Machine Learning (2024) 113:4761–4784
https://doi.org/10.1007/s10994-021-06095-3

Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data

Ekaterina Loginova[1] · Wai Kit Tsang[1] · Guus van Heijningen[2] · Louis-Philippe Kerkhove[2] · Dries F. Benoit[1]

Received: 1 March 2020 / Revised: 17 May 2021 / Accepted: 15 July 2021 /
Published online: 18 November 2021

Abstract

The emergence of cryptocurrency markets has drastically changed how online transactions are conducted and provide a new investment opportunity. This study contributes to the literature on directional cryptocurrency price returns prediction by expanding the set of meaningful features extracted from textual data with sentiment analysis and comparing their usefulness across multiple data sources. In contrast to previous studies, we use fine-grained topic-sentiment features. More specifically, aspect-based sentiment analysis models, JST and TS-LDA, are implemented to incorporate joint topical-sentiment features and the degree of text subjectivity. We collected, and make available, a dataset, which consists of data scraped from Reddit, Bitcointalk and CryptoCompare sources, to demonstrate that proposed features lead to interpretable topics and an improvement in predictive performance.

Keywords Cryptocurrency · Sentiment analysis · Reddit · Financial forecasting

Editors: Tim Verdonck, Bart Baesens, María Óskarsdóttir and Seppe vanden Broucke.

✉ Ekaterina Loginova
ekaterina.loginova@ugent.be

Wai Kit Tsang
waikit.tsang@ugent.be

Guus van Heijningen
gvanheijningen@crunchanalytics.be

Louis-Philippe Kerkhove
lkerkhove@crunchanalytics.be

Dries F. Benoit
dries.benoit@UGent.be

[1] Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium

[2] Crunch Analytics, Rodelijvekensstraat 28/bus 002, 9000 Ghent, Belgium

1 Introduction

Since its inception in 2008, Bitcoin is increasingly used in online payments and promises more secure financial transactions due to the system's decentralised control as opposed to centralised banking systems (Nakamoto, 2019). The presence of multiple cryptocurrencies, or "altcoins", increases the complexity of the market dynamics. Their popularity has led to a surge in the sharing of cryptocurrency information on social media and other online platforms (Kim et al., 2016). The relevance of Twitter and Reddit for predictive performance has been shown (Garcia & Schweitzer, 2015; Phillips & Gorse, 2017). Still, many other data sources, such as forums and news, remain under-explored.

The present study adds to the research on directional returns prediction for cryptocurrency by comparing three topic-sentiment feature extraction approaches on the forum, and news data scraped from Reddit, Bitcointalk and CryptoCompare in terms of the predictive performance for directional returns of Bitcoin. We make the dataset, which contains alternative cryptocurrencies as well, publicly available.[1] The newly created textual features lead to a higher degree of interpretability and improved performance. They are thus potentially beneficial for investors, as the prediction results can inform an investment strategy in algorithm trading (the construction of which is outside of the scope of this paper).

The rest of the paper is structured as follows: an overview of the literature is given in Sect. 2 and the data is described in Sect. 3. The methodological framework is outlined in Sect. 4 and the findings are presented in Sect. 5. Finally, limitations and implications are discussed in Sect. 6.

2 Literature review

Price dynamics in the cryptocurrency market. Compared to traditional financial assets, investing in cryptocurrencies does not appear safe (Chuen et al., 2017). Its valuation is highly dependent on many factors such as mining costs (Hayes, 2017), network structure and market effects (Kondor et al., 2014) and peer influence of traders (Krafft et al., 2018), inhibiting the transparency of its valuation as a currency (Yermack, 2015). Moreover, the cryptocurrency market is often considered volatile and prone to the occurrence of bubbles in the price dynamics, specifically in the case of Bitcoin (Gerlach et al., 2018). Many theories have already been formulated and tested to shed light on the complexities of the cryptocurrency market, such as evolutionary dynamics inspired by ecological models (ElBahrawy et al., 2017), wavelet coherence analysis (Phillips & Gorse, 2018a) or birth and death models (Wu et al., 2018), but they are not conclusive in disentangling the price dynamics of the market.

While machine learning techniques for stock market prediction are quite successful (Chang et al., 2009; Huang et al., 2005; Kannan et al., 2010; Sheta et al., 2015), a limited number of sources have focused on alternative cryptocurrencies (Alessandretti et al., 2018) other than Bitcoin (Jang & Lee, 2018; McNally et al., 2018; Jiang & Liang, 2017). In this study, we also focus on Bitcoin due to an already large number of other experimental conditions, such as feature combinations and machine learning models that were used; however, the dataset we collect contains alternative cryptocurrencies as well.

[1] https://github.com/edloginova/cryptodata.

Predictions of cryptocurrency prices. Predicting volatility (Andersen et al., 2003), which is a measure of price fluctuations, has been shown to have a significant impact on investment strategies (Fleming et al., 2003). Some studies have already attempted to predict cryptocurrency prices using machine learning. Guo and Antulov-Fantulin (2018), for example, studied the ability to make a short-term prediction of Bitcoin price fluctuations using machine learning methods and Amjad and Shah (2017) developed a trading strategy based on Bitcoin price prediction by using historical time series prices. In line with previous research (Shintate & Pichl, 2019; Valencia et al., 2019), we predict directional returns.

Accuracy is the most used evaluation measure in previous studies on directional returns (Bollen et al., 2011; Xing et al., 2018). This metric, however, could be very misleading, especially in unbalanced datasets. In financial markets, so-called bear or bull markets exist where there is a tendency for stocks to either always move upwards or downwards (Coudert & Raymond, 2011; Maheu & McCurdy, 2000). Depending on the period from which the dataset is extracted, it can be very straightforward to obtain a high accuracy if the market is moving 90% of the time in an upwards direction by always predicting a positive direction for all observations (Sun et al., 2009). As most previous studies do not report on the class imbalance and evaluate accuracy only, we argue that the true model performance is difficult to assess. Other metrics could be more appropriate such as the area under the ROC curve (AUC), which plots sensitivity versus 1-specificity (Elrahman & Abraham, 2013; He & Garcia, 2008; He & Ma, 2013). The benchmark dataset we provide will also allow for more direct and fair comparisons of predictive performance, as suggested by Nassirtoussi et al. (2014).

Textual data sources. Cryptocurrency information is spread through various social media outlets, such as Twitter, Wikipedia, or Reddit forums. Social media holds much value for predicting future events and changes (Schoen et al., 2013) by reflecting the sentiment of socio-economic phenomena and public opinions (Gonzalez-Bailon et al., 2010). Local media coverage has furthermore been evidenced to be a strong predictor for local trading (Engelberg & Parsons, 2011). Network effects are significant in online communities, and individual members and their contributions are important for the diffusion of information (Panzarasa et al., 2009). The intrinsic motivation, shared goals, and social trust among users drive innovative knowledge sharing (Hau & Kim, 2011). Similarly to online communities, social trust has also strongly influenced the rise of cryptocurrencies (Maurer et al., 2013). Detailed information, for example, is shared on the web in response to cryptocurrency price fluctuations and trade volumes, which enable users to make more informed buying/selling decisions (Fleder et al., 2015; Kim et al., 2016). Many cryptocurrencies are traded online after consulting online forums (Grinberg, 2012; Maurer et al., 2013). Tweets that increase the polarisation of sentiment have been found to positively influence the price of Bitcoin (Garcia & Schweitzer, 2015). Also, activity on Reddit has been known to indicate the spread of epidemic-like investment ideas (Phillips & Gorse, 2017), which have been beneficial for the detection of cryptocurrency price bubbles (Phillips & Gorse, 2018a). Moreover, on the basis of newspaper articles, textual information can predict the market and firm valuations (Tetlock, 2007). Indeed, news such as the fluctuations in the cryptocurrency prices and announcements on a cryptocurrency impact investment decisions (Phillips & Gorse, 2018b). However, news articles have not been extensively used in cryptocurrency market prediction yet (Phillips & Gorse, 2018a), and their topical or advanced sentiment features have not been explored.

Each platform is highly specialized in providing its particular kind of content (factual, subjective, etc.) and interacting in a specific manner with their audience, either via short messages (Twitter), carefully written articles (news) or online posts, which can vary in length from short replies to more elaborate texts (forums and Reddit). Each outlet uniquely influences investors and traders. The combination of different data sources can lead to more informed price predictions (Lamon et al., 2017), but it is not entirely clear how each social medium influences the final price. Until now, the literature has not pointed out how features from different data sources affect predictive performance for cryptocurrency prices. Additionally, the area of market prediction, and even more so cryptocurrency, suffers from the lack of high-quality datasets (Nassirtoussi et al., 2014). Past studies have primarily focused on Bitcoin while using Twitter or Reddit as data source (Garcia et al., 2014; Karalevicius et al., 2018; Kristoufek, 2013, 2015; Yelowitz & Wilson, 2015). Given the wide range of models and feature combinations explored, for the sake of clarity, the experiments in this study also focus only on Bitcoin prices as the target variable, but in the collected data, other cryptocurrencies feature as well.

2.1 Text features

Topical features. Topic modelling is a text mining technique that extracts the most prominent topics and their accompanying keywords, resulting in a conceptual overview of the corpus without going through the time-consuming process of manually sifting through the texts (Blei et al., 2003; Lee & Seung, 1999). The information about topics discussed in social media has been shown to influence the market movement (Phillips & Gorse, 2018b; Kim et al., 2017). More precisely, Phillips and Gorse (2018b) retrieved information about the temporal occurrence of various topics by using dynamic topic modelling. The authors show how particular topics tend to precede certain types of price movements, showing the relevance of topic models in cryptocurrency forecasting. Nevertheless, only Kim et al. (2017) applied topic modelling for directional Bitcoin prediction, using a basic approach that assumes a single topic in every document and without taking the sentiment into account. In this study, we investigate if using more recent and more realistic aspect-based sentiment models which extract both topic and sentiment without assuming a single topic per document improves predictive performance.

Basic sentiment features. For stock market price prediction, daily variations in Twitter mood significantly correlate with daily changes in Dow Jones Industrial Average closing values (Bollen et al., 2011). Similarly to financial markets, Twitter and cryptocurrency markets are intricately related to each other (Fry & Cheah, 2016). Analysing user sentiment has been demonstrated to be relevant for predicting virtual currency value fluctuations (Kim et al., 2015). In its simplest and most widely used form, sentiment analysis concerns the polarity of the entire text: whether it is positive or negative. The level of subjectivity expressed by the author is another important feature, yet it has not been addressed so far (Abraham et al., 2018; Jain et al., 2018).

Targeted sentiment features. Basic sentiment classification methods naively assume that each document relates to only one topic (Pang et al., 2008), even though documents can relate to several of them in reality. To overcome this issue, aspect-based sentiment analysis methods jointly extract objects of interest and their corresponding sentiment. This method can be illustrated with the following example. Consider the sentence “markets are too manipulated, but the community is helpful”. The system should determine that the sentiment about the market aspect is negative, while it is positive about the community aspect. In other words, sentiment and topics interact with each other and considering both simultaneously can be beneficial (Riloff et al., 2003). From a user’s perspective, some topics should be discarded as irrelevant while others should be detected in relation to the sentiment for a better understanding. Aspect-based sentiment analysis has been applied on stock prediction (Nguyen & Shirai, 2015), but not on cryptocurrency. Modelling sentiment and topics at the same time did not earn that much attention yet in a financial context (Xing et al., 2018). However, it is a promising research direction, as subjectivity and topic-related sentiment scores allow investors to make more informed transactions since when a prediction is made, the forecast can be linked back to topics and sentiment.

2.2 Contribution

Our contribution can be summarised as followed:

1. We address the research gap of applying aspect-based sentiment analysis (JST and TS-LDA) on textual data for cryptocurrency directional returns prediction. We also include polarity and subjectivity scores, as well as LDA topics. All the different feature configurations are explored and compared in classification experiments.
2. We demonstrate that extracted features increase the performance when predicting directional returns of Bitcoin. In contrast to previous research, this study measures the performance of the models using both ROC AUC and accuracy, and we also report the class balance.
3. We show that extracted topics are interpretable and provide a more fine-grained insight than traditional LDA. We have invited several investors to provide their opinion on the topics and included their remarks in the discussion section.
4. Our dataset combines multiple data sources, including diverse textual sources: financial data from CryptoCompare, search queries frequency from Google Trends and textual data from forums, Reddit and news. As such, we bridge the research gap concerning the sentiment and topical analysis of news data for cryptocurrency prediction.
5. We release our dataset to facilitate the experiments by other research teams. The dataset covers multiple cryptocurrencies and a longer time frame than many earlier works, as the need for it was highlighted by multiple researchers (Li et al., 2018; Phillips & Gorse, 2018b).

3 Data

Historical price data and search trends

We perform experiments on 768 days from 20 February 2017 to 06 April 2019. Financial indicators were fetched from an API provided by cryptocompare.com, a monitoring platform for the cryptocurrency market. The data consist of the daily opening and closing price, high-low and volume of several cryptocurrencies. While the platform covers nearly 1500 cryptocurrencies, we only handle the top five cryptocurrencies based on market capitalisation for data collection, and from those five coins, this study focuses only on Bitcoin in its experiments. We leave the other currencies for future studies because of the large number of factors already involved in our experimental setup. Market capitalisation data is extracted from coinmarketcap.com using an official API. Coins that were re-branded (had their name changed) or did not have sufficient coverage during the research time period

Table 1 Description of the textual datasets

Source	Number of documents	Number of unique tokens	Mean length (tokens)	Type
Reddit	2,637,346	1,285,431	27	Forum comments
CryptoCompare	79,768	195,122	77	News headlines
Bitcointalk	1,643,314	1,723,076	88	Forum comments

were excluded. The search frequency data is obtained from Google Trends for the full research period via a Python module Pytrends.[2]

Texts There are three sources of textual data in this study: Reddit, CryptoCompare and Bitcointalk.

The first source is a popular online discussion platform. It contains multiple subreddits, each focusing on a specific topic. The number of Reddit posts in the dataset consists of around 2 million comments from a subreddit on cryptocurrency.[3] The data is fetched through Pushshift.io.[4] This API was preferred over Reddit's official API, as it allows gathering data over a specific time range. In addition to the body of the comments, meta-variables are gathered: thread titles, the voting scores on comments and threads, comment and post indices and comments' parent indices. The tree structure of any thread can be reconstructed by index matching. The second source is the CryptoCompare news aggregator, from which short news titles are extracted with the official API. The third and final source is Bitcointalk,[5] one of the oldest and largest forums on cryptocurrencies. It features multilingual subforums and threads on alternative cryptocurrencies. A custom web-scraper has been developed to retrieve the forum threads. The statistics of corpora are indicated in Table 1. Extracted texts are highly domain-specific, containing abbreviations and slang.

4 Methodology

The goal is to predict directional returns, which are calculated using closing prices. An upward movement in the closing price corresponds to the positive class in this case, and no movement or a downward movement is considered as the negative class. The problem is thus binary classification [similar to Valencia et al. (2019) and Shintate and Pichl (2019)]. The dataset is only slightly imbalanced with 55.4% positive directions and 44.6% negative. Figure 1 gives an overview of the fairly complicated combination of data sources and algorithms that we propose.

[2] github.com/GeneralMills/pytrends.

[3] Reddit.com/r/CryptoCurrency/.

[4] pushshift.io.

[5] https://bitcointalk.org/.

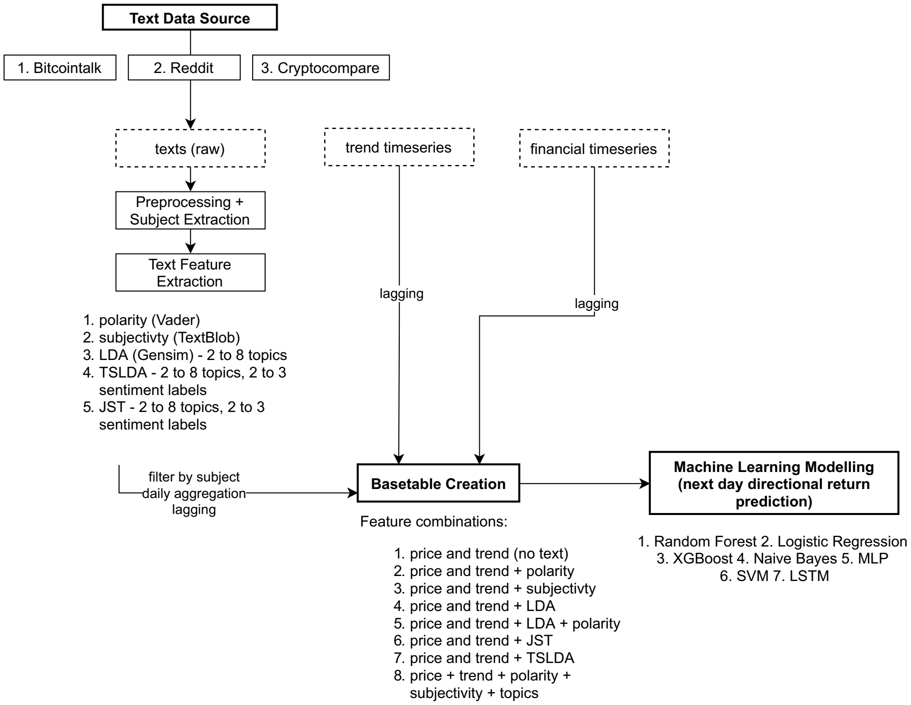
Text Data Source
1. Bitcointalk
2. Reddit
3. Cryptocompare
texts (raw)
trend timeseries
financial timeseries
Preprocessing + Subject Extraction
Text Feature Extraction
lagging
lagging
1. polarity (Vader)
2. subjectivty (TextBlob)
3. LDA (Gensim) - 2 to 8 topics
4. TSLDA - 2 to 8 topics, 2 to 3 sentiment labels
5. JST - 2 to 8 topics, 2 to 3 sentiment labels
filter by subject
daily aggregation
lagging
Basetable Creation
Machine Learning Modelling (next day directional return prediction)
Feature combinations:
1. price and trend (no text)
2. price and trend + polarity
3. price and trend + subjectivty
4. price and trend + LDA
5. price and trend + LDA + polarity
6. price and trend + JST
7. price and trend + TSLDA
8. price + trend + polarity + subjectivity + topics
1. Random Forest 2. Logistic Regression
3. XGBoost 4. Naive Bayes 5. MLP
6. SVM 7. LSTM

Fig. 1 The entire experimental setup. Factors that differ between experiments are highlighted in bold (e.g., data source, features used and model used)

4.1 Textual features

Tokenisation and part-of-speech (POS) tagging are done using NLTK[6] Python library (Loper & Bird, 2002). Before extracting topics, we remove stopwords and punctuation and lemmatize words. URLs, usernames, numbers, currency symbols and emojis are converted to special tokens (e.g., “#emoji#”). Common contractions (such as “’m”) are expanded to their full form.

Subject extraction Finding the correct subject of each comment is a crucial task for the analysis. The aim is to maintain data quality while preserving a sufficient amount of data for further analysis. The eventual subject extraction pipeline is constructed in the following way. From the list of 50 largest coins, all the cryptocurrency names and ticker symbols are obtained. Some cryptocurrencies are usually referred to by one of the words within their longer name. For instance, ICON Project is generally discussed as ICON. Therefore, the names that consist of multiple words are split and put into a separate list. Within this list, words that exist in other cryptocurrency names as well, like the words “token” or “coin”, lead to duplicates and are thus deleted from the list as they can not be used as unique identifiers. The part-of-speech tags are used for filtering out the nouns to match them with cryptocurrency names and ticker symbols. This results in a list of subjects for each observation,

[6] nltk.org.

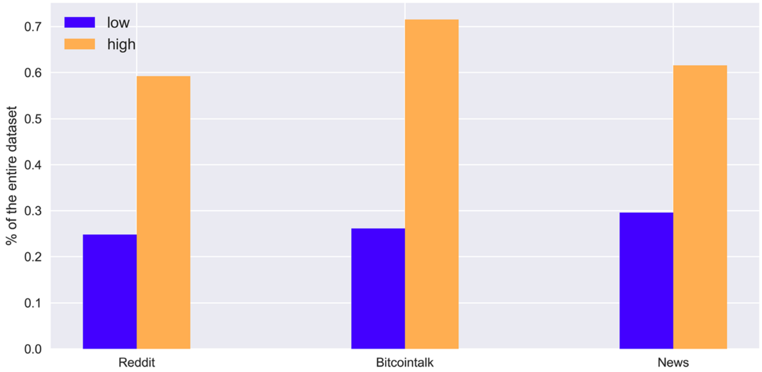

Fig. 2 The distribution of polarity labels assigned by VADER library

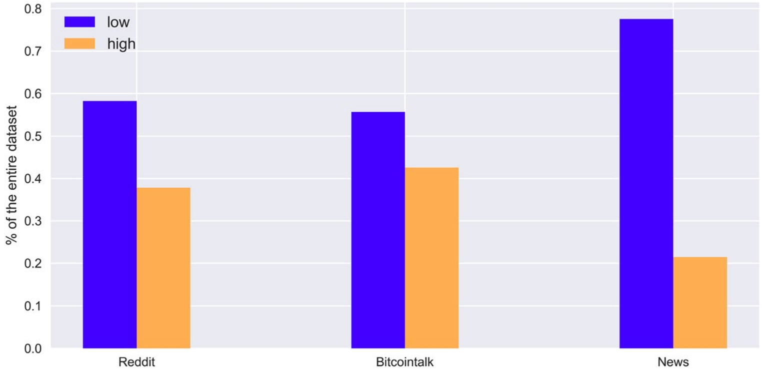

Fig. 3 The distribution of subjectivity labels assigned by TextBlob library

which is associated with the comment or post title that is analysed. Some observations can contain none of the cryptocurrency names from the list, where others contain multiple of them. When the subjects are extracted, the comment tree structure that is present on Reddit is used to further assign subjects to comments that did not identify a particular subject. In other words, the subjects that are discussed in a thread higher up the hierarchy are likely to be the point of discussion for comments that follow underneath. This assumption is used to assign subjects to comments where no subject is obtained from the comment itself. A disadvantage of this method is that it is possible to incorrectly classify comments by assuming they discuss the same subject as their parent comment does, while this was not the case. We have manually annotated a subset of 88 random texts to estimate the accuracy of this rule-based approach. The precision score is 0.95, and the recall is 0.89.

Sentiment The lexicon-based approach VADER (Hutto & Gilbert, 2014) is used to extract a compound polarity score. This library is selected due to its high performance on short informal texts, which constitute a large part of our dataset. The score ranges from − 1 (most negative) to 1 (most positive). The distribution of sentiment scores is reported in Fig. 2. The sentiment is identified for the entire text, and in combination with the subject label from the previous step, we have a coarse-grained alignment sentiment-target. This serves as a baseline method to compare with more advanced target-based sentiment

Table 2 Most salient words for LDA topics

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Remove	Bitcoin	Please	People	Market
Exchange	Blockchain	Question	Crypto	Coin
Wallet	Token	Concern	Think	Price
Account	Project	Action	Really	Bitcoin
Country	Company	Contact	Would	Crypto
Address	Network	Perform	Right	Money
Tax	Currency	Moderator	Thanks	Month
Tether	Block	Comment	Great	Value
Transfer	Platform	Argument	Point	Worth
Funds	System	Submission	Wrong	Amount
Credit	Smart	Click	Pretty	Buy
Regulation	Technology	Thread	Could	Years
Ledger	Business	Post	Someone/anyone/everyone	Buying
Transaction	Decentralize	Contribute	Understand	Dollar
Trade	Partnership	Cryptowikis	Though	Investment
		Content	Look	Start
		Bias	Still	Million
		Verge	Saying	Trading/trade

methods. In addition to polarity, we use TextBlob[7] library to extract the measure of subjectivity expressed by the text (Fig. 3). This aspect has been mostly overlooked in previous studies.

Topics Topical information is extracted per text by using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) (Gensim implementation[8]). The texts are tokenised and preprocessed beforehand: words are lemmatised, common contractions are expanded, URLs, emojis and numbers are converted into special tokens, stopwords and short words are removed. We experiment with the number of topics, trying 2 to 8 topics. As can be seen from Table 2 (produced with pyLDAvis), topics tend to gravitate towards the transaction descriptions (Topic 1), business aspect (Topic 2), knowledge contribution and discussion (Topic 3), personal opinions and conversations (Topic 4), and economic aspect (Topic 5).

In order to evaluate the concepts identified by the models, we contacted investors in cryptocurrencies with the request to double-check the topics. They agreed with the interpretations we provide, although they mention that there is a certain overlap in some of the topics, as well: for example, “country” and “regulation” are more used for ICO’s to restrict some investors from taking white paper or investing money and from exchanges to restrict some accounts doing “Know Your Customer” (KYC).

[7] https://textblob.readthedocs.io/en/dev/.

[8] https://radimrehurek.com/gensim/.

4.2 Other features

Apart from textual features, we use financial data and normalised Google Trends search frequencies to construct lagged variables. The lag is 7 days, and the features are averaged with a rolling window of 1 day (3 days lag was also tried, but it did not lead to an improvement), so that for each day, we obtain, for example, 7 lagged values for the "return" variable.

4.3 Additional processing

The final challenge is to aggregate textual features to make them compatible with these financial and trend features, essentially time series. With the aggregation of individual comments, there will be an inevitable loss of information. It can be alleviated by considering finer granularity than a day, for instance, an hour. However, the financial and Google Trends data APIs did not allow this as their most detailed data was at daily intervals. When the sentiment data is aggregated into daily observations, the resulting features are the total number of comments, the sum of positive comments and the sum of negative comments. For polarity, the threshold is 0 (so comments with the polarity score less than 0 are seen as negative), and for subjectivity, it is 0.5.

4.4 Aspect-based sentiment analysis

In the previous section, we have assumed that each text covers one or more topics, and we have extracted the sentiment of each text. However, in such a setup, if the text contains two topics, and one is mentioned negatively, while the other positively, we would assign a neutral score to both of them. It is possible to alleviate this issue and receive more fine-grained sentiment information by using aspect-based sentiment analysis (ABSA) models.

ABSA involves three steps: aspect identification, aspect mention extraction, and sentiment classification. Aspect identification can be either unsupervised or supervised if the list of target aspects are given. In the latter case, filtering methods based on the part of speech tagging or topic modelling techniques are used. Mention extraction involves correctly determining the aspect corresponding to a given text fragment. Finally, from the rest of the fragment, the sentiment expressed towards the found aspect is classified. The most popular machine learning methods for this task are support vector machines and neural networks. Nonetheless, sentiment lexicon-based approaches are also widely used. The existing supervised ABSA models have decent performance, as demonstrated by SemEval competitions (Pontiki et al., 2016). However, they are not easily transferable even within the original competition's domain, and as such, we decided against pretraining models on SemEval datasets. Due to the lack of labelled domain-specific corpora, we focus on unsupervised methods. We investigate joint models, which are mostly modifications of LDA. Some of them include a time aspect to reflect the evolution of sentiment on a given topic.

JST. Lin and He (2009) proposed a joint/sentiment topic model (JST), based on LDA (Blei et al., 2003), which detects sentiment and topic simultaneously from text. The JST model has the advantage of being fully unsupervised and can hence be applied to domains for which there are no labelled corpora. JST outputs joint topic-sentiment word distributions.

TS-LDA. While JST is not specially tailored for stock price movement, Topic Sentiment Latent Dirichlet Allocation (TS-LDA) was built to predict stock price movement using sentiments on social media (Nguyen & Shirai, 2015). TS-LDA not only simultaneously captures topic and sentiment but also applies it on multiple stocks consisting of many transaction dates. Contrary to JST, which does not distinguish between topic word and opinion word distributions, TS-LDA estimates different opinion word distributions per sentiment for each topic. By doing so, TS-LDA manages to determine which opinion words express positive or negative sentiment.

The scores for the topics are extracted with implementations provided by the authors of respective papers, then aggregated over all the comments per day and time-lagged in the same way as financial and trend predictors (7 days lag with one-day rolling window).

4.5 Predictive models

We compare a wide range of standard machine learning approaches for binary classification: Naive Bayes, Logistic Regression, Support Vector Machines, Random Forest (sklearn implementations (Pedregosa et al., 2011)). We have also implemented a Long-short term memory recurrent neural network (LSTM) and a Multi-Layer Perceptron (MLP) in Keras (Chollet et al., 2015) (with batch normalisation and Scaled Exponential Linear Unit activation function), which is trained using Nesterov Adam optimiser (Kingma & Ba, 2014).

80% of the data was allocated for the train-validation set and 20% for the final test. To account for the rather limited number of examples and slight class imbalance, we used the dropout, early stopping, class weights option of sklearn models (which penalised loss function according to the class proportion in the dataset), and SMOTE over-sampling technique.

For machine learning models, parameters were tuned with randomised search (300 iterations) and fivefold nested time-series cross-validation. For deep learning models, we performed grid-search over hyperparameters: 16, 32 and 64 units, 1 or 2 layers (MLP), 0.3, 0.5 and 0.8 dropout rate.

5 Experiments

In this section, we investigate the extracted topics and evaluate the predictive performance gain on directional returns prediction for Bitcoin. The overall pipeline, from the data collection step to the experiments, is presented in Fig. 1.

5.1 Topics

For LDA, we tried 2 to 8 topics. For JST and TS-LDA, combinations of 2 to 8 topics and 2 to 3 sentiment labels were extracted. Those settings can be seen as hyperparameters.

As a baseline approach for extracting topical information, we used LDA. The examples of the most salient words extracted by a 5 topic LDA model are presented in Table 2. Broadly speaking, topic 1 appears to concern transaction details, containing more technical and legal terms. Topic 2 is related to the networking aspect, while topic 3 deals with the community aspect, the interaction between comment authors. Finally, topic 4 presents a more general communicative aspect, and topic 5 relates to market and investments.

Table 3 Examples of JST topics (most salient words included)

State	Transactions	Emotions	Prognosis	Security
Exchange	Block	Lol	#num#	Wallet
Bank	Miner	Na	#currency#	Use
Money	Segwit	Gon	Price	Key
Currency	Network	Sh*t	Year	Address
Company	Fork	F*ck	Month	Private
Government	Node	#emoji#	reach	Safe
Country	Core	Oh	Worth	Hack
Tax	Fee	Moon	Time	Account
China	Chain		Ago	Secure
State	Transaction			Security
Dollar	Mb			Hacker
Say	Size			
Market dynamics	Opinions/knowledge	Advertisement	Law	Altcoins
Market	Like	Bounty	Government	Coin
Price	Know	Project	Country	Project
Long	Think	Ico	Control	Ethereum
Time	Want	Campaign	Ban	Ico
Bull	Good	Scam	Currency	Eth
Going	News	Join	Tax	Invest
Run	Understand	Airdrop	Money	Token
Crash	Try	Token	Exchange	Potential
Bubble	Need	Signature	China	Altcoins
Happen	Way	Social	Illegal	Market
Term	Come	Medium	Bank	Ripple
Money	Lot	Team	Cryptocurrency	Long
Bear	Bad	Participate	Regulation	Best
Future		Forum	Activity	Neo
		Icos		Future
		Platform		Investment

Examples of most salient words per topic-sentiment combination extracted by JST can be seen in Table 3. On the most general level, the methods converges to broad topics of social interaction in context of knowledge exchange (“life”, “know”, “help”, “want”) and opinions (“think”, “good”, “risk”, often in combination with “invest(ment)”), government/world outlook (“government”, “country”, “bank”, “china”), security (“hacker”, “secure”, “account”, “private”), laws (“tax”, “illegal”, “regulation”, “ban”) and technical details of transactions (“network”, “segwit”, “core”, “fork”). Interestingly, JST is also capable of retrieving important named entities (China, Satoshi), as well as jargon (“fud”, “pump”, “bull”) and emotionally intense words (such as obscene vocabulary or emojis). It also takes a time aspect into account, which is one of the most consistent topics, extracted both as a separate topic that includes month names and a more market-related one that includes words “year”, “ago”, and “time”.

Table 4 Examples of TS-LDA topics (most salient words included)

Transaction	Trade	State & law	Wiki	Market dynamics	Advertisement
Dash	Exchange	Bitcoin	Remove	Buy	Project
Fee	Wallet	Country	Cryptocurrency	Market	Coin
Block	Coin	Government	Wiki	Price	Token
Transaction	Buy	People	Cryptowikis	Sell	Blockchain
Miner	Use	Think	Bot	Think	Team
Network	Fee	Currency	Question	People	Think
Need	Coinbase	Crypto	Moderator	Time	Use
Mining	Binance	Use	Contact	Going	Company
Node	Btc	Ban	Action	Bitcoin	Product
Time	Account	Money	Concern	Money	Platform
Btc	Transfer	News	Perform	Alt	Know
Fork	Crypto	Tax	Rule	Cap	Market
Want	Time	World	Message	Month	Ico
Coin	Trade	China	Submission	Know	Need
Core	Money	Cryptocurrency	Comment	Exchange	New
Use	People	Control	Click	Go	Working
Increase	Send	Accept	Flaired	Hold	Going

However, the interpretation of sentiment classes, as opposed to the topics, is elusive. They do not correlate with polarity or subjectivity scores extracted by VADER. While sometimes we can assume negative sentiment is extracted (such as with swear words), it does not seem to hold over multiple topics (as also noted by the investors we consulted). Similarly, topics extracted by JST are not always easily interpretable when we are interested in the sentiment aspect. For instance, topic 0 appears to cover the international aspect and government regulations when combined with Sentiment 1, but no relevant words are retained in combination with Sentiment 2.

We also explore TS-LDA. In contrast to JST, TS-LDA outputs not only joint distributions but also the most important words for each topic separately (Table 4). The topics appear to be slightly more targeted versions of LDA topics (as indicated by possible interpretations we provide in the table).

Overall, we can see that the retrieved clusters of words overlap with the ones previously found in the literature, e.g. mining, transaction, security, investment, wallet, blockchain found by Kim et al. (2017), which used TM-LDA. JST and TSLDA provide slightly more nuanced versions of the same concepts. However, there are also some differences: it appears that JST is the only model that focuses on domain-specific terms and time aspects, and the wiki concept extracted by TS-LDA is also new compared to the previous studies.

5.2 Classification

In this subsection, we evaluate how the textual features discussed in the previous section can improve the prediction of the upward or downward market movement for Bitcoin.

Table 5 Possible feature configurations and corresponding labels used in tables reporting the predictive performance

Label	Feature configuration
No text	Trend and financial features only
Topic	Features based on topic models (JST/TS-LDA/LDA)
Polarity	Features based on binary sentiment
Subjectivity	Features based on subjectivity
All	Trend, financial, topic-based, polarity-based and subjectivity-based features

Table 6 This table explains the source of topic, polarity and subjectivity features for each topic model

Topic model	Sentiment features	Model
LDA	Topic	LDA
LDA	Polarity	Vader
LDA	Subjectivity	TextBlob
TS-LDA	Topic	TS-LDA
TS-LDA	Polarity	TS-LDA
TS-LDA	Subjectivity	TextBlob
JST	Topic	JST
JST	Polarity	JST
JST	Subjectivity	TextBlob

For LDA, polarity features are extracted from VADER, while TS-LDA and JST output them along with topic scores. Subjectivity is always extracted with TextBlob

There are several factors taken into account in the following experiments: which features are included, e.g. financial or text-based ones, which data source is used, and which machine learning method is applied.

Features The influence of the type of features on the predictive performance is investigated with financial & trend features, topical features and the polarity/subjectivity sentiment.

The methodology to incorporate sentiment analysis in the predictive features leads to six feature configurations. They are explained in Table 5. The source of each sentiment feature per topic model is explained in Table 6. In short, joint (JST and TS-LDA) and separate (LDA + VADER) topic-sentiment models are compared. For LDA, polarity features are extracted from VADER, while TS-LDA and JST output them along with topic scores. Subjectivity is always extracted with TextBlob.

We use the average length of comments as a default textual feature that is added to all configurations (except "none", in which exclusively non-textual features are used). For all topical models, we need to tune the number of topics, and for joint models (JST/TS-LDA) also the number of sentiment labels. For the sake of clarity, we provide only the best score across all tested configurations when reporting results. More details about performance across different hyperparameter configurations for these models are reported in Appendix A.1.

Data source The models are assessed across three data sources, namely Bitcointalk, Reddit and Cryptocompare data, each containing different types of textual data: forums, discussions and news, respectively. The raw texts differ in both quality and publication frequency.

Table 7 Comparison of maximum ROC AUC for different feature configurations (we add one feature type at a time in each column) and datasets (B—Bitcointalk, R—Reddit, CC—CryptoCompare)

	No text	Polarity	Subjectivity	Topic LDA	Topic JST	Topic TSLDA	LDA + V	JST	TS-LDA
B									
LR	0.46	0.50	0.51	0.51	*0.51*	0.50	0.50	*0.51*	0.50
RF	0.48	0.50	0.50	*0.53*	0.52	0.50	0.50	0.51	0.50
XGB	0.49	0.50	0.51	*0.52*	0.51	0.51	0.50	0.51	0.51
NB	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*
MLP	0.47	0.52	0.51	0.52	**0.58**	0.53	0.51	**0.58**	0.53
SVM	0.45	0.50	*0.53*	0.51	0.51	0.52	0.50	0.51	0.52
LSTM	0.55	0.55	0.50	0.53	0.55	*0.56*	0.53	0.54	*0.56*
R									
LR	0.46	0.51	0.44	0.49	0.50	*0.52*	0.49	0.49	*0.52*
RF	0.48	0.50	0.48	0.50	0.49	*0.51*	0.50	0.49	*0.51*
XGB	0.49	0.50	0.49	0.50	0.50	*0.51*	0.50	0.50	*0.51*
NB	0.49	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*
MLP	0.47	0.53	0.53	0.53	*0.54*	0.53	0.53	*0.54*	0.53
SVM	0.45	*0.51*	0.47	0.51	0.51	*0.51*	0.51	0.48	0.49
LSTM	0.55	0.53	0.51	0.53	**0.56**	0.56	0.51	**0.56**	0.56
CC									
LR	0.46	*0.51*	0.49	0.50	*0.51*	0.50	0.50	0.48	0.45
RF	0.48	*0.52*	0.49	*0.52*	0.50	0.49	*0.52*	0.50	0.49
XGB	0.49	*0.51*	0.50	*0.51*	0.50	0.50	*0.51*	0.50	0.50
NB	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*	*0.49*
MLP	0.47	0.52	0.52	0.52	**0.56**	0.56	0.52	**0.56**	0.56
SVM	0.45	*0.49*	0.48	0.49	0.49	0.49	0.49	0.44	0.43
LSTM	0.55	*0.55*	0.52	0.53	0.54	0.54	0.51	0.53	0.54

For topical features, the maximum across all tested configurations of the number of topics/number of sentiment labels is given. The best result per model and dataset combination is highlighted in italic. The best result per dataset (before rounding) is highlighted in bold

Predictive algorithm The final performance also depends on the machine learning methods, five of which are tested here: Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP) and Long-Short Term Memory neural networks (LSTM).

Results We evaluate the benefit of different textual features by adding them one-by-one to a baseline model that uses only financial and trend predictors. The test ROC AUC scores are reported in Table 7.

The first important finding is that our proposed approach improves the state-of-the-art in directional Bitcoin return prediction. This can be seen by comparing the column topic LDA (a similar approach as in Kim et al. (2015)) with the columns to the right of it (containing the results of the proposed advanced text models). The ROC AUC increases with at least 3% points, which is impressive given the notoriously difficult problem of directional return prediction.

In the case of topical features, only the best performance across all tested values for the number of topics is shown; the same is true for the number of sentiment labels for ABSA models (so that each score in, e.g. JST column is the best score across all tried configurations of hyperparameters).

It should be noted that when we add topical features, we add not a single feature but a set of them, e.g., five topic scores if there are five topics. We also separately consider topic scores and joint topic and sentiment scores extracted with ABSA for the sake of a clearer comparison with purely topical features extracted by LDA.

Overall, the added JST and TS-LDA features either improve the predictive performance compared to LDA or LDA in combination with Vader or remain on par with them on all dataset and model combinations. Using only topical information from ABSA models leads to better or similar performance than using both topic and sentiment. As we can see from Table 7, the best result on Bitcointalk is 0.58 (obtained with JST, compared to 0.52 with LDA), on Reddit, it is 0.56 (obtained with JST, compared to 0.53 with LDA), and on CryptoCompare, it is also 0.56 (obtained with JST, compared to 0.52 with LDA + Vader). Moreover, JST typically outperforms TS-LDA, though the scores are very close. We provide a more detailed analysis below.

Adding only basic topical information (using LDA) generally leads to slight improvement on all datasets: on Bitcointalk, the most noticeable gain is with SVM (from 0.45 to 0.51), on Reddit with MLP (from 0.47 to 0.53), and on CryptoCompare also with MLP (from 0.47 to 0.52).

Combining polarity scores with LDA topics does not lead to improvement, which proves the need for a more advanced approach for extracting sentiment and topical information simultaneously. At the same time, compared to it, TS-LDA features do increase the performance when using LSTM on all datasets. However, it is JST topical-sentiment information that leads to the most robust and noticeable gain: from 0.51 to 0.58 for MLP models on Bitcointalk, and it performs on par or better than other models on CryptoCompare and Reddit datasets.

The predictive performance of the extracted features varies over the datasets. The best results are obtained on the Bitcointalk dataset, with a ROC AUC of 0.58, as shown in Table 7. One possible explanation for a difference across data sources is that the gain in the predictive performance depends on the text length and posting frequency. When there is much noise due to the brevity of the comments (153 characters on average for Reddit) or the low frequency of posts (518 posts on average per month for CryptoCompare), while the Bitcointalk dataset has longer comments (515 characters on average) and higher frequency of posts (53,355 on average, more than 100 times the frequency of the CryptoCompare dataset).

Contrary to our expectations, the addition of subjectivity scores does not always lead to an improvement. We observe that it can both increase (e.g., for SVM on Bitcointalk) or decrease the performance (e.g., for LSTM on Bitcointalk), even though in the majority of cases (17 experiments out of 21) it improves. One possible explanation for why it has a more pronounced effect on Bitcointalk than Cryptocompare is the effect of the distribution of the subjectivity labels—as was illustrated in Fig. 3, on Bitcointalk, the imbalance is the lowest, while for CryptoCompare, it is the highest. Further research is required to investigate whether that specialised models determining subjectivity in online text data would perform better, compared to a strict lexicon-based approach that TextBlob utilises. Regarding the polarity scores, we can observe that adding them alone leads to a consistent small improvement on most datasets and predictive models.

Concerning the link between topical features and performance, we have run a preliminary analysis by investigating feature importance scores of XGBoost models. Using all features, we observed that JST features seem to occur much more frequently in the top twenty most important features for the model. More precisely, on the Reddit dataset, out of 14 possible combinations of hyperparameters, JST features appear in the top twenty features 6.3 times on average (7 times they are the top feature), while TSLDA only 2.7; similarly, 4.3 versus 0.35 for Bitcointalk and 3.2 versus 0.29 for CryptoCompare. Thus, there indeed appears to be a connection between extracted topics and market movement.

This result indicates that fine-grained sentiment and topic features can improve predictive performance compared to traditionally used text polarity and LDA topics. Thus, we argue that ABSA features are a promising direction in feature engineering. Moreover, as ABSA models capture the interaction between topic and sentiment dimensions, they also allow a more detailed insight, as shown in the previous section.

6 Conclusion

This study proposes a forecasting methodology to predict directional returns for Bitcoin using aspect-based sentiment analysis for automated feature engineering. We contribute to the literature by applying aspect-based sentiment analysis techniques (JST, TS-LDA) and exploring the benefit of using subjectivity scores on a new dataset that includes several text data sources.

While previous studies have focused almost exclusively on Twitter and Reddit (Abraham et al., 2018; Phillips & Gorse, 2017), our novel dataset includes news and forum data scraped from Bitcointalk and CryptoCompare as well as Reddit. The dataset contains five popular cryptocurrencies (Bitcoin and alternative cryptocurrencies) and is available online.

Proposed models predict directional returns of Bitcoin using topic and sentiment features, which indicate whether a particular topic and sentiment are present in a specific comment. The extracted topics provide a more fine-grained insight than traditional LDA, as shown in the interpretation analysis we give (with the comments provided by cryptocurrency investors). This study thus increases the interpretability of the cryptocurrency model and serves as a step toward understanding why certain predictions are made, which is not always possible when complex models are used for cryptocurrency forecasting (Phillips & Gorse, 2018a).

The predictive performance experiments are carried out on a wide range of machine learning models with different feature and hyperparameter configurations explored on three datasets that contain different types of text data. The added JST and TS-LDA features either improve the predictive performance compared to LDA or remain on par with it in most experiments.

The proposed aspect-sentiment analysis features could be used as an element in the algorithmic trading approach, given that they improve predictive performance (thus giving us a better estimation of the future market situation) and provide insight into model predictions, with a finer level of granularity than using traditional features such as sentiment polarity or LDA.

A limitation imposed by the data is that cryptocurrency data sources do not always provide lengthy comments, and it can be useful to look into other online platforms that have noisy data with high-frequency posts or other sources that create high-quality posts but at a lower frequency to investigate the influence of these properties in more depth.

While the intrinsic difficulty of predicting complex market dynamics on a low-frequency basis leads to a somewhat low absolute performance in some experiments, we attribute it mainly to the size of the dataset and focus on relative improvement provided by the models.

The added value of the textual features appears to depend on the nature of the text. No comparative study has been carried out so far on the properties of the data sources and their impact on predictive performance. To make daily directional forecasts for cryptocurrency, a high frequency of comments appears to be a requirement. Although popular forums such as Reddit are often consulted more frequently for cryptocurrency investment, the length and quality of the comments play a lesser role in the extraction of sentiment polarity. This study shows that sentiment polarity is valuable given that that the dataset provides information on a highly frequent basis (Reddit vs CryptoCompare) and that a longer length of comments (Bitcointalk vs Reddit) also can be valuable. A suggestion for researchers would be to use aspect-based sentiment analysis on alternative datasets instead of the most popular data sources such as Twitter or Reddit.

A Appendix

A.1 Hyperparameter optimization in LDA, JST and TS-LDA

The number of topics and sentiments in JST and TS-LDA was optimized on the basis of their cross-validation accuracy. For LDA, from 2 to 8 topics were tried. For each of the joint topic models, the best set of number and topics was optimized by ranging the models from 2 to 8 topics and varying the sentiments from 2 to 3 as shown in Tables 8, 9 and 10 for Bitcointalk, Reddit and News datasets respectively.

Table 8 ROC AUC over different combinations of hyperparameters for ABSA models on the Bitcointalk dataset (using all features)

JST			TS-LDA			LDA	
# topic	# sentiment	ROC AUC	# topic	# sentiment	ROC AUC	# topic	ROC AUC
2	2	0.544	2	2	0.537	2	0.528
2	3	0.514	2	3	0.498	3	0.525
3	2	0.518	3	2	0.511	4	0.544
3	3	0.52	3	3	0.497	5	0.526
4	2	0.515	4	2	0.515	6	0.528
4	3	0.51	4	3	0.498	7	0.526
5	2	0.505	5	2	0.514	8	0.52
5	3	0.513	5	3	0.535		
6	2	0.543	6	2	0.517		
6	3	0.523	6	3	0.507		
7	2	0.507	7	2	0.494		
7	3	0.5	7	3	0.492		
8	2	0.506	8	2	0.524		
8	3	0.506	8	3	0.513		

Only the maximum score over all machine learning models is shown

Table 9 ROC AUC over different combinations of hyperparameters for ABSA models on the Reddit dataset (using all features)

JST			TS-LDA			LDA	
# topic	# sentiment	ROC AUC	# topic	# sentiment	ROC AUC	# topic	ROC AUC
2	2	0.513	2	2	0.514	2	0.516
2	3	0.511	2	3	0.51	3	0.515
3	2	0.506	3	2	0.526	4	0.515
3	3	0.518	3	3	0.531	5	0.514
4	2	0.548	4	2	0.55	6	0.547
4	3	0.508	4	3	0.515	7	0.522
5	2	0.508	5	2	0.515	8	0.509
5	3	0.522	5	3	0.512		
6	2	0.51	6	2	0.524		
6	3	0.518	6	3	0.512		
7	2	0.506	7	2	0.517		
7	3	0.513	7	3	0.516		
8	2	0.515	8	2	0.512		
8	3	0.505	8	3	0.518		

Only the maximum score over all machine learning models is shown

Table 10 ROC AUC over different combinations of hyperparameters for ABSA models on the CryptoCompare dataset (using all features)

JST			TS-LDA			LDA	
# topic	# sentiment	ROC AUC	# topic	# sentiment	ROC AUC	# topic	ROC AUC
2	2	0.543	2	2	0.543	2	0.517
2	3	0.511	2	3	0.517	3	0.514
3	2	0.503	3	2	0.504	4	0.547
3	3	0.512	3	3	0.509	5	0.504
4	2	0.51	4	2	0.512	6	0.526
4	3	0.533	4	3	0.536	7	0.505
5	2	0.507	5	2	0.506	8	0.518
5	3	0.499	5	3	0.5		
6	2	0.503	6	2	0.5		
6	3	0.512	6	3	0.509		
7	2	0.545	7	2	0.545		
7	3	0.533	7	3	0.522		
8	2	0.537	8	2	0.538		
8	3	0.541	8	3	0.545		

Only the maximum score over all machine learning models is shown

A.2 Predictive performance in terms of accuracy

See Tables 11 and 12.

Table 11 Comparison of accuracy for different feature configurations (we add one feature type at a time in each column)

	No text	Polarity	Subjectivity	Topic LDA	Topic JST	Topic TSLDA	LDA + V	JST	TS-LDA
B									
LR	0.46	0.50	0.51	0.51	*0.51*	0.50	0.50	*0.51*	0.50
RF	0.48	0.51	0.51	0.51	*0.52*	0.50	0.50	0.51	0.50
XGB	0.51	0.52	0.51	0.52	*0.52*	0.51	0.51	0.52	0.51
NB	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*
MLP	0.47	0.52	0.52	0.51	*0.55*	0.52	0.51	*0.55*	0.52
SVM	0.50	*0.52*	0.51	0.51	0.52	*0.52*	0.51	0.52	0.52
LSTM	0.55	0.55	0.50	0.53	0.56	**0.56**	0.53	0.54	**0.56**
R									
LR	0.46	*0.52*	0.44	0.51	0.50	*0.52*	0.51	0.49	0.52
RF	0.48	0.50	0.49	0.50	0.50	*0.50*	0.49	0.49	*0.50*
XGB	0.51	0.51	0.51	0.51	0.51	*0.52*	0.51	0.51	*0.52*
NB	*0.51*	0.49	0.49	0.49	0.49	0.49	0.49	0.49	0.49
MLP	0.47	*0.55*	0.50	0.53	0.52	*0.55*	0.53	0.52	0.52
SVM	0.50	0.51	0.49	0.50	0.51	*0.51*	0.50	0.50	0.51
LSTM	0.55	0.53	0.51	0.54	**0.57**	0.56	0.51	**0.57**	0.56
CC									
LR	0.46	*0.51*	0.49	*0.51*	0.51	0.50	*0.51*	0.48	0.46
RF	0.48	0.51	0.48	*0.51*	0.50	0.49	0.51	0.50	0.49
XGB	0.51	*0.51*	0.51	*0.51*	0.51	0.51	*0.51*	0.51	0.51
NB	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*	*0.51*
MLP	0.47	0.52	0.52	0.51	*0.54*	0.54	0.51	*0.54*	0.54
SVM	0.50	*0.51*	0.50	0.50	0.50	0.50	0.50	0.50	0.49
LSTM	0.55	**0.55**	0.52	0.54	0.55	0.55	0.51	0.54	0.55

B—Bitcointalk, R—Reddit, CC—CryptoCompare. For topical features, the maximum across all tested configurations of the number of topics/number of sentiment labels is given. The best result per model and dataset combination is highlighted in italic. The best result per dataset (before rounding) is highlighted in bold

Table 12 Comparison of accuracy for different topic/sentiment models (with all features used)

	Bitcointalk			Reddit			CryptoCompare		
	TS-LDA	JST	LDA + V	TS-LDA	JST	LDA + V	TS-LDA	JST	LDA + V
LR	0.50	0.49	0.50	0.49	0.51	0.50	0.51	0.50	0.50
RF	0.51	0.50	0.52	0.49	0.49	0.48	0.51	0.49	0.50
XGB	0.52	0.52	0.52	0.51	0.51	0.51	0.52	0.52	0.52
NB	0.51	0.51	0.51	0.49	0.49	0.49	0.51	0.51	0.51
MLP	**0.53**	0.52	**0.54**	**0.54**	**0.53**	0.52	**0.55**	0.55	0.53
SVM	0.51	0.52	0.52	0.51	0.51	0.51	0.51	0.50	0.50
LSTM	0.52	0.54	**0.55**	0.50	0.51	0.51	0.52	0.52	0.53

The maximum across all tested configurations of the number of topics/number of sentiment labels is given. The best result per dataset (before rounding) is highlighted in bold

Funding Not applicable.

Availability of data and material https://github.com/edloginova/cryptodata.

Declarations

Conflicts of interest Not applicable.

Code availability https://github.com/edloginova/cryptodata.

References

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review, 1*(3), 1.

Alessandretti, L., ElBahrawy, A., Aiello, L. M., & Baronchelli, A. (2018). *Machine learning the cryptocurrency market*. arXiv preprint, arXiv:180508550.

Amjad, M., & Shah, D. (2017). Trading bitcoin and online time series prediction. In *NIPS 2016 time series workshop* (pp. 1–15).

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica, 71*(2), 579–625.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8.

Chang, P. C., Liu, C. H., Fan, C. Y., Lin, J. L., & Lai, C. M. (2009). An ensemble of neural networks for stock trading decision making. In *International conference on intelligent computing* (pp. 1–10). Springer.

Chollet, F., et al. (2015). Keras. https://keras.io.

Chuen, K., David, L., Guo, L., & Wang, Y. (2017). Cryptocurrency: A new investment opportunity? *Journal of Alternative Investments, 20*(3), 16–40.

Coudert, V., & Raymond, H. (2011). Gold and financial assets: Are there any safe havens in bear markets. *Economics Bulletin, 31*(2), 1613–1622.

ElBahrawy, A., Alessandretti, L., Kandler, A., Pastor-Satorras, R., & Baronchelli, A. (2017). Evolutionary dynamics of the cryptocurrency market. *Royal Society Open Science, 4*(11), 170623.

Elrahman, S. M. A., & Abraham, A. (2013). A review of class imbalance problem. *Journal of Network and Innovative Computing, 1*(2013), 332–340.

Engelberg, J. E., & Parsons, C. A. (2011). The causal impact of media in financial markets. *The Journal of Finance, 66*(1), 67–97.

Fleder, M., Kester, M. S., & Pillai, S. (2015). *Bitcoin transaction graph analysis*. arXiv preprint, arXiv:150201657.

Fleming, J., Kirby, C., & Ostdiek, B. (2003). The economic value of volatility timing using realized volatility. *Journal of Financial Economics, 67*(3), 473–509.

Fry, J., & Cheah, E. T. (2016). Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis, 47,* 343–352.

Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of bitcoin. *Royal Society Open Science, 2*(9), 150288.

Garcia, D., Tessone, C. J., Mavrodiev, P., & Perony, N. (2014). The digital traces of bubbles: Feedback cycles between socio-economic signals in the bitcoin economy. *Journal of the Royal Society Interface, 11*(99), 20140623.

Gerlach, J. C., Demos, G., & Sornette, D. (2018). *Dissection of bitcoin's multiscale bubble history from January 2012 to February 2018*. arXiv preprint, arXiv:180406261.

Gonzalez-Bailon, S., Banchs, R. E., & Kaltenbrunner, A. (2010). *Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions*. arXiv preprint, arXiv:10094019.

Grinberg, R. (2012). Bitcoin: An innovative alternative digital currency. *Hastings Science & Technology Law Journal, 4,* 159.

Guo, T., & Antulov-Fantulin, N. (2018). *Predicting short-term bitcoin price fluctuations from buy and sell orders*. arXiv preprint, arXiv:180204065.

Hau, Y. S., & Kim, Y. G. (2011). Why would online gamers share their innovation-conducive knowledge in the online game user community? integrating individual motivations and social capital perspectives. *Computers in Human Behavior, 27*(2), 956–970.

Hayes, A. S. (2017). Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin. *Telematics and Informatics, 34*(7), 1308–1321.

He, H., & Garcia, E. A. (2008). Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering, 9,* 1263–1284.

He, H., & Ma, Y. (2013). *Imbalanced learning: Foundations, algorithms, and applications*. John Wiley & Sons.

Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research, 32*(10), 2513–2522.

Hutto, C. J., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. In *Eighth international AAAI conference on weblogs and social media*.

Jain, A., Tripathi, S., DharDwivedi, H., & Saxena, P. (2018). Forecasting price of cryptocurrencies using tweets sentiment analysis. In *2018 eleventh international conference on contemporary computing (IC3)* (pp. 1–7) IEEE.

Jang, H., & Lee, J. (2018). An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information. *IEEE Access, 6,* 5427–5437.

Jiang, Z.,&Liang, J. (2017). Cryptocurrency portfolio management with deep reinforcement learning. In *Intelligent systems conference (IntelliSys), 2017* (pp. 905–913) IEEE.

Kannan, K. S., Sekar, P. S., Sathik, M. M., & Arumugam, P. (2010). Financial stock market forecast using data mining techniques. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, p. 4).

Karalevicius, V., Degrande, N., & De Weerdt, J. (2018). Using sentiment analysis to predict interday bitcoin price movements. *The Journal of Risk Finance, 19*(1), 56–75.

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLoS ONE, 11*(8), e0161197.

Kim, Y. B., Lee, S. H., Kang, S. J., Choi, M. J., Lee, J., & Kim, C. H. (2015). Virtual world currency value fluctuation prediction system based on user sentiment analysis. *PLoS ONE, 10*(8), e0132944.

Kim, Y. B., Lee, J., Park, N., Choo, J., Kim, J. H., & Kim, C. H. (2017). When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation. *PLoS ONE, 12*(5), e0177630.

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint, arXiv:14126980.

Kondor, D., Csabai, I., Szüle, J., Pósfai, M., & Vattay, G. (2014). Inferring the interplay between network structure and market effects in bitcoin. *New Journal of Physics, 16*(12), 125003.

Krafft, P. M., Della Penna, N., & Pentland, A. S. (2018). An experimental study of cryptocurrency market dynamics. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (p. 605). ACM.

Kristoufek, L. (2013). Bitcoin meets google trends and Wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific Reports, 3,* 3415.

Kristoufek, L. (2015). What are the main drivers of the bitcoin price? Evidence from wavelet coherence analysis. *PLoS ONE, 10*(4), e0123923.

Lamon, C., Nielsen, E., & Redondo, E. (2017). Cryptocurrency price prediction using news and social media sentiment. *SMU Data Science Review, 1*(3), 1–22.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788.

Li, T. R., Chamrajnagar, A. S., Fong, X. R., Rizik, N. R., Fu, F. (2018). *Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model*. arXiv preprint, arXiv: 180500558.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 375–384). ACM, New York, NY, USA, CIKM '09. https://doi.org/10.1145/1645953.1646003.

Loper, E., & Bird, S. (2002). *NLTK: The natural language toolkit*. arXiv preprint, arXiv:cs/0205028.

Maheu, J. M., & McCurdy, T. H. (2000). Identifying bull and bear markets in stock returns. *Journal of Business & Economic Statistics, 18*(1), 100–112.

Maurer, B., Nelms, T. C., & Swartz, L. (2013). "When perhaps the real problem is money itself!'': The practical materiality of bitcoin. *Social Semiotics, 23*(2), 261–277.

McNally, S., Roche, J., & Caton, S. (2018). Predicting the price of bitcoin using machine learning. In *2018 26th Euromicro international conference on parallel, distributed and network-based Processing (PDP)* (pp. 339–343). IEEE.

Nakamoto, S. (2019). *Bitcoin: A peer-to-peer electronic cash system*. Technical report, Manubot.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications, 41*(16), 7653–7670.

Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1354–1364). Association for Computational Linguistics, Beijing, China. https://doi.org/10.3115/v1/P15-1131. https://www.aclweb.org/anthology/P15-1131.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1–2), 1–135.

Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology, 60*(5), 911–932.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.

Phillips, R. C., & Gorse, D. (2017). Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *2017 IEEE symposium series on computational intelligence (SSCI)* (pp. 1–7). IEEE.

Phillips, R. C., & Gorse, D. (2018b). Mutual-excitation of cryptocurrency market returns and social media topics. In *Proceedings of the 4th international conference on frontiers of educational technologies* (pp. 80–86). ACM.

Phillips, R. C., & Gorse, D. (2018a). Cryptocurrency price drivers: Wavelet coherence analysis revisited. *PLoS ONE, 13*(4), e0195200.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Mohammad, A. S., Al-Ayyoub M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 19–30)

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003-volume 4* (pp. 25–32). Association for Computational Linguistics.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research, 23*(5), 528–543.

Sheta, A. F., Ahmed, S. E. M., & Faris, H. (2015). A comparison between regression, artificial neural networks and support vector machines for predicting stock market index. *Soft Computing, 7*(8), 2.

Shintate, T., & Pichl, L. (2019). Trend prediction classification for high frequency bitcoin time series with deep learning. *Journal of Risk and Financial Management, 12*(1), 17.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687–719.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.

Valencia, F., Gómez-Espinosa, A., & Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. *Entropy, 21*(6), 589. https://doi.org/10.3390/e21060589.

Wu, K., Wheatley, S., & Sornette, D. (2018). Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. *Royal Society Open Science, 5*(9), 180381.

Xing, F. Z., Cambria, E., & Welsch, R. E. (2018). Natural language based financial forecasting: A survey. *Artificial Intelligence Review, 50*(1), 49–73. https://doi.org/10.1007/s10462-017-9588-9.

Yelowitz, A., & Wilson, M. (2015). Characteristics of bitcoin users: An analysis of google search data. *Applied Economics Letters, 22*(13), 1030–1036.

Yermack, D. (2015). Is bitcoin a real currency? An economic appraisal. In *Handbook of digital currency* (pp. 31–43). Elsevier.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.