Into the Etherverse: Examining Ethereum Conversations and Correlations Between Social Media Post Volume and Price Volatility Using NLP and Text Mining

Matthew McMurry

Columbia University Master of Arts in Quantitative Methods, May 2023

Introduction

Amidst the fallout from the 2008 global financial crisis, a new type of financial instrument known as cryptocurrency materialized in the form of Bitcoin, a digital asset enabled by an innovative new technology known as blockchain. The invention and unlikely success of Bitcoin inspired the emergence of other types of cryptocurrency, an asset class that has experienced unprecedented growth in recent years with a current market capitalization exceeding $1 Trillion USD, up >10,000% from just $10B USD in 2016. What began as a fringe financial experiment has blossomed into a robust, 24/7 digital economy with tens of millions of global participants, some of whom have amassed exorbitant wealth by trading these digital assets. Cryptocurrency markets themselves are notoriously volatile, with tokens routinely experiencing >50% swings in the span of days or hours and thus cryptocurrencies are extremely risky, but also potentially lucrative speculative assets that are enticing to traders, especially to those with profitable trading strategies. To give one brief example of cryptocurrency price appreciation and volatility, the price of Ether (hereinafter referred to as ETH), the native cryptocurrency of the Ethereum blockchain, is up >20,000% since its launch in 2015, yet is currently down >60% from its all time high set less than 18 months ago. At its all time high, ETH had appreciated more in the span of 6 years in pure percentage terms than Microsoft stock had over its entire 35 year history. Currently, cryptocurrency markets provide the greatest risk adjusted returns of any financial market in existence, and therefore devising methods of anticipating future price movement of cryptocurrencies is a worthwhile endeavor.

Before doing so, one must understand what cryptocurrencies are and what makes cryptocurrency markets unique. As the example of Bitcoin above alludes to, a cryptocurrency is a digital asset that relies upon cryptography and blockchain technology to facilitate its movement and prove ownership. In order to conduct a transaction and move assets from one digital wallet to another, one must request and sign the transaction using a private cryptographic key in addition to paying a small fee (often known as a network, miner, or gas fee). The request is then broadcast to a network of nodes whose job it is to verify and validate the transaction. When the transaction is verified as being valid in that it satisfies the rules of the network and the user is proven to have sufficient funds, it is bundled with other transactions into what's known as a block (the block part of blockchain), and miners or validators validate these blocks at regular intervals at which point they are sequentially added to a publicly distributed transaction ledger, known as the blockchain, where the transaction block is permanently and immutably stored. This permanent, immutable transaction ledger acts as a public consensus record that proves ownership of assets and confirms whether a transaction has taken place, and unlike fiat currencies, which require centralized entities like banks to verify and settle transactions and maintain ledgers, cryptocurrencies are trustless and decentralized in nature and do not require a central entity to verify and settle transactions.

Clearly cryptocurrencies themselves are unique, as are the markets where they are bought and sold. Cryptocurrency markets on their face closely resemble traditional financial markets. For instance a significant amount of cryptocurrency trading takes place on exchanges where a centralized entity lists tokens for trading, pools liquidity, and matches buy and sell orders for crypto tokens, similar to how traditional financial exchanges like the NYSE or NASDAQ match buy and sell orders for shares of publicly listed companies. Coinbase and Binance are two popular centralized cryptocurrency exchanges that operate in this manner, giving cryptocurrency trading a similar look and feel to trading traditional financial assets.

However, cryptocurrency markets are dissimilar from traditional financial markets in several ways. First, in addition to the centralized exchanges mentioned, cryptocurrencies are frequently traded on Decentralized Exchanges (known as DEXs), where peer-to-peer token swaps take place that are facilitated by software and algorithms rather than a centralized market maker. This means that cryptocurrency trading is divided across many different platforms versus a limited number of central exchanges in traditional markets.

The most notable feature that distinguishes cryptocurrency markets from traditional markets, however, is the lack of regulation and guidance surrounding cryptocurrencies and a subsequent lack of reliable valuation metrics. Because there are no regulations surrounding token issuance or financial reporting standards, cryptocurrencies do not have associated financial disclosures, and as such there are almost no traditional valuation metrics that can be used as guideposts for investors and traders to evaluate potential trades or investments in cryptocurrencies. Whereas participants in traditional markets can evaluate and compare entities based on mandatory quarterly financial disclosures and guidance issued by their respective companies, traders of cryptocurrencies have no such information to rely upon. For example, Price to Earnings (P/E) ratios, which are derived from earning statements, are often used when analyzing and comparing potential investments in publicly traded companies. Because cryptocurrencies do not report earnings, traders and investors are left to make decisions based on other factors. This lack of information can be viewed negatively, but the lack of standardized valuation metrics also means there exists an opportunity to gain an edge on other market participants. By developing an understanding of underlying factors that move markets or pinpointing reliable signals that indicate market moves are imminent, one can gain such an edge.

Considering the aforementioned, an interesting phenomenon can be observed when studying the behavior of cryptocurrency market participants; namely, investment decisions are often driven by conversations and popular narratives without any regard for the underlying or

intrinsic value of the assets themselves, which is often unknown. For instance, the most recent cryptocurrency bull market was arguably driven by the viral narrative that Bitcoin was a store of value not dissimilar from gold and a hedge against inflation, which caused many to subscribe to the new “digital gold” narrative and caused a digital gold rush, a frenzy that in turn caused an increase in speculation on other cryptocurrency assets. Similar to the “Bitcoin as digital gold” narrative, new narratives began to form around so-called alt-coins (alternative coins) which captured the attention of investors as they attempted to anticipate the next wave of cryptocurrency investments. Considering this anecdote, it stands to reason that studying and understanding the conversations surrounding cryptocurrencies could reveal insights that may help one anticipate price movements and investor behavior in a reliable manner.

That leads us to the primary question of this paper. If conversations and narratives are indeed a significant catalyst of investment behavior of cryptocurrency market participants, is there a meaningful way to assess these conversations that could allow one to anticipate price fluctuations, and if so, what are some measurements that could be useful to that end? This paper seeks to address these questions through an exploration of cryptocurrency conversations, specifically those regarding the Ethereum blockchain (and the cryptocurrency ETH) through a variety of methods including text mining, NLP and regression analysis.

Literature Review

Prior to addressing these questions, it is beneficial to provide an overview of the current thinking surrounding the relationship between narratives and the dynamics of cryptocurrencies and behavior of cryptocurrency market participants.

The first text that deserves mention is Nobel Prize-winning economist Robert Shiller’s 2019 book *Narrative Economics* [1] in which he argues that narratives, which have long been ignored by economists due in part to their intangible nature, have a very real and significant role in motivating human behavior, especially in the realm of economics and therefore must be factored into any substantive economic analysis. Shiller writes: *“we need to incorporate the*

contagion of narratives into economic theory. Otherwise, we remain blind to a very real, very palpable, very important mechanism for economic change, as well as a crucial element for economic forecasting. If we do not understand the epidemics of popular narratives, we do not fully understand changes in the economy and in economic behavior." What he describes regarding traditional economic activity is precisely the phenomenon that this paper observes as it relates to the motivations of many cryptocurrency market participants.

He continues: *"traditional economic approaches fail to examine the role of public beliefs in major economic events—that is, narrative. By incorporating an understanding of popular narratives into their explanations of economic events, economists will become more sensitive to such influences when they forecast the future."* Shiller later asserts that *"popular thinking often drives decisions that ultimately affect decisions, such as how and where to invest, how much to spend or save.*" Coincidentally, the first example Shiller illustrates regarding narratives and their effect on investor behavior is Bitcoin. *"These narratives surrounding Bitcoin, the most remarkable cryptocurrency in history as judged by the speculative enthusiasm for it and its market price rather than its actual use in commerce, provide an intuitive basis for discussing the basic epidemiology of narrative economics."* Clearly, Shiller believes that narratives are intertwined with the behavior of market participants and that this is especially true of those who participate in cryptocurrency markets as evidenced by those who buy and sell Bitcoin.

Bowden et al. [2] explores sentiment in relation to the behavior of a set of individual cryptocurrency traders by examining a proprietary crypto trading data set alongside Reddit forum posts using NLP and sentiment analysis and finds that "*positive changes in the level of bullishness lead to traders executing larger trades"*, supporting the notion that traders are indeed affected by general sentiment. While this study ventures into the same realm that this paper does, it focuses largely on sentiment but ignores other potential useful signals that can be derived from large text data.

Similar to Bowden, Bharghava et al. [3] uses NLP to explore investor narratives, but in the realm of traditional financial markets. This paper creates metrics from narratives that have been extracted from media coverage during COVID-19. Instead of extracting narratives using topic modeling, Bharghava uses a predefined list of "*73 narratives that potentially can affect financial markets via two channels. We start with the Journal of Economic Literature (JEL) Classification System, which was developed for use in the Journal of Economic Literature and is a standard method of classifying scholarly literature in the field of economics."* This paper also points to conversations and narratives as a useful predictors of asset returns, finding that *"investors can apply the insights from the narrative indicators to improve asset-allocation strategies. A narrative-based dynamic asset-allocation strategy significantly outperforms the equity-only, the bond-only, and the 50/50 equity/bond balance strategies."*

Borup et al. [4], like the previous examples also uses NLP to extract key terms relating to narratives and sentiment from text sourced from daily open-ended written surveys of investors during the COVID-19 pandemic. This paper finds that "*narratives contain predictive information for future excess stock and bond returns, and this predictability remains when controlling for contemporaneous information stemming from news and social media."* Once again, narratives seem to be informative in predicting returns in traditional markets, and one can imagine this dynamic would likewise apply to cryptocurrency markets.

Houlihan & Creamer [5] conducts sentiment analysis using NLP on investor text from the social forum StockTwits and juxtaposes this with stock options volume data. This paper finds that *"sentiment extracted from social media and market data are valid additional risk factors in relation to the Fama–French and Carhart models"* and that *"sentiment can be harnessed in a predictive analytics framework to realize positive residual alpha after adjusting for market effects."* This is further evidence suggesting that conversational metrics can be useful in anticipating asset prices. This paper also notes that *"Google query search volume is a strong predictor of future economic activity in various industries".* Also worth mentioning is the finding

that, aside from the substance of the messages that were studied, "*over 70% of the stocks exhibited statistically significant correlations between underlying volume and message volume."* This suggests that message volume, irrespective of message content, could be an important factor in price forecasting models. This paper will explore this notion further in relation to Reddit message volume and the price volatility of ETH.

Cobie [6], a prominent trader and public figure in the cryptocurrency community, gives further credence to the notion that conversations and narratives play an important role in cryptocurrency investing. He describes the phenomenon of narrative-driven investment in cryptocurrency markets as follows: "*Participating in crypto markets during the thrill stages of a bull-run is isomorphically more similar to playing a modern video game than it is to investing. Most competitive modern video games have an ever-evolving metagame. The metagame can be described as a subset of the game's basic strategy and rules which is required to play the game at a high level."* He proposes that simply understanding popular narratives early and the logic of why they are popular can lead to being successful in crypto investing, and anecdotally speaking this certainly seems to be the case.

Dierckx et al. [7] takes the approach of exploring narratives and their effects on traditional markets by extracting topics from financial news articles using LDA and demonstrates that quantified narratives *"are predictive of future movements in the CBOE Volatility Index for different time horizons."* In other words, narratives extracted from financial news articles can be leading indicators of future market volatility.

Azqueta-Gavaldón [8] also considers narratives from news articles but in relation to cryptocurrency prices. When considering *"narratives propagated by the media"* and cryptocurrency prices, he finds *"strong bi-directional causal relationships between narratives and cryptocurrency prices. That is, price dynamics influence the propagation of news articles describing the cryptocurrency phenomenon while, simultaneously, narratives influence price dynamics."* This is important because it suggests that narratives often feed off of price, and price

feeds into narrative so the relationship has a self-perpetuating mechanism, making it difficult to determine causation. Regardless, there certainly seems to be a strong relationship between narrative and cryptocurrency prices according to this study.

Bonaparte & Bernile [9] seek to study cryptocurrency regulation and its effect on cryptocurrency prices by constructing a *"Crypto Regulation Sentiment Index (CRSX)"* but finds that it has "*no statistically significant long-term impact on cryptocurrency prices"* yet it does have a *"large impact on cryptocurrency price volatility and trading volume."* This suggests that sentiment derived from narratives may not be the best gauge of long-term price forecasting, yet could still be useful in short-term modeling.

Harvey et al. [10] explores common valuation approaches of Bitcoin and finds *"none of these approaches are satisfactory"* in explaining the valuation of Bitcoin while emphasizing that valuing Bitcoin and other cryptocurrencies is inherently difficult due to the lack of fundamentals surrounding them. This aspect of cryptocurrencies obviates the need for reliable metrics to look to when attempting to predict future price movement, which this paper seeks to do.

Sabersky [11] examines narrative factors from social media and news articles to determine which types of information are most likely to move cryptocurrency prices. He uses AI to assign categories to news articles and determines that funding announcements and merger and acquisition announcements from news sources have the largest impact on cryptocurrency prices. He also examines the relationship between social media, specifically Twitter, and cryptocurrency prices and finds that tweets related to airdrop and listing announcements are strongly correlated with price movement.

Also vitally important is the finding that *"increases in tweets are very likely to correspond to increases in price. This is convincing evidence that tweet volume does actually have a unique effect on price after a Listing (and airdrop, and, to a small extent, partnership and staking)".* This concept of message volume being correlated with price will be one that this paper will explore in more detail.

Considering the preceding, this paper plans to accomplish two main objectives. First, to provide insight into the evolution of cryptocurrency conversations over time, and second, to examine the hypothesis that there exists a statistically significant relationship between the volume of Ethereum-related posts on social media and short term price volatility of ETH.

Data

Data associated with the Ethereum blockchain and its native token ETH have been selected as the primary focus of this paper for several reasons. Ethereum is one of the longest-running, most widely used and established blockchains in existence with over 1 billion transactions processed since inception and currently averages over 1 million transactions per day. Given these figures, Ethereum unsurprisingly garners a tremendous amount of attention and is among the most discussed blockchains aside from Bitcoin due to its status as the preeminent smart contracts platform. In fact Ethereum, first launched in July 2015, popularized smart contracts, the advent of which allowed for much more flexible and dynamic use cases than Bitcoin which primarily functions as a global monetary network and which does not feature smart contract capabilities. Because of its prominent place in the cryptocurrency space and the amount of discussion surrounding it, conclusions drawn from analyzing the Ethereum blockchain are likely also true of other blockchains, most of which are modeled after or inspired by the Ethereum blockchain. Therefore, the Ethereum blockchain and its native cryptocurrency ETH make for an excellent case study with the potential for broader implications for the rest of the cryptocurrency space.

The data that will serve as the foundation for the analysis in this paper consists of two main sources, the first of which is historical ETH price data. Because ETH trades on different centralized and decentralized exchanges, price will often vary slightly from one exchange to another at any given time, so it is important to distinguish which exchange the pricing data is derived from. In this case, this paper specifically uses the ETH-USDT daily price chart from

Binance, by far the world's largest cryptocurrency exchange by volume at the time of writing. The price chart itself is hosted by and downloaded from crypto data provider TheTie.io, and contains the following four categories of information: timestamp, price of ETH (in USDT), daily volume (in USDT), and Daily Volume Moving Average. It should be noted that USDT is what is known as a stablecoin, which is a cryptocurrency token that is designed to trade on par (1:1) with USD. The vast majority of cryptocurrency trades that trade against the dollar are conducted with USDT or another stablecoin as the trading pair as opposed to USD itself, and in practice there is essentially no difference between USD and USDT. This dataset contains daily price data from October 2017 to April 2023.

Because the stated objective of this paper is to examine the conversations surrounding cryptocurrencies and how these conversations relate to the volatility of the price of ETH, it is also necessary to obtain text data that provides a good representation of these conversations over time, both in their breadth and in their substance. Because we live in an era dominated by digital communications, social media sites such as Twitter, Facebook, and Reddit immediately spring to mind as excellent candidates from which to source cryptocurrency text data. Beyond the obvious fact that these social media sites have troves of text data and hundreds of millions of active users, the reason social media sites are great resources for an analysis of the type proposed is that users tend to offer their honest opinions on subjects in a manner that makes them highly reflective of the way people are discussing these topics in society more broadly. While any of these three social media platforms would doubtless provide ample data to work with, certain considerations must be taken into account when choosing a data source. Among the considerations, volume of data as well as availability and ease of data collection all come into play. After considering these factors, Reddit was determined to be the most reliable, accessible, and easily decipherable data source from which to draw.

For these reasons, the second dataset this paper utilizes is derived from Reddit, one the internet's most prominent and popular social media sites. Reddit is built around community

forums, known as subreddits, where people share their thoughts and opinions regarding specific areas of interest such as sports, politics, entertainment etc. Each area of interest has its own dedicated forum which makes sourcing relevant data much less complicated than sites like Twitter and Facebook, where data is not as neatly organized. Reddit users can create original posts in these forums (what Reddit calls submissions), and other users can interact by leaving comments and voting on their favorite posts, among other things. Data associated with these posts and comments is routinely aggregated and dumped in large monthly batches by the Pushshift.io API, a third-party API which focuses on archiving Reddit data. It is via these PushShift dumps that data for this paper is obtained, specifically from the following three Ethereum related subreddits: r/Ethereum, r/EthTrader and r/EthFinance. These three forums have more than 4 million subscribers combined at the time of writing, and consequently the dataset contains a wealth of text data related to the subject of interest, namely the Ethereum ecosystem.

Data was first collected individually from each of these three forums before being processed further. Each subreddit's data consists of two distinct dump files, one for submissions and another for comments, spanning each respective subreddit's history from the first post up until December 2022. It's important to note that while the vast majority of data from these subreddits is contained in these dumps it does not capture all associated forum data. For example, any post or comment that was created and subsequently deleted before monthly Pushshift API dumps would not be collected or included in the dump files. Still, the data is comprehensive and an excellent representation of each forum as a whole.

Reddit submissions and comments datasets vary somewhat in their construction as each contains its own distinct categories of metadata. The following table lists the metadata categories associated with Reddit comments. All data are string data unless otherwise noted in parentheses.

TABLE 1

Metadata Field (Comments)	Description
author	username of commenter
author flair text	optional description of user's forum membership status, etc.
body	the actual text of the comment
controversiality	Indicates whether the comment is deemed controversial (1) or not (0) (integer)
created utc	Universal Time Coordinated timestamp of the comment
distinguished	whether a post was deemed distinguished by a moderator or admin, possible values are "moderator" or "admin"
id	unique identifier for comment
link id	link id for the comment
parent id	id for the content which the comment is in reference to
score	total score for the comment, equal to upvotes minus downvotes (integer)
subreddit	name of the subreddit
subreddit id	unique identifier for the subreddit

Of these categories, the most relevant metadata for the purposes of this paper are *body*, which contains the main text of the comment, and *created utc*, which is a UTC timestamp indicating when the comment was posted, which allows us to segment the data by time period.

While Reddit submissions share some metadata categories in common with comments, they contain their own distinct set of categories. Table 2 details the metadata categories associated with submissions. All data are string data unless otherwise noted in parentheses.

TABLE 2

Metadata Field (Submissions)	Description
author	username of submission creator
author flair text	optional description of user's forum membership status, etc.
created utc	Universal Time Coordinated timestamp of the submission
id	unique identifier for comment

media	indicates the type of media included in the submission, if any
num comments	number of comments associated with submission (integer)
permalink	permanent url for the submission
score	total score for the comment, equal to upvotes minus downvotes (integer)
selftext	for text-based submissions, the actual text of the submission
subreddit	name of the subreddit
subreddit id	unique identifier for the subreddit
title	title of the submission
url	url to linked content if any

In order to keep the analysis manageable, it was necessary to limit the scope of the paper by imposing restrictions on the timeframe of study. This paper considers discussion from Reddit forums and daily ETH price data over the span of 3 years, from January 1, 2020 to December 31, 2022. This time period, while arbitrary, was chosen because it provides a good representation of market cycles and dynamics given that it encompasses an entire bull market cycle as well as a bear market cycle which crypto markets are currently in the midst of. After each individual dump was downloaded, all 6 datasets were combined into one unified dataset. This unified dataset contains all submissions and comments from 1/1/2020 to 12/31/2022 for all three subreddits, a total of 4,100,899 rows of data after dropping all data prior to 1/1/2020 in order to conform to the specified timeframe.

Table 3 presents the cumulative number of posts (submissions and comments) per quarter for all three Reddit forums mentioned above.

TABLE 3

Time Period	Number of Posts (Submissions & Comments)
1/1/2020-3/31/2020	135,577
4/1/2020-6/30/2020	104,521
7/1/2020-9/30/2020	171,850
10/1/2020-12/31/2020	180,694
1/1/2021-3/31/2021	486,001
4/1/2021-6/30/2021	791,370
7/1/2021-9/30/2021	673,955
10/1/2021-12/31/2021	716,113
1/1/2022-3/31/2022	298,300
4/1/2022-6/30/2022	247,711
7/1/2022-9/30/2022	178,503
10/1/2022-12/31/2022	116,304
Total	4,100,899

Number of combined posts (submissions and comments) per quarter for r/Ethereum, r/EthTrader & r/EthFinance

After unifying the datasets, a cleaning process was applied to remove posts that contained missing data. Specifically, a significant number of rows contained some metadata but were missing meaningful text data. The text in these rows had been deleted or removed at some point, perhaps by a moderator or by the posters themselves and contain only the words “[deleted]” or “[removed]” in the text field. Therefore, these rows were summarily removed from the dataset to aid with NLP analysis, and the total number of entries with meaningful text was reduced from 4,100,899 to 3,187,078.

Next a segmentation process was performed on the unified dataset, and numerous smaller datasets were derived from the main dataset and grouped according to time intervals based on their UTC timestamps, with the total number of segments denoted in parenthesis: Annual (3), Quarterly (12), and Monthly (36). Figure 1 plots the daily post volume from 1/1/20 to 12/31/2022.

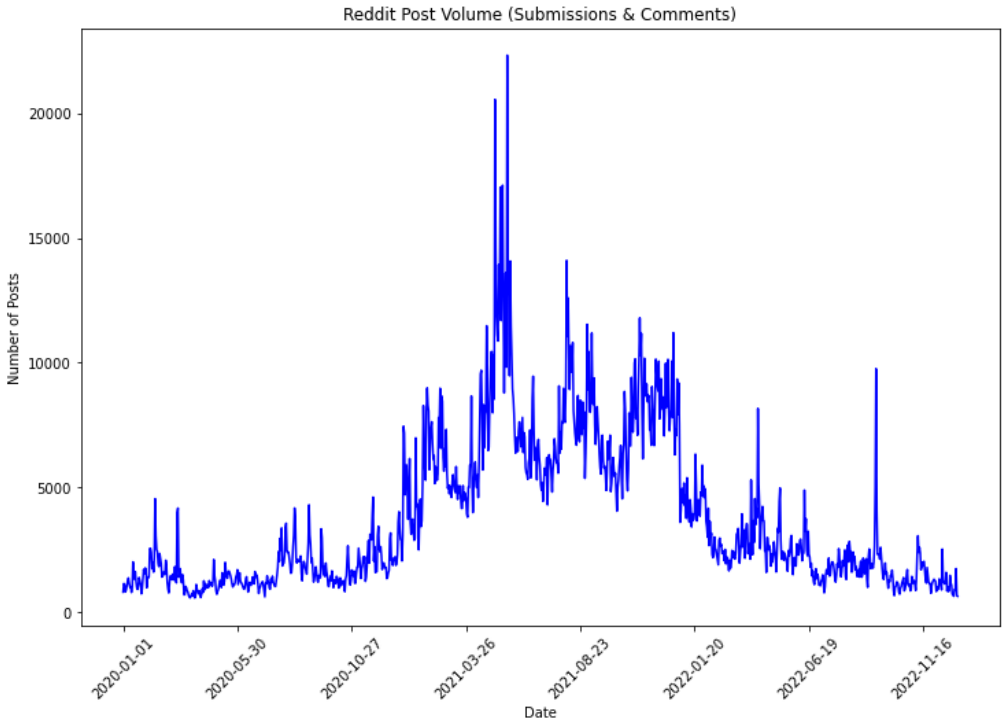

FIGURE 1

Daily post volume (submissions & comments) for r/Ethereum, r/EthTrader & r/EthFinance

These time-based segments are useful for comparing and contrasting the content of the conversations across time and for gaining insight into the evolution of popular topics, terms and phrases, all of which are essential components of the narratives that this paper argues play an important role in motivating the behavior of cryptocurrency market participants. The reasoning behind segmenting the unified dataset in this manner was to provide snapshots at different points in time, allowing for representations of conversations on different scales to hopefully provide some clarity as to which topics were driving the conversation at various points in time.

Methods

To help perform these comparisons and bring the picture into focus, NLP methods were applied to each of these segments. In order to maximize the effectiveness of these NLP methods further cleaning of the raw text data was first required, beginning with generating a new column of text from the original text where all of the text was converted to lower-case, and special characters were removed using Regex. Regex, which stands for Regular Expression, is a module built into Python and other programming languages that enables searching for patterns in text and allows for modifying or removing them altogether. Regex also was utilized for removing commonly occurring patterns that did not meaningfully contribute to the Reddit text data such as hyperlinks, which were removed from the raw text. After removal of these, manual inspection of the text was performed to identify and remove superfluous words and phrases, such as auto-generated text. One example of auto-generated text that occurred relatively frequently in the raw text data was: “this summary is auto generated by a bot and not meant to replace reading the original article.” This phrase, as the text indicates, was auto-generated as an addendum to some Reddit users’ posts, and could skew the text analysis in an undesirable manner if it weren’t removed. A few other phrases that were removed are: “your post was removed as it mentioned bitcoin in the title”; “posts that mention an alternative cryptocurrency are only permitted if they also mention ethereum or eth”; “hi, this comment is being automatically posted under your submission to facilitate the tallying of the pay2post donut penalty that r/ethtrader deducts from user donut earnings for the quantity of posts they submit”; “i am a bot, and this action was performed automatically. please [contact the moderators of this subreddit]”; and “tip this post”. These common phrases do not add value to the discussion, and hinder the effectiveness of text analysis methods by introducing unwanted noise into the data, thus necessitating their removal. Lastly, English stopwords were removed from the text using the NLTK (Natural Language Toolkit) Python package, which are words that commonly occur frequently in text but do not add much meaning. Examples of stopwords are “the”, “of”, “a”, “an”

etc. and removing them from a corpus (body of text) generally yields more coherent and better results when NLP and machine learning methods are subsequently applied.

Once this new column of cleaned text data was generated, Text Mining and NLP methods were applied. NLP, shorthand for Natural Language Processing, is a field of computer science that applies statistical and machine learning methods to written text in various ways to help make better sense of them as well as for predictive modeling. NLP is commonly used for tasks such as text summarization, topic extraction, and sentiment analysis to name just a few, and has taken on increasing importance as text data has grown exponentially in the internet age. The following flow diagram (Figure 2) illustrates the steps undertaken with respect to the Reddit data in this paper.

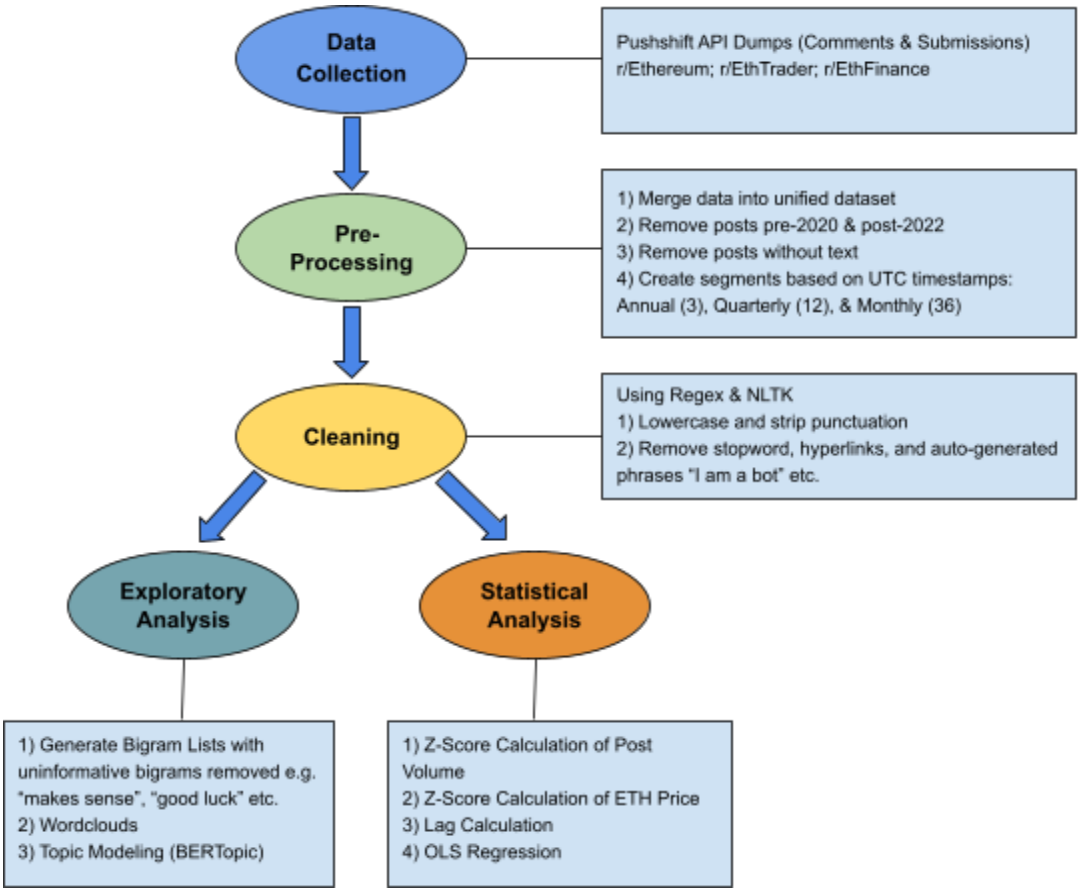

FIGURE 2

Flow-Chart Depicting Data Processing and Analysis

The first NLP method that was applied to the cleaned text column was tokenization, which is a text pre-processing step required for other NLP tasks. Tokenization is the process of breaking down a text into individual units, known as tokens, according to parameters that can be specified by the user and is useful for determining frequencies and importance of tokens within a corpus. Words, phrases, sentences, subwords, characters, and more can all be set as units of tokenization. Oftentimes words are chosen as the default tokenization unit, where each individual word becomes a token, and that is the method that was chosen for this paper. Word tokenization was performed on all text entries for each individual segment, and once the tokenization process was completed and tokens created, the structure and contents of the text for each segment was examined by way of an n-gram analysis.

N-grams is a method in NLP that breaks down a text into units of n-consecutive words, which allows for the detection of commonly occurring phrases or word pairings. Unigrams are simple one-word groupings, bigrams are two-word pairings, trigrams are three word pairings and so on. Take the following text for example: “I like to visit National Parks.” Bigrams for this text would be “I like”, “like to”, “to visit”, “visit National”, and “National Parks”. Trigrams for this text would be “I like to”, “like to visit”, “to visit National”, and “visit National Parks”. Considering these different units of word pairings can inform us about the context in which words appear and point to thematic and topical elements within a text. N-grams are often used for tasks like Named Entity Recognition, whereby identification of proper nouns within a text can be determined, such as the name of a person, organization, place of interest, etc. For example, the words “empire”, “state”, and “building” all have meanings of their own but taking trigrams into account yields “Empire State Building” which has entirely unique meaning as it refers to the famous New York City landmark.

For this text, unigrams, bigrams and trigrams were extracted from each segment separately and the most commonly occurring words/pairings (in the case of bigrams and trigrams) were tabulated. Of the three n-gram pairings, bigrams routinely yielded the most

enlightening results. Figure 3A shows a tabulation for the 30 most frequently occurring bigrams for each year while Figure 3B shows the 30 most frequent bigrams for each quarter in 2020. It should be noted that after extracting the bigrams for each segment, bigrams which do not provide meaningful context such as “lot people”, “makes sense” etc. were manually removed from Figures 3A-3B. As a result, only bigrams related to the Ethereum ecosystem were left for consideration.

Term Rank	2020	2021	2022
1	smart contract (3775)	gas fees (18380)	smart contract (4762)
2	market cap (3037)	eip 1559 (12013)	gas fees (3892)
3	eth btc (2924)	market cap (11838)	smart contracts (3198)
4	smart contracts (2779)	smart contracts (10789)	bear market (3087)
5	btc eth (2299)	smart contract (10495)	proof stake (2532)
6	bull run (2084)	buy eth (9167)	rocket pool (2506)
7	deposit contract (2066)	btc eth (8316)	post merge (2107)
8	32 eth (1960)	eth btc (8027)	market cap (2105)
9	eth price (1874)	bear market (7914)	beacon chain (1999)
10	gas fees (1797)	bull run (6321)	staked eth (1943)
11	buy eth (1764)	make money (6203)	hardware wallet (1828)
12	beacon chain (1743)	proof stake (5756)	32 eth (1725)
13	stock market (1678)	eth price (5295)	eth btc (1725)
14	gas prices (1498)	market conditions (5260)	open source (1716)
15	bull market (1487)	gas prices (5164)	ethereum short (1682)
16	eip 1559 (1435)	buy dip (4772)	removed ethereum (1675)
17	gas price (1418)	gas price (4762)	long message (1663)
18	use case (1361)	10 years (4749)	use cases (1599)
19	yield farming (1342)	store value (4577)	staking rewards (1599)
20	use cases (1217)	transaction fees (4533)	seed phrase (1568)
21	price eth (1195)	gas fee (4498)	btc eth (1544)
22	ethereum network (1150)	bull market (4471)	private key (1513)
23	make money (1145)	ethereum network (4423)	scam scam (1487)
24	store value (1101)	coinbase pro (4189)	make money (1468)
25	hardware wallet (1092)	32 eth (4185)	tornado cash (1452)
26	bear market (1078)	crypto market (3931)	eth price (1447)
27	private key (1027)	bought eth (3875)	buy eth (1440)
28	proof stake (1026)	price eth (3667)	use case (1415)
29	transaction fees (1018)	hardware wallet (3618)	ethereum network (1371)
30	open source (1006)	use cases (3601)	proof work (1365)

FIGURE 3A

Most common bigrams mentioned each year with the bigram frequency in parenthesis

Term Rank	Q1 2020	Q2 2020	Q3 2020	Q4 2020
1	smart contract (839)	smart contract (670)	smart contract (1168)	deposit contract (1175)
2	smart contracts (509)	eth btc (566)	market cap (982)	eth btc (900)
3	stock market (492)	market cap (562)	smart contracts (873)	smart contract (828)
4	eth btc (459)	smart contracts (426)	gas fees (864)	32 eth (788)
5	eth price (394)	stock market (378)	yield farming (832)	btc eth (781)
6	buy eth (391)	btc eth (371)	gas prices (831)	market cap (763)
7	bull market (386)	beacon chain (346)	eth btc (660)	smart contracts (681)
8	market cap (362)	bull run (323)	gas price (608)	beacon chain (617)
9	deposit contract (358)	eth price (305)	bull run (603)	bull run (535)
10	bull run (354)	use case (295)	btc eth (527)	eth price (451)
11	bear market (320)	buy eth (273)	eth price (489)	eip 1559 (439)
12	beacon chain (314)	use cases (264)	bull market (476)	buy eth (408)
13	use case (310)	32 eth (254)	buy eth (442)	hardware wallet (394)
14	btc eth (308)	gas limit (234)	eip 1559 (439)	proof stake (361)
15	price eth (306)	gas price (230)	32 eth (393)	eth staking (351)
16	eth dai (279)	price eth (230)	stock market (392)	gas fees (347)
17	use cases (268)	gas prices (225)	beacon chain (366)	bull market (340)
18	open source (238)	ethereum network (222)	use case (344)	staking rewards (336)
19	flash loans (234)	transaction fees (222)	transaction fees (339)	price action (322)
20	asic miners (233)	black thursday (222)	use cases (320)	private key (313)
21	32 eth (229)	eip 1559 (217)	bear market (310)	store value (301)
22	store value (229)	gas fees (203)	make money (308)	staked eth (281)
23	hardware wallet (227)	bitcoin ethereum (203)	gas costs (297)	ethereum network (272)
24	make money (226)	ethereum blockchain (199)	open source (291)	price eth (265)
25	hard fork (226)	open source (192)	impermanent loss (285)	make money (259)
26	ethereum network (223)	make money (191)	ethereum network (278)	eth staked (234)
27	defi saver (220)	rocket pool (189)	erc 20 (273)	bitcoin ethereum (231)
28	core devs (218)	store value (189)	gas fee (273)	eth locked (231)
29	gas price (212)	ethereum price (184)	l2 solutions (264)	10 years (227)
30	liquidation price (200)	private key (183)	omg network (260)	bear market (227)

FIGURE 3B

Most common bigrams mentioned in each quarter of 2020 with the bigram frequency in parenthesis

These bigram lists provide good context surrounding the most popular topics of conversation during the time periods in question. First, with reference to the annual bigrams chart, the term “gas fees”, while a popular topic of conversation in 2020 (number 10 most frequent bigram), completely dominated the conversation in 2021 (number 1 most frequent bigram). EIP 1559 (a proposal aimed at addressing gas fees) also jumped dramatically from 16th most frequent to 2nd most frequent bigram between 2020 and 2021. Also notable is the

increase in “proof stake” which references Ethereum’s shift from a proof of work to a proof of stake model for validating transactions, ascended from 12th most frequent to 5th most frequent bigram between 2021 and 2022. Similarly, examining the quarterly bigram rankings (Figure 3B) one can get a sense for the growth of certain topics or narratives from quarter to quarter. For example, “32 eth” - a reference to the amount of ETH required to become a proof of stake transaction validator - sees strong quarterly growth, ranking 21st in Q1, 13th in Q2, 15th in Q3, and 4th in Q4. Similar insights regarding other terms can be drawn by examining these charts but this paper leaves the discussion here for now. Interested readers can find monthly bigram rankings at github.com/mcm711/Reddit-NLP-Project/tree/main/Visualizations/Term%20Rankings

Another useful method for examining the evolution of conversations and highlighting key terms and phrases at different periods are Wordclouds, which were generated using the Wordcloud python package. Wordclouds are a common method of visualizing large text data, with the most frequently occurring words appearing larger than less frequent terms. Wordcloud Figures 4A-4C demonstrate the key phrases present during the three annual periods that this paper is focused on.

FIGURE 4A

2020 Annual Wordcloud (Bigrams)

FIGURE 4B

2021 Annual Wordcloud (Bigrams)

FIGURE 4C

2022 Annual Wordcloud (Bigrams)

A few features from the annual Wordclouds are worth highlighting. Several terms grew substantially relative to other terms between 2020 and 2021, and we find that some of the biggest gainers once again were “gas fees”, “eip 1559”, and “bear market”, which were quite small relative to other words in 2020, yet became prominent topics of discussion during 2021. For those who pay close attention to cryptocurrency markets, the growth in these terms should come as no surprise. As noted earlier, gas fees are the fees that users pay to have their transactions processed, and they fluctuate based on demand. During 2021, gas fees became quite expensive as demand for transaction processing peaked due to Ethereum’s increasing popularity, and EIP 1559 as mentioned earlier, is the name of a formal proposal aimed at addressing the high cost of gas fees on the Ethereum network that was implemented in August of 2021. Bear market also became a popular topic of conversation in 2021 as there were signs the market was becoming overheated and many were speculating that a bear market was on the horizon, a concern that came to fruition in 2022 as markets experienced major declines. In 2022 the most notable topics that came into prominence compared to 2021 were related to Ethereum’s switch from Proof of Work to Proof of Stake consensus which was a monumental shift in the way that Ethereum transactions were processed and validated. This event is commonly referred to as “The Merge”, when the Beacon Chain (proof-of-stake chain) merged with Ethereum’s Mainnet (the blockchain where Ethereum transactions are officially processed and recorded). The importance of The Merge is evidenced by the terms “proof stake”, “staked eth”, “32 eth”, and “post merge”. Again these topics should come as no surprise to anyone who follows the space closely and serve as confirmation that conversations on Reddit are relevant to the broader dialogue that takes place regarding Ethereum.

Clearly, Wordclouds can be useful tools for summarizing important words and phrases of social media conversations, and comparing Wordclouds over time can give a good sense of

how those conversations have shifted. This process can be replicated over shorter periods of time to get a sense for how topics and conversations are changing from quarter to quarter, month to month, or week to week; this paper has not done so for sake of brevity. But clearly these representations have value in that they allow us to organize a vast amount of text information that would be very hard to make sense of otherwise. Readers can find monthly Wordclouds at github.com/mcm711/Reddit-NLP-Project/tree/main/Visualizations/Wordclouds

Next, topic modeling was performed on each segment in an attempt to categorize and further contextualize Reddit conversations over time. Topic modeling is a branch of NLP that utilizes machine learning to extract topics from documents and attempts to group documents according to commonalities in the substance and verbiage of the written material. While there are a number of topic modeling techniques and packages that are widely used for topic modeling, particularly Latent Dirichlet Allocation (LDA), this paper utilizes BERTopic [12] topic modeling due to its flexible nature and the powerful language model it's built around. Before proceeding further, the process of topic modeling warrants explanation.

Topic modeling of all varieties involves a few common steps. First, corpuses are transformed into vector representations of the words and the frequencies with which they appear and the vectors combined into what is referred to as a document-term matrix. Because documents often contain tens of thousands of words or more dimensionality becomes an issue, and for this reason dimensionality reduction methods such as Principal Component Analysis (PCA) or Uniform Manifold Approximation and Projection (UMAP) are generally applied, along with embeddings, which are low dimensional vector representations of words that are useful in training and fitting NLP models. Corpuses can either be pre-processed by removing stopwords for example or left in their raw form before undergoing transformation into embeddings (pretrained embeddings can also be used), which can affect the results of the topic modeling depending on which model is used. Next, topic models are then trained and fitted, with each model utilizing its own clustering techniques and methods for determining topics. Lastly, once

models are fitted, topics are then assigned to each respective corpus. Since this paper uses BERTopic for topic modeling, an overview of BERTopic follows.

BERTopic is a model that uses BERT as its pre-trained language model. BERT stands for Bidirectional Encoder Representations from Transformers, and as the name suggests was developed using machine learning models known as Transformers. BERT is specifically pre-trained for natural language processing, and therefore performs very well for NLP tasks when compared to other language models.

BERTopic requires a list of documents/corpuses for its input, and tends to yield better results with documents that are left in their original form and not cleaned or pre-processed. A document-term matrix is then created which uses Term Frequency-Inverse Document Frequency (TF-IDF) to weight and normalize each vector in the document-term matrix. This TF-IDF weighted document-term matrix is then used to create a document-similarity matrix based on the cosine similarity of the document-term matrix vectors. The following formulas define how TF-IDF [13] and cosine similarity [14] are calculated.

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$
$$IDF(t) = log\frac{N}{1+df}$$
$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

[13]

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

[14]

Dimensionality reduction is applied at this stage, with UMAP being the default method, although BERTopic can accommodate other dimensionality reduction methods. The next step in the

process is clustering. BERTopic's default clustering algorithm is Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), which uses density of data to create clusters and has a mechanism which identifies and excludes noisy data points that don't belong in density-based clusters. Finally, once clusters are defined BERTopic assigns topics based on c-TF-IDF (class-based TF-IDF) to assign importance to words within each topic cluster and returns the most important words for each topic label.

The end result is human-readable topic labels which are assigned to each document. BERTopic allows for fine-tuning of almost all parameters during each stage of the process, including reducing the number of topic labels to a desired number. It also easily allows for users to visualize topics in a variety of ways. Figure 5 depicts the BERTopic modeling process [15].

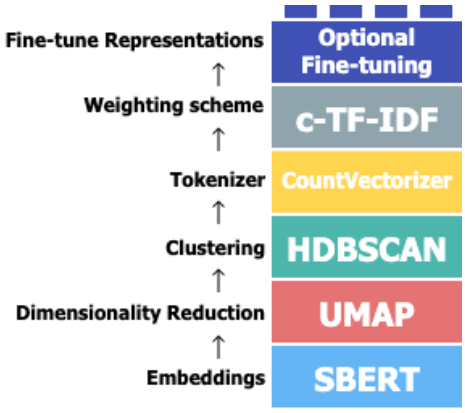

FIGURE 5

Stages of the BERTopic Modeling Process

Results of Quarterly BERTopic Topic Modeling for the year 2020 are presented in Figures 6A-6D. Each of these figures lists the 12 most common topics for that quarter as determined by BERTopic, with a Topic number assigned and a list of representative words that are most frequently occurring within that topic appearing on the left hand side of each chart.

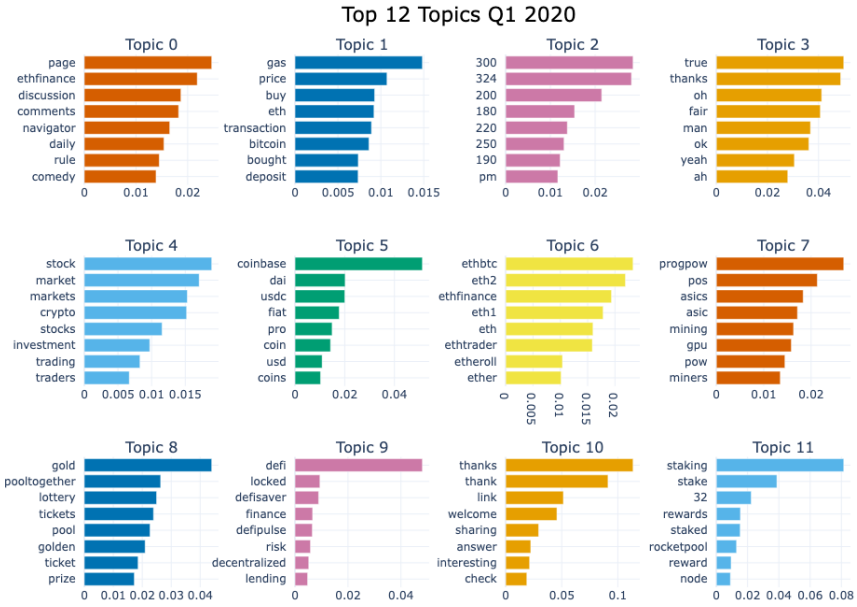

FIGURE 6A

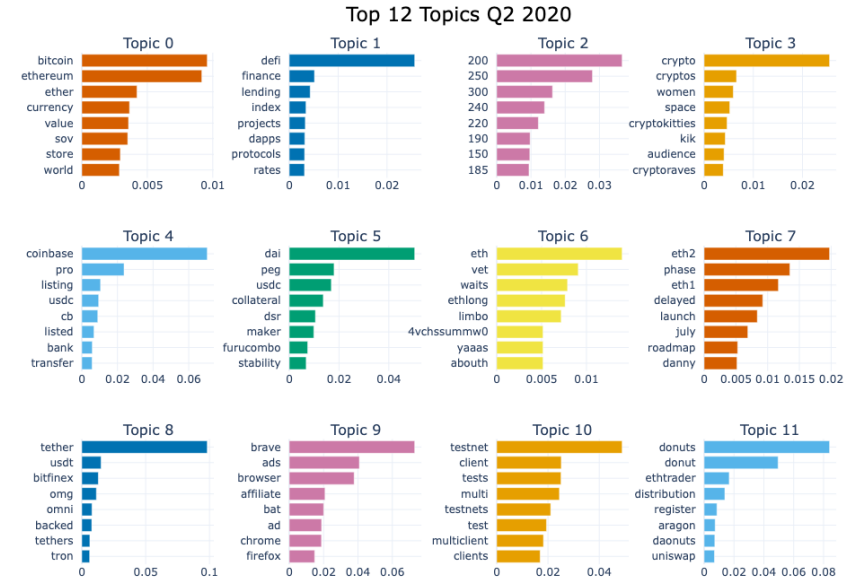

FIGURE 6B

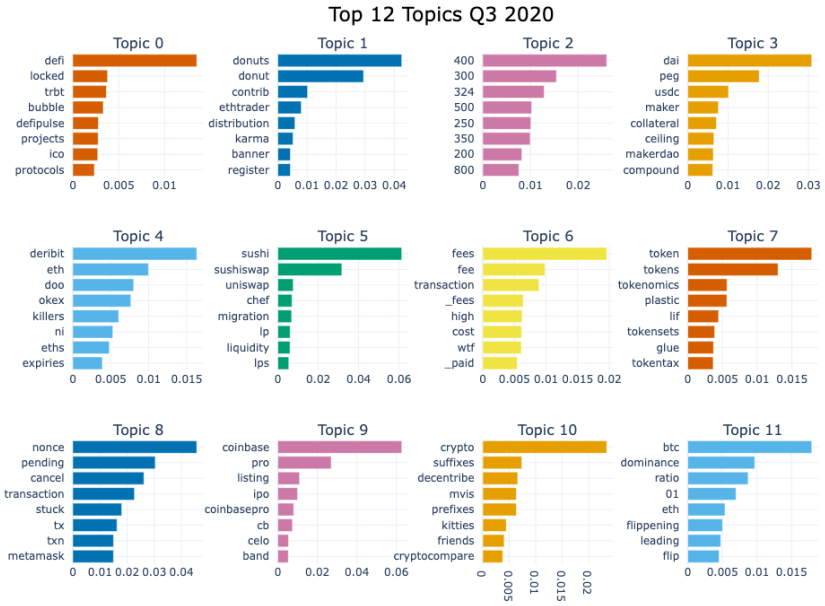

FIGURE 6C

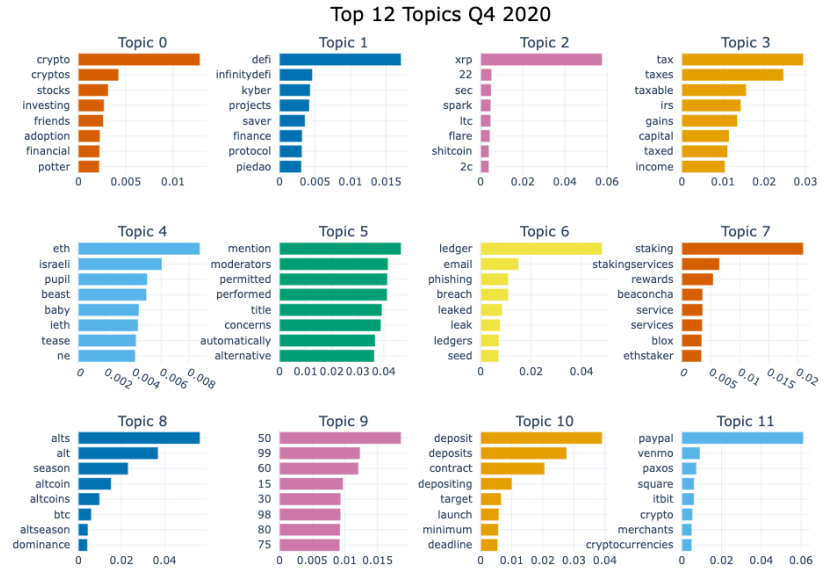

FIGURE 6D

A few notable observations can be seen from these figures. First, while BERTopic, like many topic modeling methods, is imperfect, it does a fairly good job of providing a picture of

what people are discussing in aggregate during a given period of time. Granted, there are topics like Topic 3 of Figure 6A (topic words: “true”, “thanks”, “oh”, “fair”, etc.) which are uninformative as they are likely based on a grouping of short comments of affirmation and are not providing much substance to the conversation, however, BERTopic does an admirable job of extracting and summarizing the most frequently discussed topics as well. For example Topic 1 of 6A shows that discussions of gas prices were at the forefront of conversation in early 2020, we can see that discussions of DeFi (which is industry shorthand for Decentralized Finance) exploded in Q2 2020, jumping from the 9th most discussed topic in Q1 to the 2nd most discussed topic in Q2, to the most frequently discussed topic in Q3. Other topics such as those related to “staking” and “fees” ebb and flow from quarter to quarter. Also of note are topics that only appear in a single segment, which can help give clues as to certain brief yet important events that occurred during a period of time. For example, Topic 8 in Figure 6C (topic words: “nonce”, “pending”, “cancel”, “transaction”, “stuck”, “tx”, “txn, “metamask”) are all terms related to transaction difficulties, seeming to indicate that many users had issues with pending/stuck transactions during Q3 2020 which can often happen when gas fees are extraordinarily high. Looking into why these problematic transactions occurred is beyond the scope of this paper but could be an interesting topic for further exploration. We can also safely assume that if high gas costs were responsible for transaction problems then gas fee discussions would be prominently featured during this quarter as well, and in fact Topic 6 confirms this to be the case. Comparisons of topics, and shifts in topics, could be wide ranging and the subject of an entire paper in and of itself, but we will leave the discussion here for now.

While the preceding exploratory analysis is informative of the general topics of conversation and the evolution of those conversations over time, it does not directly address the question of whether conversations are indicative of future price volatility and if they are, to what extent this is measurable. In order to answer this question directly we will test whether metrics derived from Reddit conversations related to Ethereum can be correlated with the change in the

volatility of ETH. In so doing, we follow the lead of Sabersky and others who have suggested a relationship between message volume and price fluctuation exists. Specifically, we will seek to test whether a change in Reddit Ethereum message volume is correlated with the short term volatility of ETH, which we will do by calculating and comparing Z-scores for each category of data based on 10-day moving averages. In this case Z-scores, which are a measurement of how many standard deviations a value is from a mean or expected value (the 10-day moving average), are a proxy for volatility. A Z-score of 0 indicates that the recorded value and the expected (mean) value are equivalent, while a large Z-score means there is a large deviation between recorded value and expected value and therefore indicative of volatility. The 10-day lookback period was chosen as opposed to a shorter window so that an adequate number of samples could be factored into the calculation to ideally provide an accurate average value.

The Z-Score calculation itself is rather straightforward and is derived using the following formula [15]:

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score
x = observed value
μ = mean of the sample
σ = standard deviation of the sample

[15]

In this case when calculating the Z-score for price, x is equal to the percent change in price from the previous day, μ is equal to the average change in price over the previous 10 days, and σ is equal to the standard deviation of the change in price over the previous 10 days. We perform the same calculation to get the Z-score for posts but substitute post volume for price so that x is equal to the percent change in post volume from the previous day, μ equal to the average change in post volume over the previous 10 days, and σ is equal to the standard deviation of the change in post volume over the previous 10 days. Because we are interested in

the relationship between change in post volume and the change in price of ETH irrespective of direction, we take the absolute value of each calculation. This is to account for instances where, for example, there is a sudden increase in message volume occurring shortly before a sudden decrease in ETH price, as happened in March 2020 during the early days of COVID. In such cases, there appears to be a relationship between post volume and price volatility in terms of magnitude but the direction of the changes are opposite one another. Therefore, considering the absolute value of the changes will give us a clearer picture of the general effect that changes in post volume have on price volatility.

The end result of these Z-score calculations is a dataframe with 1096 rows containing daily Z-score values for both variables spanning 1/1/20 to 12/31/22. Because we are using a 10-day lookback window, Z-score calculations are equal to 0 for the first 10 days and are calculated for the remaining 1086 days.

Results

A preliminary step in evaluating the relationship between these two variables is to examine some graphs to see whether the hypothesized relationship appears to exist. Figures 7A-7G present ETH price Z-scores vs Reddit post volume Z-scores over different time frames, from broad to more granular. Figure 7A for instance, plots the absolute values of Z-scores over the entire three year period, while Figures 7B-7D graph the relationship on an annual basis, and 7E-7G are quarterly graphs. Z-scores for price are shown by the red lines and Z-scores for post volume are shown in blue. If a relationship exists between the variables as we anticipate, we should observe a mirroring effect between the blue and red lines separated by some offset value. In theory, the closer the relationship the more the two lines should mirror one another.

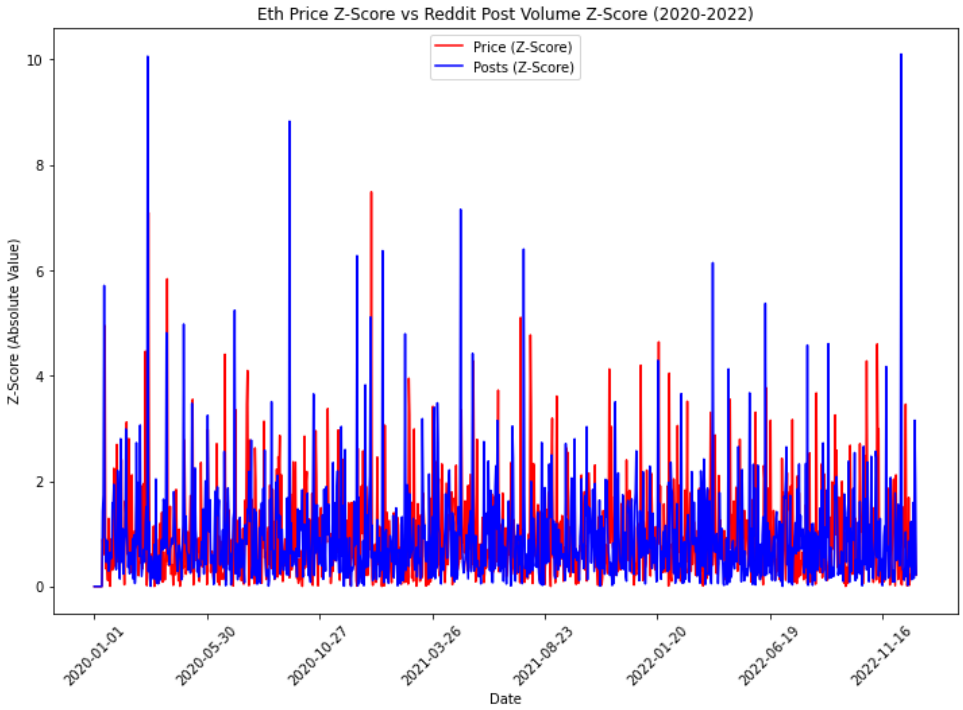

FIGURE 7A

2020-2022 ETH Price Z-Scores vs Post Volume Z-Scores

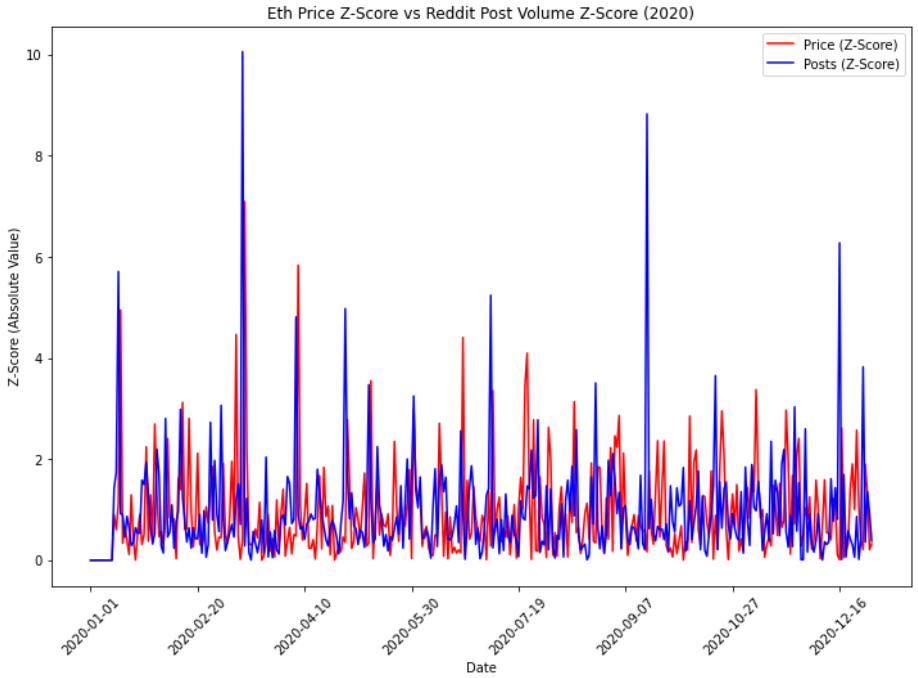

FIGURE 7B

2020 ETH Price Z-Scores vs Post Volume Z-Scores

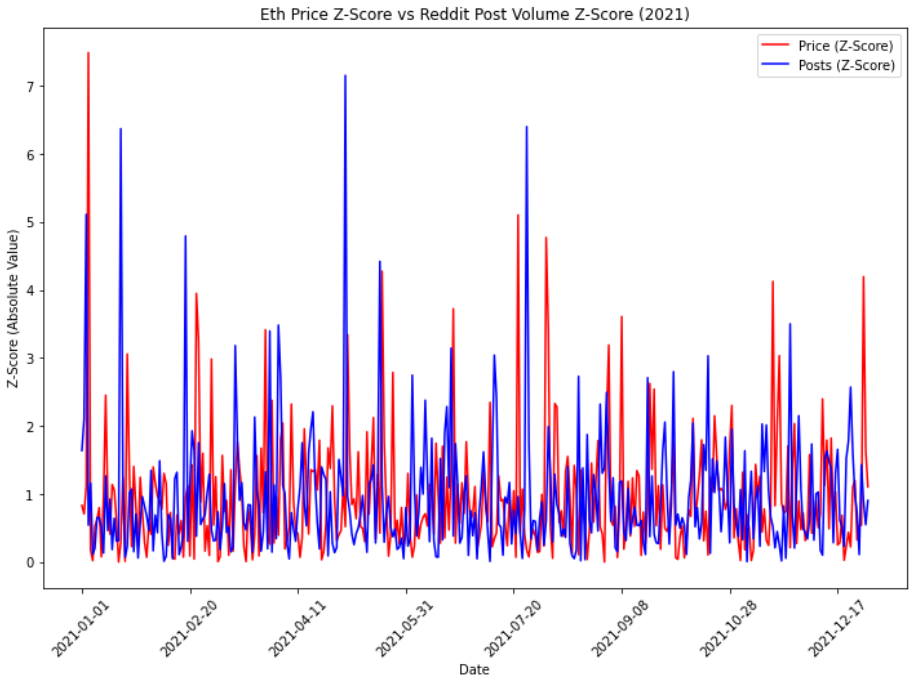

FIGURE 7C

2021 ETH Price Z-Scores vs Post Volume Z-Scores

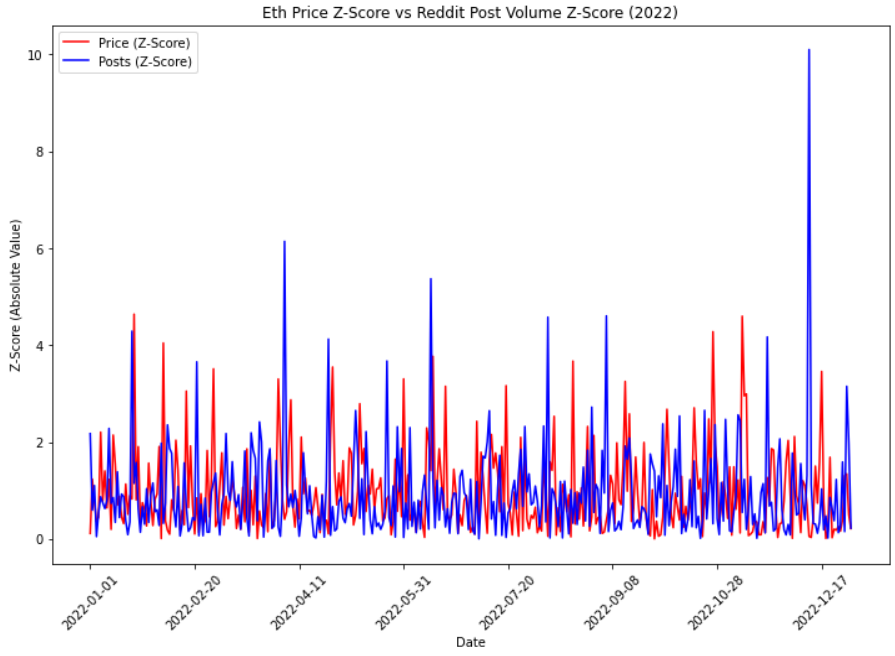

FIGURE 7D

2022 ETH Price Z-Scores vs Post Volume Z-Scores

FIGURE 7E

Quarterly ETH Price Z-Scores vs Reddit Post Volume Z-Scores (2020)

FIGURE 7F

Quarterly ETH Price Z-Scores vs Reddit Post Volume Z-Scores (2021)

FIGURE 7G

Quarterly ETH Price Z-Scores vs Reddit Post Volume Z-Scores (2022)

Upon examination, the graphs of larger time frames certainly hint that a relationship exists between the two variables, and when we look at the quarterly graphs the relationship looks to be unmistakable. It appears that the shapes of the post volume (blue) lines and the price (red) lines are of similar shape and magnitude with a slight lag separating the two, and fluctuations in post volume tend to precede fluctuations in price. While this does not hold true in

all cases (there are instances where price fluctuations precede fluctuations in post volume) it looks to be generally true, and this aligns with our expectations. The question then becomes, if a relationship exists as it appears to from the graphs, what is the correlation of the relationship and how significant is it in statistical terms.

To answer this, we first determine the lag between the two sets of variables by running *correlation_lags* from scipy's *signal* package in Python which returns a value of 1 unit, indicating that the optimal lag between the two variables is 1 day. We then offset the Z-score for ETH price by 1 day to align the two columns of data, and then run an Ordinary Least Squares Regression to evaluate the relationship in statistical terms. The OLS results are as follows:

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.088
Model:	OLS	**Adj. R-squared:**	0.087
Method:	Least Squares	**F-statistic:**	105.9
Date:	Thu, 04 May 2023	**Prob (F-statistic):**	8.85e-24
Time:	15:16:03	**Log-Likelihood:**	-1398.6
No. Observations:	1096	**AIC:**	2801.
Df Residuals:	1094	**BIC:**	2811.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.6925	0.036	19.236	0.000	0.622	0.763
zposts_absolute_value	0.2683	0.026	10.293	0.000	0.217	0.319

Omnibus:	362.613	**Durbin-Watson:**	1.917
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1274.511
Skew:	1.591	**Prob(JB):**	1.75e-277
Kurtosis:	7.216	**Cond. No.**	2.49

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The results of this univariate linear regression confirms that there is a highly statistically significant relationship between the Post Volume Z-Score and the ETH Price Z-Score (factoring

in the 1-day lag), with a t-value of 10.293 and a p-value < .001, meaning that this result is almost certainly not the result of chance. What the coefficient for the Post Volume Z-score tells us is that for a 1 point increase in the absolute value of the Post Volume Z-Score, there is a corresponding .268 point increase on average in the absolute value of the ETH Price Z-Score. Furthermore, the R-squared of .088 indicates that 8.8% of the variation in ETH Price Z-scores is explained by Post Volume Z-scores. This result aligns with what the graphs seem to indicate, and taken in tandem, is strong evidence in favor of our hypothesis that there is a statistically significant relationship between post volume in Reddit forums and price volatility of ETH.

It is important to reiterate that this finding relies on some preconditions and assumptions. The first precondition is that this result is derived from Z-Score calculations based on a 10-day lookback period. While similar results might be observed if a different lookback window were to be chosen, it is also quite possible they would not. It also assumes that this relationship between post volume and price volatility will continue to be present when sourcing from the same dataset going forward. However, it's always possible that Reddit user behavior or the dynamics that influence the volatility of ETH may begin to change, in which case this finding may not hold true or the strength of the relationship may begin to weaken moving forward. With these assumptions and preconditions taken into account however, we can say unequivocally that the hypothesized relationship is present and is highly statistically significant.

Discussion

The key implication of this finding is that it suggests that there is value to be gained by evaluating cryptocurrency conversations, as they can be potential indicators of future price volatility. Given this result, it's not a stretch to argue that conversations could be driving price action in many cases, although causation is always difficult to definitively determine. This result furthers the notions that the field of narrative finance puts forth, and suggests further explorations of this type are warranted.

Before discussing future research, it is important to address the limitations of the conclusions that can be drawn from this study. Although Reddit data is both plentiful and comprehensive, like any dataset it is far from perfect. The biggest potential drawbacks with relying solely on Reddit data are that online community forums are self-selecting, and may not be representative of the broader conversation. This could potentially lead to bias, since people who are motivated to join and participate in certain Reddit forums may be more homogenous than the population at large. For instance, people who join cryptocurrency forums may tend to be more educated on the technology, more bullish on the markets, and possess more favorable attitudes and opinions toward cryptocurrencies than an average person chosen at random. Furthermore, the variety of topics discussed in these communities may be artificially constrained, as certain topics may tend to be more favored by certain groups than other groups. Additionally, each forum has certain rules and norms that users are expected to abide by, limiting what types of discussions and posts are allowed. For example in r/EthTrader, conversations about Bitcoin that do not also include a discussion of Ethereum are banned and removed. This paper has attempted to mitigate these issues by using not one but three separate Reddit forums, which theoretically should provide a more diverse population of users and discussion topics than a single forum would. To better avoid these issues, a platform like Twitter could also be considered alongside Reddit as Twitter may provide a better representation of the general moods and opinions surrounding cryptocurrencies at any given time. For the purposes of this paper, Twitter was not a viable option because of data collection restrictions imposed on its API.

These aforementioned limitations mean that it would be unwise to draw overly broad conclusions from this paper. However, being conscious of these limitations does not prevent us from drawing impactful conclusions from the data nonetheless. Even if the relationship between post volume and ETH volatility do not apply more broadly, it is entirely valid and quite valuable to say that Reddit conversations from these forums are highly correlated with ETH price

volatility. In other words, finding a strong and reliable pattern from a data source that has limitations does not diminish the importance of the pattern or relationship. What we can state definitively is that further exploration of cryptocurrency conversations on social media platforms is warranted, especially considering the lack of traditional methods for evaluating cryptocurrencies as extensively detailed in the introduction and literature review sections.

Further research and exploration could go in a multitude of directions. It may be worth developing metrics via the text mining and NLP methods utilized in this paper, perhaps by calculating the relative proportions and growth rates from segment to segment of key terms, phrases, or topics for further time-based quantitative comparisons. One could also apply the same techniques to a more diverse or larger set of data, or to other cryptocurrencies, to see whether the results hold true more broadly, or to determine which cryptocurrency forums are most reliably correlated with price volatility. Also, because this paper focuses on the magnitude of price volatility without considering the direction of the movement, the next logical step would be to evaluate factors that are predictive of the direction of price movement. Sentiment analysis of the text data may be useful in this regard. Additionally, if exploration leads to the discovery of other repeatable signals and patterns, these could be effectively incorporated into price forecasting models. What is clear is that more work can be done as it pertains to examining cryptocurrency conversations, and that the type of analysis this paper explores may be a step in the right direction in terms of developing frameworks for understanding cryptocurrency market dynamics and behavior of cryptocurrency market participants.

Conclusion

This paper examines social media posts related to the Ethereum blockchain and the cryptocurrency ETH in an attempt to develop an understanding of popular narratives and conversations over time. The lack of traditional metrics for valuing cryptocurrencies is discussed and the concept of narrative economics and its potential usefulness as it pertains to

cryptocurrencies is detailed. This paper argues that understanding cryptocurrency narratives could be useful in explaining cryptocurrency market dynamics and predicting price volatility. Text mining, NLP, and machine learning methods are applied to distill text data from over 4 million posts sourced from three popular Ethereum Reddit forums into meaningful insights. Posts were first collected and cleaned from each forum and then combined into one large dataset before being split into time-based segments for exploratory data analysis. The most frequent bigrams for different time segments were extracted and ranked, and Wordclouds were created to compare and contrast conversational topics over time. BERTopic was used for topic modeling for further insight into how conversations have shifted and evolved over time. Each of these exploratory methods yielded useful insights and were adequate in helping summarize and contextualize topics of conversation that would be otherwise difficult to definitively determine.

Considering the lack of traditional metrics and corresponding methods for assessing the value of cryptocurrencies like ETH, this paper also examines whether metrics related to conversations can be used to predict price movement. This paper specifically tests whether changes in post volume could be useful in predicting short term ETH price volatility. Daily post volume data was derived from the main Reddit dataset while daily ETH price information was sourced from TheTie.io, and Z-scores for both variables were calculated based on a 10-day lookback period, and the absolute values compared. After an adjustment for lag, the results of an OLS regression show that there is a highly statistically significant relationship between changes in Z-scores of post volume and ETH price. This finding seems to substantiate the notion that metrics derived from social media forums like Reddit can be useful indicators of ETH price volatility, and suggest that further exploration in this area may yield similarly promising results.

References Cited

1. Shiller, Robert J. “Narrative Economics: How Stories Go Viral and Drive Major Economic Events” Princeton University Press (2019)
2. Bowden et al. Journal of International Financial Markets, Institutions & Money, “Sentiment and trading decisions in an ambiguous environment: A study on cryptocurrency trader” 80, 101622 (2022)
3. Bhargava et al. “Quantifying Narratives and Their Impact on Financial Markets” Journal of Portfolio Management (2023)
4. Borup et al. “Quantifying Investor Narratives and Their Role during COVID-19”, SSRN (2020)
5. Houlihan & Creamer, 2017. "Can Sentiment Analysis and Options Volume Anticipate Future Returns?," *Computational Economics*, Springer; Society for Computational Economics, vol. 50(4), pages 669-685, December
6. Cobie, “Trading the Metagame”, *Substack* (2021)
7. Dierckx et al. (2021). “Quantifying News Narratives to Predict Movements in Market Risk.” 10.1007/978-3-030-66891-4_12
8. Azqueta-Gavaldón, Andrés, "Causal inference between cryptocurrency narratives and prices: Evidence from a complex dynamic ecosystem", Physica A: Statistical Mechanics and its Applications, Volume 537, 2020, 122574, ISSN 0378-4371
9. Bonaparte & Bernile, “A new “Wall Street Darling?” effects of regulation sentiment in cryptocurrency markets”, Finance Research Letters, Volume 52, 2023, 103376, ISSN 1544-6123
10. Harvey et al. “An Investor’s Guide to Crypto” (September 1, 2022).
11. Sabersky, Erik, “What Actually Moves the Prices of Cryptocurrencies?” The Tie (February, 11, 2021)

12. Grootendorst, Maarten, “Neural topic modeling with a class-based TF-IDF” arXiv 2203.05794 (2022)
13. Chen, Kinder “Introduction to Natural Language Processing - TF-IDF”, Medium (2021) https://kinder-chen.medium.com/introduction-to-natural-language-processing-tf-idf-1507e907c19
14. Varun, “Cosine similarity: How does it measure the similarity, Maths behind and usage in Python”, Towards Data Science (2020) https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db
15. Z-Score Formula https://www.ztable.net/z-score/