0.1041 p = 0.0478	r = -0.0962 p = 0.0676	r = -0.1090 p = 0.03818	r = -0.0269 p = 0.6099
IRC	abs price change	r = -0.0679 p = 0.1975	r = -0.0876 p = 0.0962	r = -0.0919 p = 0.0807	r = -0.1001 p = 0.0570	r = -0.0877 p = 0.0961
IRC	abs volume change	r = -0.1068 p = 0.0424	r = -0.1402 p = 0.0076	r = -0.1142 p = 0.0298	r = -0.0854 p = 0.1047	r = -0.0078 p = 0.8817

Table 24. Correlation table of negative sentiment from source at time *t* with metric at *t+n*. Cells show Pearson correlation coefficient (r) and p-value (p)

Source	Metric	t + 1	t + 2	t + 3	t+4	t+5
News	% price change	r = -0.1159 p = 0.0274	r = -0.1119 p = 0.0333	r = -0.1062 p = 0.0434	r = -0.1110 p = 0.0346	r = -0.0783 p = 0.1369
News	% volume change	r = 0.09119 p = 0.0833	r = 0.1479 p = 0.0048	r = 0.2074 p = 0.001>	r = 0.2212 p = 0.001>	r = 0.1696 p = 0.0012
News	abs price change	r = -0.0727 p = 0.1674	r = -0.0561 p = 0.2869	r = -0.0512 p = 0.3316	r = -0.0566 p = 0.2830	r = -0.0273 p = 0.6045
News	abs volume change	r = 0.0982 p = 0.0621	r = 0.1572 p = 0.0027	r = 0.2260 p = 0.001>	r = 0.2383 p = 0.001>	r = 0.2008 p = 0.001>
Forum	% price change	r = -0.3128 p = 0.001>	r = -0.3422 p = 0.001>	r = -0.3040 p = 0.001>	r = -0.2481 p = 0.001>	r = -0.2322 p = 0.001>
Forum	% volume change	r = 0.1534 p = 0.0034	r = 0.2072 p = 0.001>	r = 0.2139 p = 0.001>	r = 0.1982 p = 0.001>	r = 0.2112 p = 0.001>
Forum	abs price change	r = -0.194 p = 0.001>	r = -0.2032 p = 0.001>	r = -0.1687 p = 0.0013	r = -0.1195 p = 0.0229	r = -0.1048 p = 0.0463
Forum	abs volume change	r = 0.2799 p = 0.001>	r = 0.4074 p = 0.001>	r = 0.4205 p = 0.001>	r = 0.4119 p = 0.001>	r = 0.4054 p = 0.001>
Reddit	% price change	r = -0.1764 p = 0.0007	r = -0.2695 p =0.001>	r = -0.2814 p =0.001>	r = -0.2254 p = 0.001>	r = -0.1871 p = 0.001>
Reddit	% volume change	r = -0.0252 p = 0.6316	r = -0.0421 p =0.4244	r = -0.0225 p =0.6686	r = 0.0214 p = 0.6850	r = 0.0527 p = 0.3178
Reddit	abs price change	r = -0.1742 p = 0.0009	r = -0.2309 p =0.001>	r = -0.2321 p =0.001>	r = -0.1874 p = 0.001>	r = -0.1607 p = 0.0021
Reddit	abs volume change	r = -0.0675 p = 0.1999	r = -0.0313 p =0.5534	r = 0.0509 p =0.3334	r = 0.0824 p =0.1174	r = 0.1019 p = 0.0527
IRC	% price change	r = 0.0274 p = 0.6039	r = 0.0556 p = 0.2914	r = 0.0637 p = 0.2267	r = 0.0683 p = 0.1945	r = 0.0697 p = 0.1856
IRC	% volume change	r = 0.0778 p = 0.1396	r = 0.1322 p = 0.0118	r = 0.1295 p = 0.0137	r = 0.1230 p = 0.0191	r = 0.0278 p = 0.5979
IRC	abs price change	r = 0.0432 p = 0.4126	r = 0.0627 p =0.2341	r = 0.0743 p = 0.1582	r = 0.0738 p = 0.1610	r = 0.0688 p = 0.1913
IRC	abs volume change	r = 0.1203 p = 0.0221	r = 0.1422 p = 0.0067	r = 0.1251 p = 0.0172	r = 0.0776 p = 0.1408	r = -0.0074 p = 0.8877

Table 25. Correlation table of cumulative (negative + positive) sentiment from source at time *t* with metric at *t+n*. Cells show Pearson correlation coefficient (r) and p-value (p)

Source	Metric	t + 1	t + 2	t + 3	t+4	t+5
News	% Change Price	r = -0.1259 p = 0.0165	r = -0.0886 p = 0.0922	r = -0.0867 p = 0.0994	r = -0.1001 p = 0.0570	r = -0.1014 p = 0.0538
News	% Change Volume	r = -0.0562 p = 0.2865	r = -0.1272 p = 0.0154	r = -0.1453 p = 0.0056	r = -0.1237 p = 0.0186	r = -0.0939 p = 0.0744
News	Abs change price	r = -0.1306 p = 0.0131	r = -0.1038 p = 0.04849	r = -0.1070 p = 0.0419	r = -0.1147 p = 0.0291	r = -0.1056 p = 0.0447
News	abs change volume	r = -0.0562 p = 0.2865	r = -0.1272 p = 0.0154	r = -0.1453 p = 0.0056	r = -0.1237 p = 0.0186	r = -0.0939 p = 0.0745
Forum	% Change Price	r = -0.2908 p = 0.001>	r = -0.3273 p =0.001>	r = -0.2692 p = 0.001>	r = -0.2460 p = 0.001>	r = -0.2517 p = 0.001>
Forum	% Change Volume	r = -0.0328 p = 0.5339	r = -0.0085 p = 0.8727	r = 0.0085 p = 0.8715	r = -0.0538 p = 0.3071	r = -0.0569 p = 0.2808
Forum	Abs change price	r = -0.2610 p = 0.001>	r = -0.3149 p = 0.001>	r = -0.3133 p = 0.001>	r = -0.2593 p = 0.001>	r = -0.2314 p = 0.001>
Forum	abs change volume	r = 0.0137 p = 0.7956	r = 0.0209 p = 0.6912	r = 0.0247 p = 0.6383	r = -0.0352 p = 0.5043	r = -0.0414 p = 0.4325
Reddit	% Change Price	r = 0.0876 p = 0.0961	r = 0.1507 p = 0.0040	r = 0.1741 p = 0.001>	r = 0.1512 p = 0.0039	r = 0.1218 p = 0.0205
Reddit	% Change Volume	r = 0.0006 p = 0.9916	r = 0.0167 p = 0.7511	r = -0.0155 p = 0.7679	r = -0.0274 p = 0.6043	r = -0.0464 p = 0.3791
Reddit	Abs change price	r = 0.0931 p = 0.0770	r = 0.1239 p = 0.0183	r = 0.1426 p = 0.0066	r = 0.1320 p = 0.0120	r = 0.1156 p = 0.0278
Reddit	abs change volume	r = 0.0213 p = 0.6869	r = -0.0016 p = 0.9760	r = -0.0736 p = 0.1625	r = -0.0670 p = 0.1841	r = -0.0753 p = 0.1527
IRC	% Change Price	r = -0.0676 p = 0.1988	r = -0.0923 p = 0.0793	r = -0.0895 p = 0.0889	r = -0.1081 p = 0.0397	r = -0.0924 p = 0.0790
IRC	% Change Volume	r = -0.0383 p = 0.4675	r = -0.0759 p = 0.1497	r = -0.0644 p = 0.2213	r = -0.0913 p = 0.0833	r = -0.0245 p = 0.6428
IRC	Abs change price	r = -0.0823 p = 0.1181	r = -0.1009 p = 0.0552	r = -0.0993 p = 0.0590	r = -0.1136 p = 0.0307	r = -0.0962 p = 0.0675
IRC	abs change volume	r = -0.0895 p = 0.0892	r = -0.1295 p = 0.0136	r = -0.0984 p = 0.0613	r = -0.0858 p = 0.1031	r = -0.0190 p = 0.7185

Table 26. F-statistics and probability values for each null hypothesis across different lag-lengths on news sentiment.

News	**Lag: 1**		**Lag: 2**		**Lag: 3**		**Lag: 4**		**Lag: 5**	
Null Hypothesis:	F -Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.
CUMULATIVESENT does not Granger Cause ABSCHANGEAVG	0.26389	0.6078	1.53485	0.2169	4.86419	**0.0025**	3.59847	**0.0068**	2.82964	**0.0161**
ABSCHANGEAVG does not Granger Cause CUMULATIVESENT	0.0169	0.8966	0.29422	0.7453	0.46331	0.7081	0.3604	0.8367	0.43122	0.8268
NEGATIVESENT does not Granger Cause ABSCHANGEAVG	3.6834	0.0557	3.79035	**0.0235**	3.46922	**0.0164**	3.06871	**0.0166**	2.84926	**0.0155**
ABSCHANGEAVG does not Granger Cause NEGATIVESENT	0.01306	0.9091	0.0114	0.9887	0.00315	0.9998	1.51E-01	9.62E-01	0.14615	0.98
POSITIVESENT does not Granger Cause ABSCHANGEAVG	0.26194	0.6091	0.23901	0.7875	2.20257	0.0875	1.66986	0.1564	1.34317	0.2455
ABSCHANGEAVG does not Granger Cause POSITIVESENT	0.15042	0.6984	0.80314	0.4487	0.70401	0.5502	4.66E-01	7.61E-01	0.40729	0.8437
ABSVOLUMECHANGE does not Granger Cause CUMULATIVESENT	4.37414	**0.0372**	2.59164	0.0763	1.87451	0.1335	1.50159	0.2012	1.10998	0.3547
CUMULATIVESENT does not Granger Cause ABSVOLUMECHANGE	4.54561	**0.0337**	2.49553	0.0839	1.73494	0.1595	1.11E+00	3.52E-01	1.35268	0.2417
PERCCHANGEAVG does not Granger Cause CUMULATIVESENT	0.09473	0.7584	0.2669	0.7659	0.62463	0.5995	0.46815	0.7591	0.40119	0.8479
CUMULATIVESENT does not Granger Cause PERCCHANGEAVG	1.00894	0.3158	1.52376	0.2193	4.4906	**0.0041**	3.13936	**0.0148**	2.482	**0.0316**
PERCCHANGEVOLUME does not Granger Cause CUMULATIVESENT	3.94398	**0.0478**	2.92	0.0551	1.74456	0.1575	1.6056	0.1723	1.31085	0.2587
CUMULATIVESENT does not Granger Cause PERCCHANGEVOLUME	2.60242	0.1076	2.60959	0.075	0.94994	0.4166	0.82142	0.5122	0.75853	0.5803
ABSVOLUMECHANGE does not Granger Cause NEGATIVESENT	3.52525	0.0612	5.18246	**0.006**	2.97718	**0.0316**	2.15452	0.0737	2.25054	**0.049**
NEGATIVESENT does not Granger Cause ABSVOLUMECHANGE	1.75862	0.1856	4.02007	**0.0188**	3.09471	**0.027**	1.46392	0.2127	0.9675	0.4376
PERCCHANGEAVG does not Granger Cause NEGATIVESENT	0.21906	0.64	0.12899	0.879	0.07601	0.9729	0.25833	0.9045	0.20456	0.9605
NEGATIVESENT does not Granger Cause PERCCHANGEAVG	5.41362	**0.0205**	4.59487	**0.0107**	3.70029	**0.012**	3.16422	**0.0142**	2.80145	**0.017**
PERCCHANGEVOLUME does not Granger Cause NEGATIVESENT	0.74294	0.3893	2.25658	0.1062	1.19638	0.311	1.04755	0.3825	1.32121	0.2544
NEGATIVESENT does not Granger Cause PERCCHANGEVOLUME	3.42616	0.065	3.2877	**0.0385**	2.19979	0.0878	2.47172	4.43E-02	1.87269	0.0984
POSITIVESENT does not Granger Cause ABSVOLUMECHANGE	7.7243	**0.0057**	5.72	**0.0036**	3.93152	**0.0088**	1.8415	0.1204	1.53708	0.1776
ABSVOLUMECHANGE does not Granger Cause POSITIVESENT	8.34727	**0.0041**	4.1187	**0.017**	2.42944	0.0651	2.07E+00	8.47E-02	1.39471	0.2256
POSITIVESENT does not Granger Cause PERCCHANGEAVG	0.06235	0.803	0.08844	0.9154	2.1425	0.0946	1.51381	0.1976	1.24207	0.2889
PERCCHANGEAVG does not Granger Cause POSITIVESENT	0.54049	0.4627	1.07E+00	0.3427	0.93026	0.4262	6.87E-01	6.01E-01	0.46431	0.8028
POSITIVESENT does not Granger Cause PERCCHANGEVOLUME	6.48723	**0.0113**	5.8258	**0.0032**	2.59539	0.0523	1.31094	0.2655	1.12198	0.3483
PERCCHANGEVOLUME does not Granger Cause POSITIVESENT	4.60202	**0.0326**	3.39966	**0.0345**	1.44377	0.2298	1.43359	0.2223	1.21789	0.3002

Table 27. F-statistics and probability values for each null hypothesis across different lag-lengths on Reddit sentiment.

Reddit	**Lag 1**		**Lag 2**		**Lag 3**		**Lag 4**		**Lag 5**	
Null Hypothesis:	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.
ABSCHANGEAVG does not Granger Cause CUMULATIVESENT	0.70582	0.4014	0.52859	0.5899	0.48504	0.6929	0.43334	0.7845	0.45817	0.8073
CUMULATIVESENT does not Granger Cause ABSCHANGEAVG	6.39111	**0.0119**	4.4292	**0.0126**	3.21015	**0.0232**	2.41033	**0.049**	2.59565	**0.0254**
ABSVOLUMECHANGE does not Granger Cause CUMULATIVESENT	0.85451	0.3559	3.18711	**0.0425**	2.83765	**0.038**	2.11369	0.0787	1.82367	0.1075
CUMULATIVESENT does not Granger Cause ABSVOLUMECHANGE	2.60587	0.1073	1.27288	0.2813	0.74941	0.5233	5.82E-01	0.6759	0.77466	0.57
PERCCHANGEAVG does not Granger Cause CUMULATIVESENT	3.58836	0.059	2.67925	0.07	1.99731	0.1141	1.72286	0.1444	1.48623	0.1936
CUMULATIVESENT does not Granger Cause PERCCHANGEAVG	4.46509	**0.0353**	3.20789	**0.0416**	2.33744	0.0734	1.98E+00	**9.64E-02**	2.52124	**0.0293**
PERCCHANGEVOLUME does not Granger Cause CUMULATIVESENT	0.13444	0.7141	3.93493	**0.0204**	3.04045	**0.0291**	2.19576	0.0691	2.15307	0.0588
CUMULATIVESENT does not Granger Cause PERCCHANGEVOLUME	0.28072	0.5966	0.09559	0.9089	1.36177	0.2543	1.18E+00	3.19E-01	1.12452	0.3469
NEGATIVESENT does not Granger Cause ABSCHANGEAVG	4.71936	**0.0305**	4.47285	**0.0121**	3.09613	**0.027**	2.4132	**0.0488**	3.36257	**0.0056**
ABSCHANGEAVG does not Granger Cause NEGATIVESENT	2.35269	0.126	3.24368	**0.0402**	3.29003	**0.0208**	2.50177	**0.0422**	2.08423	0.0669
POSITIVESENT does not Granger Cause ABSCHANGEAVG	6.18646	**0.0133**	4.50	**0.0118**	3.28056	**0.0211**	2.41738	**0.0484**	2.83936	**0.0158**
ABSCHANGEAVG does not Granger Cause POSITIVESENT	1.07406	0.3007	1.03711	0.3555	1.04604	0.3722	0.8076	0.5209	0.75611	0.582
ABSVOLUMECHANGE does not Granger Cause NEGATIVESENT	4.33377	**0.0381**	6.27653	**0.0021**	4.7308	**0.003**	3.61384	**0.0067**	2.68446	**0.0214**
NEGATIVESENT does not Granger Cause ABSVOLUMECHANGE	3.71005	0.0549	2.63771	0.0729	2.674	**0.0472**	1.69463	0.1507	1.27807	0.2728
PERCCHANGEAVG does not Granger Cause NEGATIVESENT	6.59101	**0.0107**	6.85821	**0.0012**	5.91462	**0.0006**	4.42879	**0.0017**	3.50763	**0.0042**
NEGATIVESENT does not Granger Cause PERCCHANGEAVG	2.62708	0.1059	3.58289	**0.0288**	2.22601	0.0849	1.9451	0.1025	3.95459	0.0017
PERCCHANGEVOLUME does not Granger Cause NEGATIVESENT	1.2512	0.2641	11.5687	**0.001>**	8.21694	**0.001>**	6.62051	**0.001>**	5.09271	**0.001>**
NEGATIVESENT does not Granger Cause PERCCHANGEVOLUME	0.08815	0.7667	0.24533	0.7826	1.35721	0.2557	0.47393	7.55E-01	0.65546	0.6575
ABSVOLUMECHANGE does not Granger Cause POSITIVESENT	1.93662	0.1649	4.11	**0.0172**	3.21604	**0.023**	2.44426	**0.0464**	2.00366	0.0776
POSITIVESENT does not Granger Cause ABSVOLUMECHANGE	3.06341	0.0809	1.53543	0.2168	1.25662	0.2891	8.44E-01	4.98E-01	0.88645	0.4902
PERCCHANGEAVG does not Granger Cause POSITIVESENT	4.75851	**0.0298**	4.31289	**0.0141**	3.19138	**0.0238**	2.50096	**0.0423**	2.03919	0.0727
POSITIVESENT does not Granger Cause PERCCHANGEAVG	4.0785	**0.0442**	3.15E+00	**0.0441**	2.2363	0.0837	1.89E+00	1.11E-01	2.87951	0.0146
PERCCHANGEVOLUME does not Granger Cause POSITIVESENT	0.37805	0.539	6.38436	0.0019	4.92423	0.0023	3.69986	**0.0058**	3.27451	**0.0067**
POSITIVESENT does not Granger Cause PERCCHANGEVOLUME	0.22396	0.6363	0.10169	0.9033	1.59389	0.1905	1.11915	0.3472	1.19273	0.3123

Table 28. F-statistics and probability values for each null hypothesis across different lag-lengths on forum sentiment.

Forum	**Lags: 1**		**Lags: 2**		**Lags: 3**		**Lags: 4**		**Lags: 5**	
Null Hypothesis:	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.	F-Statistic	Prob.
CUMULATIVESENT does not Granger Cause ABSCHANGEAVG	5.37386	**0.021**	4.78089	**0.0089**	2.97202	**0.0318**	2.49198	**0.0429**	1.99967	0.0781
ABSCHANGEAVG does not Granger Cause CUMULATIVESENT	12.6036	**0.0004**	6.21084	**0.0022**	4.24911	**0.0057**	3.06509	**0.0167**	2.30655	**0.0441**
NEGATIVESENT does not Granger Cause ABSCHANGEAVG	3.87527	**0.0498**	3.91498	**0.0208**	3.9053	**0.0091**	3.45839	**0.0087**	3.09085	**0.0096**
ABSCHANGEAVG does not Granger Cause NEGATIVESENT	0.69437	0.4052	1.5076	0.2228	1.16715	0.3222	0.71713	0.5807	0.71631	0.6115
POSITIVESENT does not Granger Cause ABSCHANGEAVG	0.12882	0.7199	0.43865	0.6453	1.38445	0.2472	1.86008	0.117	1.5792	0.1652
ABSCHANGEAVG does not Granger Cause POSITIVESENT	2.1043	0.1478	1.59362	0.2046	1.07648	0.359	0.96918	0.4244	1.55367	0.1726
CUMULATIVESENT does not Granger Cause ABSVOLCHANGE	0.10768	0.743	0.06604	0.9361	0.04997	0.9852	0.01054	0.9998	0.02391	0.9997
ABSVOLCHANGE does not Granger Cause CUMULATIVESENT	0.00256	0.9597	0.00521	0.9948	0.43183	0.7303	0.39869	0.8096	0.41958	0.8351
NEGATIVESENT does not Granger Cause ABSVOLCHANGE	7.40256	**0.0068**	8.76391	**0.0002**	5.48111	**0.0011**	3.21989	**0.0129**	2.66738	**0.0221**
ABSVOLCHANGE does not Granger Cause NEGATIVESENT	8.68505	**0.0034**	6.10067	**0.0025**	5.10827	**0.0018**	4.05783	**0.0031**	3.55953	**0.0037**
POSITIVESENT does not Granger Cause ABSVOLCHANGE	7.82675	**0.0054**	7.91068	**0.0004**	4.61634	**0.0035**	2.78488	**0.0266**	1.8877	0.0958
ABSVOLCHANGE does not Granger Cause POSITIVESENT	8.95978	**0.0029**	5.82	**0.0033**	6.37787	**0.0003**	5.41249	**0.0003**	4.72053	**0.0003**
PERCCHANGEAVG does not Granger Cause CUMULATIVESENT	13.1291	**0.0003**	6.41404	**0.0018**	4.40507	**0.0047**	3.14797	**0.0146**	2.2683	**0.0474**
CUMULATIVESENT does not Granger Cause PERCCHANGEAVG	6.51658	**0.0111**	5.01994	**0.0071**	3.05352	**0.0285**	2.59322	**0.0364**	2.11397	**0.0633**
PERCCHANGEVOLUME does not Granger Cause CUMULATIVESENT	0.6129	0.4342	1.4204	0.243	2.42364	0.0656	1.90812	0.1086	1.48926	0.1926
CUMULATIVESENT does not Granger Cause PERCCHANGEVOLUME	2.11021	0.1472	1.39583	0.249	1.33726	0.262	1.00886	0.4028	0.70573	0.6195
PERCCHANGEAVG does not Granger Cause NEGATIVESENT	1.25703	0.263	2.25454	0.1064	1.76	0.1545	1.06752	0.3724	1.2053	0.3062
NEGATIVESENT does not Granger Cause PERCCHANGEAVG	2.70047	0.1012	3.73282	**0.0249**	3.9802	**0.0082**	3.53612	**0.0076**	3.11343	**0.0092**
PERCCHANGEVOLUME does not Granger Cause NEGATIVESENT	16.8569	**0.0001>**	11.3981	**0.0001>**	9.26046	**0.0001>**	4.49869	**0.0015**	3.49251	**0.0043**
NEGATIVESENT does not Granger Cause PERCCHANGEVOLUME	0.62045	0.4314	1.27	0.2818	1.44031	0.2308	1.30184	0.2689	0.87589	0.4973
POSITIVESENT does not Granger Cause PERCCHANGEAVG	0.73893	0.39	0.88	0.42	1.68	0.17	2.08	0.0823	1.72002	0.1292
PERCCHANGEAVG does not Granger Cause POSITIVESENT	1.12856	0.2888	1.18909	0.3057	0.72517	0.5375	0.64918	0.6278	1.27905	0.2723
POSITIVESENT does not Granger Cause PERCCHANGEVOLUME	4.58277	**0.03**	2.09129	0.13	0.60195	0.61	0.27157	0.8962	0.07298	0.9962
PERCCHANGEVOLUME does not Granger Cause POSITIVESENT	11.2085	**0.0009**	6.77737	**0.0013**	6.40415	**0.0003**	5.15945	**0.0005**	3.92417	**0.0018**