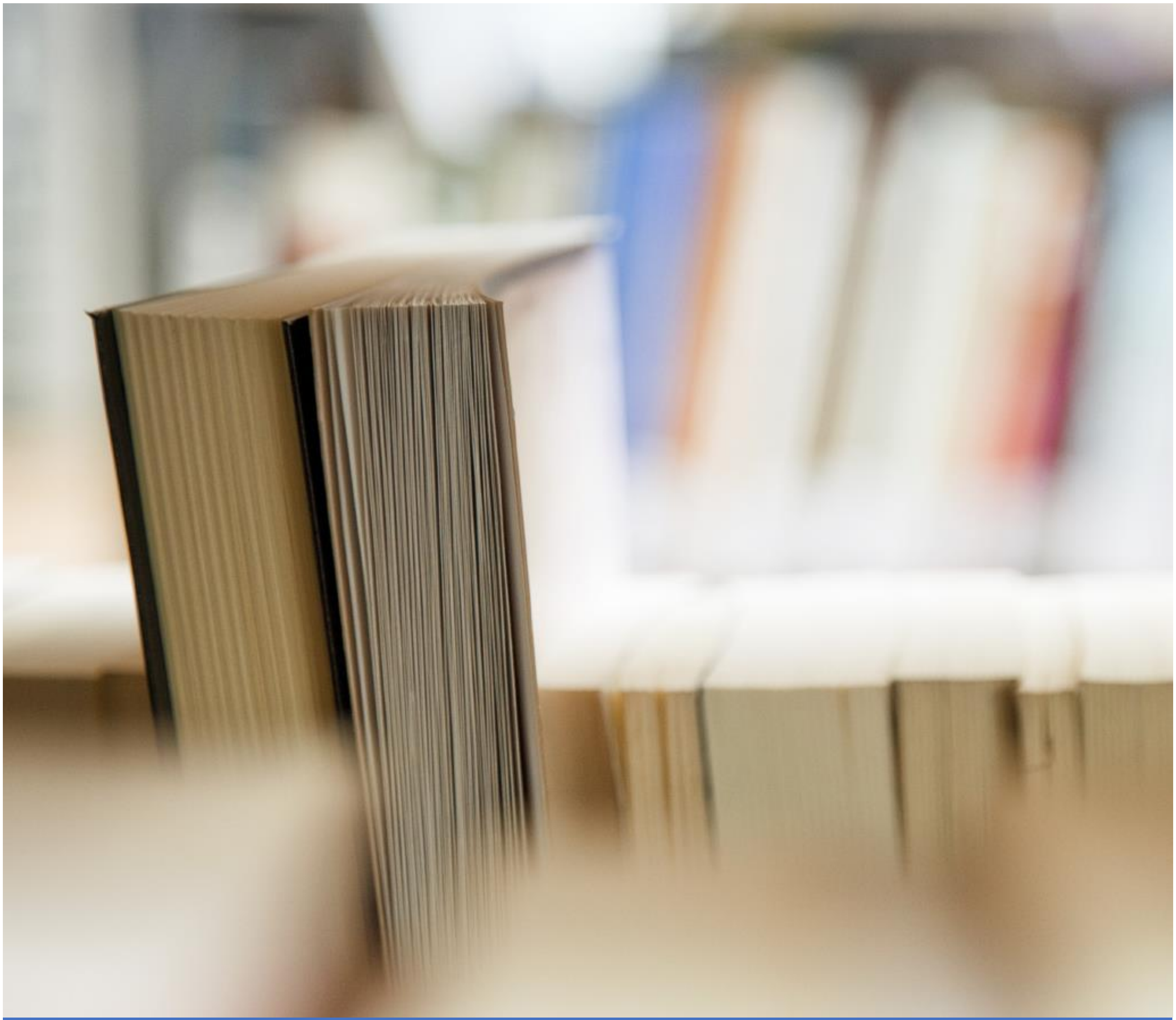


Text Analytics and Natural Language Processing (NLP)

A3: Business Insight Report

Naga Sai Venkata Surya RaviTeja Challa

10/2/21

Professor: Thomas Kurnicki

Introduction

A cryptocurrency is a digital-assets that are designed based on blockchain technology where it is used as a medium of exchange where the coins can be owned by any individuals which are stored in a secured ledger using cryptography. Bitcoin(BTC) being the heart of all the cryptocurrency coins trading at around $45,000 and the base coin Ethereum(ETH) trading at $1700 and Litecoin (LTC) which is a replica of bitcoin are the three most famous and high market cap coins to be noted in the cryptocurrency market.

Three articles were taken for this analysis to convert unstructured data to structured data, one article for each crypto coin. The reason for this analysis to check the different and common factors that could drive each crypto coin. To find patten of words that could positively and negatively effect the coins. The frameworks used for this analysis are tokenization, finding out the positive-negative effecting words, word-cloud analysis, sentiment analysis, frequency and correlograms of the three articles.

Analysis

The analysis started with loaded the 3 articles text files into r-studio and then tokenizing each article.

The most frequently appeared words for the Bitcoin dataframe were**- Bitcoin, 2013, currency, money, financial, network and transactions.**

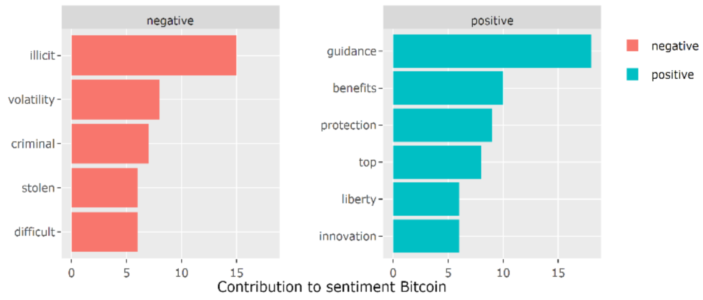
The most frequently appeared words for the Ethereum dataframe were**- Ethereum, network, smart, digital, blockchain, currency, contracts.**

The most frequently appeared words for the Litecoin dataframe were**- Litecoin, bitcoin, transactions, network, digital, blockchain, compared, lower.**

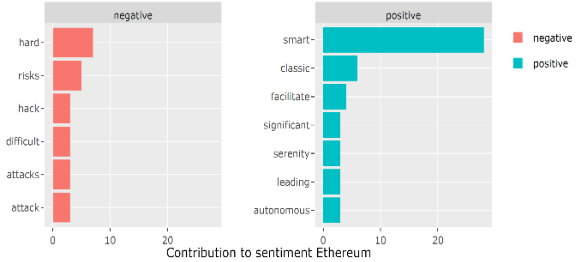
The most frequent words that appeared in all the 3 articles together were**- Currency, network, digital, blockchain and transactions.**

This shows the common features that these 3 cryptocurrency coins are related to that all the 3 coins are **currencies**, they are in their **digital** forms, the technology used is **blockchain** and their **transactions** take place for all the 3 crypto coins.

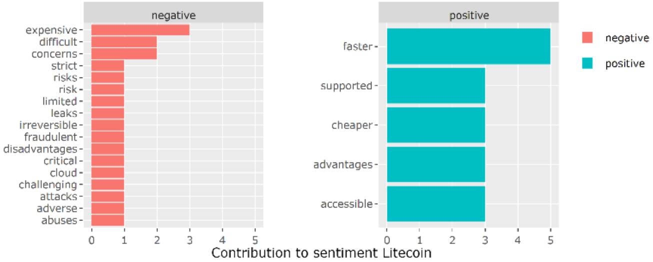
In the sentiment analysis for bitcoin dataframe, we can analysis that the words **guidance, benefits, protection, top, liberty** and innovation fall under positive whereas the **volatility, criminal, stolen** and difficult are under negative. This shows that their are **concern about the security and protection** of bitcoin.



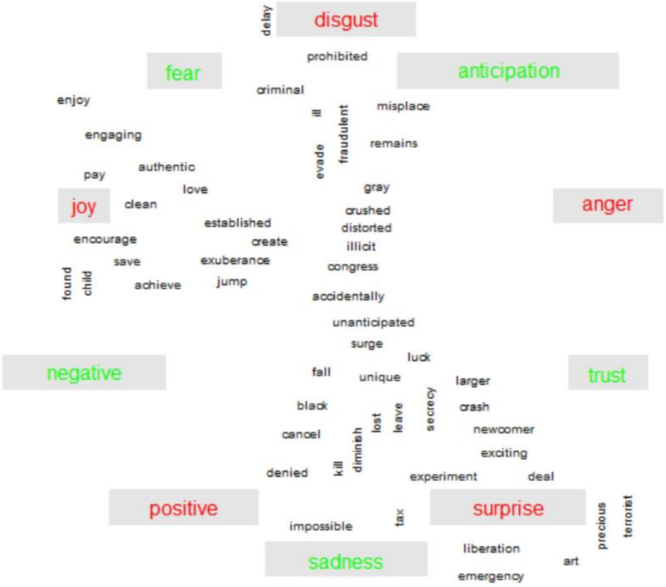
The Ethereum dataframe analysis shows that words like classic, significant, leading and autonomous are positive and risks, hard, difficult and attack are negative. This shows that Ethereum could be leading in its platform and also autonomous as well as significant that other coins. But its also seems like high in risk and more attacks than the other 2 coins.



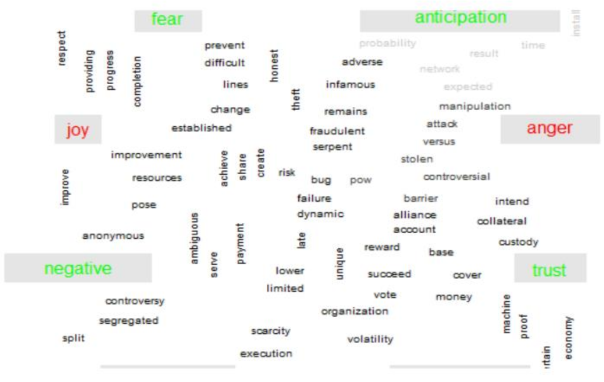
Litecoin is faster, cheaper and has more advantages and easily accessible than other 2 coins whereas its also has high risk, limited and security issues based on the negative sentiment of the coin.

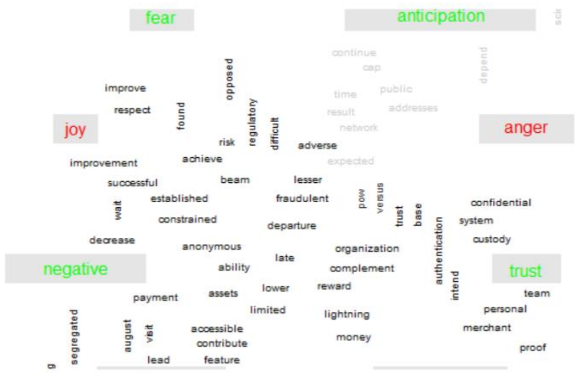


In the word cloud analysis for the bitcoin dataframe shows that most of the words are towards the surprise, joy, sadness and disgust. We can say that the investors who invest into bitcoin are joy as they might be profiting from their investment and also sad who might be the once who are in loss for their investment. Some words are leaning towards disgust and surprise which tells us that criminal effects are effecting the coins reputation and surprise for the way the bitcoin is increasing its value in the market.



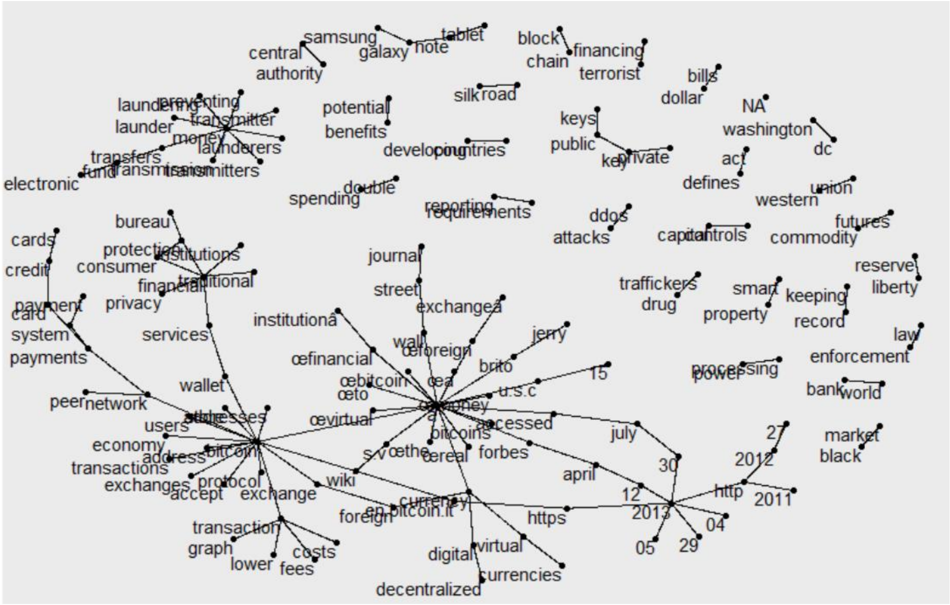
The word-cloud analysis for the Ethereum and Litecoin are more spread out compared to bitcoin which shows that these 2 coins have their positivity and negativity and, also wider investors maybe investing into these 2 coins. The words are spread around joy and as well as anger which shows that the group of people who invest might be different from bitcoin.



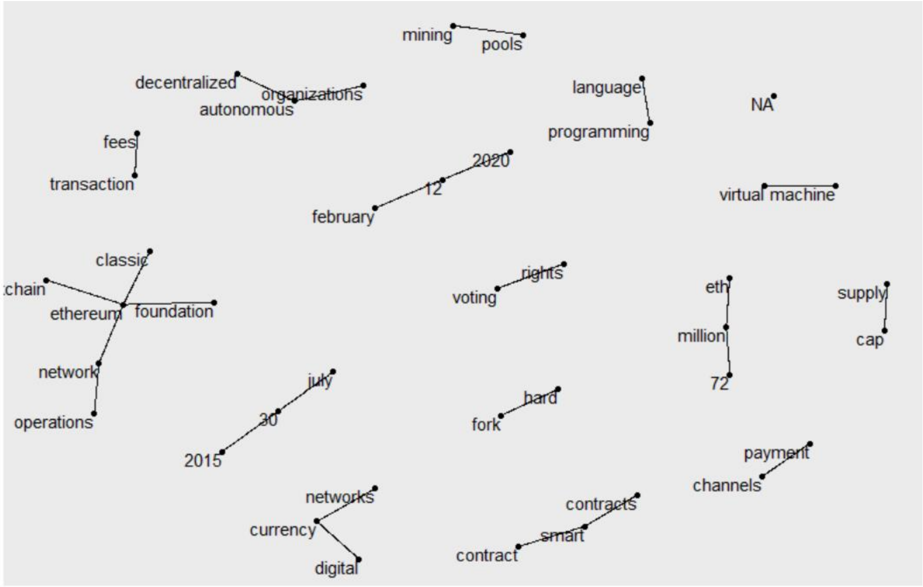


Bigram analysis

The bigram analysis of the bitcoin dataframe shows that most of the words are connected with others where most sentences are connected with most repeated words. Words like bitcoin and economy, currency are most interlinked with the other words in the sentence. Money is linked to mostly launder, transmitter and transfer words which shows that launder related to bitcoin is spoken more in the article. Lower fee, transactions, services are interlinked to bitcoin which shows the factors of bitcoin.

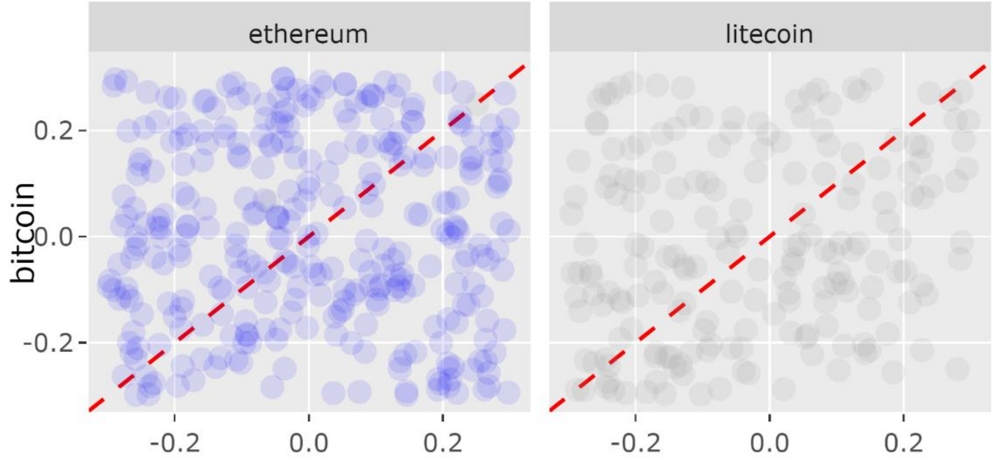


Ethereum bigram analysis shows that the words are little spread out and not mostly repeated like the bitcoin bigram. Most of them seems to be dates like 12-feb-2020, 30-july-2015 could be the dates that are most important for Ethereum coin. These could be that Ethereum coin was high at its value on those days or maybe least at its value on those dates. Ethereum is linked to words like classic that could be because Ethereum has a replica coin Ethereum classic and blockchain is the technology which it uses as the base platform. Ethereum has created a platform for most of the coins. The fees and transactions are also linked which shows that its fee is high or transaction time is high.



Correlogram Analysis

The analysis was done based on bitcoin to Ethereum and Litecoin where the frequent words are correlating of bitcoin to Ethereum and Litecoin and we can see that most words are not on the line but they are more scattred around the plot. Ethereum and Litecoin looks similar to each other when they are compared with bitcoin.



Conclusion

The 3 articles which are taken for analysis are related to cryptocurrency and the mean of Ethereum at the end of the analysis shows that its 0.62 and 0.59 for Litecoin. The 3 articles are closely related and based on crypto and blockchain technology.

Code-

```
#############################

#####Calling Libraries#######

#############################

library(textdata)

library(tidytext)

library(tidyverse)

library(dplyr)

library(janeaustenr)

library(wordcloud)

library(textdata)

library(gutenbergr)

library(reshape2)

library(textreadr)

library(scales)

library(plotly)

library(igraph)

library(ggraph)

library(tm)

library(RColorBrewer)

############################

#####Calling Dataframes#####

############################
```

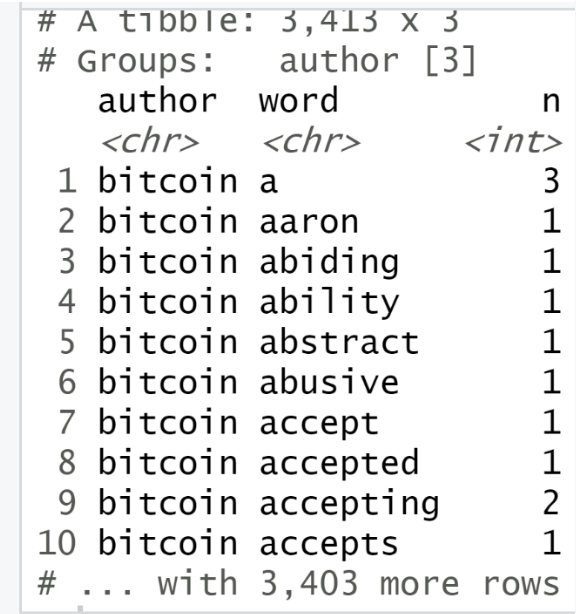
```
# Reading bitcoin
file_1 <- read_document(file="C:/Users/Teja/Documents/NLP/bitcoin/bitcoin.txt")
bitcoin <- c(file_1)
bitcoin <- data_frame(line=1, text=bitcoin)

# Reading ethereum
file_2 <- read_document(file="C:/Users/Teja/Documents/NLP/ethereum/ethereum.txt")
ethereum <- c(file_2)
ethereum <- data_frame(line=1, text=ethereum)

# Reading litecoin
file_3 <- read_document(file="C:/Users/Teja/Documents/NLP/business_insights/litecoin.txt")
litecoin <- c(file_3)
litecoin <- data_frame(line=1, text=litecoin)

#############################################
#############Tokenization####################
#############################################

# Tokenizing bitcoin
bitcoin_token <- bitcoin %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
bitcoin_token

# Tokenizing ethereum
ethereum_token <- ethereum %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
ethereum_token

# Tokenizing litecoin
litecoin_token <- litecoin %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word, sort = TRUE)
litecoin_token



##########################################
#######Combining dataframes###############
##########################################

# Combining into single data frame
crypto <- bind_rows(mutate(bitcoin_token, author="bitcoin"),
                 mutate(ethereum_token, author= "ethereum"),
                 mutate(litecoin_token, author="litecoin"))%>%#closing bind_rows
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author)
crypto

########################################
#####Sentiment Analysis#################
########################################
```

```
# Bing Sentiment Analysis using Plotly
# bitcoin
bitcoin_senti <- bitcoin %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
bitcoin_bing <- bitcoin_senti %>%
  group_by(sentiment) %>%
  top_n(5) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment Bitcoin", x=NULL)+
  coord_flip()
bitcoin_bing <- ggplotly(bitcoin_bing)
bitcoin_bing

# Ethereum
ethereum_senti <- ethereum %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
ethereum_bing <- ethereum_senti %>%
  group_by(sentiment) %>%
  top_n(5) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment Ethereum", x=NULL)+
  coord_flip()
ethereum_bing <- ggplotly(ethereum_bing)
ethereum_bing

# litecoin
litecoin_senti <- litecoin %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
litecoin_bing <- litecoin_senti %>%
  group_by(sentiment) %>%
  top_n(5) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment Litecoin", x=NULL)+
  coord_flip()
```

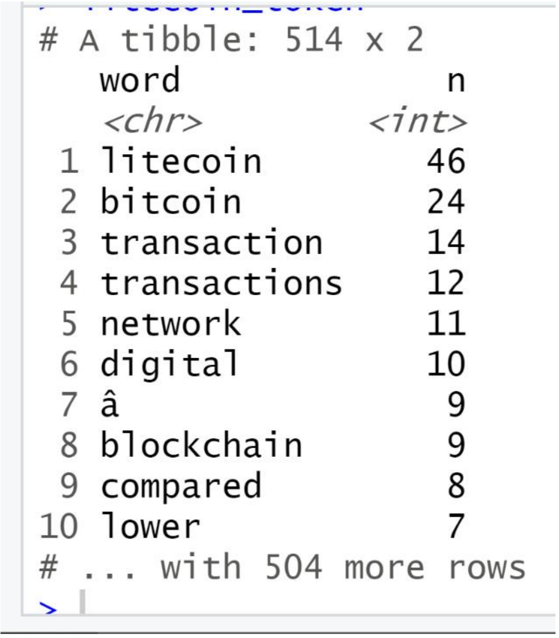
```r
litecoin_bing <- ggplotly(litecoin_bing)
litecoin_bing

# NRC Sentiment Analysis - Word Cloud
# bitcoin
bitcoin_senti_nrc <- bitcoin %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
bitcoin_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%
  acast(word ~sentiment, value.var="n", fill=0) %>%
  comparison.cloud(colors = c("grey20", "gray80"),
                    title.colors=c("red","green"),
                    max.words=100, fixed.asp=TRUE,
                    scale=c(0.6,0.6), title.size=1, rot.per=0.25)

# ethereum
ethereum_senti_nrc <- ethereum %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()
ethereum_senti_nrc %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort=TRUE) %>%

  acast(word ~sentiment, value.var="n", fill=0) %>%

  comparison.cloud(colors = c("grey20", "gray80"),

                   title.colors=c("red","green"),

                   max.words=100, fixed.asp=TRUE,

                   scale=c(0.6,0.6), title.size=1, rot.per=0.25)

# litecoin

litecoin_senti_nrc <- litecoin %>%

  unnest_tokens(word, text) %>%

  anti_join(stop_words) %>%

  inner_join(get_sentiments("nrc")) %>%

  count(word, sentiment, sort=T) %>%

  ungroup()

litecoin_senti_nrc %>%

  inner_join(get_sentiments("nrc")) %>%

  count(word, sentiment, sort=TRUE) %>%

  acast(word ~sentiment, value.var="n", fill=0) %>%

  comparison.cloud(colors = c("grey20", "gray80"),

                   title.colors=c("red","green"),

                   max.words=100, fixed.asp=TRUE,

                   scale=c(0.6,0.6), title.size=1, rot.per=0.25)



# Creating Bigrams and plotting the networks

# bitcoin

bitcoin_bigrams <- bitcoin %>%

  unnest_tokens(bigram, text, token = "ngrams", n=2) %>%

  separate(bigram, c("word1", "word2"), sep = " ") %>%

  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE)
bitcoin_bigrams

bitcoin_bigram_graph <- bitcoin_bigrams %>%
  filter(n>2) %>%
  graph_from_data_frame()
ggraph(bitcoin_bigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)

# ethereum
ethereum_bigrams <- ethereum %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE)
ethereum_bigram_graph <- ethereum_bigrams %>%
  filter(n>2) %>%
  graph_from_data_frame()
ggraph(ethereum_bigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1)

# Litecoin
litecoin_bigrams <- litecoin %>%
  unnest_tokens(bigram, text, token = "ngrams", n=2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  count(word1, word2, sort = TRUE)
litecoin_bigrams

litecoin_bigram_graph <- litecoin_bigrams %>%
  filter(n>2) %>%
  graph_from_data_frame()
ggraph(litecoin_bigram_graph, layout = "fr") +
  geom_edge_link()+
  geom_node_point()+
  geom_node_text(aes(label=name), vjust =1, hjust=1, scale=c(0.6,0.6))

# Creating single data frame with all tokens along with frequency proportions
frequency <- bind_rows(mutate(bitcoin_token, author="bitcoin"),
              mutate(ethereum_token, author= "ethereum"),
              mutate(litecoin_token, author="litecoin")
)%>%#closing bind_rows
  mutate(word=str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n/sum(n))%>%
  #select(-n) %>%
  spread(author, proportion) %>%
  gather(author, proportion, `ethereum`, `litecoin`)
frequency
```

```
# Plotting a correlogram using Plotly
plotting_graph <- ggplot(frequency, aes(x=proportion, y=bitcoin,
                         color = abs(bitcoin- proportion)))+
  geom_abline(color="red", lty=2)+
  geom_jitter(aes(text=paste("word: ", word)), alpha=.1, size=2.5, width=0.3, height=0.3)+
  #geom_text(aes(label=word), colour="gray20", alpha=1) +
  #scale_x_log10(labels = percent_format())+
  #scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "blue", high = "blue")+
  facet_wrap(~author, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "bitcoin", x=NULL)
plotting_graph <- ggplotly(plotting_graph)
plotting_graph

# Creating a Document Term Matrix (DTM)
crypto_dtm <- crypto %>%
  group_by(author) %>%
  cast_dtm(author, word, n)
crypto_dtm

# Sentiment Analysis using AFINN

ethereum_afinn <- ethereum_token %>%
  inner_join(get_sentiments("afinn"))%>%
  summarise(mean(value))
ethereum_afinn

litecoin_afinn <- litecoin_token %>%
  inner_join(get_sentiments("afinn"))%>%
  summarise(mean(value))
litecoin_afinn

# bitcoin
bitcoin_afinn <- bitcoin_token %>%
  inner_join(get_sentiments("afinn"))%>%
  summarise(mean(value))
bitcoin_afinn
```

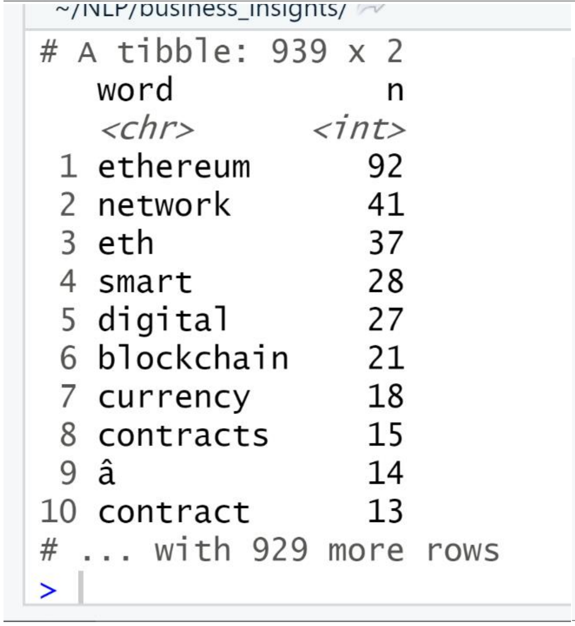
Output-



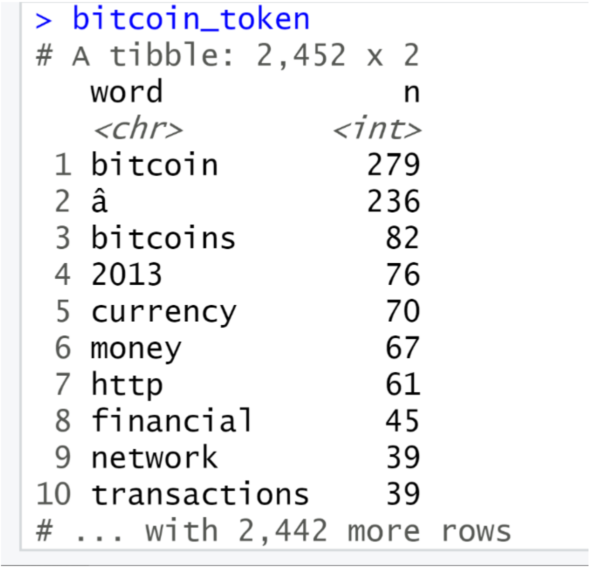
```
# A tibble: 3,413 x 3
# Groups:   author [3]
   author  word          n
   <chr>   <chr>     <int>
 1 bitcoin a             3
 2 bitcoin aaron         1
 3 bitcoin abiding       1
 4 bitcoin ability       1
 5 bitcoin abstract      1
 6 bitcoin abusive       1
 7 bitcoin accept        1
 8 bitcoin accepted      1
 9 bitcoin accepting     2
10 bitcoin accepts       1
# ... with 3,403 more rows
```



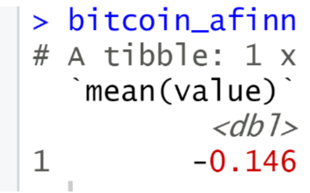
```
> litecoin_token
# A tibble: 514 x 2
   word             n
   <chr>        <int>
 1 litecoin        46
 2 bitcoin         24
 3 transaction     14
 4 transactions    12
 5 network         11
 6 digital         10
 7 â                9
 8 blockchain       9
 9 compared         8
10 lower            7
# ... with 504 more rows
>
```



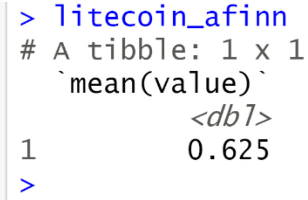
```
~/NLP/business_insights/
# A tibble: 939 x 2
   word           n
   <chr>      <int>
 1 ethereum      92
 2 network       41
 3 eth           37
 4 smart         28
 5 digital       27
 6 blockchain    21
 7 currency      18
 8 contracts     15
 9 â             14
10 contract      13
# ... with 929 more rows
>
```



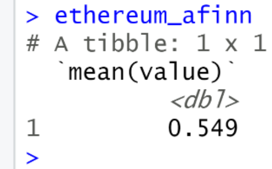
```
> bitcoin_token
# A tibble: 2,452 x 2
   word             n
   <chr>        <int>
 1 bitcoin        279
 2 â              236
 3 bitcoins        82
 4 2013            76
 5 currency        70
 6 money           67
 7 http            61
 8 financial       45
 9 network         39
10 transactions    39
# ... with 2,442 more rows
```



```
> bitcoin_afinn
# A tibble: 1 x 1
  `mean(value)`
          <dbl>
1        -0.146
```



```
> litecoin_afinn
# A tibble: 1 x 1
  `mean(value)`
          <dbl>
1         0.625
>
```



```
> ethereum_afinn
# A tibble: 1 x 1
  `mean(value)`
          <dbl>
1         0.549
>
```