FIRST DRAFT VERSION: Published on 10th of January 2021

Sentiment Analysis on the Fundamental Drivers of Cryptoasset Value

Dr. Lilian Wazanre and Kazeem Adio

10th of January 2021

Abstract

This paper presents the Sentiments Analysis carried out on drivers of the intrinsic value of cryptoassets. We refer to these value drivers as Real Digital Asset (RDA) Attributes. We begin by collecting thousands of cryptoasset tweets and news article per RDA attributes and labelled them according to their positive, negative, neutral, mixed sentiments. The resulting data was used to train an RDA Attribute sentiment classifier with FastText embeddings. The classifier achieves a weighted score of 87%. Using our RDA Attribute sentiments classifier, we determine on an 0 to 1 point scale - the degree of positivity of sentiments per attribute per cryptoasset.

1 Introduction

There are currently over 8,000 crypotseests traded on the circa 34,000 by nearly 250 million active users. Since the inception of Bitcoin in 2009, cryptoassets have been flaunted as an alternative to the traditional fiat money. Most cryptoassets are decentralised, meaning that they are controlled and regulated by a network of computers all over the world who have no affiliation with one another. Therefore, no person, company, bank, or government can control them. There are two types of decentralised cryptoasset networks, permissioned and permissionless. Permissioned network are restricted by policy to a specific group of participants, while permissionless networks such as Bitcoin, Litecoin, Ethereum, etc are open to anyone to participate thereby making them borderless. Despite their inherent libertarian principles and innovative use cases, behaviour consistent with speculative trading accounts for the majority of use cases for cryptoassets. This has created risks for users and investors at all levels. A challenging question for most users of cryptoassets is what is their intrinsic value?

We consider a response with the Friedman notion that users of cryptoassets value the ability to engage in trustless, decentralised, censorship resistance, transactions. Trivially we derive several other attributes that a cryptoasset must have in order to be of value to users. These attributes include decentralisation, good network technology, peer-to-peer transaction, etc. We refer to these value drivers as Real Digital Asset (RDA) Attributes. The RDA Index methodology aggregates these RDA Attributes into four groups, namely: Business Ecosystem Stability (E), Digital Utility (U), Technology Efficiency (T), and Sentiments (S) on the first three core attributes E,U,T. Detailed analysis of the RDA Attributes is outside the scope of this paper. We present a broad view of the attributes as follows:

1.1 Real Digital Asset Index Attributes

RDA attributes are found more in some cryptoassets than in others which enable those cryptoassets to attract and retain value.

1.1.1 Business Ecosystem Stability (E)

The Business Ecosystem Stability component represents resilience and the capability to maintain a stable state of socio-economic equilibrium that keeps the cryptoasset alive and valuable. This group of attributes reflects the value brought to the token by the strength of the team, community, investors and the influence of competitors and regulators.

1.1.2 Digital Utility (U)

Much of the speculation around the value of digital assets stems from their promise as investment vehicles, and this area is subject to intense regulatory scrutiny from securities agencies. However, for other tokens, at least a proportion of their value is intrinsic to their actual use case within a payment or software application ecosystem. A pure utility token may still rise in value if the ecosystem for which it is developed undergoes growth and demand increases, however under this group of attributes we examine the extent to which the token has inherent value due to its use cases both on-chain and off-chain.

1.1.3 Technology Efficiency (T)

All cryptocurrencies may be regarded as innovative technologies, however with a 13 year history now to draw upon, different technologies may be regarded as having distinguishable track records of effectiveness. As a range of different consensus mechanisms and mainnets proliferate at this time, we compare a diverse range of factors to yield evaluations of how efficient and secure are the specific technologies in use.

1.1.4 Sentiments (S)

The Sentiments attributes seeks to measure the level of sentiments that persists in favour of the asset's fundamentals i.e. ecosystem stability, utility, and the underlying technology. In an era when hype has had far too much influence on cryptoasset valuation and pricing, RDAi is keen to ensure sentiment is viewed with the appropriate lenses as another important factor alongside the preceding groups of core RDA attributes. The role of sentiment on the intrinsic value of cryptoassets cannot be downplayed particularly in the social endorsement of an asset's intrinsic worth.

1.2 Calculating Intrinsic Value of a Cryptoasset

From the above, we argue that the existence of RDA Attributes in a cryptoasset determines its reliability as a store of value and consequently its use as a currency. To model the intrinsic value of a cryptoasset, A, we specify a mapping $(E, U, T, S) \rightarrow r(A)$ that quantifies the intrinsic value of the asset as a function of its EUTS attributes.

1.3 Calculating the Intrinsic Value of a Cryptoasset resulting from Sentiments

We perform sentiment analysis on the three EUT group of RDA Attributes to determine a measure of the intrinsic value resulting from sentiments. We define appropriate set of key words to represent the EUT RDA Attributes and perform sentiment analysis on those key words per to determine the popularity of each attribute and compute the sentiments per attribute per asset

Cryptoasset Universe We analyse the top 2,000 cryptoassets by market capitalisation from Coinmarketcap. These are the most popular assets that have attracted over 800 billion dollars as at the 8th of January 2021.

Keywords Each attribute within each group (E, U, T) is represented by a set of keywords. There are 26 set of key words in total. The choice and source of keyword Appendix provides detail on the set of keywords that were tracked. The goal of the sentiment analysis on the keywords is to determine the weight of each RDA Attributes - the level of consideration people give to each attribute and attribute groups as a whole and to measure the sentiments per each RDA attribute per cryptoasset. Both measures are in turn used in the further calculation of the intrinsic value of each cryptoasset within the RDA Index. For details on the RDA Index methodology please refer to RDA Index - A fundamentally weighted index for the cryptoassets [?]. The following steps were followed to achieve our goal:

- we collect Twitter and sentences from News articles that specifically mention a crypto asset and a keyword that we re tracking.
- By keeping track of the Keywords, we evaluate the sentiment

- We provide the sentiment score of each keyword that is used to calculate the sentiment score of an attribute
- we provide a sentiment score per asset for each attribute that is in being tracked across social media
- As the attributes used to track cryptoasset value can change and new ones proposed, we allow for the flexibility to change these keywords without having to re-develop the sentiments analysis module.

In the following sections, first we present our data collection effort for consolidating social media and news data related to cryptoassets and the RDA attributes. In Section 3, we present our sentiment analysis model and compare our model with other off the shelf sentiment analysis models.

2 Data

Sentiment analysis is the task that associates for each record a sentiment score that indicates the level of positivity or negativity of that record. For our task, we want to measure two main sentiment scores: We use score and point synonymously in this paper.

- sentiment score per attribute across all assets
- sentiment score per asset per attribute

To achieve this, we collect Twitter and News data from the Internet (see Section 2.1). We then find out which attribute is being referred to in each of the Twitter and News records that we collected (see Section 2.1.3). To train our sentiment analysis model, we manually annotated a sample of the Tweet and News that we collected with the attribute label and sentiment label (see Section 2.2).

2.1 Data collection

We used two main sources of social media data to collect data relating to cryptocurrencies: Twitter[1] and News articles.

2.1.1 Twitter

To build out Twitter dataset, we collected Tweets using the searchtweets[2] and Tweepy[3] APIs. The searchtweets API is as a wrapper for the Twitter premium and enterprise search APIs and supports up to last 30-day Searches and full archive searches on Twitter posts. On the other hand, Tweepy is a wrapper for the Twitter API and supports up to last 7-days searches on Twitter posts.

We only wanted to collect relevant tweets, that is, tweets that are about cryptocurrencies and that refer to any of our 26 crypto attributes. We used three (3) search query strategies to retrieve relevant tweets:

1. **Querying using the asset name:** The asset name, symbol and slug were used in the search query. We also included cryptocurrency related hashtags (Example 2.2) in the search query in order to minimize collection of irrelevant tweets. Below is a sample query for 'bitcoin' asset:

   Example 2.1 *Bitcoin (#bitcoin OR #BTC) (#crypto OR #cryptocurrency OR etc.)*

   This translates to: search for tweets with the asset name 'bitcoin' and either hashtags '#bitcoin' or '#btc' and any of the crypto related hashtags (Example 2.2).

2. **Querying using crypto based hashtags:** The search query was one of the crypto related hashtags Below are example crypto related hashtags that were used as part of the search query:

[1] https://twitter.com/home
[2] https://pypi.org/project/searchtweets/
[3] http://docs.tweepy.org/en/v3.9.0/api.html

Example 2.2 *#crypto, #CryptocurrencyNews, #cryptocurrency, #cryptocurrencies, #coinbase, #cryptotrading, #cryptonews, #cryptoworld, #cryptotrader, #cryptomining, #cryptoinvestor, #binaryoptions*

3. **Querying using word combinations from the crypto attributes:** The search query was a combination of words/phrases relating to the attributes that are being tracked. The below example shows a query that retrieves tweets relating to "Technical White paper" attribute:

   Example 2.3 *(whitepaper OR white paper OR design document) (#crypto OR #cryptocurrency OR etc.)*

   This translates to tweets containing the keywords 'whitepaper' or 'white paper' or 'design document' and any of the crypto hashtags.

In addition, we filtered out retweets and restricted our search to only tweets English language.

We use Tweet Preprocessor API [4] to clean the tweets, that is, remove mentions, hashtags, retweets and urls from the tweet. We then automatically determined which crypto asset mentioned in the tweet. We came up with a simplified logic for determining which crypto asset a tweet is referring to:

1. If the tweet has only one hashtag and the hashtag is an asset name or slug to a given asset, we assign the tweet as referring to that asset (see Table 1 example 1).

2. If after stripping the tweet of mentions, hashtag and urls, we find a word in the tweet that is in the list of hashtags and is either an asset name or slug, we assign the tweet as referring to that asset (see Table 1 example 2).

Example	Original Tweet	Cleaned Tweet	Hashtags	Asset
1	Each #bitcoin or fraction thereof represents a claims to resources on the world's largest, most decentralized and most secure blockchain network. Any suggestion that such a claim has no intrinsic use value is silly. Like it or not, bitcoin has intrinsic value.	each or fraction thereof represents a claims to resources on the worlds largest most decentralized and most secure blockchain network any suggestion that such a claim has no intrinsic use value is silly like it or not bitcoin has intrinsic value	#bitcoin	Bitcoin
2	Hurry! 9.44% direct arbitrage in #WINGS. Buy WINGS in BTC market from #Bittrex and sell it on #GateIo in ETH market, you can make a maximum profit of 11200.82KRW.#crypto #arbitrage #koinknight https://t.co/Oulmxymn40	hurry direct arbitrage in buy *wings* in btc market from and sell it on in eth market you can make a maximum profit of krw	#WINGS, #Bittrex, #GateIo, .#crypto #arbitrage, #koinknight	Wings

Table 1: Determining the main asset referred to in a tweet.

Tweets are generally short, so each tweet is a record in the Twitter dataset. Only those tweets that mention a cryto asset name or a hashtag of a crypto asset were added as part of the Twitter dataset.

2.1.2 News articles

To build the News dataset, we collected news articles using 3 main News APIs: a general News API[5] and two cryptocurrency specific news APIs, CryptoControl[6] and Cryptopanic[7].

[4] https://pypi.org/project/tweet-preprocessor/

[5] https://newsapi.org/

[6] https://cryptocontrol.io/en/developers/apis

[7] https://cryptopanic.com/developers/api/

The general news api allows for retrieval of news articles from various sources though the news articles are not specific to cryptocurrencies. One needs to search for mentions of asset names from the available sources. In order to retrieve relevant news articles, we restricted the search query to retrieve news articles that contain an asset name and the prefix *crypto*.[8]

CryptoControl news api API allows for retrieval of tweets, reddit posts and news articles related to crypto assets. CryptoControl api uses the asset slug as the search query. Each article comes annotated with the asset name being mentioned in the article. The returned json object only contains titles of the news articles and source links to the article. Therefore, we used the source links to scrap the news text.

Cryptopanic news api allows for retrieval of titles and news snippets of the original news articles related to crypto assets. Cryptopanic api uses the asset symbol as the search query. Similar to CryptoControl api, each news snippet comes annotated with the asset names that are being referred to in the snippet. The returned json object also only contains titles of the news articles and source links to the article snippet. We used the source links to scrap the news text.

Each news article was split into sentences. Only those sentences that mention a crypto asset name, symbol or slug was added as unique records in the News dataset.

2.1.3 Relating crypto attributes to data records

As mentioned in Section 1, we want to associate each record, Tweet or sentence from news articles, to the Real Digital Asset (RDA) Attributes that is being referred to. This allows for the calculation of the overall sentiment score of the RDA Attribute.

For each RDA attribute that we are tracking, we have defined fine grained phrases that represent the various ways in which users talk about the crypto attributes. The phrases we used to for each crypto attribute are provided in Appendix A.

We experimented with several techniques to find out the attribute being mentioned. We experimented with topic models using Biterm topic model (Yan et al. [2013]). We encountered data data sparsity issues where the words used to describe the attributes were missing in the final vocabulary learned by the topic model. In the end, we employed fuzzywuzzy (Cohen [2020]) string similarity measure to determine the similarity between RDA attribute phrases (see Appendix A) used to describe the attributes and the data records. We used a 100% threshold for the partial ratio fuzzy score as a cutoff to make sure that attribute phrases are present in the data records. We used fuzzywuzzy to find all records that mention any of the phrases associated with a crypto attribute. If a record mentions a phrases defined for a given crypto attribute, that record is associated to the crypto attribute.

2.2 Data Annotation

We selected approximately 6000 records for annotation. We selected approximately 200 tweets per attribute. The tweeter record included tweets collected between March 2020 and August 2020, while the news sentences were extracted from news text published from as far back as 2018 to August 2020.

Each record was annotated with the attribute that is being referred to in the record (or None otherwise), and the sentiment score. We use the 26 attributes as labels and 4 levels of sentiment values (positive, negative, neutral and mixed) as labels . 3 annotators independently annotated the records.

2.2.1 Annotation results

About 2291 tweets and 3056 sentences referred to at least one of the RDA attributes. In the end, about 3284 records were relevant, that is, the records mentioned an asset and also referred to an attribute (see Table 2). Approximately 1657 were annotated as positive, 1259 and neutral, 261 as negative and 107 as mixed. Table 3 shows the number of tweets and sentences that were labeled with the respective keywords.

[8] The developer version truncates the news articles. One needs the premium version to retrieve the full articles.

Sentiment	**Number of records (tweets+news)**
positive	1657
neutral	1259
negative	261
mixed	107
total	3284

Table 2: Annotated Data

Keyword	**No. of records**
Privacy	605
community voting right	342
technical whitepaper	253
opensource software code	237
lightening network	236
market pairs number	226
digital use case	190
transaction speed	168
gold price correlation	157
team bio public	132
public network (permissionless network)	122
pending transactions	96
independent blockchain	70
limited coin supply	69
silver price correlation	67
market capitalisation	66
staking rewards	61
commodity backed DeFi	58
atomic swap	53
quantum resistance	26
number of validation nodes	25
price volatility	14
age of coin	5
Token velocity	4
customer support	2

Table 3: No of annotated records per attribute

3 Sentiment Analysis Modeling

We hypothesize that users of cryptocurrencies talk on social media about the various assets and attributes relating to those assets. We also hypothesize that in their discussions, they will speak positively or negatively about an asset with respect to a given attribute. We use the annotated data from Section 2 to train a sentiment analysis model that will predict the positivity, negativity or neutrality of records.

In this Section, we details on how we built the sentiment analysis model and compare the results of our model with two off-the-shelf sentiment analysis models. The results show that our model outperforms the two off-the-shelf models, thus providing a benchmark model for sentiment analysis of crypto related social media data.

3.1 Sentiment Analysis Module

First, we perform vectorization of the datasets using FastText (Joulin et al. [2016]) for input to the machine learning module. We use the Fasttext classifier (Joulin et al. [2016]) to build the Sentiment analysis model.

Joulin et al. [2016] used FastText to build sentiment analysis models for various datasets and their results show that models build using FastText perform often on par with other deep learning models while taking a relatively shorter amount of time to train. FastText using bag of n-grams features to take into account word order, thus improving the model performance

3.2 Evaluation

We evaluated the performance of out model against 2 available sentiment classifiers: TextBlob (Loria [2018]) and Vader (Hutto and Gilbert [2014]). Vader[9] is a "lexicon and rule-based sentiment analysis tool" that was developed specifically to help in sentiment analysis of social media data. On the other hand, TextBlob[10] is a Natural Language Processing (NLP) library that has an API for sentiment analysis among other NLP tasks. It has 2 main sentiment analysis implementation, one that uses patterns and the other that is based on an NLTK[11] classifier trained on movie reviews.

Table 4 shows how our model compared against the other two models. Even with only about 3000 annotated data, our model achieved 87% weighted F-Score, as compared to TextBlob (0.51%) and Vader (0.52%).

Model	**Precision**	**Recall**	**F-Score**
Our Model	**0.87**	**0.87**	**0.87**
Vader	0.52	0.52	0.52
TextBlob	0.51	0.50	0.51

Table 4: Results: Comparison of our model agains two off-the-shelf models.

3.3 Deriving RDA attribute and asset weights

As part of the RDA index, we came up with two weights that rely on sentiment analysis: popularity score and average sentiment score.

Popularity score Popularity score estimates the popularity of an asset or attribute based on the number of tweets collected in a given time period that refer to the asset or attribute. We use equation 1 to estimate asset popularity, where T is the total number of collected tweets for a given period and t_i is the total number of tweets that refer to the given asset over the same period.

$$asset\_popularity = \frac{t_i}{T} \tag{1}$$

Likewise, we use equation 2 to estimate asset popularity, where K is the total number of collected tweets for a given period and k_i is the total number of tweets that refer to the given crypto keyword over the same period.

$$keyword\_popularity = \frac{k_i}{K} \tag{2}$$

Sentiment weights The average sentiment score estimates the overall sentiment score of an asset or attribute for a given period of time.

For each RDA attribute, we use equation 3 to calculate the sentiment score: we take the sum of the sentiment scores, s, of all tweets that refer to the attribute and divide by the total number of tweets that refer to the attribute , T. To get the final sentiment weight for an attribute, we weight the sentiment score with the popularity score from equation 3.

[9] https://pypi.org/project/vaderSentiment/

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

[11] https://www.nltk.org/

$$sentiment\_asset = \frac{\sum_1^n S}{T} \tag{3}$$

For each asset, we calculate the sentiment score per keyword using equation 4, where we sum the sentiment scores of all tweets referring to a given attribute for that asset, $S_k a$ and divide by the number of tweets that refer to the asset, T_a. We weight the sentiment score using the asset popularity score from equation 1.

$$sentiment\_asset = \frac{\sum_1^n S_{ka}}{T_a} \tag{4}$$

3.4 Discussion

Sentiment analysis is data dependent, as show from the results. Systems trained on one dataset may not perform well on a different dataset. Our model performs well in comparison to the rule-based Vader model and also the model trained on movie reviews. To prevent loop holes in the data collection and avoid spammers, we choose only verified source domains to collect News articles. For future work, we recommend choosing tweets only from specific tweeter handles in order to avoid spamming. Though we also foresee that it would be non trivial to authenticate Twittter handles and detect spammers.

4 Conclusion

This paper in part addressed the problems of fundamental analysis and valuation of cryptoassets. In particular, this paper presented the development of the Sentiments Analysis aspect of the RDA Index. We methodologically described our data collection effort to collect social media data (Tweets and News articles) that refer to crypto assets and any of our RDA attributes. We described how we built an annotated dataset of Tweets and News articles by labeling each record with the RDA attribute mentioned in the record and a sentiment label, ether positive, negative, neutral or mixed. We presented a benchmark model for Sentiment analysis of crypto related social media texts. Our model outperformed 2 off-the-shelf model.

A Appendix A

ID	Attributes	Attribute Phrases
E1	market pairs number	[paired, pairs, market pair, trading pair, currency pair, crypto pair]
E2	team bio public	[a team of, team bio, development team, team biography, team profile, linkedin profile, strong team, experienced advisers, unparalleled team, nice work team]
E3	technical whitepaper	[whitepaper, white paper, technical whitepaper, business whitepaper, design document]
E4	opensource software code	[opensource, open-source, opensource code, open-source code, public code, public code repository]
E5	market capitalisation	[market capitalisation, market valuation, market value, market volume]
E6	limited coin supply	[deflationary, circulating supply, maximum supply, limited supply]
E7	age of coin	[genesis block, first block, start of project, project start date, age of cryptocurrency, creation date]
E8	price volatility	[inelastic, inelasticity, price inelasticity, inelastic price, price elasticity, elastic price, price volatility, change in price]
E9	community voting right	[community voting, voting right, governance]
E10	staking rewards	[staking reward, dividends, proof of stake]
E11	customer suppor	[telegram channel support, telephone support, customer care support, customer care, web chat support, contact center, contact us, email support]
E12	public network (permissionless network)	[permissioned, permissionless, decentralised, permissioned network, permissionless network, public decentralisation, publicly decentralised]
E13	number of countries of validation	[geography of validation nodes, geographic distribution, geography, transaction location, validation location, validation countries, country nodes, number of countries, decentralisation footprint, decentralised countries, country of validators]
U1	token velocity	[velocity, token velocity, transaction volume]
U2	privacy	[privacy, private transactions]
U3	gold price correlation	[gold, gold price, gold price correlation]
U4	silver price correlation	[silver, silver price, silver price correlation]
U5	digital utility usecase	[digital utility, use case, digital use case]
U6	backed by tangible asset	[commodity backed, backed by gold, backed by security, gold standard, backing, asset backing, fidiciuary]
T1	network congestion	[congestion, network congestion, pending transactions, awaiting confirmation]
T2	transaction speed	[speed, transaction per second, confirmation speed, transaction speed]
T3	independent blockchain	[blockchain network, blockchain protocol, independent chain, independent blockchain]
T4	quantum resistance	[quantum, quantum resistance, quantum resistant]
T5	number of validation nodes	[nodes, number of validation nodes, validation nodes]
T6	lightening network	[lightening network, lightning network, layer 2, layer 2 network, layer two]
T7	atomic swap	[atomic swap, atomic exchange, htlc, hash time lock contract, onchain settlement, onchain settlement, on-chain exchange, onchain exchange]

References

Adam Cohen. Fuzzywuzzy using python., February 2020. URL `https://pypi.org/project/fuzzywuzzy/`.

C.J. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. 01 2014.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*, 2016.

Steven Loria. textblob Documentation. *Release 0.15*, 2, 2018.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. pages 1445–1456, 05 2013.