**Analyzing Cryptocurrency Tweets**



***BSAN-6200-Section-01****: Text Mining & Social Media Analytics*

Author(s):   
**Kiejstut Bunikiewicz, Neil Wagner**

*You may complete this project either individually or with a partner (that is, a group of two). However, each one of you must submit the final project files on your individual Brightspace for grading. In this case, please clearly indicate both partners’ names on each file. Both partners will receive the same grade.*

29 April 2021

**Table of Contents**

[**Background 1**](#_heading=h.gjdgxs)

[**Objective 1**](#_heading=h.30j0zll)

[**PROJECT SUMMARY 1**](#_heading=h.3dy6vkm)

[**DATA PROCESSING**](#_heading=h.1t3h5sf) **2**

[**Data COLLECTION 2**](#_heading=h.4d34og8)

[**Data CLEANING 3**](#_heading=h.2s8eyo1)

[**DATA ANALYSIS 3**](#_heading=h.17dp8vu)

[**Specifics 4**](#_heading=h.1fob9te)

[**General Patterns 8**](#_heading=h.3znysh7)

[**Recommendations 11**](#_heading=h.2et92p0)

[**For Managers 11**](#_heading=h.tyjcwt)

**For Investors 12**

*[ Click on Table of Contents above, and press F9, to automatically refresh ]*

# Background

## Objective

The goal of this report is to analyze tweets mentioning cryptocurrency and predict whether a tweet is positive or negative. Through our analysis, we will identify which words are most prevalent in tweets mentioning cryptocurrency. This information will serve as a guide for our marketing recommendations for those involved in the crypto business. Furthermore, we will determine whether Twitter could be a useful tool in cryptocurrency investment. Our findings in this project can be used to help cryptocurrency investors better understand public reception of cryptocurrency. Lastly, we will provide recommendations for individuals and businesses on how to make better informed decisions to maximize their profit.

## Business and problem overview/project summary

Investing in cryptocurrency has become incredibly popular within the past decade. According to Coinbase co-founder Fred Ehrsam, roughly 10 percent of the American population now owns cryptocurrencies (*Number of Americans Who Own Crypto Might Be Much Higher Than You Think*, 2021). This means that approximately 33 million Americans have purchased crypto at some point in time, a surprisingly large number. Ehrsam’s popular app Coinbase allows users to buy and sell various types of cryptocurrencies. One could use the app to convert one cryptocurrency to another, or to send and receive cryptocurrency to and from other users. As of January 2021, the app has already surpassed 40 million users.

Bitcoin, the most popular cryptocurrency, has approximately 100 million owners worldwide. “Never before have young people been able to change economic classes so quickly,” states Bitcoin investor Erik Finman, who is already a millionaire at the age of 19 (*19-year-old bitcoin millionaire: Here’s how much you should invest in cryptocurrencies*, 2018). Many investors, including Finman, believe that bitcoin is the safest cryptocurrency to invest in. Furthermore, as mentioned above, bitcoin is also the most popular cryptocurrency. With this information in mind, our analysis will focus mainly on bitcoin.

There are various myths and uncertainties regarding cryptocurrency. For instance, many believe that digital currencies don’t have any value, aren’t secure, or that may be a scam. We will explore these questions in our analysis and hope to provide a better understanding of the reality of crypto investment.

For our project, we used machine learning to predict whether a tweet about cryptocurrency is positive or negative in sentiment. We analyzed the results to identify any patterns and significant findings. Our results are intended to help cryptocurrency users better understand the relationship between tweets and cryptocurrency and what Twitter users are saying about cryptocurrency, which will lead to insight that can be used to improve investment decisions.

# Data processing

**Data Collection**

We used Dataset ID #39 on Tweets mentioning cryptocurrency for this project. The tweets are from the month of July 2017. The dataset contains 1,048,576 rows and 5 columns. The columns are: Date, Username, Tweet, Polarity, and Sensitivity.

Our initial plan was to compare the tweets with cryptocurrency values. We used a Gemini Exchange dataset that provides the open, high, low, close, and volume Bitcoin data. This dataset was downloaded online and can be found at this following link: <https://www.cryptodatadownload.com/data/gemini/>. We selected the hourly data and edited the Excel workbook to only contain data from July 2017. Unfortunately, however, we ran into errors when trying to match the date and time columns.

As a result, we switched our focus to creating a model that will be able to predict whether a cryptocurrency tweet has positive or negative sentiment. As mentioned above, the tweets are from Dataset ID #39.

**Data Cleaning**

A variety of data cleaning steps were taken before beginning our analysis. The Twitter username text and Tweet text had ‘b’ as the first character, followed by the actual text in quotation marks (e.g., b'According to Google Trends, South Africa has the most online searches concerning bitcoin in the world'). These strings were cleaned by using the “utf-8” decoder function in python.

The next step in cleaning the data was to drop the duplicate tweets. This was done by using the drop\_duplicates function in python. A total of 46,911 duplicate rows were dropped. We also adjusted the date column, as it was originally formatted to show both the date and time in a single column. Using the Text to Columns feature in Excel, we separated this data into new columns (date and time). The dataset did not come with column headers, so we added them in python.

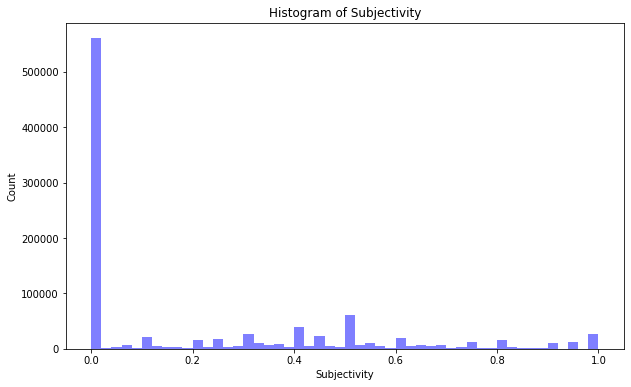
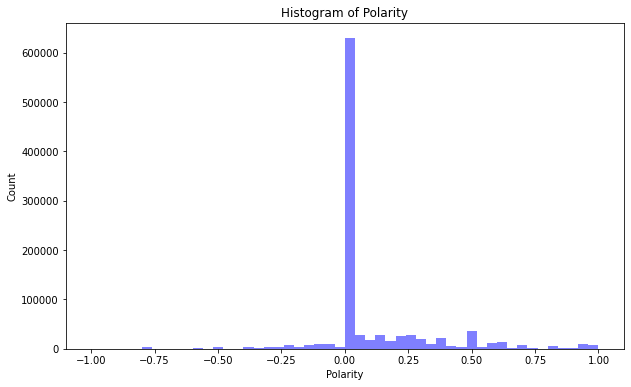
Various functions were created to remove unnecessary components from a tweet. For example, we wanted to remove any links, users, hashtags, and audio/video labels. Furthermore, we created a function to return a tokenized representation of the words in lemma form. We then appended the cleaned tweets into a list and created a new column labeled “Cleaned Tweet Texts.”

# Data analysis

Various word clouds were created with the dataset. The word clouds are useful for our analysis because they provide a graphical representation of the word frequency. The larger the word in the visualization, the more common it is in the document. Identifying the most frequent words from the tweets lead to a better understanding of what cryptocurrency users are interested in, which is very useful information for marketing and analytical purposes.

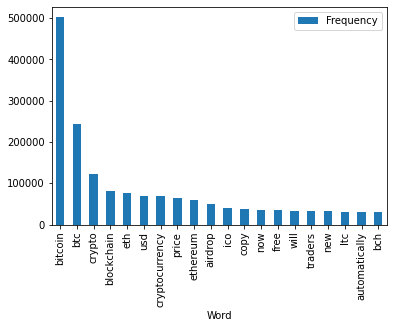
We also created a frequency chart to show a different representation of the word frequency within the tweets. The frequency chart displays the distribution of the most common words through the use of a bar graph visualization, providing a different perspective to the analyst. Not as many words appear on this chart compared to the word cloud, however the labeled y-axis provides a value to each word frequency.

A polarity histogram and a subjectivity histogram was also included in our analysis. The polarity values plotted in a histogram help highlight the overall sentiment (positive or negative) toward the subject. The polarity is a float which lies in the range of [-1, 1], where -1 means a negative statement and 1 means a positive statement. The subjectivity, on the other hand, refers to personal opinion, emotion, or judgment. The subjectivity score lies in a range of [0, 1]. The polarity histogram shows that a majority of the tweets in the dataset have a slightly positive sentiment. The subjectivity histogram shows that a majority of the tweets are objective. These two could be related, as objective statements do not often use strong emotional language.



*Figures 1 & 2. Histogram of Polarity and Histogram of Subjectivity*

## Specifics

We created a frequency chart with the cleaned text to show a visual representation of the most common words. The graph shows that bitcoin is by far the most frequent term. Other terms such as ethereum, airdrop, and automatically, are also present in the frequency chart. ICO is another term that appears on the chart, which is an acronym for initial coin offerings. ICOs are a form of cryptocurrency that businesses use in order to raise capital. Through ICO trading platforms, investors receive unique cryptocurrency tokens in exchange for their monetary investment. This is another example of Twitter being used as a means of business communication. The frequency chart (figure 2) can be seen below.

*Figure 2. Frequency chart*

A word cloud was created for positive tweets, which produced another interesting finding. Two popular terms that appeared are checkout and site checkout. This shows that crypto is being bought and sold through Twitter, which gives traders another convenient option aside from the traditional apps and websites commonly used.



*Figure 3. Positive tweets word cloud*

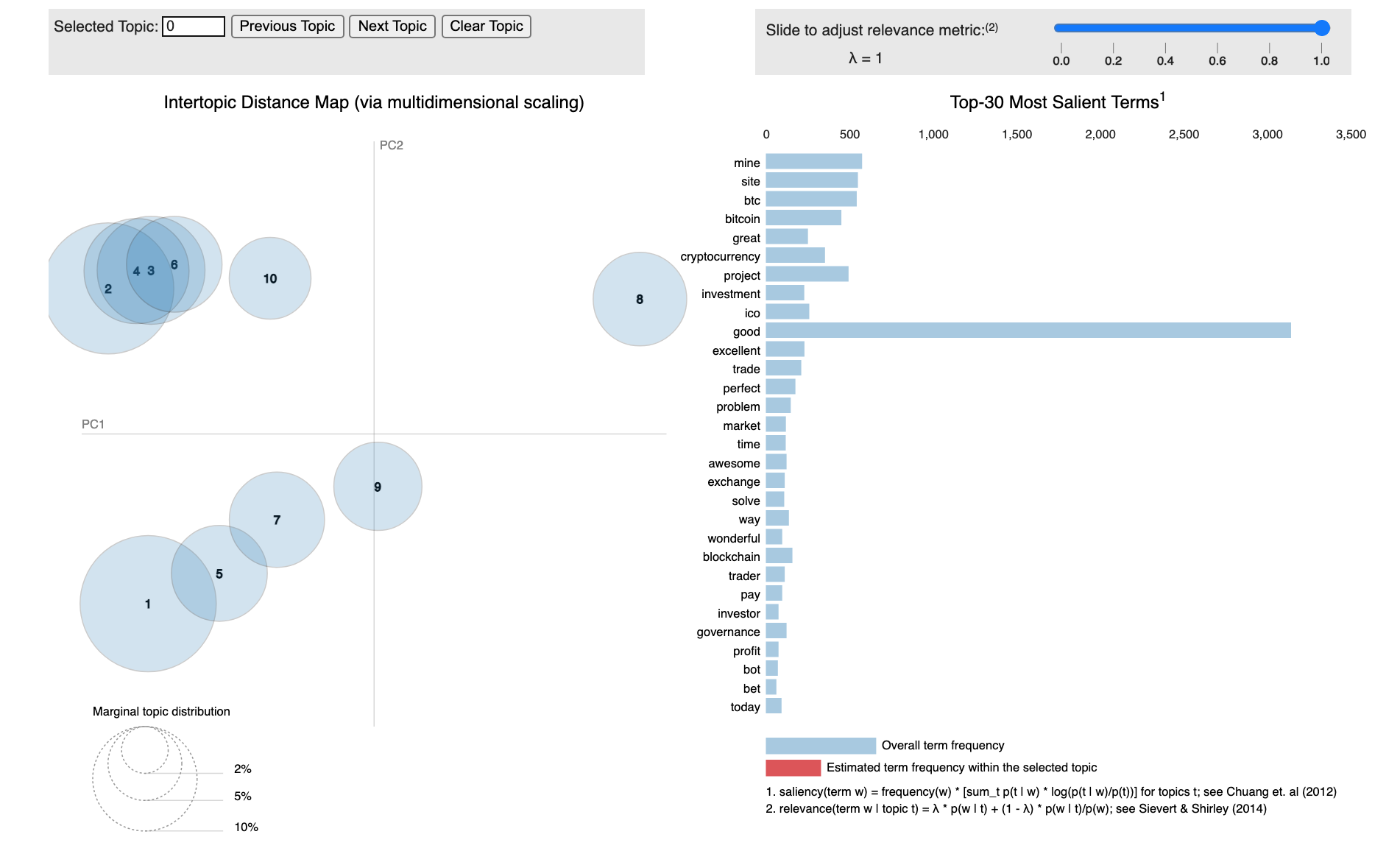
We also implemented two intertopic distance maps via multidimensional scaling to analyze each topic. One graph was created for the positive tweets and another graph for the negative tweets. The most salient term from the positive tweet graph is the word “good,” which could be a result of the level of confidence and satisfaction users experience with their bitcoin investment. Other terms include “great,” “excellent,” “perfect,” and “wonderful,” among others.

Figure 4. Intertopic Distance Map for Positive Tweet

The negative tweets map provided useful insight as well. For example, Topics 6 and 7 included a few alarming words such as “fraud,” “scam,” and “worthless.” This could indicate a presence of illegitimate traders, or possibly traders who are unhappy with their investment. These issues could also be coming from a lack of proper knowledge to effectively invest in a cryptocurrency.

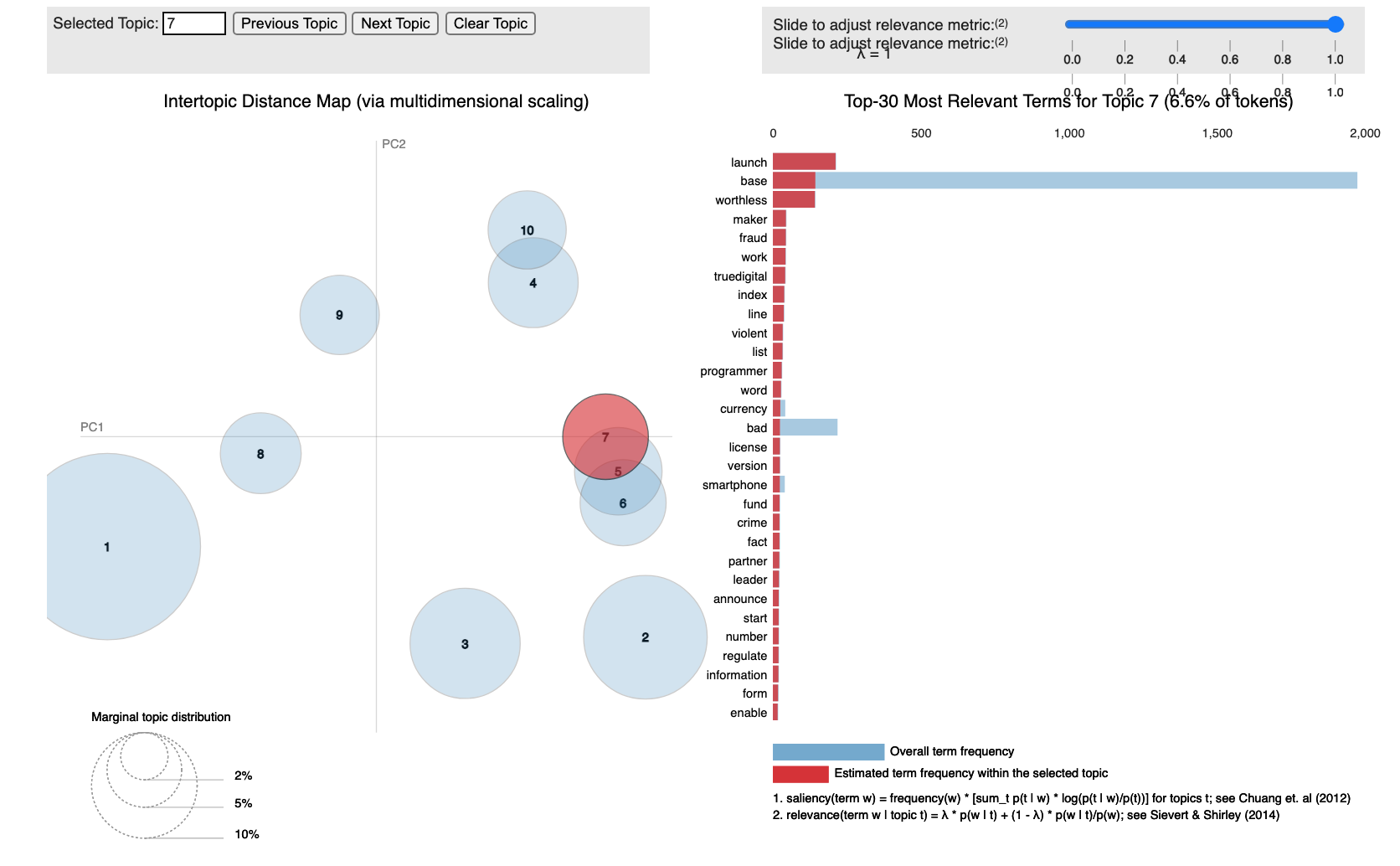


Figure 5. Intertopic Distance Map for Negative Tweets

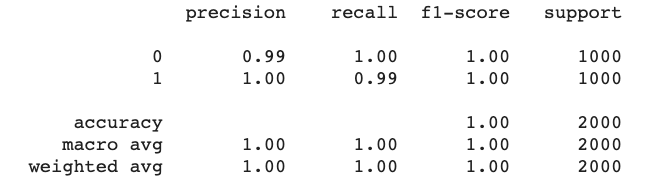
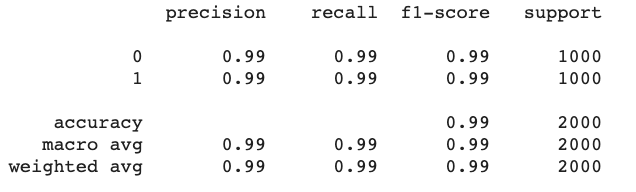
We also implemented logistic regression and Naive Bayes models to predict the likelihood of a tweet being positive or negative. First, we defined functions to create the prediction models from scratch. The results, however, did not seem reliable and parts of the code did not properly execute. As a result, we decided to use CountVectorizer to rerun the Naive Bayes and Logistic Regression models, which ultimately gave us better results. Below are screenshots of the precision, recall, and f1-score results for the Naive Bayes and Logistic Regression models.

Figure 6a. Naive Bayes results Figure 6b. Logistic Regression results

Both of these models were extremely effective classifiers with near-perfect scores. The logistic regression is slightly more effective than the Naive Bayes. However, these results are slightly odd because they are almost perfect, which is unusual for this model. This high degree of effectiveness may come from the fact that the sample included the 5000 most positive rows and the 5000 most negative rows in a dataset of over 1 million rows, which means that the positive and negative were very distinct.

## General Patterns

Multiple World Clouds were created to identify the most common words in the tweets. One Word Cloud was formed including all words from the cleaned text and another Word Cloud was formed with the words from the original text. Despite some similarities, the two Word Clouds displayed different results.

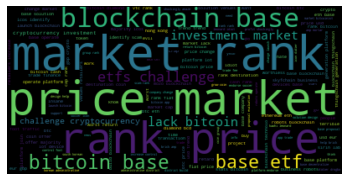


Figure 5a. World cloud with cleaned text Figure 5b. World cloud with original text

The most common word in both clouds is Bitcoin, which exemplifies how popular this specific cryptocurrency is. This should not come to a surprise, considering that Bitcoin’s value is the highest of all cryptocurrencies at over $60,000 a share. Another popular word from both graphs is the word “automatically.” Upon further research, it was discovered that this refers to the crypto trading bots that are automated to buy and sell cryptocurrencies at the correct time. This software can be an extremely useful tool for those looking into investing in cryptocurrency.

Another common term in both Word Clouds is Ethereum. Ethereum is an open-source, blockchain-based, decentralized software platform used for its own cryptocurrency, ether, which is on the rise. According to Yahoo Finance, Ethereum has already surpassed Bitcoin in terms of node count, transaction count, and total transaction fees (Sriram 2021). Contrary to Bitcoin, however, many analysts believe that the best way to make money using Ethereum is to buy ether and wait for it to increase in value over time. This means that if you are a long-term investor who won’t lose sight over short-term losses, Ethereum could be a good investment for you.

The term airdrop is also present in both graphs. In the cryptocurrency business, an airdrop is a marketing stunt that involves sending coins or tokens to wallet addresses in an attempt to promote awareness of the new virtual currency. This is an interesting finding that shows how crypto users are turning to Twitter as a means of marketing and communication.

Below is the world cloud for negative tweets. This word cloud displays the most common words for the tweets with negative sentiment.

# 

Figure 6. Negative tweets word cloud

After running the word clouds, we ran GloVe vectoring on positive and negative tweets in order to cluster the documents and get a better understanding of the word usage. Although the document clustering attempt with the GloVe vectors were not very useful, we attempted to gain some meaning by using TSNE graphs, which translate the several hundred vectors used in GloVe into a two dimensional graph. Although a majority of words were clustered closely, the TSNE graphs did reveal that certain words in both the positive and negative samples were unusual compared to the other words.



Figure 7a. Positive TSNE graph

Although a portion of the graph is difficult to read, we can still use the isolated words to draw conclusions. For example, we can determine from this graph that words such as “bang” and “idle” are unusual in comparison to the rest of the positive words. Although we don’t understand why the words are unusual (since we don’t know which vectors are being used), they could be interesting for further study.

# 

Figure 7b. Negative TSNE graph

Similarly, our negative TSNE graph is difficult to interpret, however it can still be useful to examine the isolated words above. The graph shows that words such as “doomsday” and “consensus” are unusual in negative tweets, which could be a significant finding to consider for further analysis.

# Recommendations

## For Cryptocurrency Exchange Managers

* **Create educational videos/content** - Crypto trading apps (e.g., Coinbase) can create educational videos or social media content educating their users and audience of the safety and security of trading through their apps. The Intertopic Distance Map (figure 5) shows that one of the most relevant terms in Topic 6 is the word “scam.” Furthermore, relevant terms in Topic 7 include the words “worthless,” “fraud,” and “bad.” This may be the result of people who have had bad experiences with crypto trading in the past and are unaware of the safe options that exist. As a result, we recommend that the legitimate apps and websites for crypto trading create content (such as videos, blog posts, etc.) that will educate the public on their safe and secure services.
* **Promote on social media -** Our analysis showed the level of engagement displayed on social media (specifically Twitter) from crypto users worldwide. As a result, we recommend that any businesses involved in crypto trading use social media as a means of promoting their services and engaging with their users. Social media is a great way to reach out to others, especially for a business, and we recommend managers take advantage of this opportunity.
* **Communicate clearly -** Some words appear in both the positive and negative word clouds. However, certain words only appear in the negative word clouds or LDA topic models. For example “challenge” appears frequently in negative Tweets but infrequently in positive ones. Due to the high degree of distinction between negative and positive tweets, the words that appear frequently in negative Tweets have a high relationship with negative sentiment. Even if the words seem harmless, they may carry a deeper meaning for crypto investors that could raise red flags. Managers should communicate with investors carefully by avoiding words that carry negative sentiments.

**For Crypto Investors**

* **Engage with Twitter community** - We found that many crypto users are sharing their thoughts and predictions via Twitter. We recommend that investors take advantage of the large crypto community on Twitter for extra opinions and insight. Many users share their experiences with different cryptocurrencies as well as helpful tips to consider when investing. As a result, investors can learn more about the trends and behavior of certain cryptocurrencies by simply interacting with Twitter users.
* **Automate your investing** - Cryptocurrency is a 24/7 global market. This can be further exemplified by looking at the timestamps for the tweets, which were spread out throughout all times of the day. People are trading and investing in crypto at any given moment and the market is constantly changing. As a result, it could be very useful to automate your investing strategy. By observing the world cloud, one may notice the word “bot” or “trading bot.” Trading bots will automatically make investment decisions for you, which can be an extremely useful tool to identify smart investment decisions one may otherwise not realize.

**Work Cited**

Bica, John. “Tweet Topic Modeling: Cleaning and Preprocessing Tweets.” *Medium*, Towards AI, 6 Apr. 2021, pub.towardsai.net/tweet-topic-modeling-part-2-cleaning-and-preprocessing-tweets-e3a08a8b1770?gi=8fd624898214.

CryptoDataDownload, and CryptoDataDownload on. Gemini Data, www.cryptodatadownload.com/data/gemini/.

“How Many Bitcoin Users Are There?” *How Many People Use & Own Bitcoins? (2021)*, www.buybitcoinworldwide.com/how-many-bitcoin-users/.

Mashalkar, Atharva. “Sentiment Analysis Using Logistic Regression and Naive Bayes.” *Medium*, Towards Data Science, 7 Mar. 2021, towardsdatascience.com/sentiment-analysis-using-logistic-regression-and-naive-bayes-16b806eb4c4b.

“Number of Americans Who Own Crypto Might Be Much Higher Than You Think: Headlines: News.” *CoinMarketCap*, U.Today, 31 Jan. 2021, coinmarketcap.com/headlines/news/number-of-americans-who-own-crypto-might-be-much-higher-than-you-think.

Sriram, Samyuktha. “Ethereum Could Overtake Bitcoin, Messari Analyst Says.” *Yahoo! Finance*, Yahoo!, 19 Mar. 2021, finance.yahoo.com/news/ethereum-could-overtake-bitcoin-messari-165845569.html.

Umoh, Ruth. “19-Year-Old Bitcoin Millionaire: Here's How Much You Should Invest in Cryptocurrencies.” *CNBC*, CNBC, 11 Oct. 2018, www.cnbc.com/2018/02/02/bitcoin-millionaire-says-this-is-how-much-to-invest-in-cryptocurrency.html.