Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

In recent years, Bitcoin, a digital currency, is becoming more incredibly popular among investors. As shown in figure 1, the price of bitcoin surged dramatically from USD400 in 2016 to USD46,500 in 2021. Therefore, the purpose of this analysis is to better understand the recent context of bitcoin in form of tweets and news articles for new investors in the crypto market.

Part I: Data collection

There are two main sources used to gather text related to bitcoin. The first source is Twitter where the bitcoin word contains in tweets. The second source is bitcoin news articles from various sites such as Yahoo Finance, Cointelegraph, Reddit, and Bloomberg which were collected by using Google Search. Most of these sites are well-known among financial and bitcoin communities. However, there is a limitation in Google Search which is the maximum length of the results. Therefore, for some articles, the headers were cut-off.

Part II: Data structuring

Based on the data gather, there are two analyses that will be performed. The first framework is the sentiment analysis on the recent news articles which aimed to analyze the bitcoin market in terms of attitudes expressed in texts such as positive, negative, and uncertainty. Loughran dictionary will be used in this analysis since it contains the sentiments that better describe and more reasonable within the context of financial data such as litigious, uncertainty. Another framework used in the analysis is term frequency and inverse document frequency (TF-IDF) from Twitter and all news articles which will help find the important words from each source so that investors can focus on before making a decision.

Part III: Analysis

According to the sentiment analysis, there is a relatively well-balanced between positive and negative coverage of the recent news articles, as shown in figure 2. Although there would seem to be an upward trend in investing in bitcoin, which can frequently see in the current headlines such as profitability and leading words, there are also significant factors which investors should be concerned about, especially for those who are risk averse also known as conservative investors. This can clearly see that volatility is the most frequent word in both negative and uncertainty sentiments. Another consideration is the regulation within litigious sentiment which potentially be one of the main concerns since there is no physical bitcoin, no banks, or government back up. Therefore, there might be an issue with legal claims.

In another analysis from the TF-IDF framework, there are several aspects that investors should know before investing in bitcoin as shown in figure 4. The most important words from each source are as follows:

- **Hash Rate:** bitcoin’s network measuring unit indicates three aspects; healthy, power, and profitability.
- **Ether:** another digital currency that can be traded via online exchanges.
- **Elon Musk:** Tesla’s CEO whose company recently bought $1.5 billion in bitcoin and planning to accept bitcoin as a form of payment.
- **Guggenheim:** an asset management firm that has over USD230 billion AUM. The company is also interested in investing in bitcoin.

- **Cannabis:** based on external research, cryptocurrencies like Bitcoin is one of the payment solutions in the Cannabis industry. This is caused by the difficulty in traditional payment within banking industries. Hence, cannabis entrepreneurs utilize bitcoin to solve this issue.

Part IV: Conclusion

With all things considered, bitcoin has undoubtedly been an attractive market as can obviously see from an increase in the stock price over time as well as the positive words in news headlines. Although many companies and investors are talking about bitcoin these days, there are good reasons to be concerned about risks, particularly risk averse investors. As seen in the sentiment analysis, there are uncertainty, constraints, and regulation associated with bitcoin, so investors should study the information and do the risk-return tradeoff before making any decision.

Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

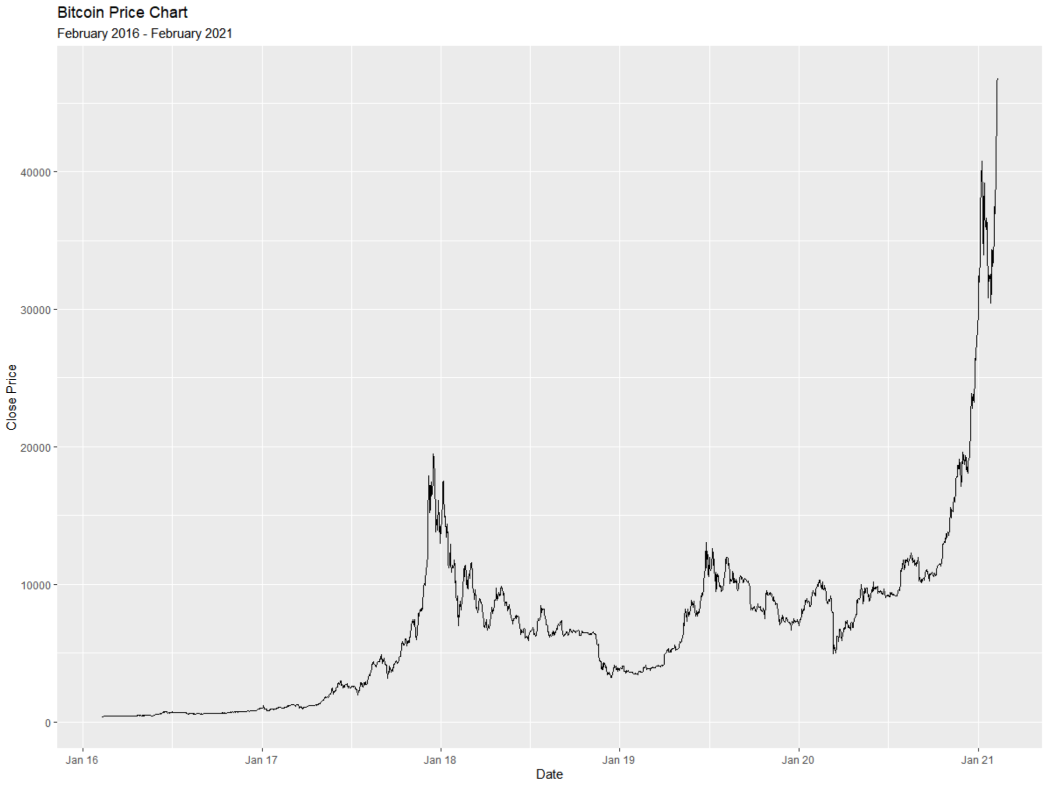
<u>Appendix</u>

R code with R output

```
################################################
## Text Analytics and NLP: Bitcoin
################################################

#load necessary packages
library(dplyr)
library(tidytext)
library(tidyverse)
library(stringr)
library(ggplot2)

#load CSV data
library(textreadr)
library(readr)

##############################
## Extract Bitcoin Price
##############################

#load Bitcoin price
btc_price <- read_csv("./datasets/csv/BTC-USD.csv")

#create new dataset without missing data
btc_price_nona <- na.omit(btc_price)

#create line chart
ggplot(btc_price_nona,
       aes(x = Date, y = Close)) +
   geom_line() +
   labs(x = 'Date',
        y = "Close Price",
        title = "Bitcoin Price Chart",
        subtitle = "February 2016 - February 2021" ) +
   scale_x_date(date_breaks = "year",
               date_labels = "%b %y")
```

<u>Output</u>: Figure 1: Bitcoin Price Chart

Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

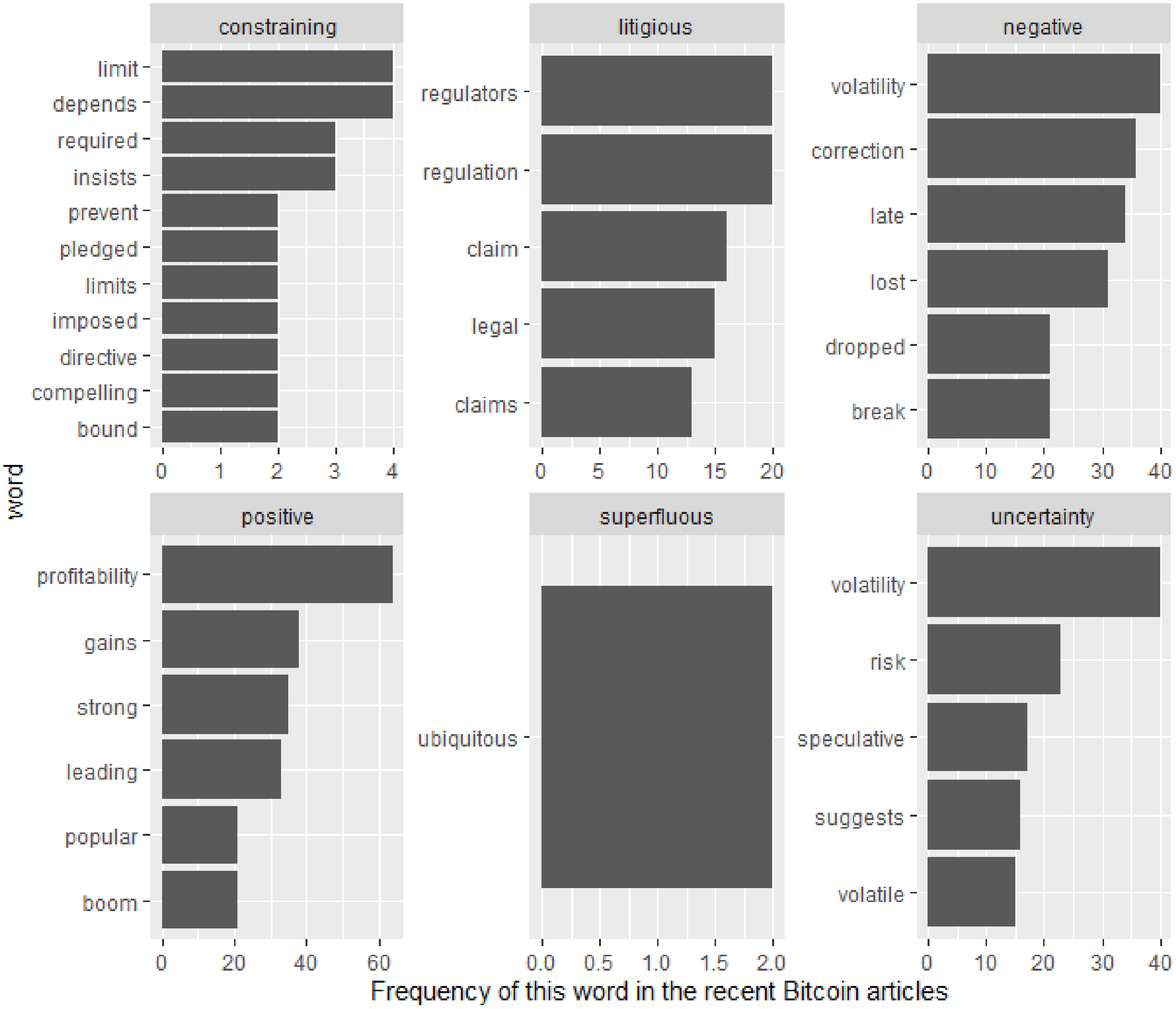


```
#########################################
## Analysis I: Sentiment analysis
#########################################

##############################
## Extract News articles
##############################

## Load news articles
library(pdftools)

# importing all PDF files
setwd("................/datasets/pdf") #specify local path!!

nm <- list.files(path="................/datasets/pdf")

my_pdf <- do.call(rbind, lapply(nm, function(x) paste(pdf_text(x), collapse = " ")))

# change column name to text
colnames(my_pdf) <- c("text")

# create new object to store news
bitcoin_news <- data.frame(line = 1:28, text = my_pdf)

## Working with stop words
# create custom stop words to remove uninformative words
```

Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

```
custom_stop_words_btc <- tribble(
  ~word,    ~lexicon,
  "crypto",    "CUSTOM",
  "cryptocurrency",    "CUSTOM",
  "cryptocurrencies",  "CUSTOM",
  "https",    "CUSTOM",
  "bitcoin",  "CUSTOM",
  "t.co",     "CUSTOM",
  "btc",      "CUSTOM",
  "1",        "CUSTOM",
  "3",        "CUSTOM",
  "5",        "CUSTOM",
  "24",       "CUSTOM",
  "10",       "CUSTOM",
  "100",      "CUSTOM",
  "yahoo",    "CUSTOM",
  "finance",  "CUSTOM"
)

# create new object for custom stop words
stop_words_btc <- stop_words %>%
  bind_rows(custom_stop_words_btc)

library(stringr)

# create vector for news sources
news_sources <- c('bitcoin','bitcoin','bitcoin',
                  'bloomberg','bloomberg','bloomberg',
                  'coindesk','coindesk','coindesk',
                  'cointelegraph','cointelegraph','cointelegraph',
                  'ft','ft','ft',
                  'medium','medium',
                  'newsbtc','newsbtc',
                  'reddit','reddit','reddit',
                  'wsj','wsj','wsj',
                  'yahoo','yahoo','yahoo')

tidy_news <-  bitcoin_news %>%
  add_column(sources = news_sources) %>%
  unnest_tokens(word, text) %>%
  filter(!str_detect(word, "\\d+")) %>%
  anti_join(stop_words_btc, by = "word")

# visualize
tidy_news %>%
  count(word) %>%
  inner_join(get_sentiments("loughran"), by = "word") %>%
  group_by(sentiment) %>%
  top_n(5, n) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n)) +
  geom_col() +
  coord_flip() +
  facet_wrap(~ sentiment, scales = "free") +
  ylab("Frequency of this word in the recent Bitcoin articles")
```

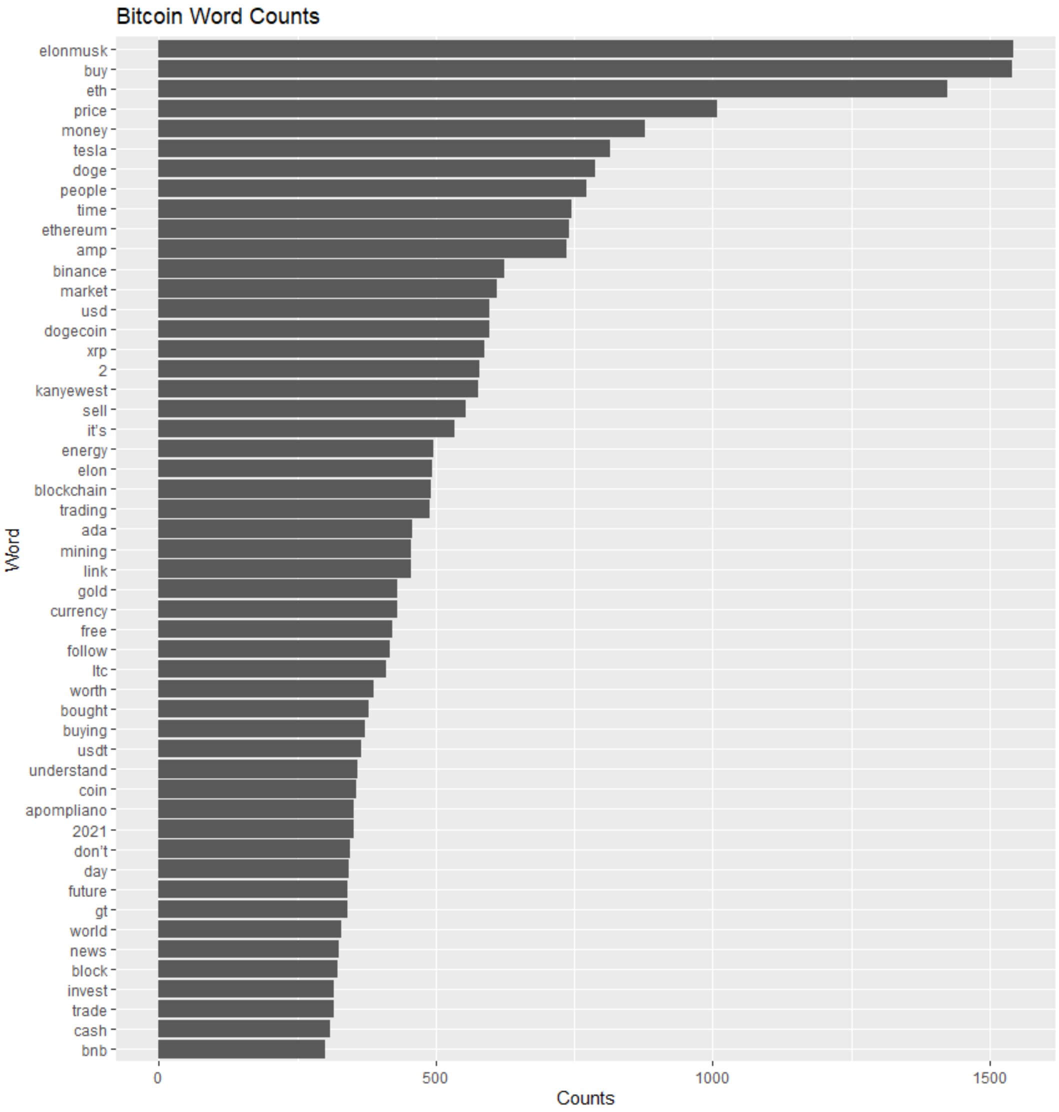
Output: Figure 2: The most common words in Bitcoin news articles associated with each sentiment in the Loughran lexicon.



```
############################################
## Analysis II: TF-IDF
############################################

library(rtweet)

#Keyword: BITCOIN or BTC
bitcoin <- search_tweets("bitcoin OR btc", n = 18000, include_rts = FALSE, lang = "en")
```

```
# create vector for twitter
add_twitter <- tibble(sources = "twitter")

# find the most common words across bitcoin tweet
tidy_bitcoin <- bitcoin %>%
    select(user_id, status_id, created_at, screen_name, source, text) %>%
    add_column(add_twitter) %>%
    unnest_tokens(word, text) %>%
    anti_join(stop_words_btc, by = "word")
word_counts_bitcoin <- tidy_bitcoin %>%
    count(word, sort = TRUE) %>%
    filter(n > 200) %>%
    mutate(word2_bitcoin = fct_reorder(word, n))

# create bar plot
ggplot(word_counts_bitcoin,
       aes( x = word2_bitcoin, y = n)) +
    geom_col() +
    coord_flip() +
    labs(
        title = "Bitcoin Word Counts",
        x = "Word",
        y = "Counts"
    )
```

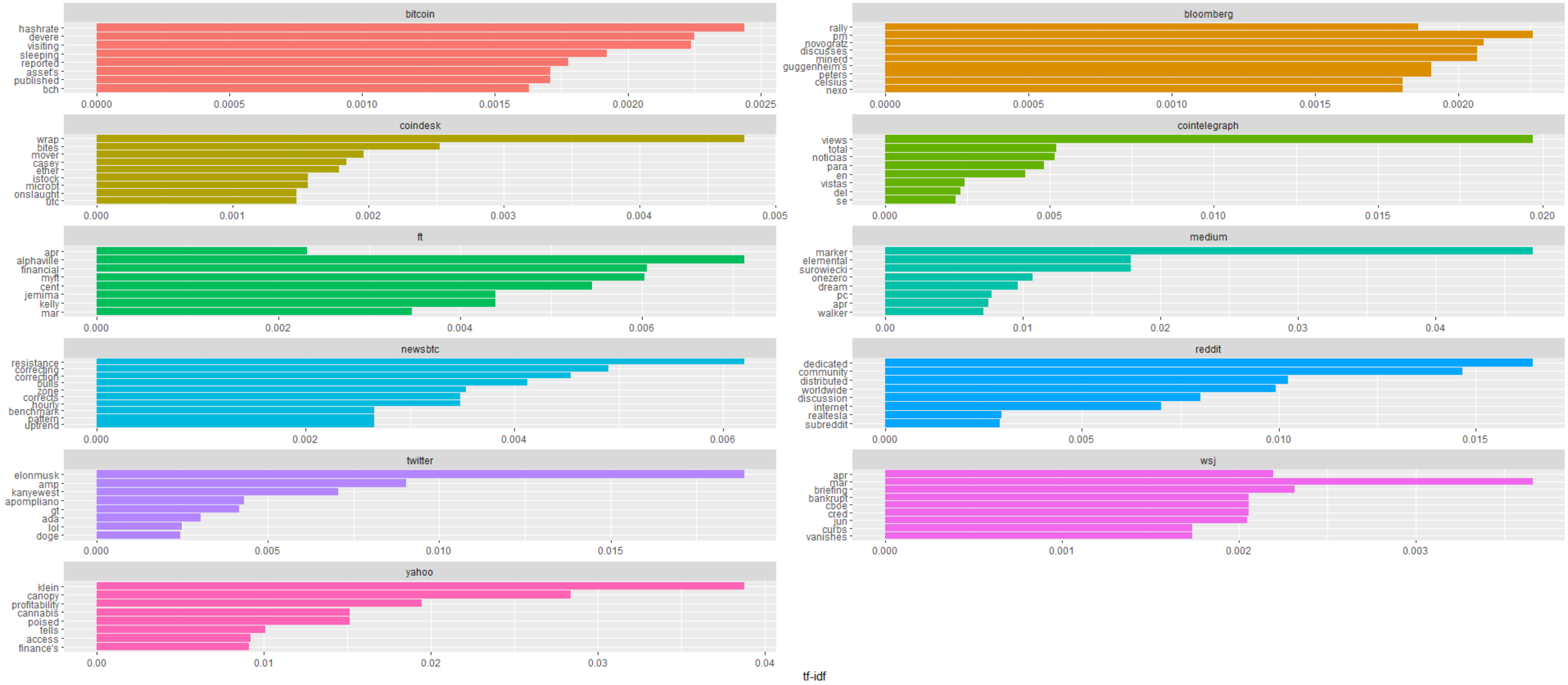
Output: Figure 3: The most common words in Bitcoin tweets.


Bitcoin Word Counts
elonmusk
buy
eth
price
money
tesla
doge
people
time
ethereum
amp
binance
market
usd
dogecoin
xrp
2
kanyewest
sell
it's
energy
elon
blockchain
trading
ada
mining
link
gold
currency
free
follow
ltc
worth
bought
buying
usdt
understand
coin
apompliano
2021
don't
day
future
gt
world
news
block
invest
trade
cash
bnb
Word
0
500
1000
1500
Counts

Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

```
# create custom stop words to remove uninformative words for TF-IDF
custom_stop_words_btc_idf <- tribble(
   ~word,    ~lexicon,
   "bitcoin.com",    "CUSTOM",
   "news.bitcoin.com",    "CUSTOM",
   "site:bitcoin.com",  "CUSTOM",
   "bloomberg",    "CUSTOM",
   "site:bloomberg.com",  "CUSTOM",
   "coindesk",    "CUSTOM",
   "site:coindesk.com",  "CUSTOM",
   "cointelegraph",    "CUSTOM",
   "site:cointelegraph.com",  "CUSTOM",
   "times",    "CUSTOM",
   "site:ft.com",  "CUSTOM",
   "medium",    "CUSTOM",
   "site:medium.com",  "CUSTOM",
   "reddit",    "CUSTOM",
   "site:reddit.com",  "CUSTOM",
   "newsbtc.com",    "CUSTOM",
   "site:newsbtc.com",  "CUSTOM",
   "newsbtc",    "CUSTOM",
   "journal",  "CUSTOM",
   "site:wsj.com",  "CUSTOM",
   "finance",  "CUSTOM",
   "street",  "CUSTOM" ,
   "wall",  "CUSTOM",
   "shutterstock",  "CUSTOM",
   "pst",  "CUSTOM",
   "el",  "CUSTOM",
   "de",  "CUSTOM",
   "sobre",  "CUSTOM",
   "la",  "CUSTOM",
   "su",  "CUSTOM",
   "los",  "CUSTOM",
   "votes",  "CUSTOM",
   "comments",  "CUSTOM",
   "ta",  "CUSTOM",
   "ft",  "CUSTOM"
)

# create new object for custom stop words
stop_words_btc_idf <- stop_words_btc %>%
   bind_rows(custom_stop_words_btc_idf)

# combine twitter and news articles data
news_tweet <- bind_rows(mutate(tidy_bitcoin),
              mutate(tidy_news))

news_words <- news_tweet %>%
   anti_join(stop_words_btc_idf, by = "word") %>%
    count(sources, word, sort = TRUE) %>%
    ungroup()

# sum total words by sources
total_words <- news_words %>%
    group_by(sources) %>%
    summarize(total = sum(n))

sources_words <- left_join(news_words, total_words)

# calculate term frequency
freq_by_rank <- sources_words %>%
    group_by(sources) %>%
    mutate(rank = row_number(),
           `term frequency` = n/total)

sources_words <- sources_words %>%
    bind_tf_idf(word, sources, n)  %>%
    select(-total) %>%
    arrange(desc(tf_idf))

#visualize
sources_words %>%
    arrange(desc(tf_idf)) %>%
    filter(!str_detect(word, "\\d+")) %>%
    mutate(word = factor(word, levels = rev(unique(word)))) %>%
    group_by(sources) %>%
    # filter(n<50) %>%
    top_n(8) %>%
    ungroup %>%
    ggplot(aes(word, tf_idf, fill = sources)) +
    geom_col(show.legend = FALSE) +
    labs(x = NULL, y = "tf-idf") +
    facet_wrap(~sources, ncol = 2, scales = "free") +
    coord_flip()
```

Output: Figure 4: Highest TF-IDF words in each source.


bitcoin
hashrate
devere
visiting
sleeping
reported
asset's
published
bch
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
bloomberg
rally
pm
novogratz
discusses
minerd
guggenheim's
peters
celsius
nexo
0.0000
0.0005
0.0010
0.0015
0.0020
coindesk
wrap
bites
mover
casey
ether
istock
microbt
onslaught
utc
0.000
0.001
0.002
0.003
0.004
0.005
cointelegraph
views
total
noticias
para
en
vistas
del
se
0.000
0.005
0.010
0.015
0.020
ft
apr
alphaville
financial
myft
cent
jemima
kelly
mar
0.000
0.002
0.004
0.006
medium
marker
elemental
surowiecki
onezero
dream
pc
apr
walker
0.00
0.01
0.02
0.03
0.04
newsbtc
resistance
correcting
correction
bulls
zone
corrects
hourly
benchmark
pattern
uptrend
0.000
0.002
0.004
0.006
reddit
dedicated
community
distributed
worldwide
discussion
internet
realtesla
subreddit
0.000
0.005
0.010
0.015
twitter
elonmusk
amp
kanyewest
apompliano
gt
ada
lol
doge
0.000
0.005
0.010
0.015
wsj
apr
mar
briefing
bankrupt
cboe
cred
jun
curbs
vanishes
0.000
0.001
0.002
0.003
yahoo
klein
canopy
profitability
cannabis
poised
tells
access
finance's
0.00
0.01
0.02
0.03
0.04
tf-idf

Text Analytics and Natural Language Processing (NLP): Bitcoin
Pimkarn Mekpruksawong

Sources

Frankenfield, J. (2021, January 30). Bitcoin. Retrieved from https://www.investopedia.com/terms/b/bitcoin.asp

Bitcoin USD (btc-usd) STOCK historical prices & data. (2021, February 10). Retrieved February 10, 2021, from https://finance.yahoo.com/quote/BTC-USD/history?period1=1455062400&period2=1612915200&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

Chen, J. (2020, September 29). What it means to be risk-averse. Retrieved February 11, 2021, from https://www.investopedia.com/terms/r/riskaverse.asp

SoFi. (2021, January 25). Bitcoin hash rate and why it matters. Retrieved February 11, 2021, from https://www.sofi.com/learn/content/bitcoin-hash-rate/

Some Bitcoin words you might hear. (n.d.). Retrieved February 11, 2021, from https://bitcoin.org/en/vocabulary

Reiff, N. (2020, August 28). Bitcoin vs. Ethereum: What's the difference? Retrieved February 11, 2021, from https://www.investopedia.com/articles/investing/031416/bitcoin-vs-ethereum-driven-different-purposes.asp

Aaron, M. (2020, March 24). We implemented cryptocurrency in our cannabis business. here's what we found. Retrieved February 11, 2021, from https://www.greenentrepreneur.com/article/347739