Zhuo *et al. Financial Innovation* (2024) 10:60
https://doi.org/10.1186/s40854-023-00501-6

Financial Innovation

REVIEW Open Access

How are texts analyzed in blockchain research? A systematic literature review

Xian Zhuo[1*], Felix Irresberger[2] and Denefa Bostandzic[1,3]

*Correspondence:
xian.zhuo@hhu.de

[1] Manchot Graduate School, Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany
[2] Durham University Business School, Durham University, Mill Hill Lane, DH1 3LB Durham, Germany
[3] Department of Management and Entrepreneurship, Witten/Herdecke University, Alfred-Herrhausen-Str. 50, 58448 Witten, Germany

Abstract

This paper provides a systematic literature review of text analysis methodologies used in blockchain-related research to comprehend and synthesize existing studies across disciplines and define future research directions. We summarize the research scope, text data, and methodologies of 124 papers and identify the two most common combinations of these dimensions: (1) papers that focus on specific cryptocurrencies tend to apply sentiment analysis to instant user-generated content or news articles to discover the correlations between public opinion and market behavior, and (2) studies that examine the broad concept of blockchain with text data from documents published by companies tend to apply topic modeling techniques to explore classifications and trends in blockchain development. We discover five major research topics in the academic literature: relationship discovery, cryptocurrency performance prediction, classification and trend, crime and regulation, and perception of blockchain. Based on these findings, we highlight three potential research directions for researchers to select topics and implement suitable methodologies for text analysis.

Keywords: Blockchain, Text analysis, Systematic literature review, Machine learning algorithm, Topic modeling, Sentiment analysis

Mathematics Subject Classification: C10, C80, O30

Introduction

Blockchain technology and its economics have attracted considerable attention from academic researchers. The total volume of research has increased dramatically, with the proportion of empirical studies growing gradually in recent years (Casino et al. 2019; Xu et al. 2019; Frizzo-Barker et al. 2020). Data availability is often a primary obstacle in empirical studies in emerging research areas, such as blockchain, where it is not clear which alternative data sources should or can be used for quantitative analysis. Owing to its nature, a blockchain primarily comprises numerical data such as on-chain transactions by users or network (value) metrics, trading activity, price data of cryptoassets, or financial reports of the few available companies, most of which are readily available in the public blockchain. However, these datasets can be complemented by text data to obtain more data from consortiums and private blockchains, thus expanding the research span and deriving additional relevant insights.

Given the decentralized nature of the public blockchain ecosystem, there are limited compulsory disclosures or official platforms representing the comprehensive information of single blockchain projects that can serve as sources of blockchain-related information. Alternative sources of textual data play a vital role for different parties in gathering information and making decisions within a blockchain network. For example, the sentiments of a crowd (via news, social media, or other text sources) may be a more relevant reference for investment in the blockchain ecosystem than in corporations. Such data can affect the market, influence investors' decisions, and provide an impetus for blockchain development. Researchers can make use of texts in blockchain-related contexts to obtain information in the data from more perspectives (i.e., explore not only the metadata describing the data but also the actual content of the data) and make inferences that cannot be made before with only numbers.

Therefore, in this study, we focus on providing an overview of text analysis methodologies and data sources as they pertain to blockchains, which differ from the text-based analyses of corporations. There is no consensus on the type of text data that should or could be used to analyze a specific blockchain network or project; therefore, our systematic overview helps alleviate this concern.

Several types of blockchain-related text data are publicly available. First, blockchain is a frequent topic in news articles reporting, with subtopics including the performance of cryptocurrencies and the latest developments in the technology. Second, because of the technical nature of blockchain technology, online platforms or forums such as Twitter, GitHub, and Reddit have been actively used by different groups (e.g., investors and developers) to express their opinions and share and track new developments (Mendoza-Tello et al. 2018). Blockchain startups also use social media for marketing. Third, blockchain project whitepapers provide key information (e.g., technical and marketing) to potential investors and are the primary method for understanding project details (Cohney et al. 2019).

In all these cases, manual examination of large-scale text content is exceptionally labor-intensive and time-consuming, if not impossible. Hence, computer-based text analysis is essential. Researchers across disciplines have provided guidelines for using such type of approaches. Grimmer and Stewart (2013), for example, illustrate the promise and the pitfalls of text analysis for political science. Günther and Quandt (2016) give a comprehensive overview of text analysis methods useful in digital journalism research. Studies in economics and finance have addressed the advantages and disadvantages of different methodologies (Loughran and McDonald 2016; Cong et al. 2021; Gentzkow et al. 2019).

Such reviews have not been conducted in blockchain-related research areas, despite the close connection between blockchain technology and multiple text datasets. Therefore, we argue that it is necessary to use a transparent approach and an academic standpoint to synthesize the current knowledge in the literature to better understand the relevance and potential of text analysis. In this study, we conduct a systematic literature review by examining published and unpublished academic literature, focusing on text analysis associated with blockchain topics across disciplines. We provide the fundamental principles and relevant sources of text analysis methodologies and connect the relationships of research scopes, text data, and methodologies to provide researchers with a

reference for choosing suitable combinations of the above elements with respect to their research question at hand. We then pinpoint the specific research topics studied in the literature and propose directions for future research. This review serves as a guide for researchers from different disciplines interested in conducting blockchain-related text analysis studies.

Research methodology

We conduct a systematic review of the academic literature on blockchain-related research using text analysis. Research in this area has expanded because of the rapid development of blockchain technology. However, because of the interdisciplinary nature of blockchain research, research perspectives vary starkly, posing difficulties in searching for and gathering knowledge beyond a single field. We focus on computer-based text analysis used in blockchain research to comprehend and synthesize studies across disciplines that utilize text analysis as a primary or ancillary methodology. We aim to gain knowledge from the existing literature in this area and discover future research opportunities. We adopt the guidelines of Siddaway et al. (2019) and the PRISMA statement (Liberati et al. 2009; Moher et al. 2009; Page et al. 2021a, b).

Definition of research questions

The first stage of a systematic review involves defining research questions that guide subsequent actions. We propose the following research questions to achieve the objectives of our review:

RQ1 Which research scope, text data, and methodology are used to conduct text analysis in the blockchain area?

Both blockchain and text analysis are broad concepts. This question is designed to identify the specific scope of the studies (e.g., cryptocurrency,[1] smart contract[2]), the text data being analyzed (e.g., social media posts and news), and specific methodologies or techniques used to perform the analyses (e.g., sentiment analysis). We aim to bridge and highlight the connections between these elements in each study. This will assist researchers in selecting the appropriate data and methodologies for their research.

RQ2 What topics are addressed using text analysis in current literature?

The research questions determine how the research develops, and text analysis is one of the methods used to serve the purposes of a study. Regardless of whether text analysis is used alone or as part of a broader analysis, we intend to provide an interdisciplinary

[1] The first use case for blockchains is the creation of cryptocurrencies (e.g., Bitcoin), where Nakamoto (2008) proposed a design for a decentralized payment system in which all transactions are stored in transparent blocks, and transactions are validated through a consensus protocol. The idea is to build trust through protocols and operate the system without authority (i.e., a trusted third party).

[2] A smart contract is essentially a computer-coded contract on blockchain that is automatically executed when the contract terms are met. This increases the enforceability of business contracts without the involvement of a trusted third party (Cong and He 2019).

overview of the topics and research questions addressed in the existing literature, and illustrate how text analysis contributes to the study of these topics.

RQ3 What are the research gaps and promising future research topics?

Based on the findings of our review, we identify understudied areas and future research opportunities using text analysis in blockchain research. This allows researchers to recognize promising research topics and specify the methodologies (and data) they can use.

Literature search and selection

Initial keyword searches were conducted on May 24, 2022, followed by updated searches on August 23, 2022, to find relevant studies. We chose the Web of Science (WoS) and Scopus databases to cover publications indexed in academic databases. As text analysis in blockchain research is relatively new, some studies may not have been published. Therefore, we also performed a keyword search of the Social Science Research Network (SSRN) to distinguish unpublished papers (e.g., working and discussion papers) (Garanina et al. 2021). Subsequently, backward snowballing of the articles obtained through keyword searches was performed to identify additional articles.

For a comprehensive result, our query keywords encompassed not only *blockchain* and *text analysis* but also synonyms and multiple specific topics relevant to the area. Relevant words from blockchain included *blockchain, cryptocurrency, stablecoin*,[3] crypto token, smart contract, initial coin offering (ICO), security token offering (STO), and initial exchange offering (IEO),[4], and *non-fungible token (NFT)*[5] Keywords from text analysis included *text analysis, textual analysis, text analytics, topic modeling, natural language processing (NLP), word embedding, sentence embedding, bag of words* , and *sentiment analysis*. We also used asterisks (*) and quotation marks ("") to eliminate the impacts of plural forms, hyphens, or spelling variations. A description of our keyword-selection process and a complete list of keywords are included in the Appendix.

Keywords were searched in the title, abstract, and keywords.[6] The exact query is as follows:

(blockchain* OR cryptocurrenc* OR stablecoin* OR "crypto token*" OR "smart contract*" OR "initial coin offering*" OR "security token offering*" OR "initial exchange offering*" OR "non*fungible token*") AND ("text* analysis" OR "text analytics" OR

[3] Stablecoins are cryptocurrencies designed to be price-stable by pegging their values to a specific asset (or a basket of assets), making them a better medium of exchange than typical cryptocurrencies. The most common peg is to the US dollar.

[4] ICO is an alternative way of financing projects or startups by creating and issuing tokens on a blockchain and selling them to raise funds. IEOs can be seen as an ICO supervised by cryptocurrency exchange platforms: the project goes through due diligence before commencing the sale, which gives investors more assurance about the validity and success of the project. STOs are tokenized digital securities and are sold in security token exchanges. They are classified as securities and are subject to rigorous vetting before issuance.

[5] NFTs differ from other tokens by its non-fungibility. A token can represent ownership of a specific item (e.g., painting, land) and is not interchangeable with other tokens because it has unique (digital) properties encoded in the smart contract that creates it.

[6] For WoS, we also searched in Keywords Plus. It is a feature of WoS that returns the articles in results if the words or phrases in our search appear frequently in the titles of these articles' references, but not in the title of the article itself. By doing this, we also collected articles that have the potential to be relevant to our topic but did not have the keywords placed in the article.

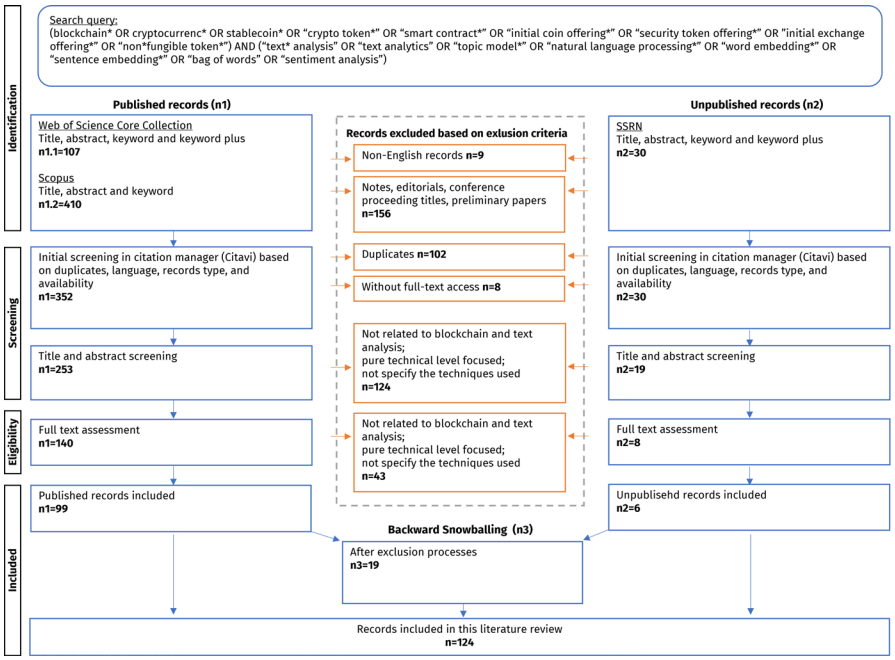

Fig. 1 The flowchart of the literature selection phases

"topic model*" OR "natural language processing*" OR "word embedding*" OR "sentence embedding*" OR "bag of words" OR "sentiment analysis")

The details of the literature search and selection process are presented in Fig. 1. Search queries in the two databases returned 517 records. First, we screened the metadata of the articles to remove articles that were (1) non-English articles, (2) notes, editorials, conference proceedings titles, and preliminary papers, (3) duplicates, and (4) without full-text access. We screened the titles and abstracts to remove articles based on our content-based exclusion criteria. To obtain relevant articles from multiple perspectives, we did not set inclusion/exclusion criteria by discipline. Alternatively, we checked the content of the articles and only excluded an article if (1) it did not contain information related to both blockchain and text analysis, (2) it focused purely on the technical aspect of blockchain, or (3) it did not specify the specific text analysis techniques used. After the above screening, 140 articles remained for full-text assessment, and we applied the exclusion criteria again and obtained 99 published articles. Our search on SSRN initially returned 30 articles. We removed 24 articles based on our exclusion criteria, leaving six unpublished articles. Subsequently, we conducted backward snowballing on 105 articles included in the keyword searches (i.e., we went through the references of the included articles) to find additional articles that did not appear in the keyword searches. This process yielded nineteen additioanl 19 papers. A total of 124 studies were included in the literature review.

Descriptive results

This section reports the descriptive results of the papers, including publication trends, keyword networks, and citation rankings.

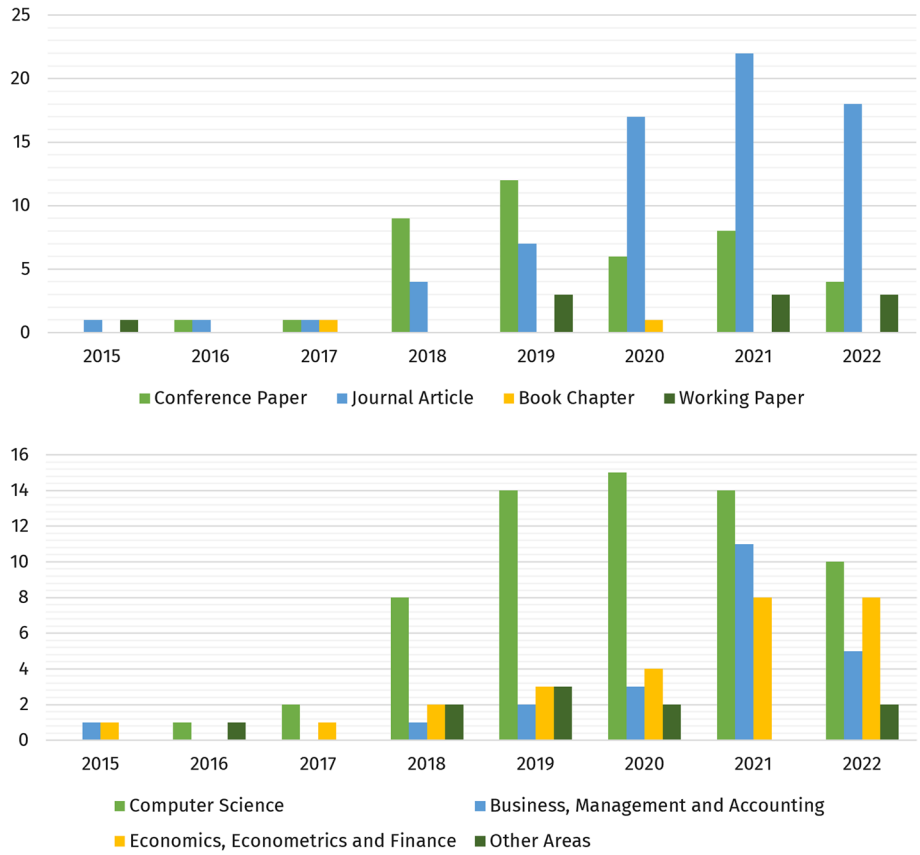

Fig. 2 The types and research areas of the publications in each year

Publication trend

Figure 2 depicts the number of papers on a yearly basis subject to article type and research area. Although we did not set any timeframe restrictions in our keyword search, the first blockchain paper using text analysis appeared in 2015, 6 years after the birth of the Bitcoin blockchain (Nakamoto 2008). The total number of papers published annually has been increasing, indicating the growing interest in and recognition of text analysis as a methodology for blockchain-related research. Until 2019, conference proceedings were the main channels through which related papers were published; however, from 2020 onward, the number of papers published in journals began to increase. For several years, computer science papers have largely dominated the topic, which can be explained by the entry requirements for coding skills in many machine learning-based text analyses. Nevertheless, later years saw a growing number of papers from business-, economics-, and finance-related fields. Studies from other areas, such as social sciences and multidisciplinary studies, have also contributed to this topic. The number of papers in most of these areas remains limited. However, the growing diversification of research areas indicates that interest has begun to spread from computer science to these areas.

We analyzed the network of papers' keywords (see Fig. 3).[7] The size of the nodes reflects the frequency, the connection between the nodes indicates the co-occurrence

[7] We cleaned the keywords of the papers before conducting the network analysis to eliminate the effects of the plural form, abbreviation, and spelling variation, etc.

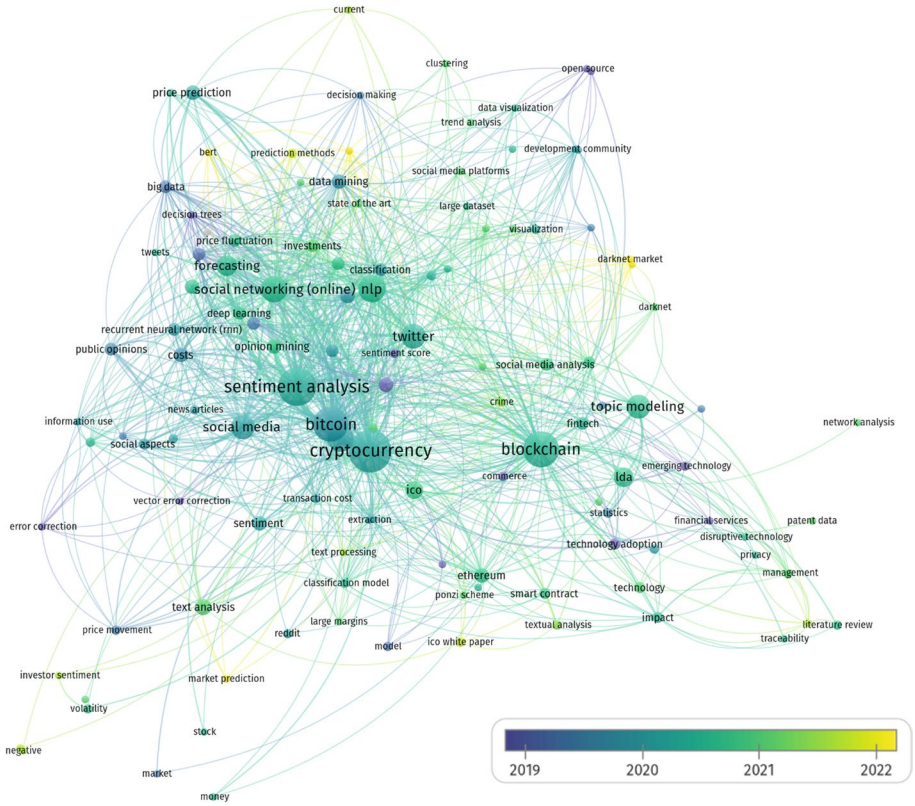

Fig. 3 The keyword frequency and co-occurrence networks

of keywords in a paper, and the color of the nodes indicates the average year in which the keyword appears. The most common keywords are the three blockchain concepts: *Bitcoin*, *cryptocurrency*, and *blockchain*. Bitcoin had the earliest average occurrence and was associated with crime (e.g., crime, DarkNet market), social media (e.g., social networking, Twitter), and sentiment (e.g., opinion mining and sentiment analysis). Cryptocurrency is associated not only with crime but also with financial activities (e.g., financial services and investments), classification, and clustering (e.g., recurrent neural networks, deep learning, and topic modeling). The keyword blockchain tends to co-occur with specific applications (e.g., commerce and FinTech), topic modeling, and relationship analysis (e.g., network and trend analyses). Different keyword associations imply that the different scopes of topics within a blockchain are related to distinct economic activities and analyses. Individual text analysis-related keywords are mentioned less frequently; however, they appear in each blockchain scope. Sentiment analysis tends to go together with Bitcoin and cryptocurrency, whereas topic modeling and the corresponding keywords connect closely to cryptocurrency and blockchain.

Citation ranking

Citation analysis helps identify the impact and common concerns of papers. However, one problem with using citations as an indicator of impact is that older papers have longer periods of citation accumulation. Thus, to offset this problem, we ranked the

Table 1 The top 10 most cited papers by total citation and citation per year

Paper	Citation	CPY	Data and period	Methodology	Summary
Polasik et al. (2015)	431	54	Nexis database: English-language news mentioning Bitcoin 04.2011– 03.2014	Lexicon-based sentiment analysis (Henry's finance-specific dictionary)	This study examines the determinants of Bitcoin price and the drivers of its success. Using sentiment in newspaper articles as one of the variables, it discovers that the negative mentions of Bitcoin lead to a price drop, while exhortatory pieces increase the Bitcoin price.
Kim et al. (2016)	268	38	Comments and relevant replies in three cryptocurrency online communities: Bitcointalk, Forum ethereum, Xrpchat 12.2013–02.2016	Lexicon-based sentiment analysis (VADER)	This study uses the contents on three cryptocurrency communities to predict the price and number of transaction fluctuations. The sentiment of posts, the number of posts/replies, and the number of views of posts are used to perform Granger causality test on each currency for a time lag of 1–13 days. The results show that positive comments affect the price fluctuations of Bitcoin, whereas Ethereum and Ripple are influenced by negative comments.
Mai et al. (2018)	208	42	Bitcointalk 01.2012–12.2014 Twitter: hastag Bitcoin 09.2014–12.2014	Lexicon-based sentiment analysis (LM lexicon)	This study investigates the impacts of social media on Bitcoin price. It separates the users into two groups, (1) the silent majority of users and (2) the vocal minority, and examines the impacts of these two groups, respectively. It finds that Bitcointalk has a more substantial impact than Twitter, and the silent minority exerts a more significant effect on future Bitcoin prices.
Georgoula et al. (2015)	170	21	Twitter: keywords and hashtags Bitcoin, BTC, and Bitcoins 10.2014–01.2015	Machine leaning-based sentiment analysis (Support Vector Machines)	This study sheds light on the factors determining the price of Bitcoin in the short- and long-run. It adds Twitter sentiment into conventional prediction model. Specifically, it constructs a Twitter sentiment measure using SVMs and finds that sentiments have a positive short-run impact on Bitcoin prices.
Abraham et al. (2018)	164	33	Twitter: hashtags Bitcoin and Ethereum 03.2018–05.2018	Lexicon-based sentiment analysis (VADER)	This study uses a linear model for predicting price changes of Bitcoin and Ethereum utilizing Twitter sentiment, tweet volume and Google Trends data. The results indicate that Twitter sentiment tends to be positive regardless of price direction and is, therefore, not a feasible predictor of price changes.
Kraaijeveld and de Smedt (2020)	132	44	Twitter: hashtags including following nine cryptocurrencies: Bitcoin, Ethereum, XRP, Bitcoin Cash, EOS, Litecoin, Cardano, Stellar and TRON 06.2018–08.2018	Lexicon-based sentiment analysis (VADER, LM lexicon, and manually complied cryptocurrency-related words)	This study tests to what extent Twitter sentiment can be used to predict price returns for nine cryptocurrencies. It measures sentiments using a self-constructed lexicon and performs bilateral Granger-causality testing to find the causality. It finds the predictive power of Twitter sentiment for several cryptocurrencies.

Table 1 (continued)

Paper	Citation	CPY	Data and period	Methodology	Summary
Grover et al. (2019)	122	31	Twitter: hashtag Blockchain 01.2018–02.2018	Lexicon-based sentiment analysis (Bing)	This study explores blockchain acceptance by examining the tweet information. It combines manual content analysis and lexicon-based sentiment analysis to distinguish the topics discussed and the user opinion. The analysis shows that users are attracted by security, privacy, transparency, trust and traceability. Furthermore, blockchain benefits are more frequently discussed than its drawback.
Karalevicius et al. (2018)	120	24	Expert media news from CoinDesk, Cointelegraph, NewsBTC 05.2013–02.2016	Lexicon-based sentiment analysis (Harvard-IV General Purpose Psychological Dictionary and LM lexicon)	This study utilizes Bitcoin-related news articles to predict semi-short-term Bitcoin price movement. Integrating the sentiments of such news shows that the market initially overreacted to the news articles, resulting in multiple corrections.
Li et al. (2019)	93	23	Twitter: keywords and hashtags ZClassic, ZCL, and BTCP 01.2019–02.2019	Lexicon-based sentiment analysis (Textblob)	This study analyzes Twitter signals as a medium for user sentiment to predict the hourly price fluctuations of ZClassic. It compiles the tweets into an hourly sentiment index, creating a weighted index giving larger weight to retweets. These two indices and the raw sentiment are used as input for Extreme Gradient Boosting Regression Tree Model for prediction.
Valencia et al. (2019)	90	23	Twitter: keywords and hashtags including following four cryptocurrencies: Bitcoin, Ethereum, XRP, Litecoin 02.2018–04.2018	Lexicon-based sentiment analysis (VADER)	This study uses sentiments on Twitter as input features for multiple machine learning algorithms to predict the price movement of four cryptocurrencies. It shows that Twitter data alone can be used to predict certain cryptocurrencies.
Kim et al. (2020)	79	26	Academic papers: keyword or abstract contain "Blockchain", "Block chain", and "Block-chain" in six databases: Scopus, ScienceDirect, Web of Science, IEEE Xplore, Google Scholar, and Korean Citation Index. 01.2014 - 08.2018	Topic modeling (W2V-LSA)	This study proposes an improved method for topic modeling (W2V-LSA) and performs an annual trend analysis of blockchain-related literature. The experimental results confirmed the usefulness of W2V-LSA in terms of the accuracy and diversity of topics by quantitative and qualitative evaluation, and it can be an option for researchers using topic modeling for technology trend analysis.

papers in terms of both total citations and citations per year (CPY) (Dumay and Cai 2014) and considered the top ten papers from both criteria. Table 1 lists these papers and summarizes their text data, sample period, text analysis techniques, and brief abstracts of the papers.

Nine papers appeared on both lists; one older paper (Georgoula et al. 2015) fell short of CPY and was surpassed by a newer paper (Kim et al. 2020). The topics of high-impact papers tended to concentrate on a narrow range. Ten studies applied sentiment analysis and nine explored the predictive power of sentiment from social media platforms/news for cryptocurrency prices. Most studies focused on Bitcoin or a few altcoins with large market caps, while Kraaijeveld and de Smedt (2020) included nine cryptocurrencies, and Li et al. (2019) studied a smaller cryptocurrency called ZClassic (ZCL). One study examined the sentiments of blockchain-related tweets and found that blockchain benefits were discussed more than its drawbacks (Grover et al. 2019). The study by Kim et al. (2020) proposed a new topic modeling method and applied it to conduct a literature review on blockchain research to discover research trends. A detailed discussion is provided in Table 1.

Discussion of research questions

RQ1 Which research scope, text data, and methodology are used to conduct text analysis in the blockchain area?

In this section, we briefly introduce the scope, text data, and methodologies used in the papers and bridge the elements to identify the most used combinations. Figure 4 displays the connections among research scopes, text data, and methodologies in proportion to the number of papers.[8]

Research scope

'Specific cryptocurrency' (72 papers, 58%) is the most frequently used scope and Bitcoin in particular is the most studied cryptocurrency. To better recognize the importance of Bitcoin, we separate studies that focus exclusively on Bitcoin (40 papers, 32%) from the others. Other studies examine cryptocurrencies with large market caps, special small cryptocurrencies (Li et al. 2019; Mnif et al. 2021; Vacca et al. 2021), or a large number of cryptocurrencies to represent the market (Steinert and Herff 2018; Schwenkler and Zheng 2021).

Another substantial scope is the general concept of blockchain (26 studies, 21%). These studies treat blockchain technology and its applications as a whole and discover its uses in particular fields (e.g., supply chain management (Medhi 2020; Hirata et al. 2021; Xu and He 2022), banking (Daluwathumullagamage and Sims 2020), and accounting (Garanina et al. 2021)) and how blockchain-related topics evolve (over time) (Zhang et al. 2021a; Chousein et al. 2020; Medhi 2020; da Silva and Moro 2021; Zeng et al. 2018; Shahid and Jungpil 2020; Perdana et al. 2021).

[8] Some of the papers use various types of text data and methodologies; therefore, the sums of text data and methodology exceed the number of papers.

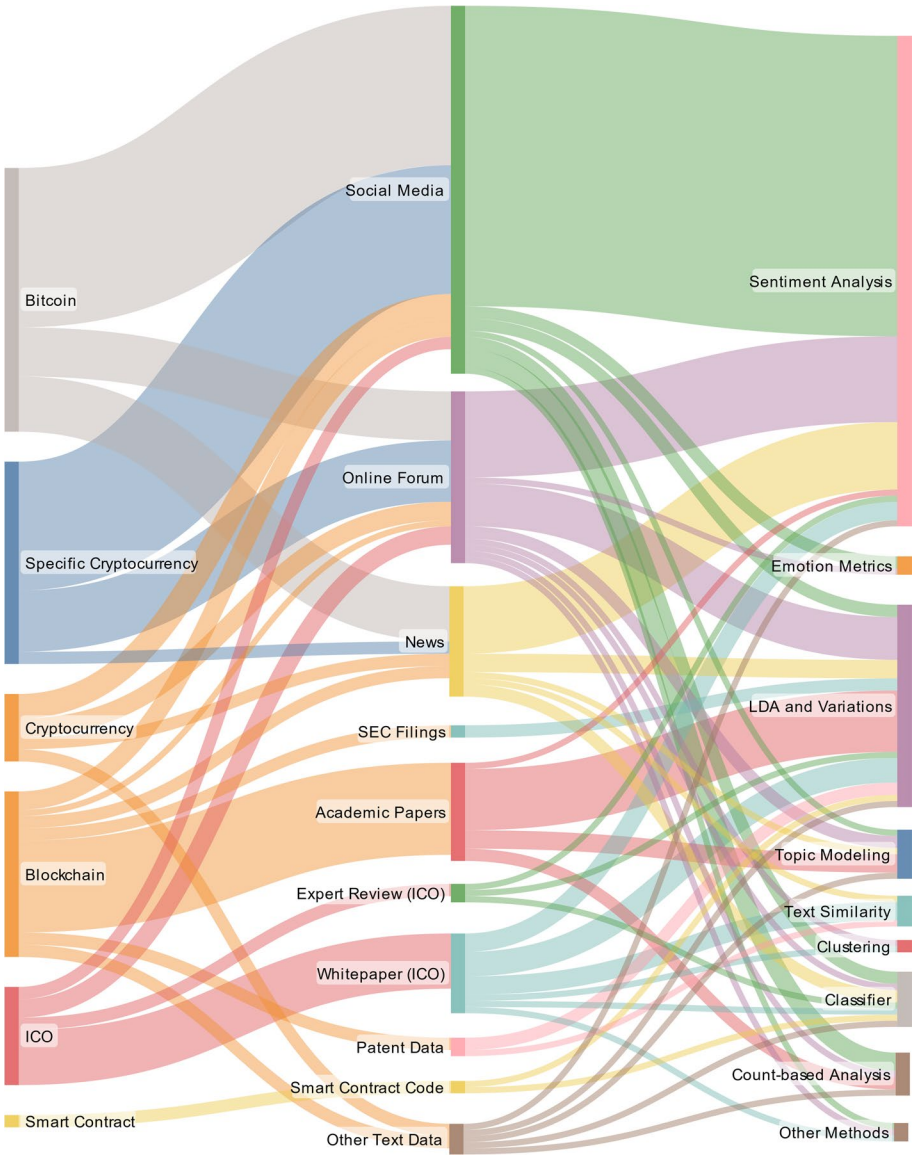

Fig. 4 The connections among research scope, text data, and the methodology

The literature also covers the scope of the cryptocurrency market as a whole (11 papers, 8.9%) (Caliskan 2020; Siu et al. 2021), ICO projects (13 papers, 10.5%) (Toma and Cerchiello 2020; Liu et al. 2021; Sapkota and Grobys 2021), and smart contract (two papers, 1.6%) (Ibba et al. 2021; Zhang et al. 2021a).

It is worth noting that, in our search, the keywords also included stablecoin, NFT, and STO, but we found no papers that used text analysis to examine these scopes. This may have resulted from the late development of these blockchain use cases. However, increasing growth in such applications has been observed in recent years (Lambert et al. 2021; Wang et al. 2021b), thus creating opportunities and the needs to address relevant research questions using text analysis.

Text data

Table 2 summarizes the text data and corresponding data sources we identify from the papers, which helps researchers navigate to the sources of their target data. We categorize texts into four groups: (1) corporate-produced documents, (2) user-generated content, (3) news, and (4) academic papers.

Corporate-produced document Corporate-produced documents utilize formal and technical languages to provide detailed information about the company or specific products and services. Despite the precise information provided by these documents, we found only 18 studies that used such texts. ICO whitepaper, which pitches the project idea and outlines the business plan, is a voluntary disclosure by the ICO project team

Table 2 The detailed information of text data sources used in the literature

	Text information	Data source	Data type *	Example paper
Unique text data in blockchain	Smart contract code	Solidity	1	Ibba et al. (2021), Zhang et al. (2021a)
	Whitepaper	https://ICOHolder.com https://ICOMarks.com https://ICORatings.com https://ICODrops.com https://FoundICO.com https://CryptoCompare.com	1	Sapkota and Grobys (2021), Thewissen et al. (2022)
	Whitepaper and Expert report	https://ICOBench.com	1, 2	Xu et al. (2021)
	Cryptocurrency community: technical and economic topics related to Bitcoin or other altcoins	https://bitcointalk.org https://www.xrpchat.com https://forum.ethereum.org	2	Kim et al. (2016), Gurdgiev and O'Loughlin (2020)
	Bitcoin abuse report	https://Bitcoinabuse.org	2	Choi et al. (2022)
	Cryptocurrency news	https://www.coindesk.com https://www.newsbtc.com https://www.fxstreet.com/cryptocurrencies/news https://cointelegraph.com https://www.cryptocompare.com https://cryptocoin.news	3	Karalevicius et al. (2018), Farimani et al. (2022)
General Data Sources contain blockchain topics (through keyword search)	Firm disclosures including 10-Ks, conference calls, etc.	SEC Filings Full list: https://www.sec.gov/forms	1	Yen and Wang (2021), Stratopoulos et al. (2022)
	Patent filing	Patent database: USPTO, EPO	1	Wang et al. (2021a), Zhang et al. (2021a)
	The required skills for job applicant	Online recruitment website	1	Ge et al. (2021)
	Terms of Services Agreement	Company website	1	Caliskan (2020)
	Social media platform: contain blockchain-related messages	Twitter, Sina Weibo, Stocktwits	2	Chen et al. (2019a), Pan et al. (2020), Huang et al. (2021)

Table 2 (continued)

	Text information	Data source	Data type *	Example paper
General Data Sources contain blockchain topics (through keyword search)	Online forums or groups: contain blockchain-related posts	Reddit, Github, Telegram, StackExchange, Discord	2	Alahi et al. (2019), Hinds-Charles et al. (2019), Nizzoli et al. (2020)
	Online forum: contains criminal/illicit topics	HackForums	2	Siu et al. (2021)
	Users' review about a specific product/service	App store	2	Voskobojnikov et al. (2021)
	Web data (news, social networking websites, forums, etc,.)	Web data monitoring: Webz.io, Notified, OpView Social Listening Platform	2, 3	Lu et al. (2017), Inamdar et al. (2019), Grassman et al. (2021)
	News articles	Newspaper channels: The Financial Times, The Economist, The Economic Times, Business Insider, The Wall Street Journal	3	Azqueta-Gavaldón (2020)
	News articles from multiple channels	News Terminals: Nexis, Refinitiv Eikon, NewsAPI, RavenPack	3	Polasik et al. (2015), Rognone et al. (2020), Anamika and Subramaniam (2022)
	Academic Papers/industry articles	WoS, Scopus, Google Scholar, Science Direct, IEEE Xplore Digital Library, ACM Digital Library, JSTOR, SSRN, Business Source Premier	4	Shahid and Jungpil (2020), da Silva and Moro (2021)

* (1) corporate-produced documents; (2) user-generated content; (3) news; and (4) academic papers

to attract potential investors (Florysiak and Schandlbauer 2022; Thewissen et al. 2022). Another example of such document is smart contract code. Although the code does not strictly belong to human language, its fixed format enables researchers to obtain information regarding the subject of the contract (Ibba et al. 2021; Zhang et al. 2021a). Blockchain-related texts can also be extracted from corporate documents, such as SEC and patent filings, through keyword searches and used to examine blockchain adoption (Yen and Wang 2021; Wang et al. 2021a; Zhang et al. 2021a; Stratopoulos et al. 2022).

User-generated content Among all text data, user-generated content was the most frequently used (85 times, 64%). This type of text features a shorter length and informal language, and generally expresses the opinions of users on a particular topic. Social media platforms offer rich resources for such texts (56 times, 42%). Specifically, most studies chose Twitter to extract text data for conducting the analyses (Patil et al. 2018; Huynh 2021; Mareddy and Gupta 2022), while others used Sina Weibo (a Chinese microblogging website) or Stocktwits (a social media platform focused on financial topics) (Chen et al. 2019a; Pan et al. 2020; Huang et al. 2021).

Compared with social media platforms, online forums often have a specific focus and attract users with shared interests; therefore, they tend to offer deeper discussions. Cryptocurrency-specific forums, such as bitcointalk, XRPChat, and Ethereum Community Forum (Kim et al. 2016; Gurdgiev and O'Loughlin 2020), have sections with distinctive topics. User discussions on topic-focused forums, such as GitHub, Reddit, and StackExchange have provided insights into the development of blockchain (Hinds-Charles et al. 2019; Bahamazava and Reznik 2022; Ortu et al. 2022). There are numerous cummnities (i.e., subreddits) within the crypocurreny framework of Reddit (e.g., r/CryptoMarkets, r/Bitcoin), and users can join the communities to share up-to-date news or express their opinions on topics. In contrast, HackForums contains posts on illicit activities (Siu et al. 2021).

News News articles are one of the most widespread and accessible types of textual data. They provide up-to-date factual information on events, and commentaries/opinions on a topic. Analyzing blockchain news on a scale allows researchers to identify the evolution and public sentiment toward the technology. For instance, multiple news channels report the upcoming Ethereum Shanghai Hard Fork, but they contain different sentiments toward the event: FXStreet (2023) neutrally introduces the updates it would bring; U.Today (2023) illustrates multiple reasons for developers to be concerned about the hard fork, while Bloomberg (2023) is comparatively optimistic about it by emphasizing that "Shanghai is expected to push more people and institutional investors to stake their coins to support the Ethereum network and earn yield."

Many studies use cryptocurrency-specific news channels (e.g., Coindesk and Cointelegraph) as their primary news data sources (Karalevicius et al. 2018; Farimani et al. 2022), whereas others search for blockchain-related news from financial newspapers (e.g., The Financial Times and The Economist) through keyword searches (Azqueta-Gavaldón 2020).

Academic paper Literature reviews assist researchers in understanding the current status of research, identifying research gaps, and guiding future research (Chakkarwar and Tamane 2019; Shahid and Jungpil 2020; Garanina et al. 2021). Unlike the standard literature, in which researchers spend time manually examining papers, the automated processing of text-analysis-assisted literature reviews enables researchers to acquire insights into a large number of papers in a specific area in a short time.

Methodology

Choosing a suitable methodology depends not only on the data characteristics but also on the research questions of the study. Our goal is not to provide a systematic classification of the methodologies, but to provide a big picture of the methodologies used in blockchain-related literature. Therefore, the methodologies presented in this section may overlap. For example, the underlying methodology of sentiment analysis can be a machine-learning-based classifier. This section outlines the principal methodologies most directly related to the research questions. In addition, we summarize the specific text analysis techniques used in the papers in Table 3 to provide supplementary details.[9]

[9] The mathematical principles of the methodologies are beyond the scope of this review, but for each methodology, interested readers can refer to the cited studies for details.

Table 3 The detailed information of text analysis techniques used in the literature

Analysis type	Sub-category	Specific technique	References	Example papers
Feature extraction	Count-based	BoW	Zhang et al. (2010)	Yen et al. (2021)
		N-Gram	Cavnar et al. (1994)	El-Masri and Hussain (2021)
		TF-IDF	Ramos (2003)	Pan et al. (2020)
		DDPWI	Proposed in the paper	Burnie and Yilmaz (2019)
	Word/Sentence embedding	Word2vec	Mikolov et al. (2013)	Kilimci (2020); Kim et al. (2020); Liu et al. (2021)
		Doc2vec	Le and Mikolov (2014)	
		GloVe	Pennington et al. (2014)	
		FastText	Bojanowski et al. (2017)	
		Affective Tweet	https://affectivetweets.cms.waikato.ac.nz	Balfagih and Keselj (2019)
		A-BiRNN	Proposed in the paper	Xu et al. (2021)
Sentiment analysis	Lexicon/rule-based	VADER	Hutto and Gilbert (2014)	Kim et al. (2016); Abraham et al. (2018)
		TextBlob	https://textblob.readthedocs.io	Jain et al. (2018); Li et al. (2019)
		Sentistrength	http://sentistrength.wlv.ac.uk	Caviggioli et al. (2020)
		SentiWordNet	Baccianella et al. (2010)	Cheuque Cerda and L. Reutter (2019)
		Alex Davies word list	Christie and Huang (1995)	Stratopoulos et al. (2022)
		Bing	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html	Grover et al. (2019); Hassan et al. (2021)
		AFINN	Nielsen (2011)	Ayvaz and Shiha (2018); Toma and Cerchiello (2020)
		LM lexicon	Loughran and McDonald (2011)	Mai et al. (2018); Dittmar and Wu (2019)
		Harvard-IV General Purpose Psychological Dictionary	Stone et al. (1966)	Karalevicius et al. (2018)
		Quantitative Discourse Analysis Package	https://www.rdocumentation.org/packages/qdap/versions/2.4.3	Sapkota and Grobys (2021)

Table 3 (continued)

Analysis type	Sub-category	Specific technique	References	Example papers
Sentiment analysis	Lexicon/rule-based	Henry's finance-specific dictionary	Henry (2008)	Mnif et al. (2021); Anamika and Subramaniam (2022)
		Pattern library	https://github.com/clips/pattern	Galeshchuk et al. (2018)
		SentimentR	https://github.com/trinker/sentimentr	Rahman et al. (2018); Chiarello et al. (2021)
		Ethical and unethical words dictionary	Constructed in the paper	Barth et al. (2020)
		63 cryptocurrency words and abbreviations	Constructed in the paper	Kraaijeveld and de Smedt (2020)
		Crypto-specific sentiment dictionary (in Chinese)	Constructed in the paper	Huang et al. (2021)
		Crypto-specific lexicon (words, emojis, informal langugage)	Constructed in the paper	Chen et al. (2019a)
	Machine leanring-based (algorithms)	Long short-term memory (LSTM)	Hochreiter and Schmidhuber (1997)	Inamdar et al. (2019); Şaşmaz and Tek (2021)
		Recurrent neural network	Goldberg (2017)	
		Random forest	Ho (1995)	
		Naïve Bayes	Jurafsky and Martin (2017)	
		Support vector machine	Boser et al. (1992)	
		Gradient boosting	Friedman (2001)	
		BERT	Devlin et al. (2018)	Bashchenko (2022); Ortu et al. (2022)
		Bidirectional LSTM	Mousa and Schuller (2017)	Han et al. (2020)
		Voting-included Algorithm	Constructed in the paper	Pant et al. (2018)
		Sentiment Graph	Constructed in the paper	Yao et al. (2019)
	Analytics Tool	Crimson Hexagon social sentiment	https://www.carahsoft.com/crimson-hexagon	Stanley (2019)
		Semantria	https://www.lexalytics.com	Caviggioli et al. (2020)
		Meaningcloud	https://www.meaningcloud.com	

Table 3 (continued)

Analysis type	Sub-category	Specific technique	References	Example papers
Sentiment analysis	Analytics Tool	StanfordCoreNLP	https://stanfordnlp.github.io/CoreNLP	Moustafa et al. (2022)
		OPView	https://www.opview.com.tw	Lu et al. (2017)
		RavenPack	https://www.ravenpack.com/products/edge/data/news-analytics	Rognone et al. (2020)
Emotion metrics		NRC-VAD Emotion Lexicon	https://saifmohammad.com/WebPages/nrc-vad.html	Toma and Cerchiello (2020)
		NRC Word-Emotion Association Lexicon	https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm	Chursook et al. (2022)
		Text2Emotion	https://shivamsharma26.github.io/text2emotion	Aslam et al. (2022)
Topic modeling	Topic modeling algorithm	LDA	Blei et al. (2003)	Fu et al. (2019); Hirata et al. (2021); Laturnus (2022)
		DTM	Blei and Lafferty (2006)	Linton et al. (2017); Lee et al. (2022)
		SentLDA	Bao and Datta (2014)	Thewissen et al. (2022)
		Joint/sentiment topic model	Lin and He (2009)	Loginova et al. (2021)
		Topic sentiment latent dirichlet allocation	Nguyen and Shirai (2015)	
		Nonnegative Matrix Factorization	Lee and Seung (1999, 2000)	Kang et al. (2020)
		Anchored Correlation Explanation	Gallagher et al. (2017)	Nizzoli et al. (2020)
		Word2vec-based Latent Semantic Analysis (W2V-LSA)	Proposed in the paper	Kim et al. (2020)
	Analytics tool	Leximancer	https://www.leximancer.com	Daluwathumullagamage and Sims (2020); Perdana et al. (2021)
Text similarity		Cosine Similarity	Kwon and Lee (2003)	Yen et al. (2021)
		Jaccard Similarity Coefficient	Jaccard (1912)	Sapkota and Grobys (2021)
		SBERT	Reimers and Gurevych (2020)	Bashchenko (2022)
Clustering		K-means clustering	MacQueen (1967)	Choi et al. (2022)
		DBSCAN clustering	Ester et al. (1996)	

Table 3 (continued)

Analysis type	Sub-category	Specific technique	References	Example papers
Classifier	Machine learning algorithm	Catboost	Prokhorenkova et al. (2018)	Chousein et al. (2020); Schwenkler and Zheng (2021)
		Random Forest	Ho (1995)	
		XGBoost	Chen and Guestrin (2016)	
		Neural network	Hashimoto et al. (2016)	
		Naïve Bayes	Jurafsky and Martin (2017)	
Readability		Flesch-Kincaid Readability	Flesch (1979)	Narman et al. (2018); Sapkota and Grobys (2021)
		Dale-Chall Readability	Dale and Chall (1948)	
		Gunning Fog Index	Gunning (1952)	
		Automated Readability Index	Senter and Smith (1967)	
		Simple Measure of Gobbledygook	McLaughlin (1969)	
		Coleman-Liau Index	Coleman and Liau (1975)	
		Linsear Write	Klare (1974)	
		AWS blockchain template	https://docs.aws.amazon.com/blockchain-templates	Stanley (2019)
Network analysis		Google knowledge graph	https://developers.google.com/knowledge-graph	Pan et al. (2020)

Text preprocessing Before conducting the actual analysis, multiple cleaning procedures should be applied to the raw text to prepare it as the input material. The necessary steps vary depending on the text condition and planned analysis. However, we identified standard preprocessing steps suitable for the majority of texts: removing special characters and punctuation, removing numbers and stopwords, lower-casing, spelling corrections, tokenization, assigning part-of-speech tags, and stemmization/lemmatization. Some raw texts require more cleaning than others. For example, texts from social media and online forums usually use informal language and emojis which can lead to misinterpretation. Papers therefore conducted additional procedures (Birim and Sönmez 2022; Critien et al. 2022): remove # and @user, remove URL links, convert emojis to words, and convert vocabulary abbreviations to words. These procedures remove redundant text, convert unrecognizable characters into valuable information, and are vital preparation steps.

Feature extraction The cleaned texts should be transferred to number representations to allow the computer to read and use for further analyses. It can also reduce computational complexity, enhance performance, and avoid the overfitting problem, making it an essential procedure in text analysis (Kou et al. 2020). This representation per se can also provides information and insight. Count-based methods are straightforward

to understand and interpret. The Bag-of-words (BoW) is one of the most widely used approaches. It represents words according to their frequency in the corpus, disregarding order and context. N-grams are extensions of BoW that breaks the corpus into a contiguous sequence of n words. It can capture more context around each word, but produces a sparser feature set than BoW. BoW and N-grams assume that words that occur more frequently are more relevant and do not always hold true. Term frequency-inverse document frequency (TF-IDF) (Salton et al. 1975) adds another metric of how rarely a word occurs across the entire corpus and assigns rarer words a higher score. Although such representations are generally used as inputs for further analysis, we identify papers that highlight frequent words and interpret them as blockchain topics (Zeng et al. 2018; Burnie and Yilmaz 2019; El-Masri and Hussain 2021). However, this method can be misleading, because count-based methods discard linguistic structures and may miss crucial text information.

Word-embedding mitigates this problem by representing words in vectors to capture their semantic and syntactic contexts in a document (Cong et al. 2021). In the vector space, the shorter the distance between two word vectors, the higher is the similarity of the words. *Word2vec* (Mikolov et al. 2013) is one of the most frequently used word embedding methods. It includes two configurations: skip-gram and continuous bag of words (CBOW). A skip-gram uses the current word to predict the surrounding words, whereas CBOW predicts the current word using its surrounding words. A generalization of *word2vec* and *doc2vec* (Le and Mikolov 2014) adds a document feature vector to the word vector to capture the semantics of the paragraphs and documents. Word-embedding techniques are not frequently used in the literature, but we found that Kim et al. (2020) and Liu et al. (2021) integrated these techniques when processing their texts. Two other word-embedding models, *GloVe* and *fastText*, were used by Kilimci (2020).

Analysis Sentiment analysis is the dominant text-analysis approach in the literature (80 times, 53%). There are two major types of sentiment analysis: lexicon/rule-based and machine learning-based (Vohra and Teraiya 2013).

Lexicon-based sentiment analysis calculates the sentiment score of a text based on the polarity of each word (i.e., positive, negative, or neutral) from sentiment dictionaries in which each vocabulary is assigned a sentiment score. Examples of well-established sentiment dictionaries include Valence Aware Dictionary for Sentiment Reasoning (VADER) (Hutto and Gilbert 2014), which is particularly suitable for social media contexts, and Loughran and McDonald sentiment lexicon (LM lexicon) (Loughran and McDonald 2011) in the finance domain. However, off-the-shelf dictionaries can sometimes generate inaccurate results because of different sentiments of the same vocabulary in different contexts (Loughran and McDonald 2011). Therefore, some researchers have developed new and additional dictionaries (e.g., new vocabularies and emojis) in blockchain contexts for higher accuracy of sentiment quantification (Chen et al. 2019a; Barth et al. 2020; Kraaijeveld and de Smedt 2020).

Machine learning-based sentiment analysis adopts machine learning classifiers to study the sentiments of texts and classify them into instinctive sentiment groups. Researchers can build a model and train their data or apply a pre-trained model (e.g., Bidirectional Encoder Representations from Transformers (BERT)) to their analysis. Compared to lexicon/rule-based sentiment analysis, it is dynamic and can better fit the research context. We identified 12 papers that adopted this approach (e.g., Patil et al. 2018; Balfagih and Keselj 2019; Inamdar et al. 2019; Aslam et al. 2022). In particular, Han et al. (2020) and Akba et al. (2021) propose and assess new models for sentiment analysis.

Sentiment analysis tools have also been utilized in academic studies (Lu et al. 2017; Stanley 2019; Caviggioli et al. 2020; Moustafa et al. 2022). Such tools develop unique algorithms and reduce the programming requirements for researchers. However, most of these tools are commercially oriented, incur high subscription fees, and lack transparency regarding their algorithms. Hence, albeit the convenience, researchers should be cautious when using such tools.

In some studies, emotion-detection metrics have been applied in conjunction with sentiment analysis to achieve more precise emotion separation. For example, the NRC-VAD Emotion lexicon has three dimensions: valence, arousal, and dominance (Mohammad 2018). This provides another layer for sentiment and can increase the quality of the analysis.

The Latent Dirichlet Allocation (LDA) and its variations were frequently chosen (33 times, 22%) for text analysis. LDA is a topic-modeling algorithm developed by Blei et al. (2003). Topic modeling can identify the patterns of vocabulary and phrases in documents (within the corpus of interest), detect the differences in their topics, and cluster the documents according to the topics discussed in the documents. LDA is one of the most popular topic-modeling algorithms. It assumes that each document in the corpus consists of a number of latent topics and that each topic is characterized by a word distribution. Each topic is presented with a list of words and their fitting possibilities. Its variations include dynamic topic models (DTM), which add temporal features to the model (Blei and Lafferty 2006) and SentLDA, which considers the boundaries between sentences and assumes that all words in a sentence are sampled from the same topic (Bao and Datta 2014). The texts used in LDA models are typically unlabeled, and the researchers' task is to choose the optimal number of topics, which is primarily determined by the perplexity and coherence scores (Blei et al. 2003; Newman et al. 2010). After narrowing down the choices for the optimal number of topics, researchers become involved and integrate their interpretations to choose the optimal number of topics for the model. Together with other topic modeling and clustering algorithms, they belong to unsupervised machine learning. Evaluations of unsupervised machine learning vary from model to model, and human judgment is often required to evaluate the model quality. Nevertheless, these models are valuable for exploring the underlying features of a text without establishing an upfront framework (Grimmer and Stewart 2013). This is especially applicable to research in blockchain, which is still understudied and has few established classifications.

In contrast, supervised machine-learning classifiers are applied to pre-labeled texts, and the texts are classified into pre-specified groups. The idea is to first manually categorize a set of documents and then train a supervised model that automatically learns how to assign categories to documents using a training set (Bao and Datta 2014). Owing to the training process, they are domain-specific and better fit the research context (Grimmer and Stewart 2013). Multiple models are often applied to the same dataset and researchers can easily compare the performance of classifiers using certain metrics (e.g., precision, recall, accuracy, F1-score) to select the best-fitting model. Nevertheless, in blockchain-related research, they are utilized much less for text data (nine times, 6%).

Bridging the elements

Figure 4 shows that the combinations of the elements are diversified depending on the purpose of the studies. Nevertheless, we observe two primarily adopted paths for text analysis in blockchain research: (a) papers studying specific cryptocurrencies tend to apply sentiment analysis to instant user-generated content or news articles to discover the correlations between public opinions/emotions and cryptocurrency market behavior, and (b) papers studying the broad concept of blockchain primarily choose official documents from companies (e.g., SEC and patent filings) and apply topic models to explore the classifications or trends in the sector.

The links among the above elements are not permanent; that is, researchers can choose combinations according to their requirements. To select effective combinations, researchers must understand the characteristics of the data, presumptions to use a particular methodology, and the questions they intend to investigate. The design should facilitate the generation of interpretable and meaningful results to answer the research questions.

RQ2 What topics are addressed using text analysis in current literature?

The data and methodologies are used to serve the purpose of the study and should be chosen depending on the research questions (Grimmer and Stewart 2013). In the following section, we summarize blockchain-related topics discussed in the existing literature that involve text analyses.

Relationship discovery

Researchers have used different text data (often combined with other variables) to identify correlations. The speculative nature and high volatility of cryptocurrencies have led to studies exploring the relationship between market fluctuations and information on online platforms. Different factors of online discussions, including the counts of specific keywords, discussions of different topics, and sentiment classes, are extracted. These factors are used as variables to test whether they are associated with cryptocurrency market activities, such as price changes and the co-movement of peer cryptocurrencies (Polasik et al. 2015; Phillips and Gorse 2018; Barth et al. 2020; Schwenkler and Zheng 2021). From more specific perspectives, studies distinguish different user groups and

vocabularies and find that content from certain groups or the presence of certain words is more closely related to changes in the cryptocurrency market (Burnie and Yilmaz 2019; Kang et al. 2020). Xie (2021) explores the relationships among online discussions and demonstrates that online communities' conflicting opinions and redundant discussions result in low trading volumes.

An ICO whitepaper, perceived as a prospectus for an initial public offering (IPO) in a less regulated way, provides information that can impact investors' decisions and, to some extent, determine the success of projects. Many dimensions of such texts influence the performance of ICO. For instance, ICO projects with higher technological sophistication shown in whitepapers are more likely to be successful and less likely to be delisted (Liu et al. 2021). Those whitepapers that are unique-that is, have more project-specific information and avoid borrowing common phrases from previous whitepapers-can lead to higher fundraising amounts and better post-ICO performance (Yen and Wang 2021; Florysiak and Schandlbauer 2022). The readability and sentiment expressed in whitepapers can also affect investors' decisions to invest in the described project (Stanley 2019; Sapkota and Grobys 2021).

For public companies that meet higher disclosure standards, blockchain-related information can be extracted from 10-K filings and used to investigate whether blockchain adoption brings value and efficiency to companies (Yen et al. 2021).

Cryptocurrency performance prediction

Forecasting has always been an important topic in cryptocurrency studies. In addition to econometric methods and statistical models for price prediction, sentiment has also been used as a predictor of market movement (Mao et al. 2011; Fang et al. 2022). The effect of sentiment on the cryptocurrency market could be magnified by the lack of traditional financial fundamentals in valuation, and vocal and active investors on social media (Corbet et al. 2018; Gurdgiev and O'Loughlin 2020). Machine learning models, especially supervised models, are often applied to use sentiment data for prediction. Sentiment is used as the sole input to a model or as a supplement to conventional variables (e.g., price, trading volume, blockchain metadata (Sebastião and Godinho 2021)).

Texts from social media are extracted, and each document is assigned a sentiment score using a sentiment analysis technique (see Table 3 for details). The scores (along with other variables) are subsequently used as inputs for the prediction models. They have predictive power for the direction of price movement (Loginova et al. 2021; Critien et al. 2022) and the short-term (e.g., hourly and daily) magnitude of price changes (Li et al. 2019; Farimani et al. 2022; Ortu et al. 2022).

The impact of social media content depends particularly on the level of information dissemination. Thus, celebrity or opinion leader posts (i.e., influencers) or discussions about them could have more power than other posts (Kang et al. 2020). Huynh (2021; 2022) quantifies the tweet sentiments of Donald Trump and Elon Musk using LM lexicon and finds that negativity in Trump's tweets leads to higher returns on Bitcoin, whereas both pessimistic and optimistic expressions from Musk have a positive effect on Bitcoin returns. Cary (2021) analyzes the tweet sentiment about Elon Musk's

performance on Saturday Night Live on 8 May 2021 and found that the negative opinion toward his performance led to the price decline of Dogecoin.

Prediction models have also been used in ICO studies. Text data variables (e.g., expert reviews and social media sentiment) and non-text variables (e.g., sale price, project duration, and expert ratings) are utilized simultaneously to predict the success of ICO projects (Xu et al. 2021; Chursook et al. 2022).

Overall, studies focusing on predicting market movements and project success constitute a large proportion of the papers in this review. However, the data and methodologies mainly follow a similar direction: applying sentiment analysis to Twitter posts and associating the respective sentiment metrics with high market capitalization cryptocurrencies.

Classification and trend

One step in understanding large-scale texts containing multiple documents is to categorize the documents and create classifications. Using clustering/topic models or classifiers, content features (i.e., the topics discussed) in documents can be extracted and used to group documents into different classifications. By adding a temporal dimension to the static classification, the classification information can provide the trends of a particular group of topics.

Such models can be valuable when applied to academic papers in literature reviews to facilitate an understanding of existing studies and identify further research. Unlike standard literature reviews, in which researchers read through papers to derive results, topic modeling-based literature reviews extract the titles and abstracts of papers and rely on algorithms to extract topics from the texts. Classification algorithms are used to understand the current state and development of blockchain research (Chakkarwar and Tamane 2019; Shahid and Jungpil 2020; Lee et al. 2022). Some studies have dived into blockchain applications within a sector (e.g., consumer trust, banking, and accounting) to facilitate researchers and practitioners in identifying future research areas and business opportunities (da Silva and Moro 2021; Daluwathumullagamage and Sims 2021; Garanina et al. 2021). Although it enables researchers to examine text content on a large scale without time-consuming manual reading, one of the drawbacks of using text analysis for literature reviews is the lack of an information screening process, during which irrelevant papers are excluded from the review.

Most papers included in this review (Xu and He (2022) is an exception) directly use all papers from the keyword search results as their input for topic models and further analyses. In this case, many irrelevant papers may be erroneously included in the models and the noise information they contain can be significant, leading to biased or inaccurate conclusions. To avoid undermining the advantages of topic modeling, researchers must carefully design the selection criteria for their dataset when performing such studies.

At a more technical level, the classification and trends of blockchain infrastructure and application design problems have also been addressed. Using texts from technique-oriented platforms (e.g., GitHub and StackExchange), some studies have observed a shift in developers' interests from mining to software development (Alahi et al. 2019;

Hinds-Charles et al. 2019). A special case involves the use of a smart contract code as an input for topic models or classifiers. Researchers can then discover the most common uses of smart contracts and identify Ponzi schemes by analyzing the code (Ibba et al. 2021; Zhang et al. 2021b). Despite the focus on technical information, such studies have implications not only for developers and computer scientists but also benefit researchers in finance and economics by, for instance, identifying investor interests and customer demands.

The evolution of the blockchain topic is often tied to unique events that affect market activity and trigger changes in investor behavior. Linton et al. (2017), for example, study how blockchain topics change during periods of significant events in the cryptocurrency world, such as the insolvency of the MtGox Bitcoin exchange in 2014 (Goldstein and Tabuchi 2014) and the hack into Bitfinex in 2016 (Baldwin 2016) (e.g., from sole 'Bitcoin trading' topics to 'security issues' or 'scams' as predominant topics in online forums). Other researchers (Daluwathumullagamage and Sims 2020; Pan et al. 2020; Bahamazava and Nanda 2022) incorporate the influence of specific events (e.g., Bitcoin halving events, the introduction of regulations, and COVID-19) into their models to better interpret the change in interest during different periods.

Crime and regulation

Illegal activities and crimes have always surrounded discussions on cryptocurrency. Many early users appraised the (pseudo)anonymity of cryptocurrency and used it as currency for illicit purchases on DarkNet. In the early stages, cryptocurrencies were suggested that cryptocurrencies contribute to improving black markets (Foley et al. 2019).

Bahamazava and Reznik (2022) and Bahamazava and Nanda (2022) explore the posts from Reddit (subreddit DarkNet) to study the criminal topic evolution and the mainstream methods to trade cryptocurrencies illegally. Crime-related texts on other channels such as Twitter, Telegram, and HackForums are also used to identify the specific illegal activities discussed (Barth et al. 2020; Nizzoli et al. 2020; Siu et al. 2021). One rich first-hand source for examining fraud from the victim's side is the reports from https://www.bitcoinabuse.com, where the victims of Bitcoin fraud share their experiences and post the original messages they received from the abusers. Choi et al. (2022) cluster these messages and find high similarity of a large number of messages, suggesting the existence of only slight modification of fraud messages and certain patterns of the language usages from Bitcoin fraud instigators. Zhang et al. (2021b) apply an improved CatBoost classifier to smart contract codes to find the common characteristics of Ponzi schemes hidden in the lines.

Although studies inspecting illegal activities have accumulated, the number of studies exploring relevant regulations remains minimal. We identified only two studies that explicitly discussed regulatory issues. In the study by Bahamazava and Nanda (2022), after discovering the preferred methods of buying cryptocurrencies for money laundering, they cross-examined anti-money laundering regulations in Italy and Russia to see if they have corresponding paragraphs to address such purchasing methods. Chousein et al. (2020) investigate how service providers of public blockchain systems communicate with their users about the influences of the EU General Data Protection Regulation

(GDPR) on their services and find a shortage of communication and transparency on GDPR compliance issues.

There are two reasons for the lack of regulation-oriented text analysis studies. First, the time lag between the introduction of regulations in different jurisdictions limits the availability of data for regulatory studies. Second, analyzing the content of regulations requires a computer program to understand the legal terms. Therefore, context-specific dictionaries are required to correctly extract information. Researchers should also have domain knowledge to interpret the results accurately, which can be challenging in many areas. Nevertheless, because understanding regulatory frameworks is essential to advance our understanding, combat blockchain crimes, and promote blockchain adoption, more research is needed from the perspective of blockchain-related regulations.

Perception of blockchain

The perception of (potential) users is crucial for the development of emerging technologies such as blockchain. Public acceptance does not merely rely on economic benefits, but also on other aspects. Studies have attempted to discover how the public perceives blockchain technology and the drivers of attitude construction. Such studies are closely associated with social and cultural factors and are, therefore, located in interdisciplinary studies, such as behavioral finance. The number of papers was not significant (seven papers) in this review; however, the questions discussed were diverse.

Blockchain was initially surrounded by suspicion and considered a questionable technology; however, its acceptance grew gradually. Users are attracted to the security, privacy, transparency, trust, and traceability offered by blockchain (Grover et al. 2019), but their adoption is still hindered by a lack of blockchain knowledge and distrust of blockchain (Yadav et al. 2021). Doubts can be removed by building channels for the public to gain knowledge about it:1) articles from the media help the public obtain more information about blockchain, which boosts further exploration of the technology and acceptance; 2) existing business problems motivate experimenting with blockchain and enhance trust (Perdana et al. 2021). Cultural background also helps shape the perceived value of blockchain. Grassman et al. (2021) conduct a comparative study between Sweden and Japan on the attitude towards autonomy that cryptocurrency brings. The principle of autonomy has a higher intrinsic value in Sweden, whereas Japan adopts a more pragmatic view of autonomy (i.e., facilitating investment prospects).

In broad-term blockchain, specific products with distinctive characteristics are viewed differently. Some studies (Caliskan 2020; Mnif et al. 2021; Bashchenko 2022) explore the perceptions of Bitcoin, Bitcoin Green, and cryptocurrency exchanges and explained the reasons for their interpretations.

RQ3 What are the research gaps and promising future research topics?

We now summarize the research gaps described in the papers and observed by us and develop future research topics to which future studies could address.

Improvement of data preparation

The quality of the input data largely determines the model output results; however, the complexity of text data makes it challenging to prepare. Many current studies merely conduct standard data preparation and omit the features of different types of text. To prevent "garbage-in-garbage-out", future research can look more deeply into the characteristics of specific texts and prepare the data in a way that fits the characteristics of the texts.

Data selection After text preprocessing, the text data should be further selected or weighted by considering the text features. This procedure is yet neglected by a substantial number of papers. For example, Twitter offers millions of short texts daily, but misinformation is omnipresent. Bots and fake accounts should not be ignored and should be separated from others (Burnie and Yilmaz 2019; Kraaijeveld and de Smedt 2020). Bashchenko (2022) divides news into two types: (a) endogenous news, which describes the past price movement; (b) fundamental news, which provides information that can have higher impacts. When using news for price prediction, endogenous news should be filtered out because it has a limited influence on future prices.

Another way to improve preparation can be achieved by setting relevance levels for the texts. Twitter accounts can be weighted according to their influence levels (e.g., number of followers, retweets, and user networks) (Jain et al. 2018; Li et al. 2019), and the influence of a patent is reflected by the number of citations.

Dictionary building Dictionaries are essential in text analysis models (e.g., sentiments and topics). However, they are generally only applicable to a specific context since vocabularies can change their meanings depending on discipline (Loughran and McDonald 2011). The impact of using an off-the-shelf dictionary in other areas can be a substance for blockchain studies, as new vocabularies and jargons have been invented in blockchain. Studies have indicated that designing a domain-specific lexicon for blockchain could potentially improve the accuracy of analysis (Balfagih and Keselj 2019; Chen et al. 2019a; Sattarov et al. 2020). existing studies primarily adopt the VADER (Hutto and Gilbert 2014) and LM lexicons (Loughran and McDonald 2011), and only a few studies have developed or integrated blockchain-specific lexicons (Chen et al. 2019a; Barth et al. 2020; Kraaijeveld and de Smedt 2020; Huang et al. 2021).

Extension to underused data and growing areas

In this review, we find a concentration of text data uses from social media, online forums, and academic papers. Simultaneously, many other documents containing valuable information are underused. Corporate-generated documents (e.g., SEC and patent filings) are not frequently utilized despite their importance in revealing corporate-level information. For instance, in finance studies, patent filings are used to identify specific FinTech categories (Chen et al. 2019b, 2022). Studies use 10-Ks for different purposes: product description sections for the new industry set according to product similarity (Hoberg and Phillips 2016), business descriptions for company's asset specificity(Chen et al. 2022), and risk disclosures for risk detection (Bao and Datta 2014; Hanley and Hoberg 2019). Corporate disclosures are versatile, and cater to multiple research purposes. One limitation of corporate disclosures is that blockchain startups have limited

mandatory disclosures. Nevertheless, future research can make greater use of such documents to gain insights into blockchain adoption strategies of established companies.

Another gap in the review is the absence of papers related to the keywords NFT, STO, IEO, and stablecoin. These are relatively new concepts in blockchain and are largely understudied. Researchers investigating these areas will contribute to a better understanding of market mechanisms. For example, potential text data in NFTs include descriptions and social media discussions of NFT items. STOs are treated as traditional securities and adhere to all rights and obligations including approved prospectuses for public offerings. IEO project whitepapers were thoroughly vetted by exchange prior to launch. Therefore, the above documents are more standardized and can be used similarly as standard corporate disclosures. Stablecoin is connected to conventional financial systems and have drawn attention to financial stability issues. News (integrated with event studies) could provide coverage from this perspective.

Regulation

Given the increasing trend of cryptocurrency in the monetary system, government policies and regulations are essential for counteracting risks, restricting illicit activities, and protecting consumers (Chokor and Alfieri 2021).

Many jurisdictions have updated or supplemented their regulatory frameworks to accommodate the existence of cryptocurrencies and other blockchain-based decentralized applications (e.g., Market in Crypto-Assets (MiCA) and Framework for International Engagement on Digital Assets). Issues such as money laundering, terrorist financing, and tax evasion have been extensively recognized and addressed. In addition, organizations such as the International Organization for Standardization (ISO) and the Financial Stability Board (FSB) are working to establish international rules and standards to promote collaboration among jurisdictions. Many proposed frameworks are still in their initial stages or awaiting implementation, and updates can be expected.

Texts used in regulation-related research are not limited to regulatory documents, but also include other texts, such as corporate disclosures related to blockchain or cryptocurrency (SEC 2022), terms of service agreements, and online discussions about regulatory terms. Future research could integrate regulatory factors into the study, examine the impact of regulations on markets in different jurisdictions (Barth et al. 2020), and observe users' perceptions of and reactions to specific regulations. This could provide insightful implications for practitioners and policymakers regarding the implementation of relevant regulations and how takers of specific regulations will adopt them.

Conclusion

The uncomplicated access and rich information in blockchain-related texts make them ideal for complementing numerical data in research. However, a comprehensive review of this topic to provide guidance for researchers is lacking.

This study addresses this issue by making several contributions to the literature. First, we provide comprehensive summaries of research scope, text data sources, and text analysis methodologies in the existing literature to guide researchers in finding

pertinent resources. Second, we go beyond individual elements and exhibit the connections between them. We conflate the above elements and display the two most frequently used combinations: (1) papers focusing on cryptocurrencies conduct sentiment analysis on posts from instant user-generated content or news articles to find the correlations between sentiment and market behavior, and (2) papers examining the concept of blockchain use formal documents to apply topic modeling to discover classifications and trends. We emphasize that it is crucial to choose appropriate combinations considering variable perspectives, such as data characteristics and research questions. Finally, we integrate blokcchain-related research areas and text analysis approaches into a joint framework. By not restricting our search to one discipline, we are able to capture the use of text analysis in non-technical blockchain studies across disciplines and provide multiple perspectives on the topic. We highlight five major research topics discussed in the literature: relationship discovery, cryptocurrency performance prediction, classification and trend, crime and regulation, and the perception of blockchain. Furthermore, by referring to individual papers and aggregated information, we uncover three future research topics that researchers can explore: improvement of data preparation, studies with underused data and growing areas, and regulation-related research.

We are aware that this review shares publication bias of literature reviews. Studies with statistically significant results are more likely to be published, leading to a publication bias (Rosenthal 1979). To alleviate the impact of bias, we searched the most comprehensive databases for peer-reviewed papers and chapters. We also included unpublished working papers on SSRN in keyword searches. Backward snowballing was conducted on the included papers to identify more papers that did not appear in the keyword searches. We believe that through our multiple procedures for identifying targeted papers, we obtained a comprehensive collection of papers for this literature review.

Despite this limitation, this study provides a timely academic-oriented review of the text analysis approaches used in blockchain research. Our detailed summaries will help researchers navigate specific text data types and methodologies. The findings of the current research landscape and suggested future directions could facilitate the selection of promising research topics and the implementation of suitable methodologies for their analyses. Overall, this review will be useful for researchers from various disciplines interested in exploring large-scale text data in blockchain-related research.

Appendix: The list of keywords for the query

The initial list of keywords with fundamental blockchain concepts based on our knowledge of the blockchain ecosystem (i.e., blockchain, cryptocurrency, smart contract, and ICO) and expand our list by sampling academic papers that include additional keywords. In this way, we build up a wider set of keywords by adding non-redundant keywords after observing keywords used in the academic literature. Our list of keywords is an intersection of keywords used in many blockchain-related papers. The complete list of keywords for the query is as follows:

Category	Search term	Keyword
Blockchain	Blockchain*	Blockchain
		Blockchains
	Cryptocurrenc*	Cryptocurrency
		Cryptocurrencies
	Stablecoin*	Stablecoin
		Stablecoins
	"Crypto token*"	Crypto token
		Crypto tokens
		Crypto-token
		Crypto-tokens
	"Smart contract*"	Smart contract
		Smart contracts
	"Initial coin offering*"	Initial coin offering
		Initial coin offerings
	"Security token offering*"	Security token offering
		Security token offerings
	linitial exchange offering*"	Initial exchange offering
		Initial exchange offerings
	"Non fungible token*"	Non fungible token
		Non fungible tokens
		Non-fungible token
		Non-fungible tokens
Text analysis	"Text* analysis"	Text analysis
		Textual analysis
	"Text analytics"	Text analytics
	"Topic model*"	Topic model
		Topic models
		Topic modeling
		Topic modelings
		Topic modellings
	"Natural language processing*"	Natural language processing
		Natural language processings
	"Word embedding*"	Word embedding
		Word embeddings
	"Sentence embedding*"	Sentence embedding
		Sentence embeddings
	"Bag of words"	Bag of words
		Bag-of-words
	"Sentiment analysis"	Sentiment analysis

Abbreviations

BERT	Bidirectional encoder representations from transformers
BoW	Bag-of-words
CBOW	Continuous bag of words
CPY	Citation per year
DTM	Dynamic topic models
FSB	Financial Stability Board
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
GDPR	General Data Protection Regulation
ICO	Initial coin offering
IEO	Initial exchange offering
IPO	Initial public offering

ISO	International Organization for Standardization
LDA	Latent dirichlet allocation
LM lexicon	Loughran and McDonald sentiment lexicon
LSTM	Long short-term memory
MiCA	Market in Crypto-Assets
NFT	Non-fungible token
NLP	Natural language processing
SEC	U.S. Securities and Exchange Commission
SSRN	Social Science Research Network
STO	Security token offering
TF-IDF	Term frequency-inverse document frequency
VADER	Valence Aware Dictionary for Sentiment Reasoning
WoS	Web of Science
ZCL	ZClassic

Acknowledgements
The authors would like to thank the anonymous reviewers and the editors for their valuable commentsand suggestions which led to great improvement of this paper through the revision process.

Authors contributions
XZ: Conceptualization, Methodology, Writing. FI: Supervision, Reviewing, Writing. DB: Supervision,Reviewing, Writing. All authors read and approved the final manuscript.

Funding
This research is supported by the Manchot Graduate School "Competitiveness of Young Enterprises"at the Heinrich-Heine-University of Düsseldorf. Funding was provided by the Jürgen Manchot Stiftung.

Availability of data and materials
The papers used in this review can be obtained using the same search terms in relevant databases. Thefull list of papers included in this review is available from the correspondence author upon reasonablerequest.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 June 2022 Accepted: 25 April 2023
Published online: 29 February 2024

References

Abraham J, Higdon D, Nelson J, Ibarra J (2018) Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Sci Rev 1(3):1
Akba F, Medeni IT, Guzel MS, Askerzade I (2021) Manipulator detection in cryptocurrency markets based on forecasting anomalies. IEEE Access 9:108819–108831
Alahi I, Islam M, Iqbal A, Bosu A (2019) Identifying the challenges of the blockchain community from Stackexchange topics and trends. 2019 IEEE 43rd Ann Comput Softw Appl Conf (COMPSAC) 1:123–128
Anamika A, Subramaniam S (2022) Do news headlines matter in the cryptocurrency market? Appl Econ 54(54):6322–6338
Aslam N, Rustam F, Lee E, Washington PB, Ashraf I (2022) Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. IEEE Access 10:39313–39324
Ayvaz S, Shiha MO (2018) A scalable streaming big data architecture for real-time sentiment analysis. In: Proceedings of the 2018 2nd international conference on cloud and big data computing, pp 47–51
Azqueta-Gavaldón A (2020) Causal inference between cryptocurrency narratives and prices: evidence from a complex dynamic ecosystem. Phys A: Stat Mech Appl 537:122574
Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the seventh international conference on language resources and evaluation (LREC'10)
Bahamazava K, Nanda R (2022) The shift of darknet illegal drug trade preferences in cryptocurrency: the question of traceability and deterrence. Forens Sci Int: Digit Investig 40:301377
Bahamazava K, Reznik S (2022) The comparative analysis of regulations in the Italian Republic and the Russian Federation against cryptolaundering techniques. J Money Laundering Control
Baldwin C (2016) Bitcoin worth \$72 million stolen from bitfinex exchange in Hong Kong. Reuters Media. Accessed 6 Nov 2022
Balfagih AM, Keselj V (2019) Evaluating sentiment c1assifiers for Bitcoin tweets in price prediction task. In: 2019 IEEE international conference on big data (Big Data), pp 5499–5506
Bao Y, Datta A (2014) Simultaneously discovering and quantifying risk types from textual risk disclosures. Manag Sci 60(6):1371–1391
Barth JR, Herath HS, Herath TC, Xu P (2020) Cryptocurrency valuation and ethics: a text analytic approach. J Manag Anal 7(3):367–388

Bashchenko O (2022) Bitcoin price factors: natural language processing approach. Available at SSRN 4079091
Birim ŞÖ, Sönmez FE (2022) Social sentiment analysis for prediction of cryptocurrency prices using neuro-fuzzy techniques. In: International conference on intelligent and fuzzy systems, pp 606–616
Blei DM, Lafferty JD (2006) Dynamic topic models. In: Proceedings of the 23rd international conference on machine learning, pp 113–120
Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022
Bloomberg: Ethereum Developers Push Ahead With Shanghai Upgrade to Enable Withdrawals (2023). https://www.bloomberg.com/news/articles/2023-01-05/ethereum-developers-push-ahead-with-update-enabling-withdrawals Accessed 01/24/2023
Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, pp 144–152
Burnie A, Yilmaz E (2019) Social media and Bitcoin metrics: which words matter. R Soc Open Sci 6(10):191068
Caliskan K (2020) Platform works as stack economization: cryptocurrency markets and exchanges in perspective. Sociologica 14(3):115–142
Cary M (2021) Down with the #dogefather: evidence of a cryptocurrency responding in real time to a crypto-tastemaker. J Theor Appl Electron Commer Res 16(6):2230–2240
Casino F, Dasaklis TK, Patsakis C (2019) A systematic literature review of blockchain-based applications: current status, classification and open issues. Telemat Inform 36:55–81
Caviggioli F, Lamberti L, Landoni P, Meola P (2020) Technology adoption news and corporate reputation: sentiment analysis about the introduction of Bitcoin. J Prod Brand Manag 29(7):877–897
Cavnar WB, Trenkle JM et al (1994) N-gram-based text categorization. In: Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, vol. 161175
Chakkarwar V, Tamane SC (2019) Quick insight of research literature using topic modeling. In: Zhang Y-D, Mandal JK, So-In C, Thakur NV (eds) Smart trends in computing and communications 2019, vol. 165, pp 189–197
Chen CY-H, Després R, Guo L, Renault T (2019a) What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble. Technical report, IRTG 1792 Discussion Paper
Chen MA, Wu Q, Yang B (2019) How valuable is fintech innovation? Rev Financ Stud 32(5):2062–2106
Chen MA, Hu S, Wang J, Wu Q (2022) Can blockchain technology help overcome contractual incompleteness? Evidence from state laws. Available at SSRN 3915895
Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining, pp 785–794
Cheuque CG, Reutter LJ (2019) Bitcoin price prediction through opinion mining. In: Companion proceedings of The 2019 World Wide Web conference, pp 755–762
Chiarello F, Belingheri P, Bonaccorsi A, Fantoni G, Martini A (2021) Value creation in emerging technologies through text mining: the case of blockchain. Technol Anal Strateg Manag 33(12):1404–1420
Choi J, Lee T, Kim K, Seo M, Cui J, Shin S (2022) Discovering message templates on large scale Bitcoin abuse reports using a two-fold NLP-based clustering method. IEICE Trans Inf Syst 105(4):824–827
Chokor A, Alfieri E (2021) Long and short-term impacts of regulation in the cryptocurrency market. Q Rev Econ Financ 81:157–173
Chousein Z, Tetik HY, Sağlam RB, Bülbül A, Li S (2020) Tension between GDPR and public blockchains: a data-driven analysis of online discussions. In: 13th international conference on security of information and networks, pp 1–8
Christie WG, Huang RD (1995) Following the pied piper: do individual returns herd around the market? Financ Anal J 51(4):31–37
Chursook A, Dawod AY, Chanaim S, Naktnasukanjn N, Chakpitak N (2022) Twitter sentiment analysis and expert ratings of initial coin offering fundraising: evidence from Australia and Singapore markets. TEM J 11(1):44–55
Cohney S, Hoffman D, Sklaroff J, Wishnick D (2019) Coin-operated capitalism. Columbia Law Rev 119(3):591–676
Coleman M, Liau TL (1975) A computer readability formula designed for machine scoring. J Appl Psychol 60(2):283
Cong LW, He Z (2019) Blockchain disruption and smart contracts. Rev Financ Stud 32(5):1754–1797
Cong LW, Liang T, Yang B, Zhang X (2021) Analyzing textual information at scale. In: Information for Efficient decision making: big data, blockchain and relevance. World Scientific, Singapore, pp 239–271
Corbet S, Meegan A, Larkin C, Lucey B, Yarovaya L (2018) Exploring the dynamic relationships between cryptocurrencies and other financial assets. Econ Lett 165:28–34
Critien JV, Gatt A, Ellul J (2022) Bitcoin price change and trend prediction through Twitter sentiment and data volume. Financ Innov 8(1):1–20
da Silva CF, Moro S (2021) Blockchain technology as an enabler of consumer trust: a text mining literature analysis. Telemat Inform 60:101593
Dale E, Chall JS (1948) A formula for predicting readability: instructions. Educ Res Bull 27(2):37–54
Daluwathumullagamage DJ, Sims A (2021) Fantastic beasts: blockchain based banking. J Risk Financ Manag 14(4):1–43
Daluwathumullagamage DJ, Sims A (2020) Blockchain-enabled corporate governance and regulation. Int J Financ Stud 8(2):1–38
Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
Dittmar R, Wu DA (2019) Initial coin offerings hyped and dehyped: an empirical examination. SSRN Electron J
Dumay J, Cai L (2014) A review and critique of content analysis as a methodology for inquiring into IC disclosure. J Intell Cap 15(2):264–290
El-Masri M, Hussain EMA (2021) Blockchain as a mean to secure internet of things ecosystems: a systematic literature review. J Enterp Inf Manag 34(5):1371–1405
Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. KDD 96:226–231

Fang F, Ventre C, Basios M, Kanthan L, Martinez-Rego D, Wu F, Li L (2022) Cryptocurrency trading: a comprehensive survey. Financ Innov 8(1):1–59
Farimani SA, Jahan MV, Fard AM, Tabbakh SRK (2022) Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. Knowl-Based Syst 247:108742
Flesch R (1979) How to write plain english: a book for lawyers and consumers, 1st edn. Harper & Row, New York
Florysiak D, Schandlbauer A (2022) Experts or charlatans? ICO analysts and white paper informativeness. J Bank Financ 139:106476
Foley S, Karlsen JR, Putniņš TJ (2019) Sex, drugs, and bitcoin: how much illegal activity is financed through cryptocurrencies? Rev Financ Stud 32(5):1798–1853
Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
Frizzo-Barker J, Chow-White PA, Adams PR, Mentanko J, Ha D, Green S (2020) Blockchain as a disruptive technology for business: a systematic review. Int J Inf Manag 51:102029
Fu C, Koh A, Griffin P (2019) Automated theme search in ICO whitepapers. J Financ Data Sci 1(4):140–158
FXStreet: Ethereum Shanghai Upgrade: guide to the ETH hard fork, unstaking and liquid staking projects (2023)
Galeshchuk S, Vasylchyshyn O, Krysovatyy A (2018) Bitcoin response to Twitter sentiments. In: CEUR workshop proceedings, pp 160–168
Gallagher RJ, Reing K, Kale D, Ver Steeg G (2017) Anchored correlation explanation: topic modeling with minimal domain knowledge. Trans Assoc Comput Linguist 5:529–542
Garanina T, Ranta M, Dumay J (2021) Blockchain in accounting research: current trends and emerging topics. Account Audit Account J 35(7):1507–1533
Ge C, Shi H, Jiang J, Xu X (2021) Investigating the demand for blockchain talents in the recruitment market: evidence from topic modeling analysis on job postings. Inf Manag 59(7):103513
Gentzkow M, Kelly B, Taddy M (2019) Text as data. J Econ Lit 57(3):535–74
Georgoula I, Pournarakis D, Bilanakos C, Sotiropoulos D, Giaglis GM (2015) Using time-series and sentiment analysis to detect the determinants of Bitcoin prices. Available at SSRN 2607167
Goldberg Y (2017) Neural network methods for natural language processing. Synth Lect Human Lang Technol 10(1):1–309
Goldstein RAM, Tabuchi H (2014) Erosion of faith was death knell for Mt. Gox. NY Times. Accessed 6 Nov 2022
Grassman R, Bracamonte V, Davis M, Sato M (2021) Attitudes to cryptocurrencies: a comparative study between Sweden and Japan. Rev Socionetwork Strateg 15(1):169–194
Grimmer J, Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit Anal 21(3):267–297
Grover P, Kar AK, Janssen M, Ilavarasan PV (2019) Perceived usefulness, ease of use and user acceptance of blockchain technology for digital transactions: insights from user-generated content on Twitter. Enterp Inf Syst 13(6):771–800
Gunning R (1952) Technique of clear writing
Günther E, Quandt T (2016) Word counts and topic models. Digit J 4(1):75–88
Gurdgiev C, O'Loughlin D (2020) Herding and anchoring in cryptocurrency markets: investor reaction to fear and uncertainty. J Behav Exp Financ 25:100271
Han S, Ye S, Zhang H (2020) Visual exploration of internet news via sentiment score and topic models. Comput Vis Med 6(3):333–347
Hanley KW, Hoberg G (2019) Dynamic interpretation of emerging risks in the financial sector. Rev Financ Stud 32(12):4543–4603
Hashimoto K, Xiong C, Tsuruoka Y, Socher R (2016) A joint many-task model: growing a neural network for multiple NLP tasks. arXiv preprint arXiv:1611.01587
Hassan MK, Hudaefi FA, Caraka RE (2021) Mining netizen's opinion on cryptocurrency: sentiment analysis of Twitter data. Stud Econ Financ 39(3):365–385
Henry E (2008) Are investors influenced by how earnings press releases are written? J Bus Commun (1973) 45(4):363–407
Hinds-Charles C, Adames J, Yang Y, Shen Y, Wang Y (2019) A longitude analysis on Bitcoin issue repository. In: 2018 1st IEEE international conference on hot information-centric networking (HotICN), pp 212–217
Hirata E, Lambrou M, Watanabe D (2021) Blockchain technology in supply chain management: insights from machine learning algorithms. Marit Bus Rev 6(2):114–128
Ho TK (1995) Random decision forests. Proc 3rd Int Conf Doc Anal Recognit 1:278–282
Hoberg G, Phillips G (2016) Text-based network industries and endogenous product differentiation. J Polit Econ 124(5):1423–1465
Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Huang X, Zhang W, Tang X, Zhang M, Surbiryala J, Iosifidis V, Liu Z, Zhang J (2021) LSTM based sentiment analysis for cryptocurrency prediction. In: International conference on database systems for advanced applications, pp 617–621
Hutto C, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web and Soc Med 8:216–225
Huynh TLD (2021) Does Bitcoin react to Trump's tweets? J Behav Exp Financ 31:100546
Huynh TLD (2022) When Elon musk changes his tone, does Bitcoin adjust its tune? Comput Econ. https://doi.org/10.1007/s10614-021-10230-6
Ibba G, Ortu M, Tonelli R (2021) Smart contracts categorization with topic modeling techniques. In: Marin B, Wautelet Y, Heng S, Assar S, Asprion PM, Morichetta A (eds) CEUR Workshop proceedings, vol. 3031, pp 64–73
Inamdar A, Bhagtani A, Bhatt S, Shetty PM (2019) Predicting cryptocurrency value using sentiment analysis. In: 2019 International conference on intelligent computing and control systems (ICCS), pp 932–934
Jaccard P (1912) The distribution of the flora in the Alpine zone. New Phytol 11(2):37–50
Jain A, Tripathi S, Dwivedi HD, Saxena P (2018) Forecasting price of cryptocurrencies using tweets sentiment analysis. In: 2018 Eleventh international conference on contemporary computing (IC3), pp 1–7

Jurafsky D, Martin J (2017) Naive bayes and sentiment classification. Speech and language processing. Standford University Press, Redwood City, pp 74–91
Kang K, Choo J, Kim Y (2020) Whose opinion matters? Analyzing relationships between Bitcoin prices and user groups in online community. Soc Sci Comput Rev 38(6):686–702
Karalevicius V, Degrande N, de Weerdt J (2018) Using sentiment analysis to predict interday Bitcoin price movements. J Risk Financ 19(1):56–75
Kilimci ZH (2020) Sentiment analysis based direction prediction in Bitcoin using deep learning algorithms and word embedding models. Int J Intell Syst Appl Eng 8(2):60–65
Kim S, Park H, Lee J (2020) Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: a study on blockchain technology trend analysis. Exp Syst Appl 152:113401
Kim YB, Kim JG, Kim W, Im JH, Kim TH, Kang SJ, Kim CH (2016) Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS One 11(8):0161197
Klare GR (1974) Assessing readability. Read Res Q 10(1):62–102
Kou G, Yang P, Peng Y, Xiao F, Chen Y, Alsaadi FE (2020) Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Appl Soft Comput 86:105836
Kraaijeveld O, de Smedt J (2020) The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. J Int Financ Mark Inst Money 65:101188
Kwon O-W, Lee J-H (2003) Text categorization based on K-nearest neighbor approach for web site classification. Inf Process Manag 39(1):25–44
Lambert T, Liebau D, Roosenboom P (2021) Security token offerings. Small Bus Econ. https://doi.org/10.1007/s11187-021-00539-9
Laturnus V (2022) What matters to crypto investors? Insights from token offerings on the Ethereum blockchain. Available at SSRN 4087795
Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196
Lee D, Seung HS (2000) Algorithms for non-negative matrix factorization. In: Leen T, Dietterich T, Tresp V (eds) Advances in neural information processing systems. vol 13. MIT Press
Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
Lee J, Zo HJ, Steinberger T (2022) Exploring trends in blockchain publications with topic modeling: implications for forecasting the emergence of industry applications. Available at SSRN 4079332
Li TR, Chamrajnagar AS, Fong XR, Rizik NR, Fu F (2019) Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model. Front Phys 7:98
Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D (2009) The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. Ann Internal Med 151(4):65
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
Linton M, Teo EGS, Chen CY, Härdle WK (2017) Dynamic topic modelling for cryptocurrency community forums. In: Härdle WK, Chen CY-H, Overbeck L (eds) Applied quantitative finance. Springer, Berlin, pp 355–372
Liu Y, Sheng J, Wang W (2021) Technology and cryptocurrency valuation: evidence from machine learning. Available at SSRN 3577208
Loginova E, Tsang WK, van Heijningen G, Kerkhove L-P, Benoit DF (2021) Forecasting directional Bitcoin price returns using aspect-based sentiment analysis on online text data. Mach Learn. https://doi.org/10.1007/s10994-021-06095-3
Loughran TIM, McDonald B (2016) Textual analysis in accounting and finance: a survey. J Account Res 54(4):1187–1230
Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J Financ 66(1):35–65
Lu H-K, Yang L-w, Lin P-C, Yang T-H, Chen AN (2017) A study on adoption of Bitcoin in Taiwan: using big data analysis of social media. In: Proceedings of the 3rd international conference on communication and information processing, pp 32–38
MacQueen J (1967) Classification and analysis of multivariate observations. In: 5th Berkeley symposium on mathematical statistics and probability, pp 281–297
Mai F, Shan Z, Bai Q, Wang X, Chiang RH (2018) How does social media impact Bitcoin value? A test of the silent majority hypothesis. J Manag Inf Syst 35(1):19–52
Mao H, Counts S, Bollen J (2011) Predicting financial markets: comparing survey, news, Twitter and search engine data. arXiv preprint arXiv:1112.1051
Mareddy S, Gupta D (2022) Analysis of Twitter data for identifying trending domains in blockchain technology. In: Smys S, Bestak R, Palanisamy R, Kotuliak I (eds) Computer networks and inventive communication technologies, vol. 75, pp 651–672
McLaughlin GH (1969) SMOG grading: a new readability formula. J Read 12(8):639–646
Medhi PK (2020) Blockchain-enabled supply chain transparency, supply chain structural dynamics, and sustainability of complex global supply chains: a text mining analysis. In: Information for efficient decision making: big data. Blockchain And relevance. World Scientific Publishing Co, Singapore, pp 273–312
Mendoza-Tello JC, Mora H, Pujol-López FA, Lytras MD (2018) Social commerce as a driver to enhance trust and intention to use cryptocurrencies for electronic payments. IEEE Access 6:50737–50751
Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
Mnif E, Lacombe I, Jarboui A (2021) Users' perception toward Bitcoin Green with big data analytics. Soc Bus Rev 16(4):592–615
Mohammad SM (2018) Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of The annual conference of the association for computational linguistics (ACL), Melbourne, Australia

Moher D, Liberati A, Tetzlaff J, Altman DG (2009) PRISMA Group: preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Internal Med 151(4):264–269
Mousa A, Schuller B (2017) Contextual bidirectional long short-term memory recurrent neural network language models: a generative approach to sentiment analysis. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 1, Long papers. Association for Computational Linguistics, Valencia, Spain, pp 1023–1032
Moustafa H, Malli M, Hazimeh H (2022) Real-time Bitcoin price tendency awareness via social media content tracking. In: 2022 10th international symposium on digital forensics and security (ISDFS), pp 1–6
Nakamoto S (2008) A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf. Accessed 10 Apr 2022
Narman HS, Uulu AD, Liu J (2018) Profile analysis for cryptocurrency in social media. In: 2018 IEEE international symposium on signal processing and information technology (ISSPIT), pp 229–234
Newman D, et al (2010) Automatic evaluation of topic coherence. In: Human language technologies: the 2010 annual conference of the north american chapter of the association for computational linguistics, pp 100–108
Nguyen TH, 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, vol. 1: Long Papers, pp 1354–1364
Nielsen FÅ (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903
Nizzoli L, Tardelli S, Avvenuti M, Cresci S, Tesconi M, Ferrara E (2020) Charting the landscape of online cryptocurrency manipulation. IEEE Access 8:113230–113245
Ortu M, Uras N, Conversano C, Bartolucci S, Destefanis G (2022) On technical trading and social media indicators for cryptocurrency price classification through deep learning. Exp Syst Appl 198:116804
Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE et al (2021) PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. bmj. https://doi.org/10.1007/s10994-021-06095-3
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (Clin Res ed) 372:71
Pan L, Feng L, Jiayin Q (2020) Adaptive evolution mechanism of blockchain community based on token-based halving event. In: 2020 Chinese automation congress (CAC), pp 6140–6144
Pant DR, Neupane P, Poudel A, Pokhrel AK, Lama BK (2018) Recurrent neural network based Bitcoin price prediction by Twitter sentiment analysis. In: 2018 IEEE 3rd international conference on computing, communication and security (ICCCS), pp 128–132
Patil AP, Akarsh TS, Parkavi A (2018) A study of opinion mining and data mining techniques to analyse the cryptocurrency market. In: 2018 3rd international conference on computational systems and information technology for sustainable solutions (CSITSS), pp 198–203
Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
Perdana A, Robb A, Balachandran V, Rohde F (2021) Distributed ledger technology: its evolutionary path and the road ahead. Inf Manag 58(3):103316
Phillips RC, Gorse D (2018) Mutual-excitation of cryptocurrency market returns and social media topics. In: 4th international conference on frontiers of educational technologies, pp 80–86
Polasik M, Piotrowska AI, Wisniewski TP, Kotkowski R, Lightfoot G (2015) Price fluctuations and the use of Bitcoin: an empirical inquiry. Int J Electron Commer 20(1):9–49
Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) Catboost: unbiased boosting with categorical features. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems, vol 31. Curran Associates, Inc.,
Rahman S, Hemel JN, Anta SJA, Al Muhee H, Uddin J (2018) Sentiment analysis using R: an approach to correlate cryptocurrency price fluctuations with change in user sentiment using machine learning. In: 2018 Joint 7th international conference on informatics, electronics and vision (ICIEV) and 2018 2nd international conference on imaging, vision and pattern recognition (icIVPR), pp 492–497
Ramos J (2003) Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, vol. 242, pp 29–48
Reimers N, Gurevych I (2020) Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813
Rognone L, Hyde S, Zhang SS (2020) News sentiment in the cryptocurrency market: an empirical comparison with forex. Int Rev Financ Anal 69:101462
Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychol Bull 86(3):638
Salton G, Yang C-S, Yu CT (1975) A theory of term importance in automatic text analysis. J Am Soc Inf Sci 26(1):33–44
Sapkota N, Grobys K (2021) Fear sells: determinants of fund-raising success in the cross-section of initial coin offerings. Available at SSRN 3843138
Şaşmaz E, Tek FB (2021) Tweet sentiment analysis for cryptocurrencies. In: 2021 6th international conference on computer science and engineering (UBMK), pp 613–618
Sattarov O, Jeon HS, Oh R, Lee JD (2020) Forecasting Bitcoin price fluctuation by Twitter sentiment analysis. In: 2020 international conference on information science and communications technologies (ICISCT)
Schwenkler G, Zheng H (2021) News-driven peer co-movement in crypto markets. Available at SSRN 3572471
Sebastião H, Godinho P (2021) Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financ Innov 7(1):1–30
SEC: Staff Accounting Bulletin No. 121 (2022)
Senter R, Smith EA (1967) Automated readability index. Technical report, Cincinnati Univ OH

Shahid MN, Jungpil H (2020) A cross-disciplinary review of blockchain research trends and methodologies: topic modeling approach. In: 53rd annual hawaii international conference on system sciences, HICSS 2020, vol. 2020, pp 4053–4059
Siddaway AP, Wood AM, Hedges LV (2019) How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Ann Rev Psychol 70:747–770
Siu GA, Collier B, Hutchings A (2021) Follow the money: the relationship between currency exchange and illicit behaviour in an underground forum. In: 2021 IEEE european symposium on security and privacy workshops (EuroS &PW), pp 191–201
Stanley M (2019) The application of behavioural heuristics to initial coin offerings valuation and investment. J Br Blockchain Assoc 2(1):7776
Steinert L, Herff C (2018) Predicting altcoin returns using social media. PloS One 13(12):0208119
Stone PJ, Dunphy DC, Smith MS (1966) The general inquirer: a computer approach to content analysis. M.I.T. Press, Oxford, England
Stratopoulos TC, Wang VX, Ye H (2022) Use of corporate disclosures to identify the stage of blockchain adoption. Account Horiz 36(1):197–220
Thewissen J, Shrestha P, Torsin W, Pastwa AM (2022) Unpacking the black box of ICO white papers: a topic modeling approach. J Corp Financ 75:102225
Toma AM, Cerchiello P (2020) Initial coin offerings: risk or opportunity? Front Artif Intell 3:18
U.Today: Ethereum (ETH): Shanghai Hard Fork Causes Concern Among Developers, Here Are Reasons (2023)
Vacca S, Costerbosa CL, Spada A, Riotta G, Uras N (2021) Investigation of coronavirus impact on blockchain and cryptocurrencies markets. In: 2021 IEEE/ACM 4th international workshop on emerging trends in software engineering for blockchain (WETSEB), pp 56–60
Valencia F, Gómez-Espinosa A, Valdés-Aguirre B (2019) Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy 21(6):589
Vohra S, Teraiya J (2013) A comparative study of sentiment analysis techniques. J Jikrce 2(2):313–317
Voskobojnikov A, Wiese O, Mehrabi Koushki M, Roth V, Beznosov K (2021) The U in crypto stands for usable: An empirical study of user experience with mobile cryptocurrency wallets. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–14
Wang J, Fan Y, Zhang H, Feng L (2021) Technology hotspot tracking: topic discovery and evolution of china's blockchain patents based on a dynamic LDA model. Symmetry 13(3):415
Wang Q, Li R, Wang Q, Chen S (2021b) Non-fungible token (NFT): overview, evaluation, opportunities and challenges. arXiv preprint arXiv:2105.07447
Xie P (2021) The interplay between investor activity on virtual investment community and the trading dynamics: evidence from the Bitcoin market. Inf Syst Front 24(4):1287–1303
Xu M, Chen X, Kou G (2019) A systematic review of blockchain. Financ Innov 5(1):1–14
Xu W, Wang T, Chen R, Zhao JL (2021) Prediction of initial coin offering success based on team knowledge and expert evaluation. Decis Support Syst 147:113574
Xu XF, He YY (2022) Blockchain application in modern logistics information sharing: a review and case study analysis. Prod Plan Control. https://doi.org/10.1080/09537287.2022.2058997
Yadav J, Misra M, Rana NP, Singh K, Goundar S (2021) Netizens' behavior towards a blockchain-based esports framework: a TPB and machine learning integrated approach. Int J Sports Mark Spons 23(4):665–683
Yao W, Xu K, Li Q (2019) Exploring the influence of news articles on Bitcoin price with machine learning. In: 2019 IEEE Symposium on computers and communications (ISCC), pp 1–6
Yen J-C, Wang T (2021) Stock price relevance of voluntary disclosures about blockchain technology and cryptocurrencies. Int J Account Inf Syst 40:100499
Yen J-C, Wang T, Chen Y-H (2021) Different is better: how unique initial coin offering language in white papers enhances success. Account Financ 61(4):5309–5340
Zeng S, Ni X, Yuan Y, Wang F-Y (2018) A bibliometric analysis of blockchain research. In: 2018 IEEE intelligent vehicles symposium (IV), vol. 2018, pp 102–107
Zhang H, Daim T, Zhang YP (2021) Integrating patent analysis into technology roadmapping: a latent dirichlet allocation based technology assessment and roadmapping in the field of blockchain. Technol Forecast Soc Change 167:120729
Zhang Y, Kang S, Dai W, Chen S, Zhu J (2021b) Code will speak: early detection of Ponzi smart contracts on Ethereum. In: 2021 IEEE international conference on services computing (SCC), pp 301–308
Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybernet 1(1):43–52

Publisher's note