Bitcoin
Text Analysis

Class of Program for Big Data Analysis-STU

Final Project

Fall 2021

By **Agnes Sythole** and **Hulnise Jean-François**

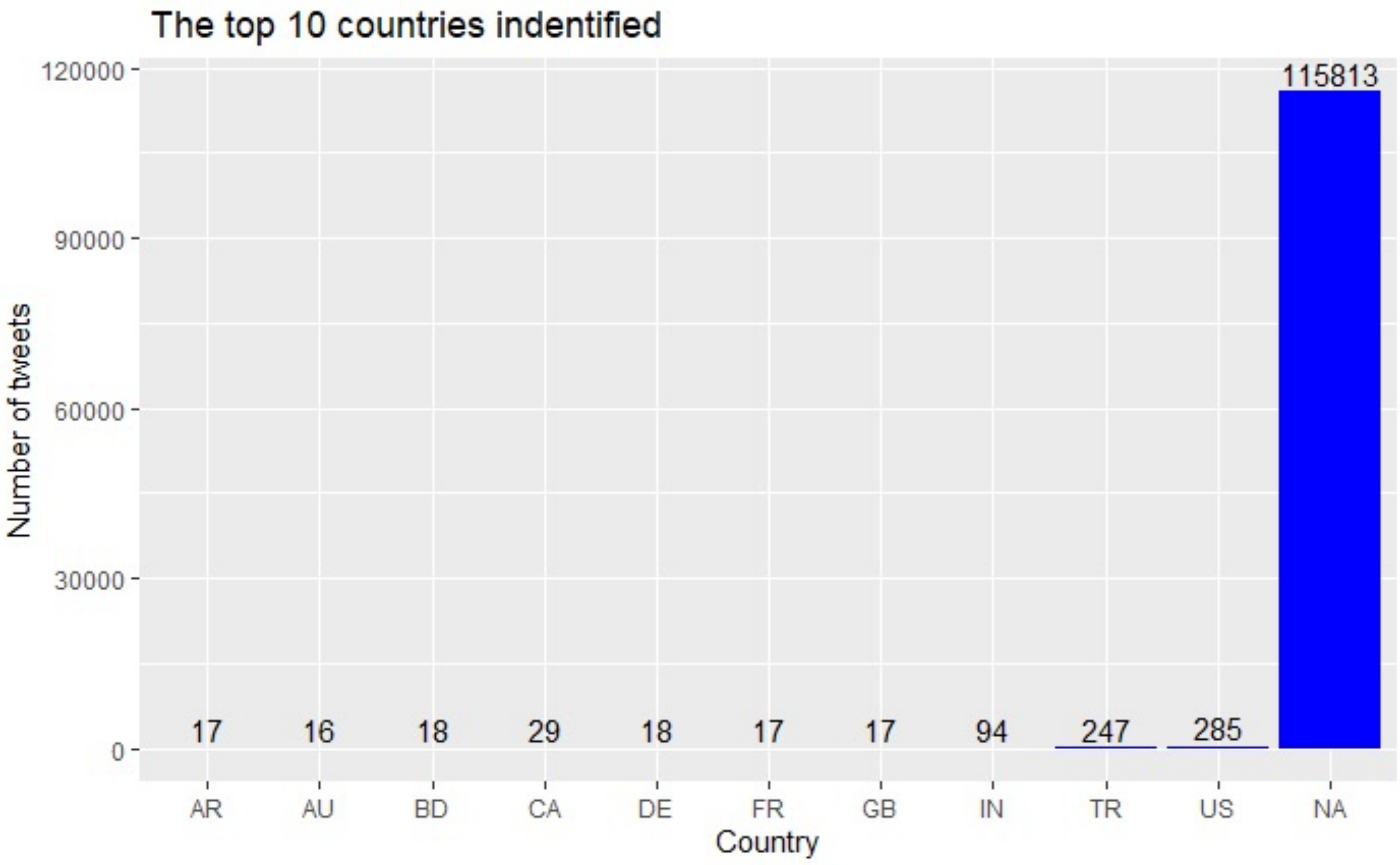
Summary

- Introduction
- Exploratory analysis
- Text Analysis
- Classification
- Clustering
- Conclusion
- Sources

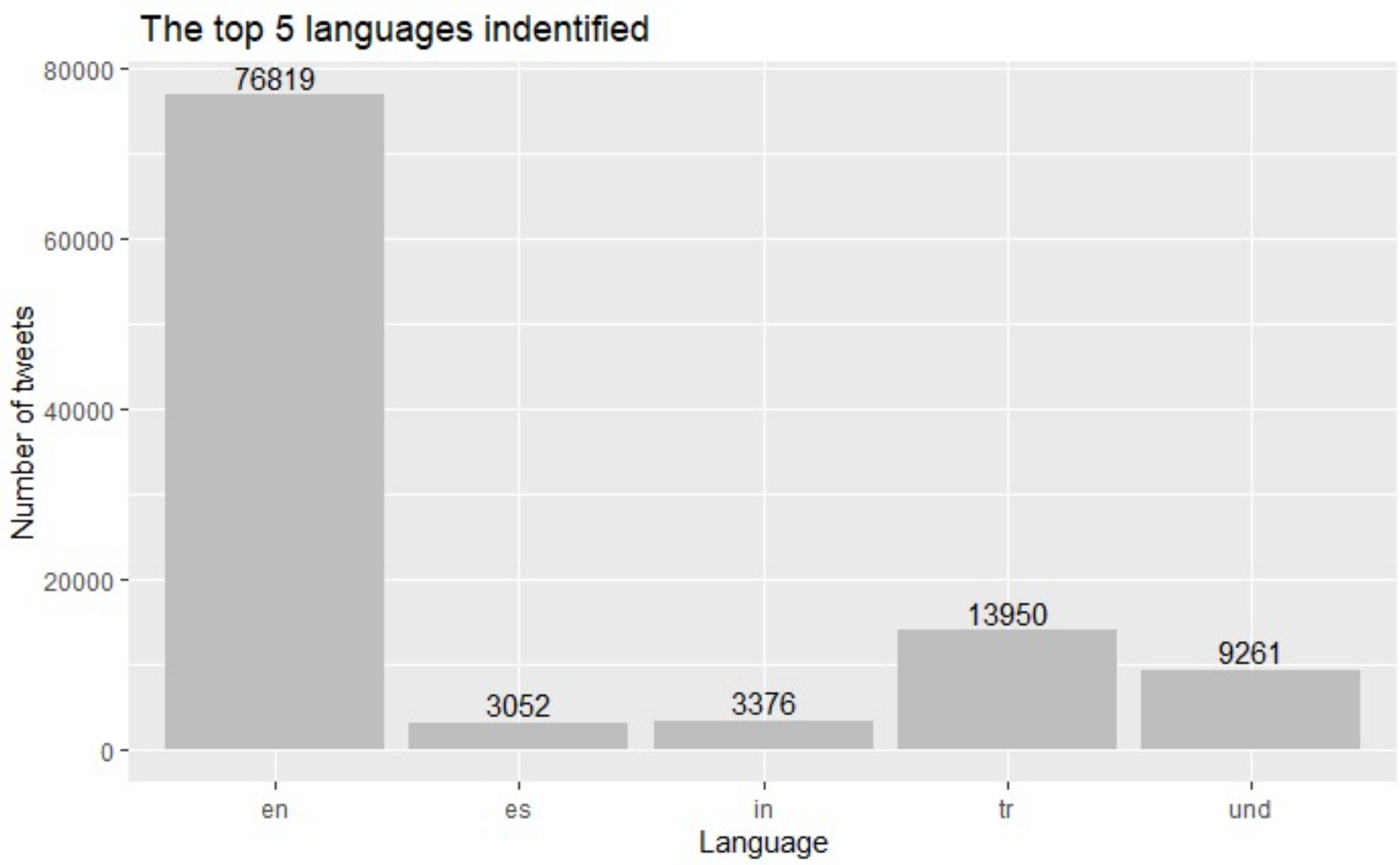
Bitcoin is a decentralized digital currency created in January 2009. Commonly abbreviated as BTC, it is known as a type of cryptocurrency because it uses cryptography to keep it secure. It offers the promise of lower transaction fees than traditional online payment mechanisms do, and unlike government-issued currencies, it is operated by a decentralized authority. There are no physical bitcoins, only balances kept on a public ledger that everyone has transparent access to. Bitcoin is not being recognized legally in most parts of the world, but it is still very popular. Despite its rapid growth and an increasing number of users, bitcoin is considered too volatile and because of its irreversibility, too risky. Our project is about finding what are the sentiments about bitcoins now and see if the people nowadays seem more eager to use it or not.

For that project we used a Twitter Developer API account to collect data from tweeter. We considered the tweets with the words bitcoin and BTC. We did text analysis of the tweets to understand better the position of the people about bitcoins, created classification models and clusters in order to identify if bitcoin have potential of becoming a more largely used currency based on the insight given by the data.

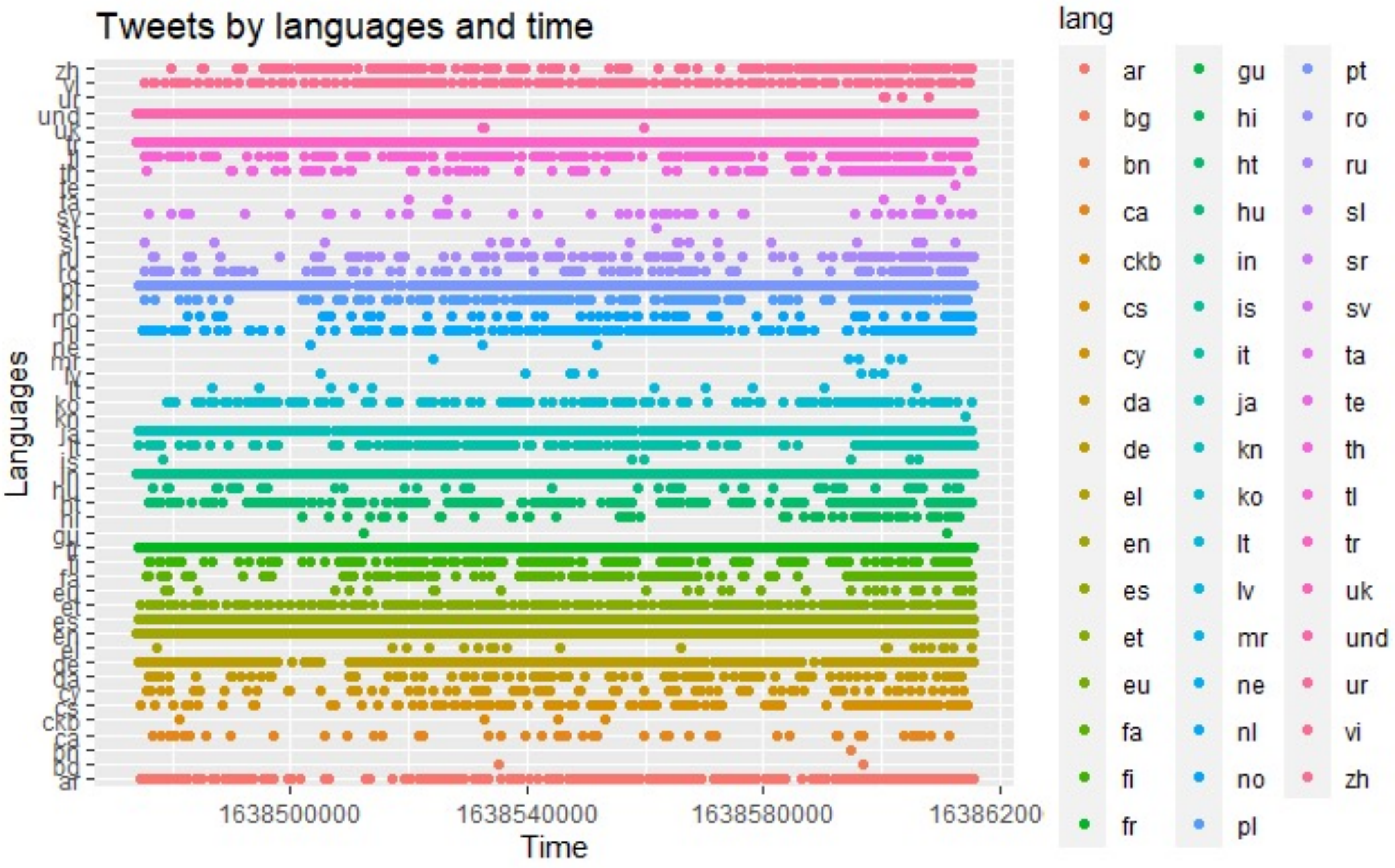
Exploratory Analysis

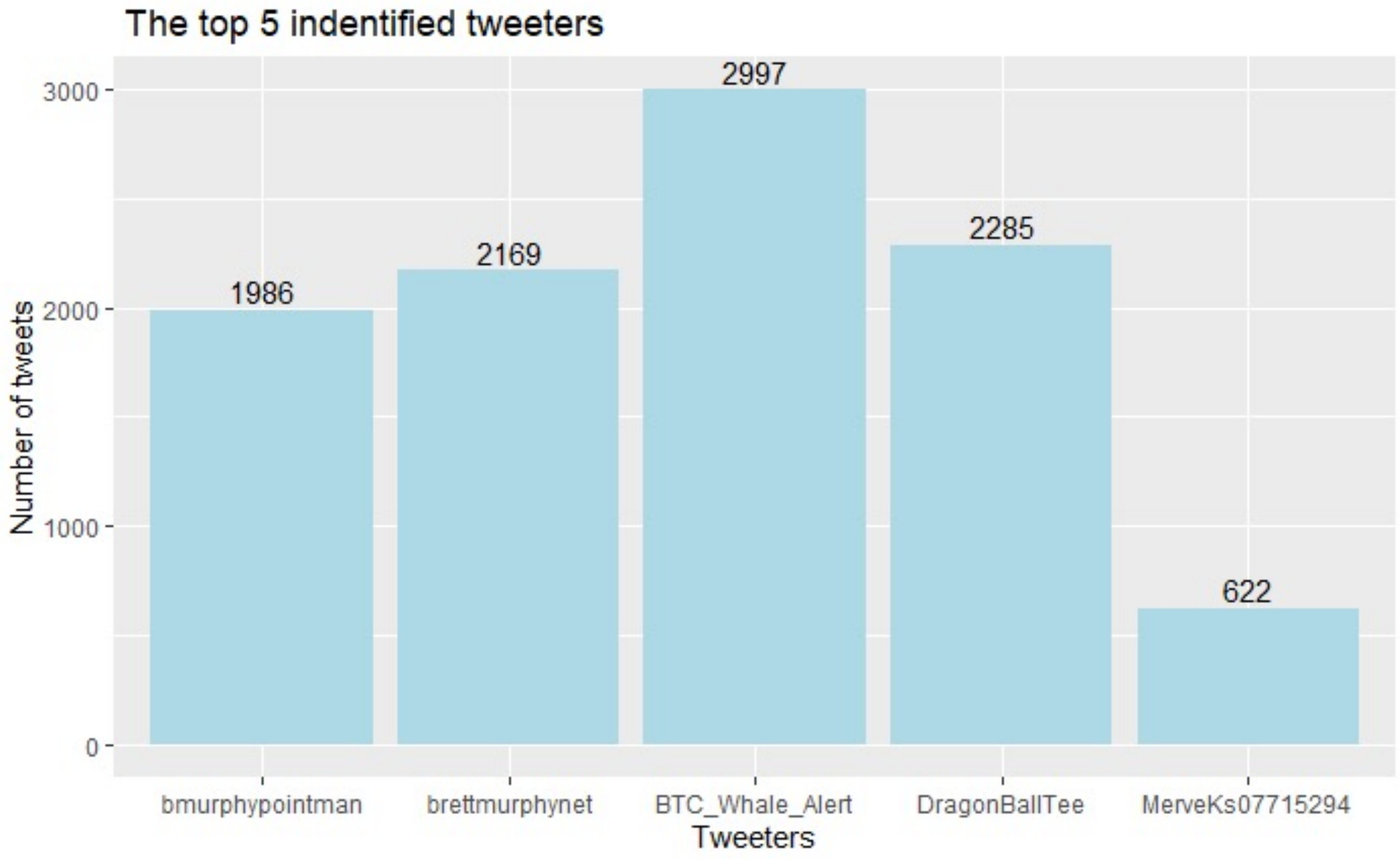


"Most of the countries where the tweeters are located are not identified in the dataset, but we can believe that many countries are represented because we have 67 countries identified.
As we can see in this visualization the top 4 countries identified are : United States, Turkey, India and Canada "

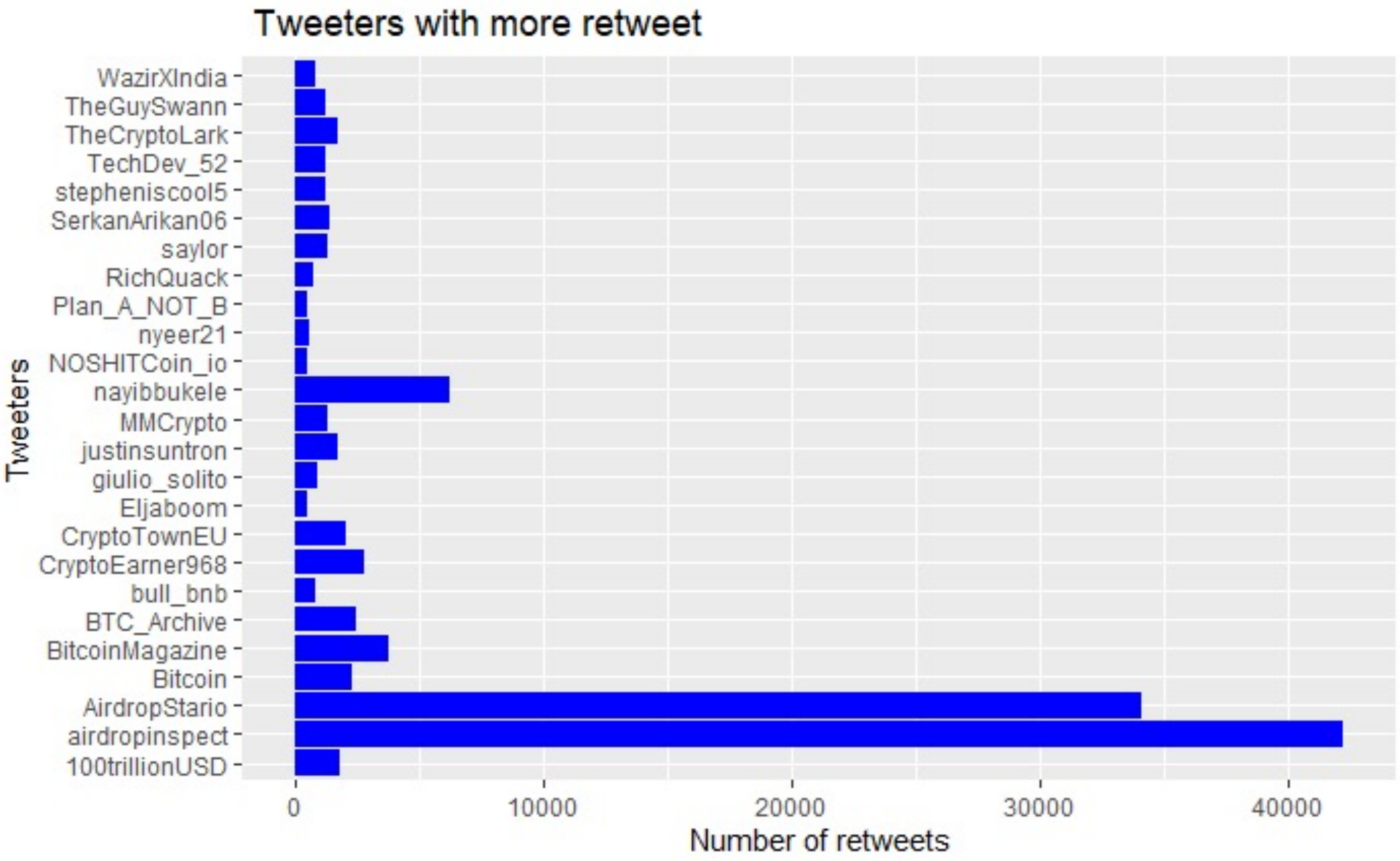


“We have 50 languages represented in the dataset and the top 2 are : English and Turkish“

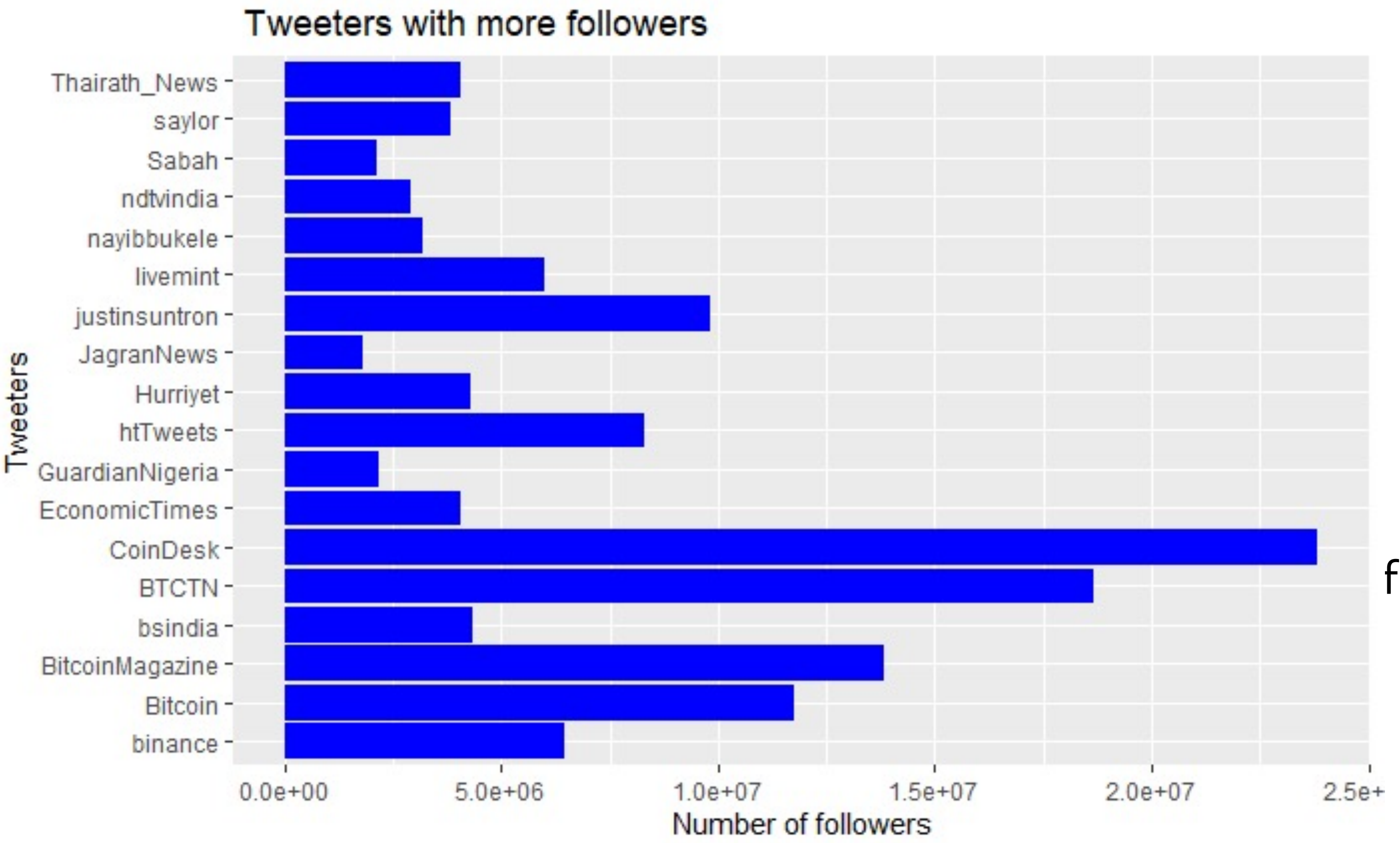

Tweets by languages and time
Languages
Time
1638500000
1638540000
1638580000
16386200
lang
ar
bg
bn
ca
ckb
cs
cy
da
de
el
en
es
et
eu
fa
fi
fr
gu
hi
ht
hu
in
is
it
ja
kn
ko
lt
lv
mr
ne
nl
no
pl
pt
ro
ru
sl
sr
sv
ta
te
th
tl
tr
uk
und
ur
vi
zh



We have 41344 different tweeters, and this visualization shows the top 5.



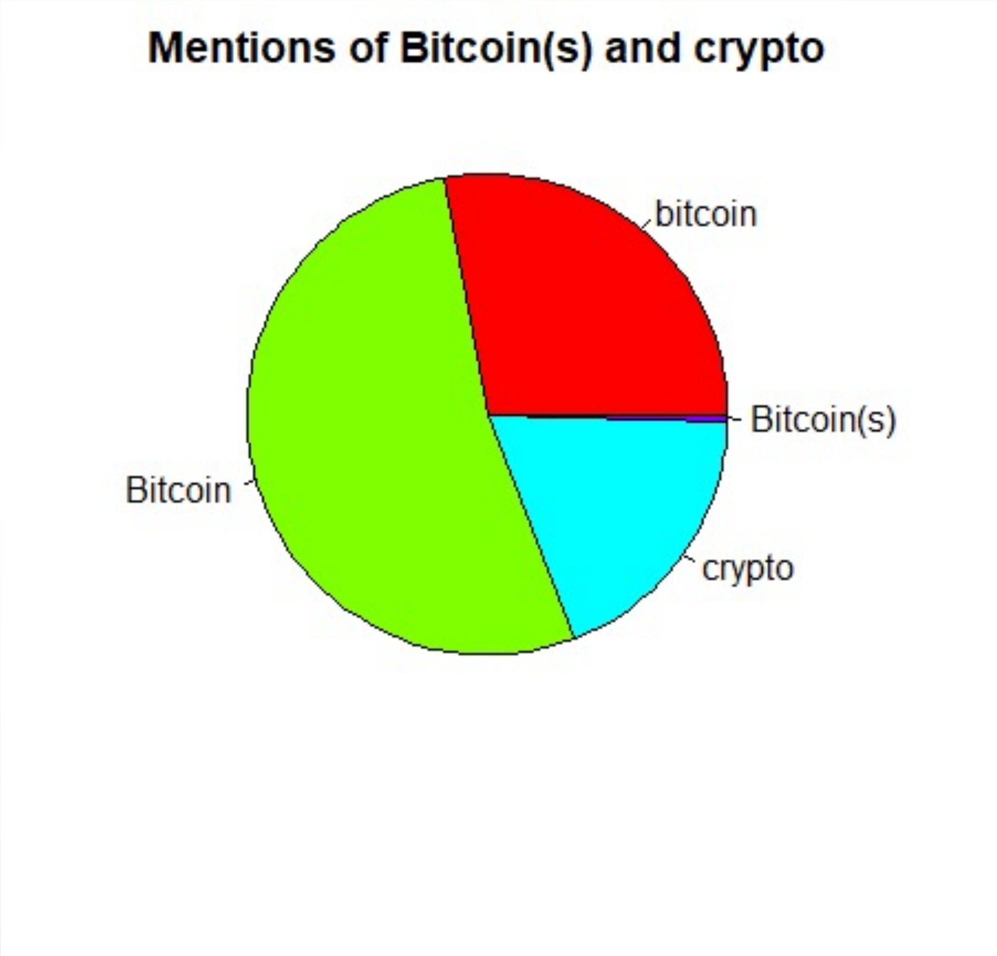
"The 3 tweeters account with more retweets are : airdropinspect , AirdropStario and nayibbukele "



"The top 3 tweeters account with more followers are : CoinDesk, BTCTN and BitcoinMagazine"

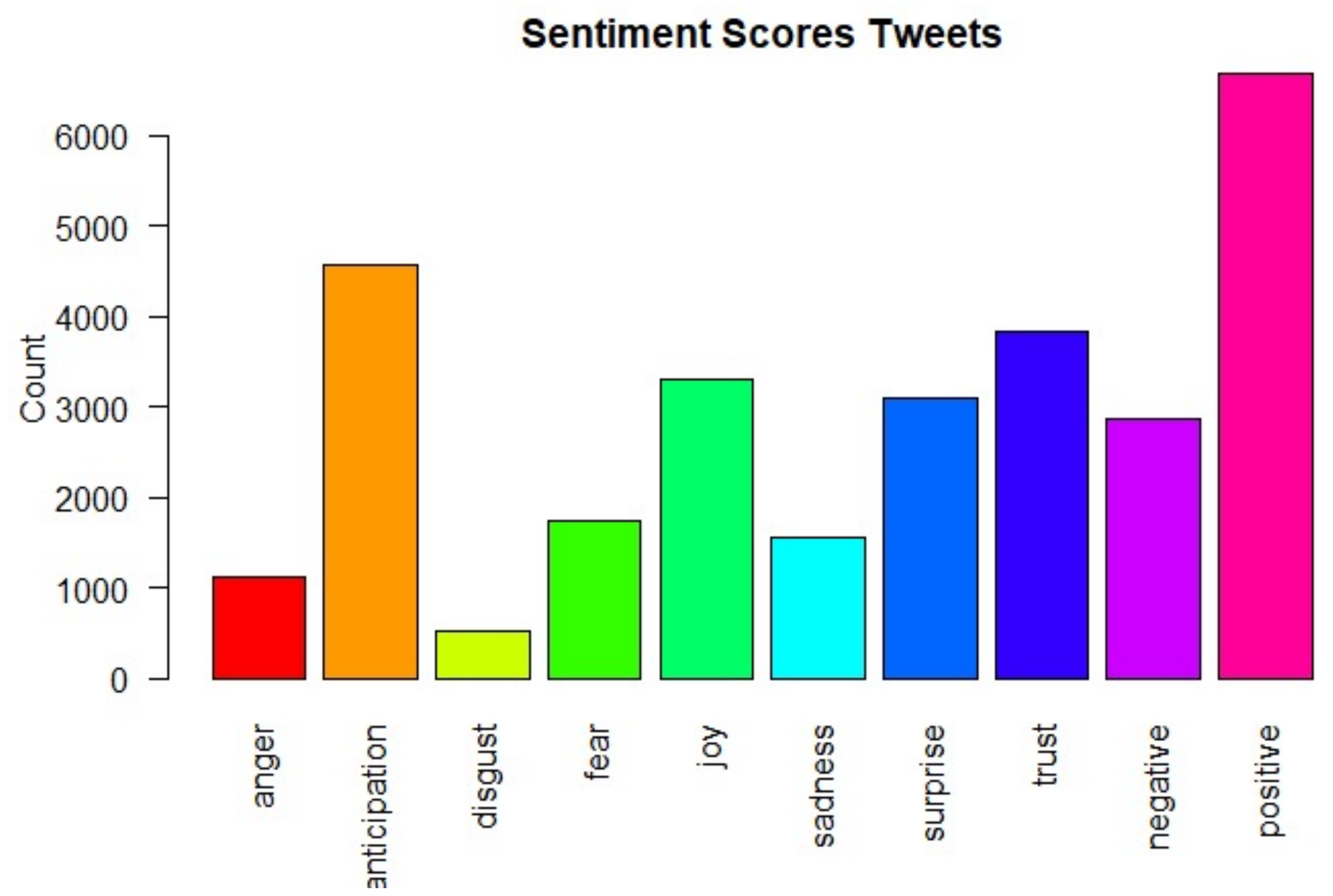
- Text Analysis

" Number of tweets with 'bitcoin(s)' or Bitcoin(s): 115196."



- Number of hashtags: 116793
- Number of mentions: 38081
- Giveaway was mentioned 776 times

- The top features are :
  - #bitcoin : 119742
  - $: 82392
  - #btc : 26273
  - #crypto : 25072
  - project : 16150
  - #cryptocurrency : 15329
  - 🚀 : 15255
  - #eth : 13283
  - btc : 12236
  - #Ethereum:11668



As we can see , there are mostly positive sentiments about bitcoin.

Classification

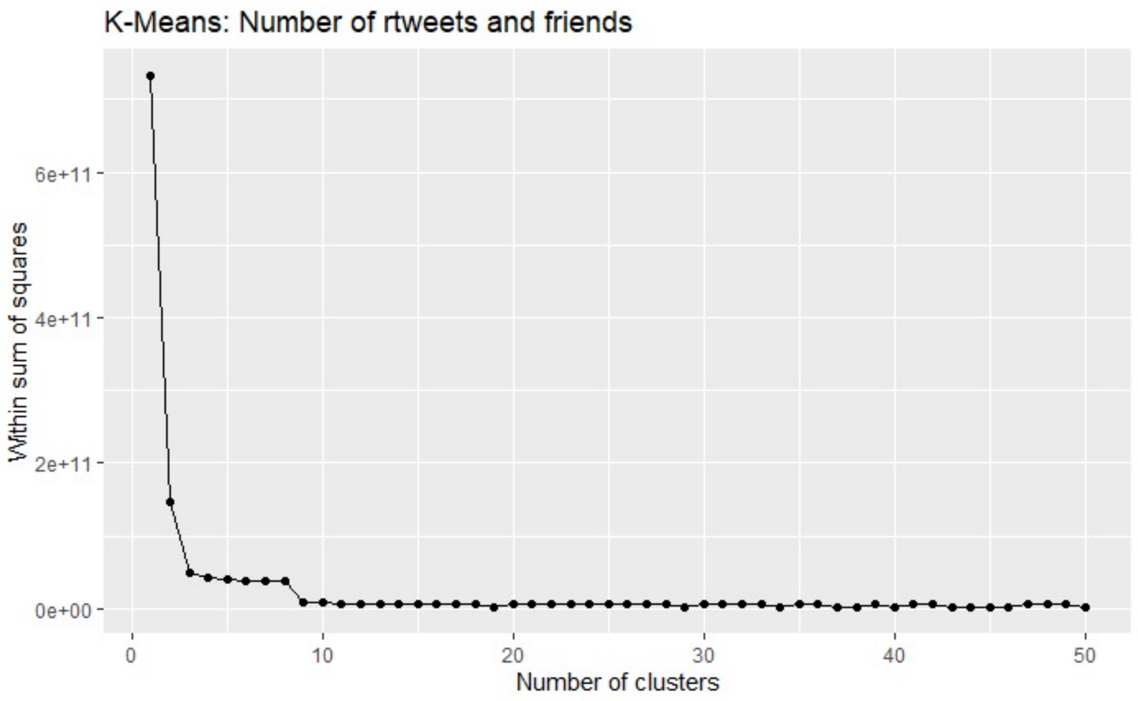
We found the following accuracy for each model created to predict number of retweets based on those elements of the dataset :created_at, screen_name, display_text_width, retweet_count, lang, protected, followers_count, friends_count , listed_count, statuses_count, favourites_count, account_created_at, verified.

- Naive Bayes Classifier Accuracy: 1
- LaPlace1 Accuracy: 0.07
- LaPlace of 3 Accuracy: 0.06

We found the following accuracy for each model created to predict number of retweets based on those elements of the dataset :retweet_count, protected, followers_count, friends_count , statuses_count, favourites_count, verified.

- The accuracy of the Decision tree is : 0.0884

Clustering



The ideal K is 3 and the ideal k's wss value is : 48633709843

Conclusion

Considering tweets in the dataset, we can say that a little part of the population know about bitcoin. They are also aware of the risks related to it’s use, for example : its volatility and its irreversibility . Nevertheless, we will not forget the advantages of using bitcoin like the low fees and the reduced risk of fraud associated to it. Our data as shown that most of the people aware of it have positive thoughts about it.

As cryptocurrencies and bitcoin are young , it will require time, efforts and awareness too for people to get to know it better. Its use will increase as more individuals get to have information about it to understand how it works and more businesses accept it as payment.

Sources

https://www.investopedia.com/terms/b/bitcoin.asp

https://www.bitcoin.com/