
# THE CORRELATION BETWEEN REDDIT SENTIMENT AND THE STRONGEST- AND WEAKEST PERFORMING CRYPTOCURRENCIES OF 2021

EMILY COPPENS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

WORD COUNT: 8239

STUDENT NUMBER

u922049

COMMITTEE

dr. Eva Vanmassenhove
dr. Grzegorz Chrupala

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2022

ACKNOWLEDGMENTS

I would like to thank dr. Eva Vanmassenhove for her guidance, feedback and encouragement right from the start of the thesis-writing period. In addition, I would like to thank my parents and my brother and sister for always supporting me.

# THE CORRELATION BETWEEN REDDIT SENTIMENT AND THE STRONGEST- AND WEAKEST PERFORMING CRYPTOCURRENCIES OF 2021

EMILY COPPENS

**Abstract**

Investing in cryptocurrencies has gained a lot of momentum over the past few years, driven by social activity on platforms such as Reddit or Twitter. This research investigates to what extent Reddit sentiment can reveal pricing trends for the strong performing and weak performing cryptocurrencies of 2019-2021. The current body of research is mostly focused on Bitcoin, stock pricing and does not show many forms of distinction between the currencies or stocks explored. This paper takes a new approach by differentiating between currencies that have shown a strong performance and currencies that have shown a weak performance from 2019-2021. A lexical-based approach was used to perform sentiment classification on unlabeled Reddit data. In choosing the most appropriate model, three lexicon models were tested. In addition, a statistical analysis was performed in order to identify correlations between pricing, polarity and the number of comments per submission. The data set consisted of scraped Reddit data and cryptocurrency market value data. It was found that VADER was the most appropriate lexicon model for social text with an accuracy of 86%. Furthermore, the sentiment class distribution was higher towards positive sentiment for stronger currencies and had a higher classification of negative sentiment for weaker currencies. Lastly, for the both the strong and weak group of currencies a correlation was found for polarity and change in price. No significant correlations were found for the weak performing currencies.

# 1 INTRODUCTION

The period from 2019 to 2021 has been quite turbulent for investors. In 2020 the term meme stock gained momentum. (Costola, Iacopini, & Santagiustina, 2021) define meme stocks as unstable stocks that fluctuate significantly based on not only financial, but also on social activity over time. A popular example of a meme stock is GameStop's stock (GME), which price skyrocketed from $2.57 to $347.51 in January (Yahya & Chiu, 2021). A lot of communication around GME took place on the Reddit form r/WallStreetBets.

The Reddit platform aggregates news, opinions and discussions about a wide variety of trending topics (Bikhchandani, Hirshleifer, & Welch, 1992). It can be seen as the Valhalla of forums for any topic, with channels and subchannels for users to interact with. Users contribute by either starting a discussion on a subreddit, commenting on a post, commenting in reply to another comment and up- or downvoting posts or comments. According to Melton, Olusanya, Ammar, and Shaban-Nejad (2021), roughly 430 million users access Reddit, aggregating a staggering 1.7 billion visits to the website per month on average [1]. The r/WallStreetBets channel was created for everyday people who wanted to invest. Ultimately, their Reddit activity resulted in an enormous influence in the surge (Yahya & Chiu, 2021). The meme stock phenomenon of fluctuating prices is not only limited to the stock market. Similar market behavior emerged for cryptocurrencies, specifically altcoins. An altcoin is short for alternative coin (cryptocurrency) and encompasses any currency alternative to Bitcoin. Bitcoin is based on an open-source code, which means that others can contribute to it or copy it to create their own altcoin (Ong, Lee, Li, & Chuen, 2015). When looking at the list of cryptocurrencies, it becomes evident how popular creating a new currency has become. There are currently over 6000 different currencies on the market [2].

Defining to what extent social media activity has is associated with a cryptocurrency's market value, has social importance. Consequently, it has quite an extensive body of research (see section: Related Work, page 4). The subject matter poses several interesting ethical discussions. Coins hyped on social media can make investors vulnerable for fraud. For instance, in pump and dump schemes, which Mirtaheri, Abu-El-Haija, Morstatter, Ver Steeg, and Galstyan (2021) describe as drawing in investors to invest in artificially inflated coins so they can sell their holdings. Their research done on cryptocurrency manipulation on social media also explains the hype is often created on popular discussion forms such as Twitter and

[1] https://www.statista.com/statistics/443332/reddit-monthly-visitors/
[2] https://www.statista.com/statistics/863917/number-crypto-coins-tokens/

Reddit. They state that even the Commodity Futures Trading Commision (CTFC) and the Securities and Exchange Commission (SEC) have issued warnings against the increasing number of fraud schemes. So, should a group be able to steer the pricing trend as much as they have done thus far? If the influence is so large, how do we protect investors from fraudulent schemes?

In addition, within the domain of this research, there is scientific importance in researching behavior in social context. Natural Language Models are constantly being created, extended and tested in order to capture the contextual meaning and sentiment of unlabeled social media conversations. In return, this aids in the understanding and modeling of human social behavior.

This research is written to provide a new perspective to the possible association between social media activity and a cryptocurrency's market value. The main topic is the correlation between sentiment and the market value of the selected cryptocurrencies. Therefore, the main research question is as follows:

> *"To what extent can Reddit sentiment reveal pricing trends for the strong performing (Binance, Bitcoin, Cardano and Solana) and weak performing (Dogecoin, Compound, SafeMoon and Tether) cryptocurrencies of 2019 - 2021?"*

However, in order to answer the research question, additional steps are to be taken. Therefore, three sub research questions have been defined in support of answering the main research question:

RQ1 *"How well do unsupervised lexicon models perform in the sentiment classification of Reddit data?".*

In order to assure the most accurate sentiment labeling of the data, three lexicon models will be compared in terms of accuracy to find which model performs better.

RQ2 *"How does the sentiment analysis differ when comparing the best- and worst performing cryptocurrencies?".*

With the intention to shift the focus from solely being on stable, well-known currencies, this research question explores any behavioral differences concerning sentiment for stronger and weaker performing currencies.

RQ3 *"To what extent do the fluctuations in sentiment correlate with the fluctuations in the market value of the cryptocurrencies?"*.

With the previous two questions forming the base, correlations between sentiment on Reddit with the market value behavior are analysed for significance.

## 2 RELATED WORK

The volatility of market value for cryptocurrencies or stock pricing has been a well-researched phenomenon over the past decade, including the influence of social activity. This section will cover a number of such relevant works extracted from the existing body of research. In order to shed adequate light on both the researched behavior on social media as well as the researched effects of social media on the market value, the works are split in two sections: Sentiment on Social Media and Sentiment on Market Value.

### 2.1 *Sentiment on Social Media*

The following articles are discussed in order to justify whether social media is an important source of community forming and information sharing in addition to finding appropriate methods for sentiment analysis. The first work to be discussed in this section was written by Knittel and Wash (2019), who researched threads on the r/Bitcoin subreddit and identified relevant topics and themes. The aim was to examine the societal impact of Bitcoin and the influence online communities have on Reddit. They stated Reddit was chosen over other platforms, as communication on Reddit is much more active than StackExchange or Bitcoin's official forums. In their research they found that the 'Bitcoiners' community felt the support and engagement on the social media platform is necessary for Bitcoin to grow.

Glenski, Saldanha, and Volkova (2019) analyzed the discussion spread between Bitcoin, Ethereum and Monero in developer interest and community interest on social media platforms. Their research differs from Knittel and Wash (2019), as their research was not exploratory. Instead, they aggregated submission data using the Reddit API. From the aggregated data they analyzed the entire commentary tree and its characteristics. It was found that users tend to be more active within one subreddit dedicated to a specific domain, as opposed to general subreddits. This indicates that choosing what subreddits to scrape is important and will have an impact on what behavioral patterns could be found. In addition, Monero went most viral, which their research stated to be correlated to the depth and breadth

of the discussions. Due to the fact that there was a correlation between the depth of discussions and virality of the coin, it was recommended to extend this research by looking at a larger variety of coin types.

The aforementioned articles suggest that communities are formed around cryptocurrencies, that it is crucial to choose relevant subreddits for the analysis, and that in some particular cases, virality or hype on Reddit could correlate with generated discussion. The following articles dive deeper in to the methods of sentiment analysis for social media.

Melton et al. (2021) used Reddit and Twitter data to compare sentiment towards COVID-19 vaccines across different cities. Although their topic differs, the approach is relevant. For the sentiment analysis, they used an unsupervised lexicon-based approach by calculating the polarity and subjectivity using the TextBlob model. In addition, they analyzed correlations between the number of comments and new cases of COVID-19. It was found that the peak of positive cases of COVID-19 was correlated to the amount of generated discussion. Their research concluded that changes in the activity on Reddit and Twitter can aid in understanding concerns and sentiment around the pandemic.

Stieglitz and Dang-Xuan (2013) used sentiment analysis to examine their hypothesis of a positive relationship between social media sentiment and its diffusion through social online networks. More specifically, their research focuses on politics on Twitter. For the sentiment analysis SentiStrength was used, a lexicon which classifies text for positive and negative sentiment using a range of -5 to 5. They then used a regression model to examine whether the sentiment is associated with the number of retweets. They found that emotionally charged tweets are more likely to spread, than neutral tweets are. In addition, there was a positive correlation to retweet quantity and retweet speed.

Dhaoui, Webster, and Tan (2017) used both a lexicon-based model as well machine learning (ML) to analyze sentiment on brand-related social media conversations on Facebook, in order to find which approach results in the highest accuracy. The lexicon that was used was the Linguistic Inquiry and Word Count which analyses common words. As a limitation, it doesn't support emoji's. For the ML approach the aggregated Facebook data was manually labeled in order to use a supervised technique. Several algorithms were included such as Random Forest, SVM, Bagging and Decision Trees. Both approaches classified text as negative, neutral and positive and their data set consisted of 850 observations. They found that the lexicon and ML results were very similar. This is contradicting other to statements from researchers such as Zhang, Gan, and Jiang (2014) that claim ML models achieve higher accuracy. In addition, the lexicon approach does not require labeled data. For this exact reason, Jurek, Mulvenna, and Bi

(2015) used a lexicon approach for their sentiment analysis on social media. The sentiment analysis was run to forecast box-office revenues for movies using Twitter data and their lexicon used was SentiWordNet. It achieved 77.3% accuracy.

Lexicons are a common choice for sentiment analysis on social media. Hutto and Gilbert (2014) took VADER, a commonly used rule-based lexicon and compared its performance to 11 state-of-the-art lexicon models. As the ground truth, they took sentiment data, which has been labeled by 20 trained human labelers, aggregated and averaged. VADER performed well and even outperformed some human raters on the ground truth. None of the other models were able to outperform VADER, even more complex models such as Support Vector Machines (SVMs). This is presumably due to its simplicity and ability to be computationally efficient.

## 2.2 *Sentiment on Market Value*

The following works provide different methods to analyze the effects social media has on the market value of cryptocurrencies.

An exploratory analysis was conducted by Wooley, Edmonds, Bagavathi, and Krishnan (2019) to analyze the relationship between Bitcoin and Ethereum and the public opinion in order to predict price movements. Their model was created as a binary model predicting a price increase or decrease in the future using network dynamics extracted from a number of manually analyzed user submissions. Dynamics include features such as the number of submissions, speed of submissions and comment trees. From their findings they concluded Reddit derived features can be helpful in deriving the market price of Bitcoin and Ethereum. In addition, Telli and Chen (2021) conducted research with a similar aim, investigating the relationship between the cryptocurrency markets of Bitcoin, Ethereum, Litecoin and XRP and the activity on the online platforms: Wikipedia and Reddit. They found that user's behavior is anti-persistent, meaning that an increase is very likely to be followed by a decrease and vice-versa. Their research suggests that investors can let this information play an important role in their decision making. Interestingly, they found that Bitcoin and alt coins share a similar anti-persistent dynamic, which could indicate that social strategies for bitcoin could be copied for altcoins.

Smuts (2019) found a positive correlation between Bitcoin and Ethereum prices and sentiment obtained from data scraped from Google Trends and Telegram. A Long-Short Term Model was used to predict price movements of cryptocurrencies with their sentiment labels and correlations. Their results backed their hypothesis that the obtained sentiment can predict price movements of cryptocurrencies. Mittal, Dhiman, Singh, and Prakash

(2019) used a variety of statistical models to predict short-term price fluctuations including a linear and polynomial regression. In contrast to the above-mentioned research, their model showed low correlations between tweet sentiment and the Bitcoin price. Kaminski and Gloor (2014) analyzed correlations between activity on Twitter and the price of Bitcoin using Pearson Correlation. They concluded that the sentiment established on Twitter was a consequence of the pricing trends. This means that the sentiment can not be used for pricing predictions but essentially, the price could predict the sentiment.

Similarly to existing research on cryptocurrencies, Sul, Dennis, and Yuan (2017) researched sentiment on Twitter in order to predict stock returns. They took approximately two years worth of data from Twitter which accumulated to ten thousand observations. ML was used for classifying the data as a positive or negative sentiment. It was analyzed whether it was linked to average daily stock returns of a variety of companies in the S&P 500, based on the diffusion (spread) of the sentiment. With their correlation analysis a significant association was established between the spread of sentiment and stock returns. Broadstock and Zhang (2019) conducted research with a similar aim, using Twitter data to test whether sentiment has pricing power on stock returns of firms on the S&P 500. They used the NRC Emotion Lexicon to classify the tweets as positive or negative. With the sentiment throughout the day, they assessed stock return fluctuations over a variety of time (1, 5 and 30 minute returns). Their results concluded significant stock return reactions to sentiment. They also found significance for broader, market-wide sentiment reactions. This means sentiment can have short-term effect on stock returns of specific firms as well as an effect on all stock returns for multiple firms over time.

Nguyen, Shirai, and Velcin (2015) built a model using social media sentiment to achieve more accurate stock price movement predictions, than models that use historical pricing. Their data was aggregated from the message boards on Yahoo Finance from 18 different stocks. People who post messages have the opportunity to label their own messages with sentiment such as 'strong buy', 'hold' or 'strong sell'. With the manually labeled data, the data that was not labeled a sentiment by the author and historical pricing of the stocks, they build their sentiment classification model. Their model was able to predict stock price movement with an accuracy over 60% for certain stocks. This is still a 2.07% improvement over models with historical pricing. It is also an indication that every stock behaves differently and they are difficult to generalize in a model. Nguyen and Shirai (2015) wrote another paper where they built a model to predict stock price movement, now using topic modeling of sentiment on social media. In their research they used Twitter data and as sentiment classification

tools OpinionFinder and Google Profile of Mood States was used. The output was either a positive or negative mood. Their stock price data came from Yahoo Finance. Similar to the previously mentioned research written by Nguyen et al. (2015), the Yahoo message board sentiment labels were used to train the model. They used the ML algorithm SVMs to build the prediction model based on historical prices, sentiment and topics extracted from the Twitter data. Their model outperformed models based solely on historical prices in accuracy by 6.07%. Chen, De, Hu, and Hwang (2011) investigated the effects of social media on the stock market by looking at stock-price changes of almost 3000 U.S common stocks, and stock opinions on both Seeking Alpha and the Wall Street Journal. The sentiment was classified as positive and negative by a word classification scheme and they used regression analysis to explore whether the sentiment had an effect on stock returns. It was found that there is an impact, and that the impact of the sentiment from Seeking Alpha was larger than the impact on the Wall Street Journal. This indicates that the platform from which the data is aggregated plays an important role in the analysis.

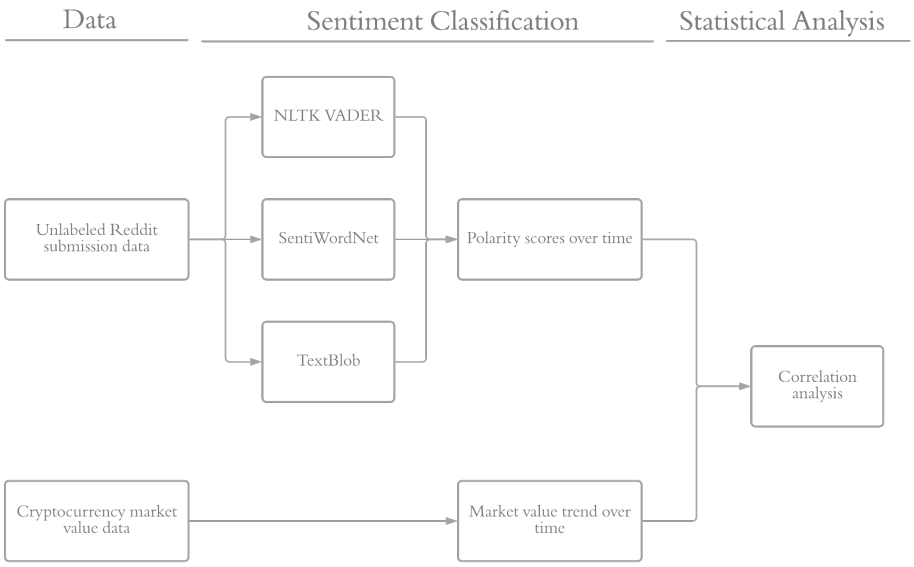
It is clear that with a phenomenon so volatile, the existing research does not currently represent a uniform conclusion on the effects of sentiment on market value. Even when a causal relationship can be established between sentiment and market value of stock or cryptocurrencies, there is still contradicting evidence on which has an effect on which.This thesis aims is to add to the existing body of research, by investigating to what extent Reddit sentiment is correlated to market value of a variety of cryptocurrencies. It differentiates itself from previous work by providing a different perspective. Namely, most research focuses on the prediction of stock price or stock price movements using Twitter. Over the past two years Reddit has become increasingly more popular as well as the trading in cryptocurrencies. In addition, contrasting to current research done on cryptocurrencies, which is mostly written on established coins such as Bitcoin and Ethereum, this research will include a variety of altcoins. These coins have been founded more recently and are thus potentially subject to more volatility. Lastly, this research distinguishes coins that have had a strong performance with coins that have had a weaker performance throughout 2019 and 2021. The theoretical exploration has influenced choices made in the methodology (further explained on page 9). Due to the fact that Reddit data is unlabeled, a similar unsupervised lexical-approach to sentiment classification will be used, inspired by Melton et al. (2021) amongst other papers in the literature review. In addition, similarly to Kaminski and Gloor (2014) the Pearson Correlation will be used for a statistical analysis on the data. Furthermore, the purpose is to take part in the exploration of extracting insights from

social behavior using raw textual data. This can in turn be used for fraud prevention, regulation and further research.

## 3 METHOD

This section will further elaborate on the various methods used in this research. It consists of data aggregation, data preprocessing, sentiment classification and a statistical analysis as illustrated in Figure 1.

Figure 1: Model Framework




### 3.1 *Unsupervised Sentiment Classification*

Crucial to this research is the sentiment classification. The difficulty lies in the fact that the data is raw, unlabeled textual data. In order to perform unsupervised learning for the sentiment classification, a lexicon modeling approach will be followed. According to Zhang et al. (2014), the accuracy is generally higher for ML approaches; however, the amount of labeled data necessary to train and evaluate the model is too large for the scope of this research. Lexicon models, on the other hand are pre-trained models designed for similar use cases. Therefore, they are are commonly used for unsupervised sentiment analysis (Zhang et al., 2014). There is a variety of developed lexicon models. This research applies three commonly used ones. This decision was made to avoid making assumptions on which model

would best apply to the submissions. Ultimately, the best performing one will be chosen based on its accuracy and a critical analysis of its confusion matrix. The three models used are: nltk Valence Aware Dictionary and Sentiment Reasoner (VADER)[3], nltk SentiWordNet[4] and Textblob[5].

VADER is a pre-trained, rule-based model which has a large list of lexical features. It searches the sentence for any words present in the VADER lexicon as stated by Bonta and Janardhan (2019). Subsequently, it returns a positive or negative score, as well as a polarity score. Meaning, to what degree is the sentence positive or negative. SentiWordNet, also a pre-trained, rule-based and much as the name indicates, assigns sentiment to every word in a sentence. It uses Part of Speech (POS) tagging to find relations between words and classifies them as either positive or negative. The positive and negative terms are aggregated and the sentence receives a class using polarity. Lastly, TextBlob is a pre-trained, rule-based lexicon model; however, TextBlob allows you to choose which corpora to import from the NLTK library. In contrast to VADER and SentiWordNet, the list of lexical features depends on what corpora is chosen. In addition, TextBlob outputs a polarity score between -1 and 1.

## 3.2 *Statistical Analysis*

Once the data set is classified, a statistical analysis is performed to assess whether the sentiment has an effect on the changes in price and vice versa. This will be done in the form of a correlation analysis. A correlation analysis assesses whether there is an association between two features and how strong that association is. The type of correlation used is the Pearson Correlation (r):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where $x^i$ is the value of the sample of the x-variable, $\overline{x}$ is the mean of all values of the x-variable, $y^i$ is the value of the sample of the y-variable and $\overline{y}$ is the mean of all values of the y-variable. The Pearson Correlation function in Python outputs two integers: the Pearson's Correlation Coefficient (PCC) and the p value. According to Schober, Boer, and Schwarte (2018), the PCC is a statistic measuring the association between variables. When data are correlated, one unit of change in one variable, will result in a unit of change in the other variable. This can be either positive or negative and therefore PCCs range from -1 to 1. The p value is the computed

[3] https://www.nltk.org/modules/nltk/sentiment/vader.html
[4] http://www.nltk.org/howto/sentiwordnet.html
[5] https://textblob.readthedocs.io/en/dev/

probability of observing the given test statistic, or one more extreme, if the null hypothesis (there is no correlation) were true.[6]

In order to ensure the right sample size for the correlation, the sample size will be computed using the following formula:

$$N = [(Z_a + Z_b)/C]^2 + 3 \tag{1}$$

Where $Z_a$ is the standard normal deviation (1.96) for the alpha value and $Z_b$ is the standard normal deviation (1.2816) for the beta value. Furthermore, C is calculated using the following formula:

$$C = 0.5 * ln(1 + r)/(1 - r)] \tag{2}$$

Where alpha is the threshold probability of rejecting the null, beta is the probability of failing to reject the null and r is the expected correlation coefficient. The chosen values for the threshold, beta and expected correlation are 0.05, 0.10 and 0.15, respectively. The chosen value for the expected correlation is relatively low due to the complex nature of the research. This formula states that in order for the correlation to differ from zero, a sample size of 463 is necessary. This method is introduced by Hulley (2007), and was designed for clinical research approaches. In addition, it is used in a variety of research papers. For instance, those written by Fadayevatan et al. (2019) and Kawanabe, Suzuki, Tanaka, Sasaki, and Hamaguchi (2018).

## 4 EXPERIMENTAL SETUP

Based on the literature review and methodology, the following experimental setup has been followed. It consists of preparation, data aggregation, data preprocessing, data annotation, sentiment classification and the statistical analysis.

### 4.1 *Pre-programming Preparation*

Before starting in Python, some decision making around the cryptocurrencies was necessary. Classifying which currencies belong to the stronger group and which ones belong to the weaker group is not as straightforward. Meme coins such as Dogecoin[7] can grow tremendously in value in a matter of days, but drop in value just as quick. Whether it has performed well depends on perspective, timing and the purpose of holding the coin (short- or long-term investing). In order to bring structure to the decision making,

[6] Information retrieved from the Tilburg University Statistics and Methodology course: Basics.pdf, page 25

[7] https://coinmarketcap.com/currencies/dogecoin/

the website CoinRanking[8] was consulted. CoinRanking makes distinctions between the best and worst performing cryptocurrencies over a specified period of time. Not surprisingly, the list changes daily, but it did give an indication as to how strong and weak performance can be assessed. Namely, by looking at the value at the beginning of the year and at the end of the year. Cryptocurrencies who's value has increased significantly compared to the beginning of the year are considered a strong performing coin. Those who's value has decreased, or peaked and significantly decreased after, are considered a weak performing coin. Therefore, the classification is based on a long-term investment. Aside from CoinRanking, there were a number of other considerations in order to ensure there is enough data. A minimum of one bilion dollars for the market cap was set, in order to ensure they are substantial enough. Furthermore, the subreddits had to have an aggregated minimum of 8K users for coins founded after 2020. With the performance information provided by CoinRanking in mind as well as the previously mentioned considerations, the cryptocurrencies were hand-picked. The strong performing coins are: Binance, Bitcoin, Cardano and Solana. The weak performing coins are: Dogecoin, Compound, Safemoon and Tether. For every cryptocurrency the dedicated subreddits, along with their number of users and year founded, have been identified and listed in Appendix A and B, pages 28 and 29 respectively. Lastly, there are large subreddits dedicated to investing in general, such as *r/Investing/* or *r/CryptoCurrency/* that could contain additional interesting sentiment. These have also been aggregated and can be found in Appendix C, page 29.

## 4.2 *Data Aggregation*

The next step was to aggregate the submission data from Reddit. In order to do so, the chosen subreddits from the tables in Appendix A and B were scraped using the Python Pushshift.io API Wrapper (PSAW) [9]. The Pushshift API was chosen as it has over four billion submissions and provides direct API access to the submission database. It also allows searches with keywords, enabling the examination of the database for submissions on specific cryptocurrencies. PSAW outputs up to 23 different features, generating an extensive, unlabeled dataset. Two functions were written in order to be able to loop through submissions. Firstly, the URL that Pushshift.io uses to search for the submissions was built. It consists of the standard first section of the URL referred to as the endpoint [10]. For this research the endpoint is: 'https://api.pushshift.io/reddit/search/submission/?title='.

[8] https://coinranking.com/coins/best-and-worst
[9] https://reddit-api.readthedocs.io/en/latest/
[10] https://github.com/pushshift/api

On this URL the query, the after and before time stamps and the subreddit was built. Every API call accesses the specified webpage and stores the data in JSON format.

Secondly, the desired data points were extracted and stored. As previously mentioned, PSAW aggregates up to 23 features. With computation time and efficiency in mind, seven features were chosen:

- Post ID
- Title
- Author
- Score (the number of upvotes of a post, minus the number of downvotes)
- Publish Date
- Total No. of Comments
- Flair (a tagging system for users to add additional context to their posts, not every user has one)

The before and after variables used to build the URL are dates translated into timestamps[11]. Every minute of the day has a unique timestamp and to ensure this research is as recent as possible the timestamps run from the day the scraping started on the 6th of October 2021, back exactly two years, to the 6th of October 2019. The query is the keyword to search for in every submission, in this case being the cryptocurrency in question and the full list of subreddits mentioned above. This setup allows for Python to loop through all cryptocurrencies for all predefined subreddits and were concatenated to one Pandas data frame ultimately consisting of 83,352 submissions.

The second data frame that needed to be aggregated was the historical data on the market value of all cryptocurrencies. These were exported[12] and loaded in to Python. All market value data frames consist of the same seven features:

- Date (daily)
- Price
- Opening Price
- Highest Price

[11] This was done using: https://www.unixtimestamp.com/index.php.
[12] All data is publicly available on Yahoo Finance

- Lowest Price
- Volume Traded
- Daily change of Price (in percentages)

Lastly, the results of data aggregation are displayed in Table 1 below. As previously mentioned, this research differentiates between newer and established coins. Therefore, the number of observations for coins such as Solana, Compound and Safemoon are smaller as there was no data on them in 2019.

Table 1: Number of Aggregated Data Observations

| | **Reddit Data** | **Market Data** |
|---|---|---|
| *Binance* | 17408 | 732 |
| *Bitcoin* | 43579 | 732 |
| *Cardano* | 2515 | 732 |
| *Solana* | 2166 | 422 |
| *Dogecoin* | 8402 | 732 |
| *Compound* | 3594 | 415 |
| *Safemoon* | 3262 | 97 |
| *Tether* | 2426 | 732 |

## 4.3 *Data Preprocessing*

With the submission data frame in place, the subsequent step is preprocessing. Some of the preprocessing steps are done for the entire data frame, and some steps vary depending on the lexicon model used. Firstly, the 'Publish Date' variable consisted of both the date and the timestamp and was therefore split. The date in dd/mm/yyyy format was saved and then further split in to year, month, week and day features using the Python datetime module[13] allowing for an easier analysis of the data at a later stage.

With regards to the market value data frame, the symbols were removed and dates were then sorted in ascending order. Due to the volatile nature of the data, any missing data were omitted as opposed to being replaced. With the same logic, outliers were not omitted nor replaced in order to capture the volatility.

[13] https://docs.python.org/3/library/datetime.html

## 4.4 *Data Annotation*

From the concatenated Reddit data frame a subset of 100 submissions was extracted using sampling with a random_state of 1. These were saved and manually labeled. The annotator labeled the submissions using scores of -1, 0 and 1 representing negative, neutral and positive sentiment, respectively. Within the subset, 15% of submissions were considered negative, 50% considered positive and 35% was labeled positive.

## 4.5 *Sentiment Classification*

The next step is to conduct the sentiment analysis with the three chosen lexicon models.

Nltk VADER: the VADER classifier works well with social text and takes emojis, elongated words and punctuation in to account. The more preprocessing the less information VADER is able to capture. Therefore, no additional steps were taken before applying the lexicon model. The SentimentIntensityAnalyzer()[14] from the VADER package was applied to every row of submissions and the compound score was stored in a new column.

Textblob: the TextBlob package was imported to Python[15] as well as the nltk corpora[16] which TextBlob uses for its sentiment classification. The lines were stripped from any punctuation. Subsequently, every word was tagged and searched for nouns. Lastly, TextBlob assigned each sentence a polarity, which was stored in a new column.

SentiWordNet: First, the text is further preprocessed by converting to lower string and removing hashtags, links, special characters, multiple spaces, single characters and stopwords. This was done using the Regular Expressions package in Python[17] and the stopwords module[18]. Subsequently, the words are lemmatized using the WordNetLemmatizer module from the nltk.stem package[19]. Every word is tokenized and tagged as either an adjective, noun, adverb and verb allowing the SentiWordNet algorithm to bring more context to the sentence. The algorithm then assigns a positive and negative score to every word and the ultimate sentiment score for each sentence is the positive minus the negative score.

As the lexicon models output a polarity they had to be rounded to -1, 0 or 1 in order to be classified as negative, neutral or positive respectively.

[14] https://www.nltk.org/api/nltk.sentiment.sentiment_analyzer.html
[15] https://textblob.readthedocs.io/en/dev/
[16] https://www.nltk.org/api/nltk.corpus.html
[17] https://docs.python.org/3/library/re.html
[18] https://www.nltk.org/_modules/nltk/corpus.html
[19] https://www.nltk.org/_modules/nltk/stem/wordnet.html

Polarities smaller than -0.2 were labeled negative, polarities between -0.2 and 0.2 were labeled neutral and polarities larger than 0.2 were labeled positive. This was only done for testing the accuracy of the model, further research was conducted using the polarities. In addition to assessing the performance of three individual lexicons, the accuracy was calculated for a majority vote. For a majority vote, the output of all three lexicon models will be calculated, and the sentiment with the most votes will be assigned to the submission.

After finalizing the list of polarities, an additional feature was created for the submission data frame: the change in polarity. This feature represents is the percentage change of the polarity, based on the previous day. A change in percentage makes it easier to comprehend whether there was a big change in sentiment from one day to another, than if one was to assess the polarities individually.

## 4.6 *Statistical Analysis*

Important to the statistical analysis was ensuring that the two data frames could be concatenated in a meaningful way. The market value data frame only consisted of 732 rows (with the exception of Solana, Compound and Safemoon), whereas the submission data frames were much longer. Therefore, the first step was to group the submission data by date, taking the daily mean of the previously calculated polarities. The same was done for the change in polarity. These have now been transformed to the daily averages. In addition, the number of comments were grouped by date and summed. This way both data frames are of equal length and could be concatenated. The full data frame used for the statistical analysis consisted of the columns shown in Figure 2.

| | Date | Price | Open | High | Low | Vol. | Change % | Vol | Day | Week | Month | Year | Polarity | Comments | Polarity_change |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **365** | 2020-10-06 | 10602.6 | 10,790.2 | 10,800.3 | 10,530.8 | 69.07K | -1.73 | 69070.0 | 6 | 41 | 10 | 2020 | 0.034418 | 127 | 3.13588 |
| **364** | 2020-10-07 | 10670.9 | 10,601.0 | 10,680.1 | 10,553.3 | 49.27K | 0.64 | 49270.0 | 7 | 41 | 10 | 2020 | 0.086171 | 210 | 4.64969 |
| **363** | 2020-10-08 | 10924.1 | 10,670.7 | 10,948.6 | 10,549.6 | 79.78K | 2.37 | 79780.0 | 8 | 41 | 10 | 2020 | 0.075887 | 1121 | 2.43725 |
| **362** | 2020-10-09 | 11054.2 | 10,923.5 | 11,103.0 | 10,836.9 | 70.95K | 1.19 | 70950.0 | 9 | 41 | 10 | 2020 | 0.060243 | 511 | 1.38466 |
| **361** | 2020-10-10 | 11298.4 | 11,053.5 | 11,475.0 | 11,053.1 | 65.81K | 2.21 | 65810.0 | 10 | 41 | 10 | 2020 | 0.089736 | 636 | 4.02681 |

Figure 2: Five rows of the full data frame

As Glenski et al. (2019) mentioned in the Related Work section (page 4), the depth of discussions are related to the virality of the coin. In addition, Stieglitz and Dang-Xuan (2013) mentioned that Twitter sentiment was positively correlated with the number of retweets. Therefore, the daily number of comments on the submissions are included in the modelling. The Pearson Correlation was calculated for two different models: Polarity

versus Change in Price and Comments versus Change in Price. This was done in Python using the Stats package from Scipy[20].

### 4.7 *Visualisation of Data*

The last step is to create interpretable visualisations of the data. This is done using the MatPlotLib package in Python[21]. The first visualization consisted of the graphing of the polarities. Secondly, the polarities were graphed against the market trend over the past year. The graphs had a double x-axis in order to examine both behaviors simultaneously. The reason for graphing the data in addition to outputting the statistic is that the Pearson Correlation merely outputs an PCC and a p value, whereas the graphs give better context as to how the data are behaving.

A list of all packages used can be found in Appendix D, page 29.

## 5 RESULTS

In this section, the results of the sentiment analysis as well as the statistical analysis are reported and supported by relevant tables and figures.

### 5.1 *Sentiment Modeling*

Firstly, the performance of the lexicon models was evaluated based on their accuracy of classifying Reddit sentiment. After consulting the performance measures, a clear difference between the lexicon models and their performance can be seen, as is displayed in Table 2, below.

[20] https://docs.scipy.org/doc/scipy/reference/stats.html
[21] https://matplotlib.org/

Table 2: Performance of the Lexicon Models

The left column represents the three applied lexicon models and the majority vote model. The column on the right represents the accuracies obtained when testing the models' output with the annotated data set.

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **SentiWordNet** | **58%** | **0.51** | **0.48** | **0.49** |
| *Class: -1* | | *0.25* | *0.20* | *0.22* |
| *Class: 0* | | *0.61* | *0.76* | *0.68* |
| *Class: 1* | | *0.65* | *0.49* | *0.56* |
| **TextBlob** | **62%** | **0.60** | **0.48** | **0.47** |
| *Class: -1* | | *0.50* | *0.07* | *0.12* |
| *Class: 0* | | *0.59* | *0.86* | *0.70* |
| *Class: 1* | | *0.72* | *0.51* | *0.60* |
| **VADER** | **86%** | **0.87** | **0.87** | **0.87** |
| *Class: -1* | | *0.93* | *0.93* | *0.93* |
| *Class: 0* | | *0.88* | *0.86* | *0.87* |
| *Class: 1* | | *0.81* | *0.83* | *0.82* |
| ***Majority Vote*** | **71%** | **0.75** | **0.60** | **0.63** |
| *Class: -1* | | *0.80* | *0.27* | *0.40* |
| *Class: 0* | | *0.67* | *0.88* | *0.76* |
| *Class: 1* | | *0.79* | *0.66* | *0.72* |

The sentiment scores of -1, 0 and 1 allocated to the submissions represent the negative, neutral and positive classes in the confusion matrices, respectively. When comparing the results on the manually annotated subset, SentiWordNet scored an accuracy of 58%. In addition the weighted precision, recall and F1-scores were 0.51, 0.48 and 0.49, respectively. Precision, Recall and F1 were particularly low for the negative (-1) class, scoring only 0.25, 0.20 and 0.22, significantly pulling down the average. This could be due to the slight imbalance of the manually annotated data set. SentiWordNet was outperformed by TextBlob, which scored an accuracy of 62%. Similarly to SentiWordNet, TextBlob has a low recall and F1 score of 0.07 and 0.12 respectively for the negative class. This significantly impacts the weighted averages and its accuracy. Contrastingly, TextBlob had a much higher precision for the negative class. The bottleneck for both models is the classification of the negative class as the neutral class. VADER on the other hand, scored a 86% accuracy, significantly outperforming the other two models. The precision, recall and F1 scores were all above 0.81 and much more balanced, indicating that the model is not impacted by the slight class imbalance. Lastly, the majority vote model scored 71%, which corresponds to the lower performance of the first two models. Therefore, the polarities were assigned to all cryptocurrencies using the VADER lexicon model.

## 5.2 *Sentiment Descriptives*

Besides looking at the confusion matrices of the models, this section will briefly cover some descriptives of the sentiment output.

Within the sentiment classification, there is an evident class imbalance, which can be seen in the bar plot in Figure 3. However, the nature of the lexicon approach is individually looking at words of a text and assigning them a score. Therefore, the imbalance should not affect the model as much as it would in ML models, for example. Interestingly, all three models have classified approximately the same number of submissions as positive. The differences lie in the negative and neutral classes. TextBlob seems to classify more submissions as neutral, whereas VADER and SentiWordNet are more implied to classify submissions as negative.

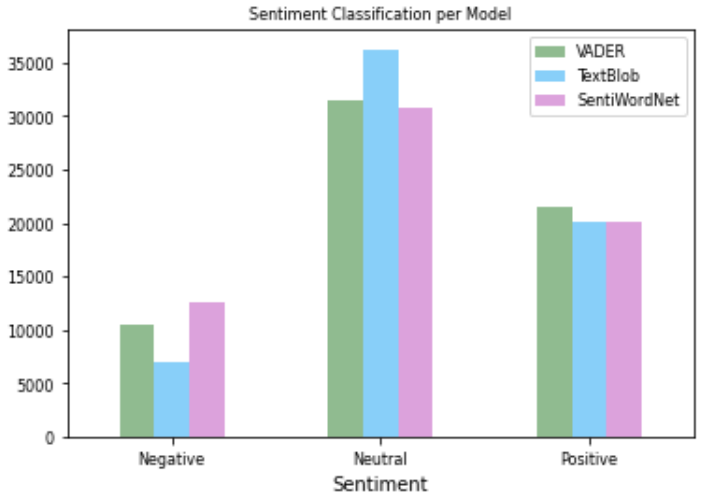



Figure 3: Distribution of Sentiment Classes per Model

The analysis represented by Figure 3 was further explored by looking at the VADER polarities per cryptocurrency. In addition, the polarities grouped for the stronger and weaker performing currencies are distinguished and analyzed. From the analysis, it can be seen that the overall trend for the negative class for all currencies is varying. For instance, only 10% of Cardano's submissions are negative, whereas 18% of submissions are negative for Binance. Both are listed in the group of stronger currencies. Similarly, from the weaker group 14% of Dogecoin's submissions were negative, whereas 25% of Compound's submission were negative.

When looking at the averages per group, the stronger currencies score a much lower percentage of negative submissions (14%) compared to the weaker currencies (20%). Moreover, the stronger currencies score a

higher percentage of positive submissions (32%) compared to the weaker currencies (28%). Naturally, the neutral class is the most balanced for all currencies. Even when looking at the differences per group, stronger currencies averaged 54% whereas the weaker currencies averaged 52%, only differentiating by 2%.

The full table can be consulted in Appendix E on page 31.

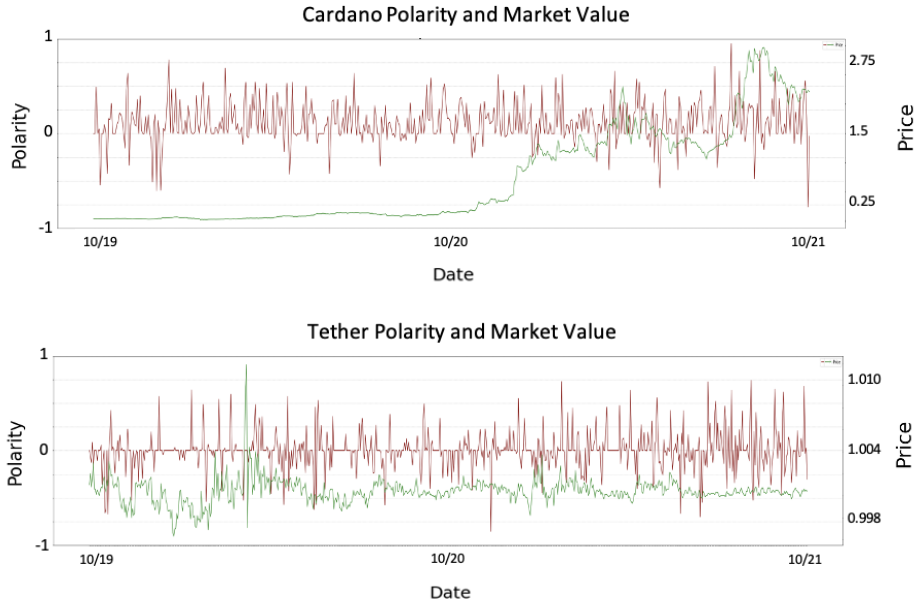
## 5.3 *Statistical Analysis*

Besides descriptives and qualitative data, it is important to visualize the analysis to get a good understanding and feel familiar with the data. Therefore, the polarities from the VADER model were individually graphed against the market value trend of every cryptocurrency. These graphs can be found in Appendix F on page 28 and Appendix G on page 29. From merely looking at the graphs, it can be seen that for the currencies that have been performing stronger in 2019-2021, the averaging polarities are a lot less extreme and tend to linger between -0.5 and 0.5 with occasional exceptions. For the weaker performing cryptocurrencies the sentiment seems to fluctuate more extreme between -1 and 1.

In order to provide guidance as to how the graphs were analysed, an example from both the stronger and weaker group were drawn. The two graphs in Figure 4 represent Cardano (strong) and Tether (weak) graphed.

Figure 4: Cardano and Tether Polarities versus Market Value

These graphs illustrate the daily averaged polarity scores (in red) and the respective market value (in green) graphed over two years. The x-axis represents the month and year, the left y-axis represents the polarity scores ranging from -1 to 1 and the right y-axis represents the market value range.




What immediately stands out from the polarity line in red, is that Tether shows a constant fluctuation between both positive and negative sentiment over the year with a clearly less promising trend in market value. This is to be expected from a weaker performing cryptocurrency. On the other hand, Cardano fluctuates towards negative sentiment with a lot less consistency and tends to remain more positive, above the 0 polarity line. The largest peaks in sentiment both positive and negative were in the last months where the market value was heavily increasing and decreasing.

Although the graphs give a different perspective to the analysis, it is clear that even with a stable currency sentiment can fluctuate significantly. Therefore, the next step is to confirm whether statistics back the previous statement. The results from the Pearson Correlation analysis can be found in Table 3 (page 22). Correlation 1 and 2 represent polarity vs change in price and comments vs change in price, respectively.

When considering a p value smaller than 0.05 as significant, there are a number of currencies where there is a significance. This is the case for Correlation 1 of the Cardano, Solana, Dogecoin and Compound currencies. For all four coins the PCC is somewhat low, scoring 0.10, 0.18, 0.11 and 0.18, respectively. Although not a strong correlation, the small p values confirm

that the correlation is not zero. As for the other individual currencies, the PCC is low and the p value is high. This means that the model can not state that the variables are correlated. For Correlation 2 the only significance was found for Bitcoin, where a PCC of 0.07 indicates a low correlation.

When grouping the strong and weak currencies, Correlation 1 has p values smaller than 0.05 for both groups. Although the PCC scores of both groups are low, interestingly the PCC score for weaker performing currencies (0.16) is twice as large as the PCC score for the stronger performing currencies (0.08). No associations were established for Correlation 2 for comments and change in price.

Table 3: Pearson Correlations

The two main columns represent the two correlation models. Every main column is subdivided in to the PCC score and the p value. The bottom two rows are the strong and weak currencies grouped together. Results marked with * indicate that the correlation is significant with p=<0.05)

| | **Correlation 1 Polarity vs Change in Price** | | **Correlation 2 Comments vs Change in Price** | |
|---|---|---|---|---|
| | **PCC** | **P value** | **PCC** | **P value** |
| **Binance** | -0.02 | 0.63 | -0.003 | 0.94 |
| **Bitcoin** | -0.005 | 0.88 | 0.07 | 0.05* |
| **Cardano** | 0.10 | 0.02* | -0.05 | 0.23 |
| **Solana** | 0.18 | 0.02* | -0.01 | 0.94 |
| **Dogecoin** | 0.11 | 0.002* | 0.003 | 0.92 |
| **Compound** | 0.18 | 0.01* | -0.01 | 0.91 |
| **Safemoon** | -0.01 | 0.86 | -0.19 | 0.06 |
| **Tether** | 0.03 | 0.36 | -0.002 | 0.95 |
| **Strong** | 0.08 | 0.00* | 0.01 | 0.61 |
| **Weak** | 0.16 | 0.00* | 0.01 | 0.77 |

# 6 DISCUSSION

## 6.1 *Research Question One*

*"How well do unsupervised lexicon models perform in the sentiment classification of Reddit data?"*

The first research question aimed to find which lexicon model best performed in labeling Reddit data. The models were evaluated in terms of accuracy and VADER outperformed SentiWordNet and TextBlob with an accuracy of 86%. Prior research differs greatly in choosing methods to label the data or do not mention the labeling process at all. With that being said,

VADER is an adequate starting point on which can be further developed. A suggestion for further research would be including cryptocurrency jargon as lexicon features. This way, popular terms such 'bullish', or 'ATH' (all-time-high) can be included in the assigning of polarity scores. In addition, most submissions are classified as neutral. This could be explained by the various topics of discussion around cryptocurrencies. For instance, discussion around the development of blockchain or other projects are mostly to inform the community or ask questions. These submissions imaginably do not always have a strong positive or negative sentiment. Therefore, further research may include topic modeling in order to ensure only relevant submissions are considered in the analysis.

## 6.2 *Research Question Two*

*"How does the sentiment analysis differ when comparing the stronger and weaker performing cryptocurrencies?"*

In order to better understand the relationship between sentiment and the market value, the sentiment is analysed with a distinction between the strong and weaker performing currencies. Intuitively, it would be expected that stronger currencies that have either grown significantly in value would have a high positive to negative ratio. For currencies that have enjoyed stable growth it would be expected that they have a larger positive and neutral to negative ratio.

When looking at the ratio's, this research has shown a larger positive to negative ratio for the stronger currencies over the past year. This quantifies the intuition previously mentioned. Furthermore, when comparing weaker currencies to the stronger group, they have a clear lower average of positive sentiment and a higher average of negative sentiment. Although both findings fall within the line of expectation, interestingly the weaker currencies still have a larger number of positive sentiment than they have negative sentiment. Generally, it can be stated that a distinction can be made between the sentiment of strong and weaker performing cryptocurrencies.

## 6.3 *Research Question Three*

*"To what extent do the fluctuations in sentiment correlate with the fluctuations in the market value of the cryptocurrencies?"*

The last research question addresses the newly introduced perspective of differentiating between cryptocurrency performance.

When analyzing correlations between price, polarity and number of comments, Correlation 1 was able to to substantiate a association for four out of eight cryptocurrencies. For Correlation 2, an association was found for only one cryptocurrency and is therefore not deemed a significant model. Furthermore, when grouping the stronger and weaker currencies, both groups find a significant correlation between polarity versus change of price. This is a contrasting difference to Correlation 2, where no significant association was found for either groups. As previous research has mentioned, social activity as well as the nature of cryptocurrency is complex. This justifies the low PCC values in the results and how this is to be expected with the volatile nature of cryptocurrencies. With patterns found beyond the general results, it is believed that the sentiment can be seen as a cog in a bigger wheel and therefore play an important part in a larger model.

### 6.4 *Limitations*

Although this research framework was set up critically, it does not come without limitations. Firstly, the manually annotated was labeled by one annotator, which can lead to a potential bias in the labels. In addition, the manually annotated data set only consisted of 100 submissions. Preferably, the labeled subset could have been larger, in order to be more thorough in the calculations of the accuracies. The above-mentioned topics were limited due to time constraints. In addition, the choice of analyzing altcoins means less data for the statistical analysis. Due to the fact that they have been founded recently, three out of eight coins have a data frame of less than 732 observations. In addition, it lead to imbalanced submission data from the web scraping where Binance and Bitcoin had significantly more observations than the other coins did. Moreover, the nature of cryptocurrencies is inherently complex as well as modeling behavior in social context. Prices could be underpinned by a number of additional features such as emotions, other external economic factors and long-term scenarios.

## 7 CONCLUSION

This paper examines to what extent sentiment on the social media platform Reddit can reveal pricing trends for the strong performing and weak performing cryptocurrencies of 2021. This was done using a data set consisting of 83,352 scraped submissions for eight currencies and the daily market

value of each individual currency from 2019 to 2021. As the Reddit data is unlabeled, a lexicon-based approach was used. In choosing the most appropriate model, three lexicon models were tested. This research shows VADER is an adequate lexicon-based tool for sentiment of unlabeled social text on which can be built further by adding relevant jargon to the lexicon. Furthermore, the sentiment class distribution showed an interesting distinction between strong and weaker performing currencies. Namely, the distribution of positive sentiment was larger for stronger currencies than it was for weaker currencies. Subsequently, weaker currencies had a larger proportion of negative sentiment than stronger currencies did. Lastly, the statistical analysis revealed low, yet significant correlations for polarity and change in price for both the strong and weak currencies grouped. No significant correlations were found for the correlation between the number of comments and the change in price for both groups. Therefore, it can be stated that Reddit sentiment is not a strong stand-alone factor in the market value trend. Nevertheless, sentiment does have potential to play a supporting role in a broader model.

## 8 DATA SOURCE / CODE

This research was conducted using data from Reddit and Yahoo Finance, as well as Python code. Statements with regards to the sources used in this paper are listed below[22]:

- "Work on this thesis did not involve collecting data from human participants or animals.
- The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.
- The author of this thesis acknowledges that they do not have any legal claim to this data or code.
- The code and full data set used in this thesis can be found on: https://github.com/emilycoppens/masterthesis.git

## 9 REFERENCES

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, *100*(5), 992–1026.

[22] This sample text is frrom the Tilburg University DS&S Masters Thesis guidelines -vF2021

Bonta, V., & Janardhan, N. K. N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1–6.

Broadstock, D. C., & Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. *Finance Research Letters*, *30*, 116–123.

Chen, H., De, P., Hu, Y., & Hwang, B.-H. (2011). Sentiment revealed in social media and its effect on the stock market. In *2011 ieee statistical signal processing workshop (ssp)* (pp. 25–28).

Costola, M., Iacopini, M., & Santagiustina, C. R. (2021). On the" mementum" of meme stocks. *arXiv preprint arXiv:2106.03691*.

Dhaoui, C., Webster, C. M., & Tan, L. P. (2017). Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*.

Fadayevatan, R., Alizadeh-Khoei, M., Hessami-Azar, S. T., Sharifi, F., Haghi, M., & Kaboudi, B. (2019). Validity and reliability of 11-face faces pain scale in the iranian elderly community with chronic pain. *Indian journal of palliative care*, *25*(1), 46.

Glenski, M., Saldanha, E., & Volkova, S. (2019). Characterizing speed and scale of cryptocurrency discussion spread on reddit. In *The world wide web conference* (pp. 560–570).

Hulley, S. B. (2007). *Designing clinical research*. Lippincott Williams & Wilkins.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

Jurek, A., Mulvenna, M. D., & Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, *4*(1), 1–13.

Kaminski, J., & Gloor, P. A. (2014). Nowcasting the bitcoin market with twitter signals. *CoRR*, *abs/1406.7577*. Retrieved from http://arxiv.org/abs/1406.7577

Kawanabe, E., Suzuki, M., Tanaka, S., Sasaki, S., & Hamaguchi, T. (2018). Impairment in toileting behavior after a stroke. *Geriatrics & gerontology international*, *18*(8), 1166–1172.

Knittel, M. L., & Wash, R. (2019). How" true bitcoiners" work on reddit to maintain bitcoin. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–6).

Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence. *Journal of Infection and Public Health*.

Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Ver Steeg, G., & Galstyan, A. (2021). Identifying and analyzing cryptocurrency manipulations in social media. *IEEE Transactions on Computational Social Systems*, *8*(3), 607–617.

Mittal, A., Dhiman, V., Singh, A., & Prakash, C. (2019). Short-term bitcoin price fluctuation prediction using social media and web search data. In *2019 twelfth international conference on contemporary computing (ic3)* (pp. 1–6).

Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1354–1364).

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, *42*(24), 9603–9611.

Ong, B., Lee, T. M., Li, G., & Chuen, D. L. K. (2015). Evaluating the potential of alternative cryptocurrencies. In *Handbook of digital currency* (pp. 81–135). Elsevier.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*(5), 1763–1768.

Smuts, N. (2019). What drives cryptocurrency prices? an investigation of google trends and telegram sentiment. *ACM SIGMETRICS Performance Evaluation Review*, *46*(3), 131–134.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, *29*(4), 217–248.

Sul, H. K., Dennis, A. R., & Yuan, L. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, *48*(3), 454–488.

Telli, Ş., & Chen, H. (2021). Multifractal behavior relationship between crypto markets and wikipedia-reddit online platforms. *Chaos, Solitons & Fractals*, *152*, 111331.

Wooley, S., Edmonds, A., Bagavathi, A., & Krishnan, S. (2019). Extracting cryptocurrency price movements from the reddit network sentiment. In *2019 18th ieee international conference on machine learning and applications (icmla)* (pp. 500–505).

Yahya, M. A., & Chiu, V. (2021). The meme stock paradox. *forthcoming in (Arizona State) Corporate and Business Law Journal (Winter 2022)*.

Zhang, H., Gan, W., & Jiang, B. (2014). Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th*

*web information system and application conference* (p. 262-265). doi: 10.1109/WISA.2014.55

# 10 APPENDICES

## APPENDIX A: STRONG PERFORMING CURRENCIES

Table 4: List of subreddits for strong-performing currencies

This table includes in what year the currency was founded, which subreddits have been created and how many users are active on the subreddit.

| Currency | Founded In | Dedicated Subreddit | Number of Users |
|---|---|---|---|
| **Binance** | 2017 | https://www.reddit.com/r/binance/ | 644K |
| | | https://www.reddit.com/r/BinanceExchange/ | 47.3K |
| **Bitcoin** | 2009 | https://www.reddit.com/r/Bitcoin/ | 3.4M |
| **Cardano** | 2017 | https://www.reddit.com/r/cardano/ | 620K |
| | | https://www.reddit.com/r/CardanoMarkets/ | 9.7K |
| | | https://www.reddit.com/r/CardanoCoin/ | 2.7K |
| | | https://www.reddit.com/r/CardanoTrading/ | 7.2K |
| **Solana** | 2020 | https://www.reddit.com/r/solana/ | 64K |
| | | https://www.reddit.com/r/SolanaNFT/ | 922 |
| | | https://www.reddit.com/r/solanatech/ | 255 |
| | | https://www.reddit.com/r/Solanax/ | 4.5K |

## APPENDIX B: WEAK PERFORMING CURRENCIES

Table 5: List of subreddits for weak-performing currencies

This table includes in what year the currency was founded, which subreddits have been created and how many users are active on the subreddit.

| Currency | Founded In | Dedicated Subreddit | Number of Users |
|---|---|---|---|
| **Dogecoin** | 2013 | https://www.reddit.com/r/dogecoin/ | 2.2M |
| | | https://www.reddit.com/r/dogecoinbeg/ | 12.6K |
| | | https://www.reddit.com/r/dogecoindev/ | 23.2K |
| | | https://www.reddit.com/r/DogeCoinFaucets/ | 2.8K |
| | | https://www.reddit.com/r/dogecoinchallenge/ | 568 |
| | | https://www.reddit.com/r/dogemarket/ | 25.1K |
| | | https://www.reddit.com/r/dogeducation/ | 11.6K |
| **Compound** | 2020 | https://www.reddit.com/r/Compound/ | 8.8K |
| **Safemoon** | 2017 | https://www.reddit.com/r/SafeMoon/ | 269K |
| | | https://www.reddit.com/r/SafeMoonBuySellAdvice/ | 9.3K |
| | | https://www.reddit.com/r/SafeMoonInvesting/ | 2.7K |
| | | https://www.reddit.com/r/SafemoonNews/ | 7.2K |
| **Tether** | 2020 | https://www.reddit.com/r/Tether/ | 8.4K |
| | | https://www.reddit.com/r/usdt/ | 610 |

## APPENDIX C: ADDITIONAL SUBREDDITS

Table 6: List of general subreddits

This table includes subreddits not specific to any currency, but used for discussion on investing topics in general.

| Subreddit | Number of Users |
|---|---|
| https://www.reddit.com/r/crypto | 209K |
| https://www.reddit.com/r/investing/ | 1.9M |
| https://www.reddit.com/r/altcoin/ | 192K |
| https://www.reddit.com/r/CryptoCurrency/ | 3.5M |
| https://www.reddit.com/r/CryptoCurrencies/ | 257K |
| https://www.reddit.com/r/CryptoMarkets/ | 604K |

## APPENDIX D: IMPORTED PYTHON PACKAGES

- import pandas as pd
- import numpy as np
- import requests

- import json
- import csv
- import time
- import datetime
- import textblob
- from textblob import TextBlob
- import string
- import nltk
- import ssl
- from nltk.stem import WordNetLemmatizer
- from nltk.corpus import stopwords
- from nltk.tokenize import word_tokenize
- import spacy
- import re
- from nltk.corpus import sentiwordnet as swn
- from IPython.display import clear_output
- import plotly.express as px
- import seaborn as sns
- import matplotlib.pyplot as plt
- import plotly
- from nltk.corpus import wordnet as wn
- from nltk.corpus import sentiwordnet as swn
- from nltk.sentiment.vader import SentimentIntensityAnalyzer
- from sklearn.metrics import confusion_matrix
- from sklearn.metrics import classification_report
- from collections import Counter
- import matplotlib.gridspec as gridspec
- from gridspec import GridSpec

## APPENDIX E: SENTIMENT CLASS DISTRIBUTION

Table 7: Class Distribution per Cryptocurrency

This table summarizes the distribution of sentiment classes per cryptocurrency, in addition to the distribution averages for the strong and weak-performing groups.

| **Cryptocurrency** | **Class Distribution** | | |
|---|---|---|---|
| *Strong* | **Negative** | **Neutral** | **Positive** |
| **Binance** | 18% | 51% | 31% |
| **Bitcoin** | 18% | 53% | 29% |
| **Cardano** | 10% | 56% | 34% |
| **Solana** | 14% | 54% | 32% |
| *avg* | *14%* | *54%* | *32%* |
| *Weak* | | | |
| **Dogecoin** | 14% | 57% | 29% |
| **Compound** | 25 % | 48% | 26% |
| **Safemoon** | 20% | 53% | 27% |
| **Tether** | 21% | 48% | 30% |
| *avg* | *20%* | *52%* | *28%* |

## APPENDIX F: STRONG CURRENCIES: GRAPHING SENTIMENT AND MARKET VALUE

The figures below consist of four graphs. The top three graphs are a visualisation of the three correlation models. The first graph represents the 'Change in Price' versus the 'Polarity' variables. The second graph represents the 'Change in Price' versus the 'Number of Comments' variables and the last graph represents the 'Polarity' versus 'Price' variables.

The graph at the bottom represents the polarity in red, and the market value in green. On the x-axis the months and year are represented, the left y-axis represents the polarity and the right y-axis represents the market value of the cryptocurrency.

Figure 5: Binance

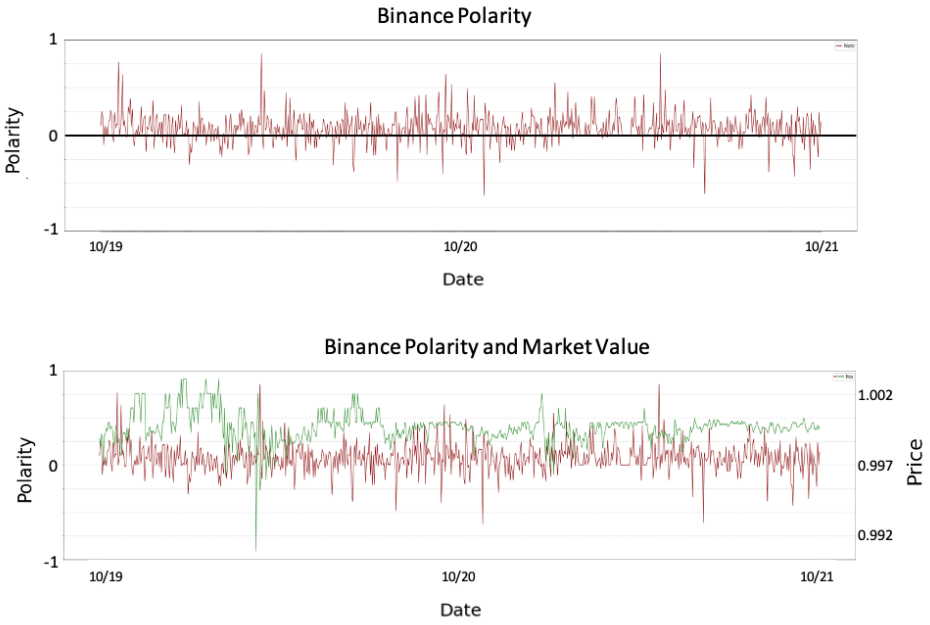



Figure 6: Bitcoin

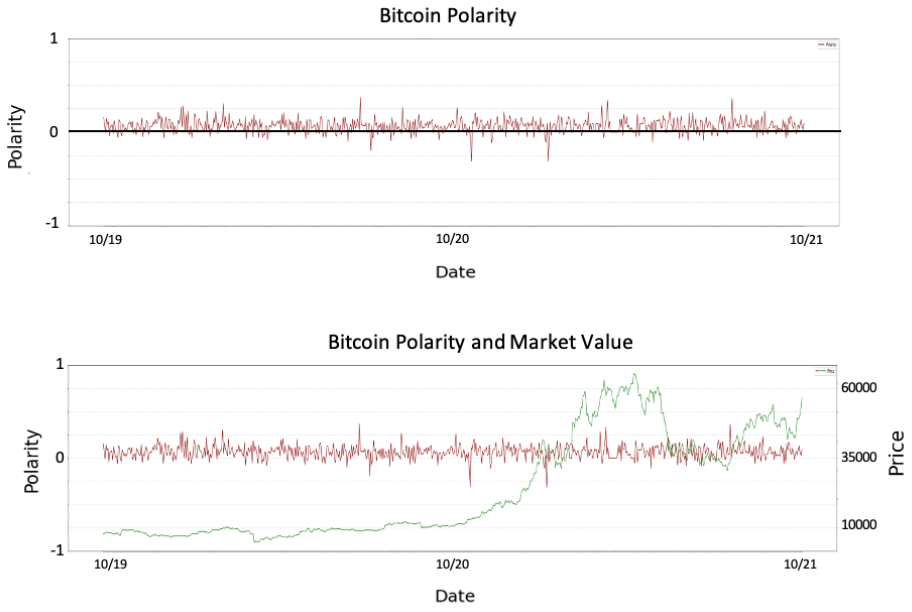


Figure 7: Cardano

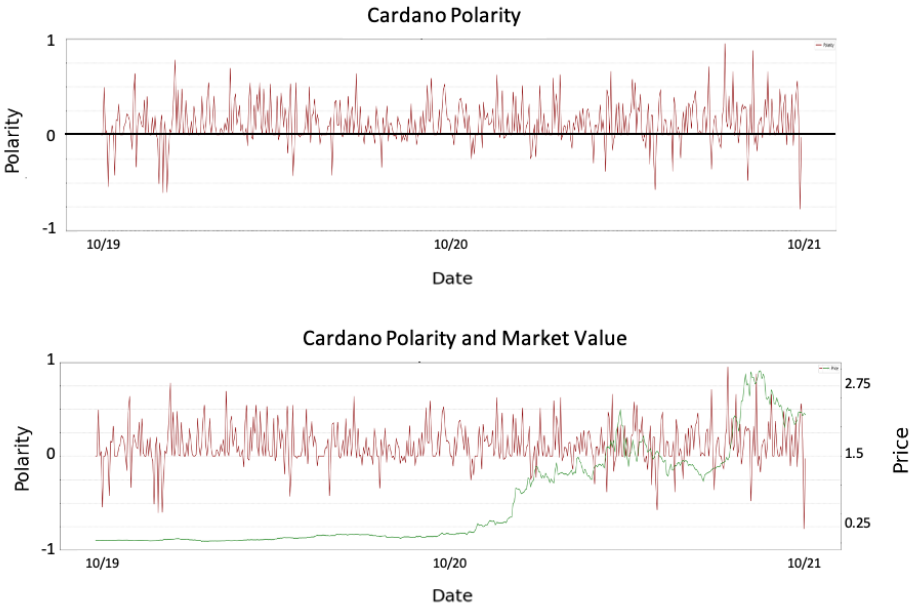
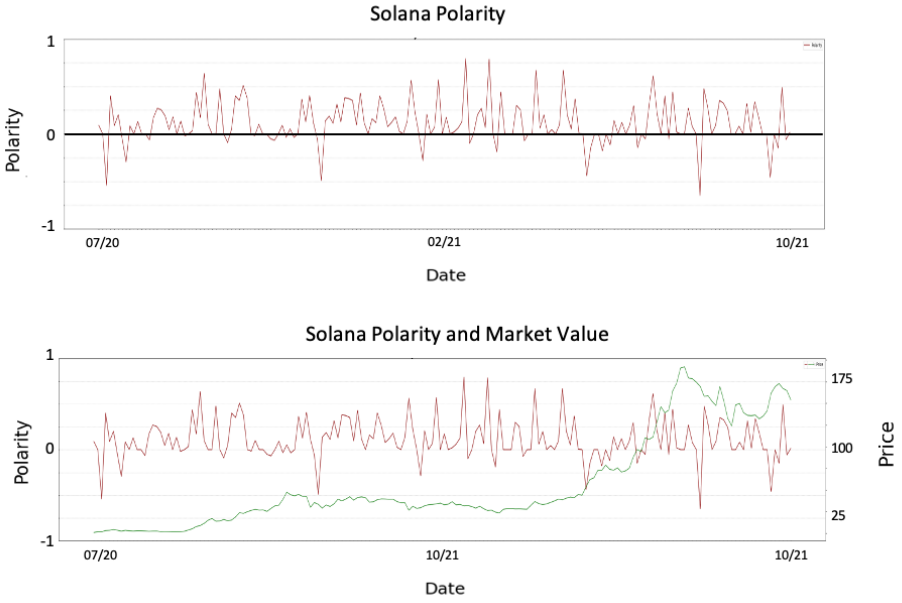



Figure 8: Solana



## APPENDIX G: WEAK CURRENCIES: GRAPHING SENTIMENT AND MARKET VALUE

The figures below consist of four graphs. The top three graphs are a visualisation of the three correlation models. The first graph represents the 'Change in Price' versus the 'Polarity' variables. The second graph represents the 'Change in Price' versus the 'Number of Comments' variables and the last graph represents the 'Polarity' versus 'Price' variables.

The graph at the bottom represents the polarity in red, and the market value in green. On the x-axis the months and year are represented, the left y-axis represents the polarity and the right y-axis represents the market value of the cryptocurrency.

Figure 9: Dogecoin

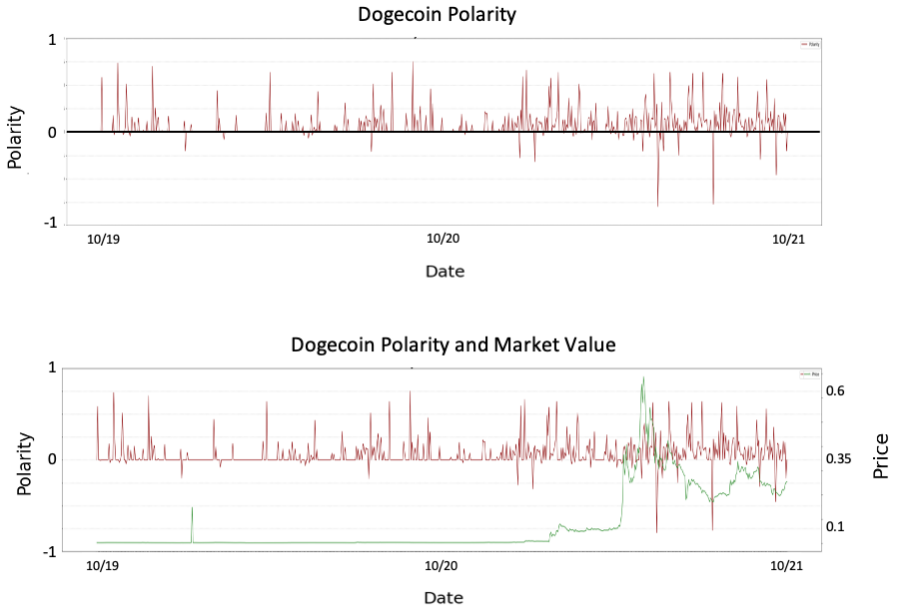


Figure 10: Compound

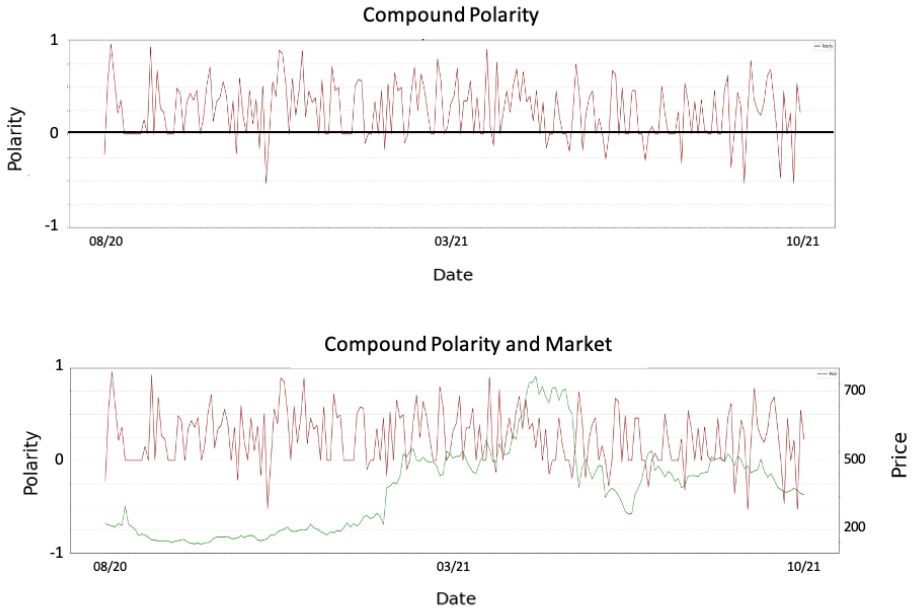



Figure 11: Safemoon

Figure 12: Tether