============= DRAFT ==============

# 2018 Brazil’s election results: Fear or real change on population political position?

Or

# Fear as a definite victory factor for small political parties.

## Abstract

On 2018 Brazil elected an almost obscure and radical president, Mr. Jair Bolsonaro, an ex-Army Capitan that have actuated very poorly on Congress for almost 30 years. Surprisedly voters decided, massively move to the right side of political spectra after more than 24 years of center-left governments. Do Brazilians really think this way? But, if yes what Brazilians? The poor? The wealth? From the North? From the South? Or at the end was only fear? And if this is the reason, other questions arose: fear of what? An “orchestrated sub continental left revolution”? The market’s collapse. So many questions aroused everywhere in the world, trying to explain what happened in Brazil. The purpose of this work will be to study what really happened during the 2 rounds of 2018 Brazil’s election, having social media as the database for that. The work will show that majority of voters has in fact a more centric political position, been in fact fear (heavily feed by fake news) that put Bolsonaro in Brasilia.

## Dataset

Personal tweets from a select number of users around Jair Bolsonaro:

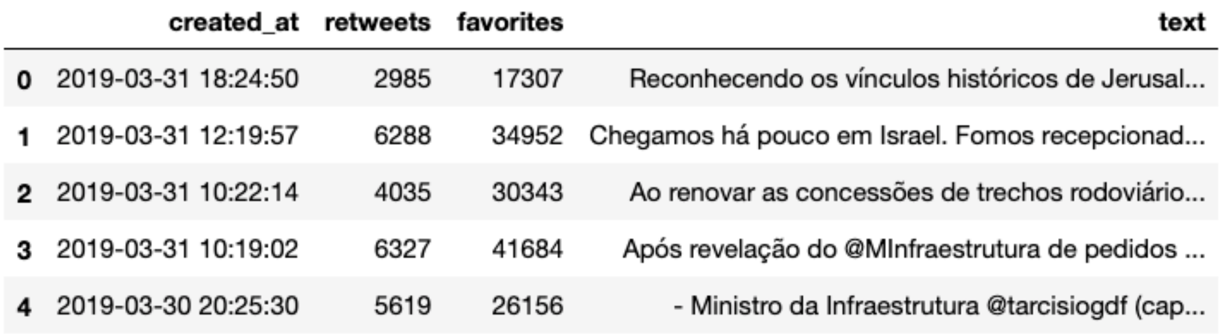
A. Bolsonaro’s party:
- Jair Bolsonaro (JB) – ex-Congressman for RJ and actual president
- Carlos Bolsonaro (CB) – Son and Rio’s city council
- Flavio Bolsonaro (FB) – Son and Senator for RJ
- Eduardo Bolsonaro (EB) – Congressman for Sao Paulo

B. Haddad’s party (Bolsonario’s oppositos on 2nd round)
- Fernando Haddad – Ex São Paulo’s mayor and left candidate (replaced Lula)
- Lula da Silva (LS) – Former president and pre-candidate from left
- Gleici Hoffmann (GH) – Congressman for Parana (ex-Senator) and PT party president

C. Journalists – Influencers:
- Felipe Moura Brasil (FM) – Journalist and consider the most influent political personality on Twitter

Each personal dataset has around 3,220 tweets. The complete raw tweet dataset was saved and as start; the following data was selected from raw tweets and saved on a CSV file:

1. Twitter screen_name (to be added later)
2. Date of tweet creation
3. # of re-tweets
4. # of favorites
5. Tweet original text



| | created_at | retweets | favorites | text |
|---|---|---|---|---|
| **0** | 2019-03-31 18:24:50 | 2985 | 17307 | Reconhecendo os vínculos históricos de Jerusal... |
| **1** | 2019-03-31 12:19:57 | 6288 | 34952 | Chegamos há pouco em Israel. Fomos recepcionad... |
| **2** | 2019-03-31 10:22:14 | 4035 | 30343 | Ao renovar as concessões de trechos rodoviário... |
| **3** | 2019-03-31 10:19:02 | 6327 | 41684 | Após revelação do @MInfraestrutura de pedidos ... |
| **4** | 2019-03-30 20:25:30 | 5619 | 26156 | - Ministro da Infraestrutura @tarcisiogdf (cap... |

➔ The individual datasets will be later concatenated on a single dataset

## Pre-processing

Each tweet text should be pre-processed in order to obtain clean tokens:

- Remove twitter Return handles (RT @xxx:)
- Remove twitter handles (@xxx)
- Remove hashtags (#xxx)
- Remove URL links (httpxxx)
- Remove special characters, numbers, punctuations
- Remove stop-words
- Stemming and lemmatization
- Part-of-Speech tagging (POS). ➔ See Note bellow
- Reduce tokens to lower case

*Note on POS: The Emotion-Lexicon to be used on this work, should be applied to verbs, substantives and adverbs. Proper Noun and personal names and places should be removed. This task will be done on a second phase.*

At end, a “clean_text” feature will be added to dataset. For example:

Text:

```
'Lula foi o 1° presidente brasileiro a visitar Israel. Mas também visitou
a Palestina, não ofendeu povos e religiões, defendeu a paz e a solução pac
ífica de dois estados, respeitando a tradição diplomática brasileira. E o
Brasil ganhou com mais respeito e comércio #timeLula https://t.co/2uonprCZ
fV'
```

Clean_text:

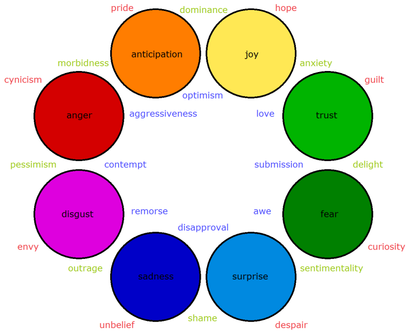
```
'lula 1° presidente brasileiro visitar israel visitar palestina ofender re
ligião defender paz solução pacífico respeitar tradição diplomático brasil
eiro brasil ganhar respeitar comércio'
```

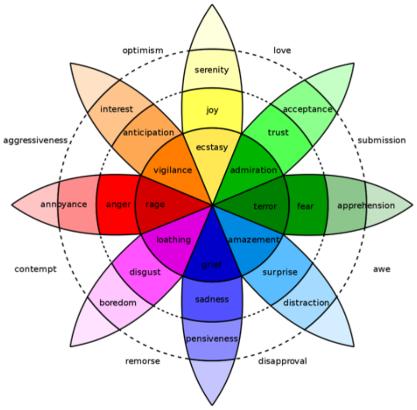
## Methodology

There is a lot of work that is based on polarity (positive/negative) analysis of tweets, usually called “Sentiment Analysis”. In this work, a more complex approach will be used, that is the extraction of “emotions” from tweets.

The model of 8 basic human emotions proposed by Plutchik in 1980 will be used:

- 'anger'
- 'joy'
- 'surprise'
- 'fear'
- 'sadness'
- 'trust'
- 'disgust'
- 'anticipation'



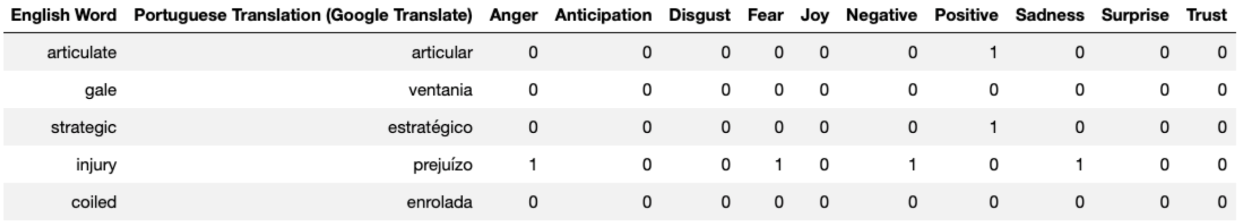


Among the methodologies that can be found to extract emotions, two can be highlighted:

1. Supervised approach using Machine Learning and labeled tweets
2. Non-Supervised method using “Lexicon”

The first attend to extract emotions and sentiment (polarity) from each cleaned tweet, will be using a non-supervised “Lexicon-based approach”. Each tweet, once cleaned and reduced to tokens, will be compared with the words on an emotion dataset, the “NRC Emolex”.

The “EmoLex”, NRC Emotion Lexicon, is a list of English words and their associations with the eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) as discussed above and also two sentiments (negative and positive). The annotations were manually done by crowdsourcing. The dataset (http://sentiment.nrc.ca/lexicons-for-research/) is in English but is also available in other languages as Portuguese (Google Translator was applied on it). The work was discussed on the paper: Saif et al, 2013: "Crowdsourcing a Word–Emotion Association Lexicon": https://arxiv.org/pdf/1308.6297.pdf



| English Word | Portuguese Translation (Google Translate) | Anger | Anticipation | Disgust | Fear | Joy | Negative | Positive | Sadness | Surprise | Trust |
|---|---|---|---|---|---|---|---|---|---|---|---|
| articulate | articular | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| gale | ventania | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| strategic | estratégico | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| injury | prejuízo | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| coiled | enrolada | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Each tweet once cleaned and reduced to tokens, are compared with the emoLex file, having each token associated to one of the emotion (8) and sentiment (2) features.

For score computation a python library for parsing and applying emotion and sentiment was used together with a the emotion dataset "NRC Emolex" (py-lex: https://github.com/dropofwill/py-lex).

Once calculated, the features are grouped, and its frequency calculated. The resultant score will be added to dataset.

Below an example of a cleaned tweet, with 22 tokens, where for each one of them, if in the Lexicon, was "marked with associated emotions (note the a single token can be associated to more than one emotion):

Clean Tweet:

```
'lula 1° presidente brasileiro visitar israel visitar palestina ofender re
ligião defender paz solução pacífico respeitar tradição diplomático brasil
eiro brasil ganhar respeitar comércio'
```

Tokens associated to emotions:

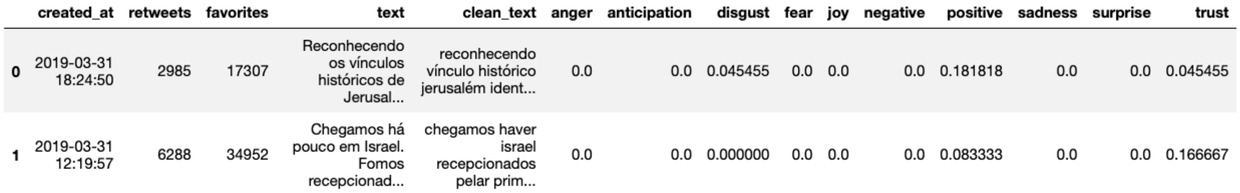
```
1.  lula
2.  1°
3.  presidente  ==>  {'trust', 'positive'}
4.  brasileiro
5.  visitar
6.  israel
7.  visitar
8.  palestina
9.  ofender   ==>  {'disgust', 'negative', 'anger'}
10. religião  ==>  {'trust'}
11. defender  ==>  {'fear', 'positive'}
12. paz
13. solução   ==>  {'positive'}
14. pacífico  ==>  {'joy', 'positive', 'trust', 'anticipation', 'surprise'}
15. respeitar
16. tradição
17. diplomático  ==>  {'trust', 'positive'}
18. brasileiro
19. brasil
20. ganhar    ==>  {'positive'}
21. respeitar
22. comércio  ==>  {'trust'}
```

And the resultant average score for each emotion:

- 'anger': 0.045454545454545456,
- 'joy': 0.045454545454545456,
- 'surprise': 0.045454545454545456,
- 'fear': 0.045454545454545456,
- 'sadness': 0.0,
- 'positive': 0.2727272727272727,
- 'trust': 0.22727272727272727,
- 'disgust': 0.045454545454545456,
- 'anticipation': 0.045454545454545456,
- 'negative': 0.045454545454545456

$$Score_{AVG} = \frac{1}{m} \sum_{i=1}^{m} Emotion_i$$

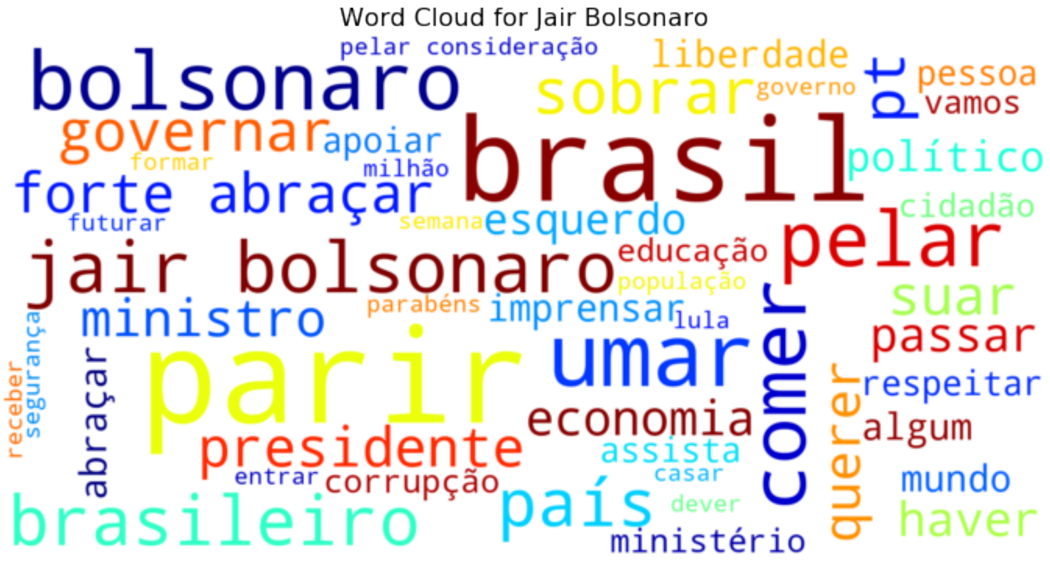
The resultant average score will be added to each dataset as a new feature:

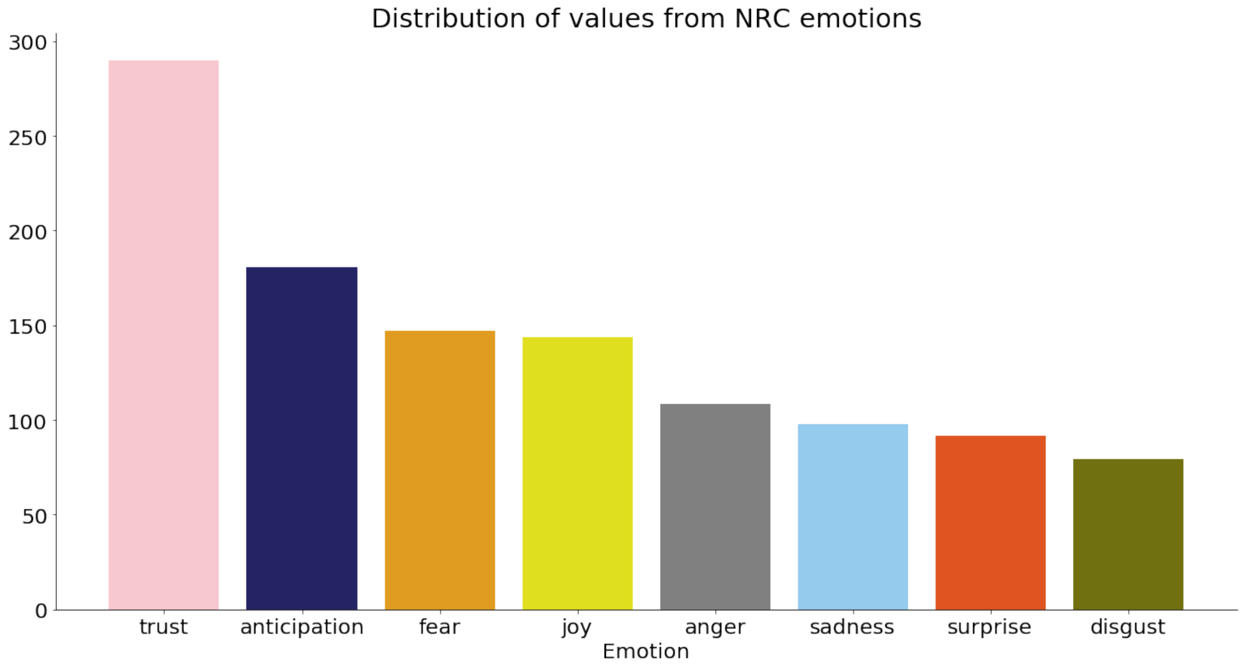


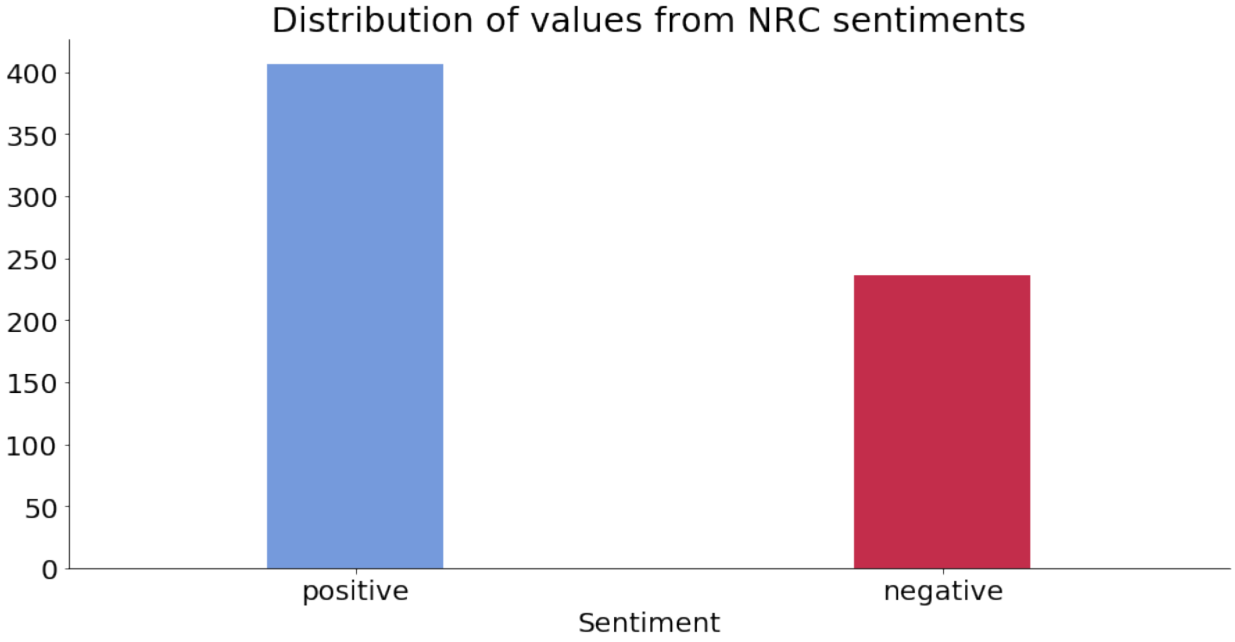
| | created_at | retweets | favorites | text | clean_text | anger | anticipation | disgust | fear | joy | negative | positive | sadness | surprise | trust |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2019-03-31 18:24:50 | 2985 | 17307 | Reconhecendo os vínculos históricos de Jerusal... | reconhecendo vínculo histórico jerusalém ident... | 0.0 | 0.0 | 0.045455 | 0.0 | 0.0 | 0.0 | 0.181818 | 0.0 | 0.0 | 0.045455 |
| 1 | 2019-03-31 12:19:57 | 6288 | 34952 | Chegamos há pouco em Israel. Fomos recepcionad... | chegamos haver israel recepcionados pelar prim... | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.083333 | 0.0 | 0.0 | 0.166667 |

## Emotion and Sentiment Analysis

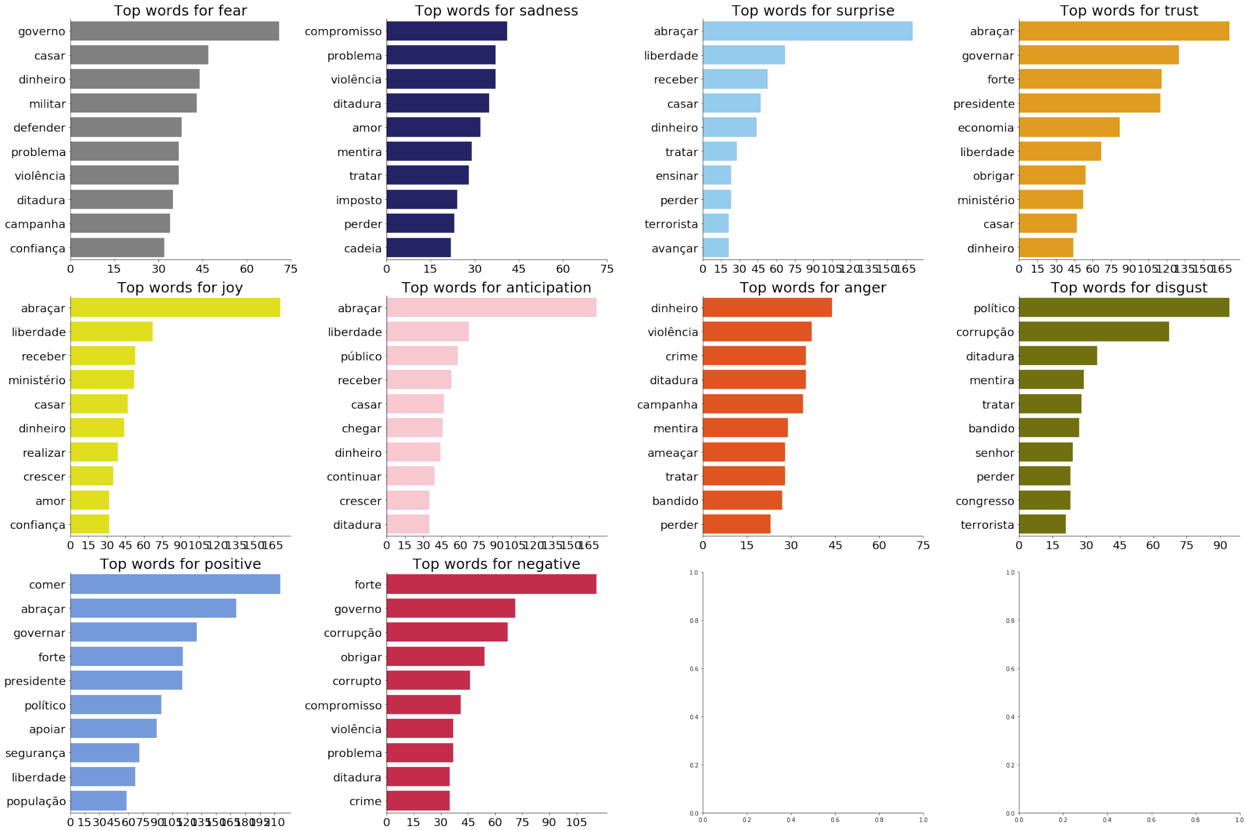
For each one of the users (screen_names), a general overview was done, having the 4 bellow graphs as a result:




Distribution of values from NRC emotions
300
250
200
150
100
50
0
trust
anticipation
fear
joy
anger
sadness
surprise
disgust
Emotion


Distribution of values from NRC sentiments
400
350
300
250
200
150
100
50
0
positive
negative
Sentiment

# Main Words Associated to each one of the emotion features



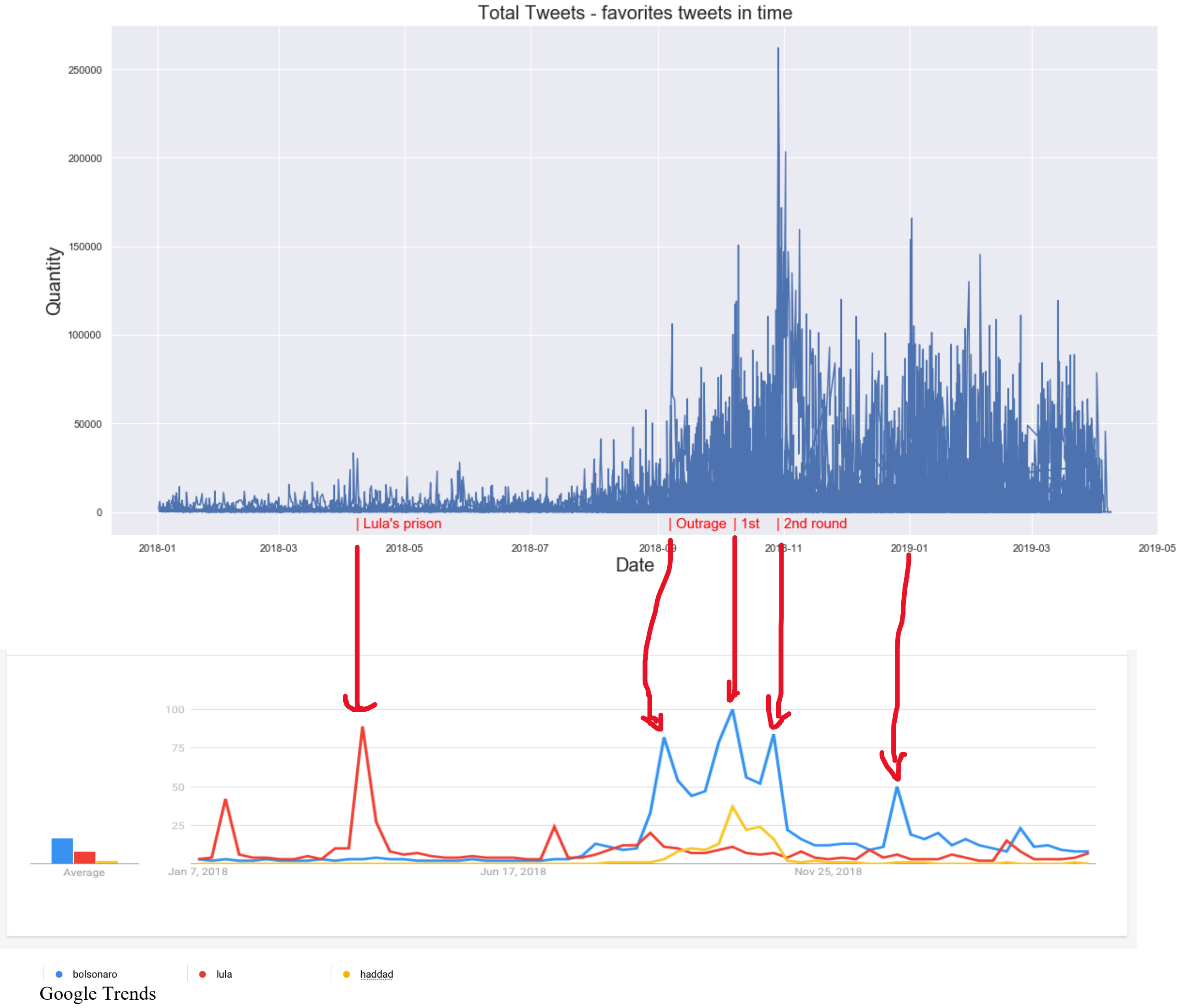
## Dataset Creation, enhancement and preliminary analysis

A. Each individual dataset was concatenated on a single file
B. Missing data was deleted (some clean text tweets resulted on NaN)
C. Tweets creation data were converted datatime type
D. Dataset was filtered for tweets older than January 2018

Final dataset resulted on around 25,000 tweets

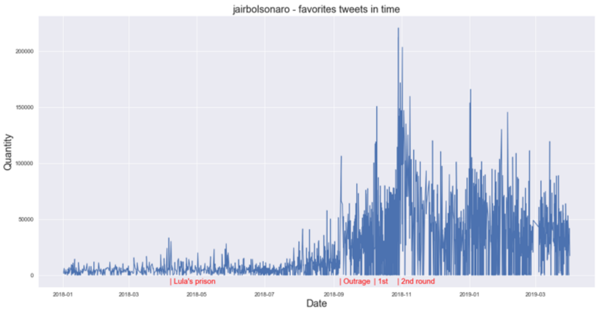
Below, an overview of total dataset tweets, showing how tweets were favorited on time. Some important events (that have a high correlation on Google Trends) were marked as:

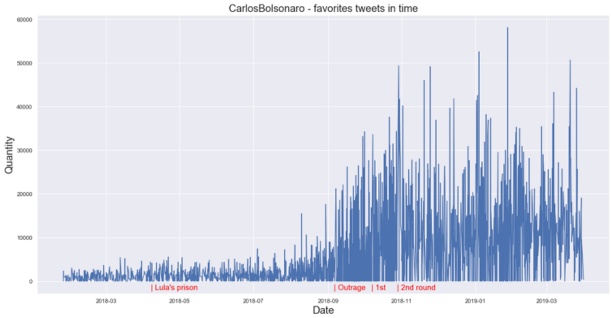
- Lula’s Prison
- Bolsonaro’s outrage
- Election 1st round
- Election 2nd round

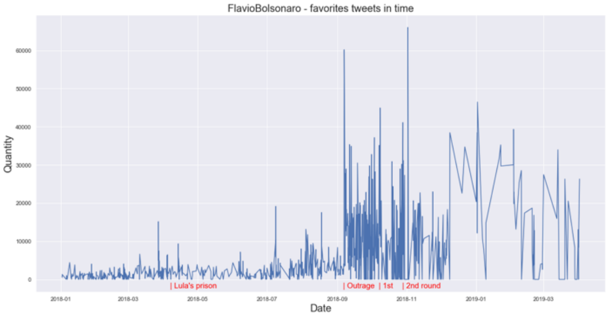


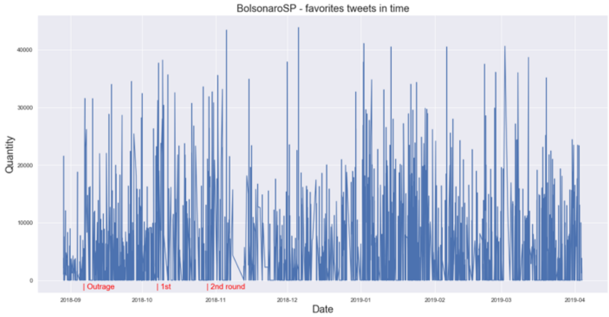
Google Trends

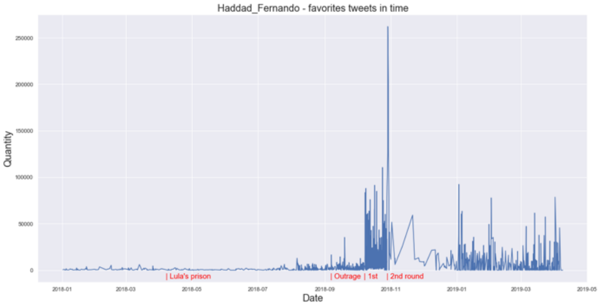
# Favorites by screen_name

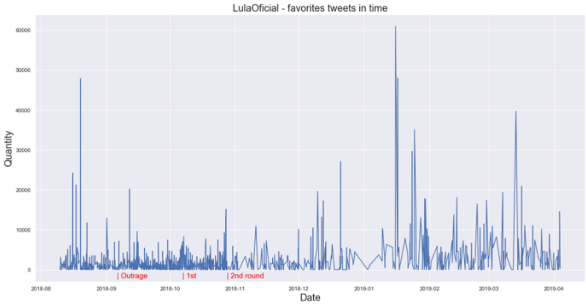

jairbolsonaro - favorites tweets in time

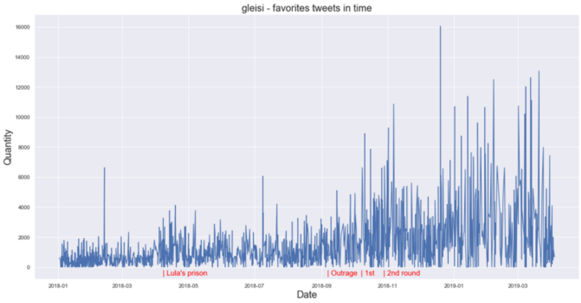

CarlosBolsonaro - favorites tweets in time

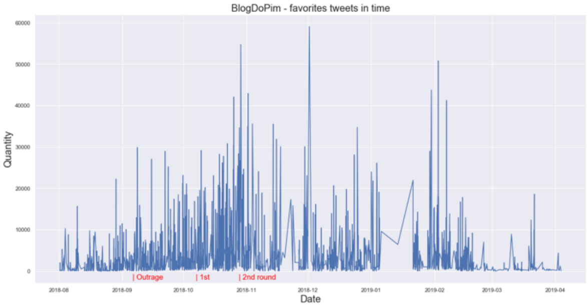

FlavioBolsonaro - favorites tweets in time


BolsonaroSP - favorites tweets in time


Haddad_Fernando - favorites tweets in time

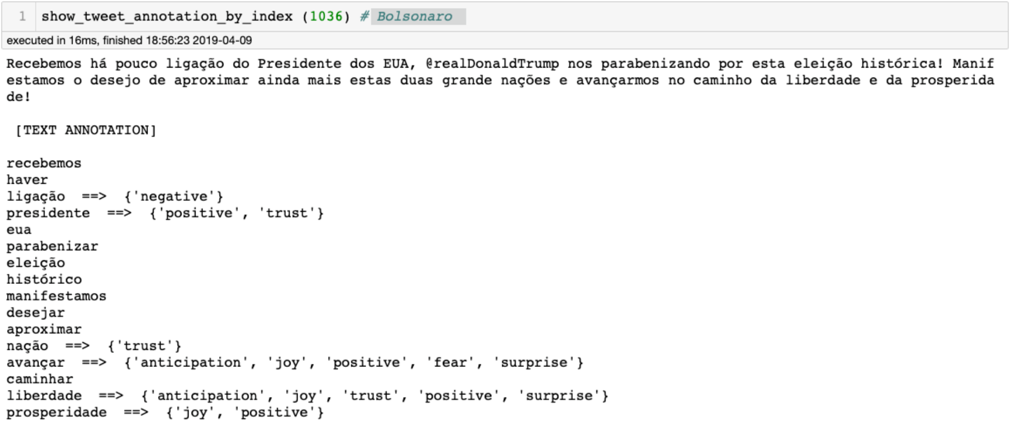

LulaOficial - favorites tweets in time


gleisi - favorites tweets in time


BlogDoPim - favorites tweets in time

A deeper analysis was done on top tweets (more than 150,000 favorites). 10 tweets were found, being 2 from Fernando Haddad and 8 from Bolsonaro.

For example:



```
show_tweet_annotation_by_index (1036) # Bolsonaro
```

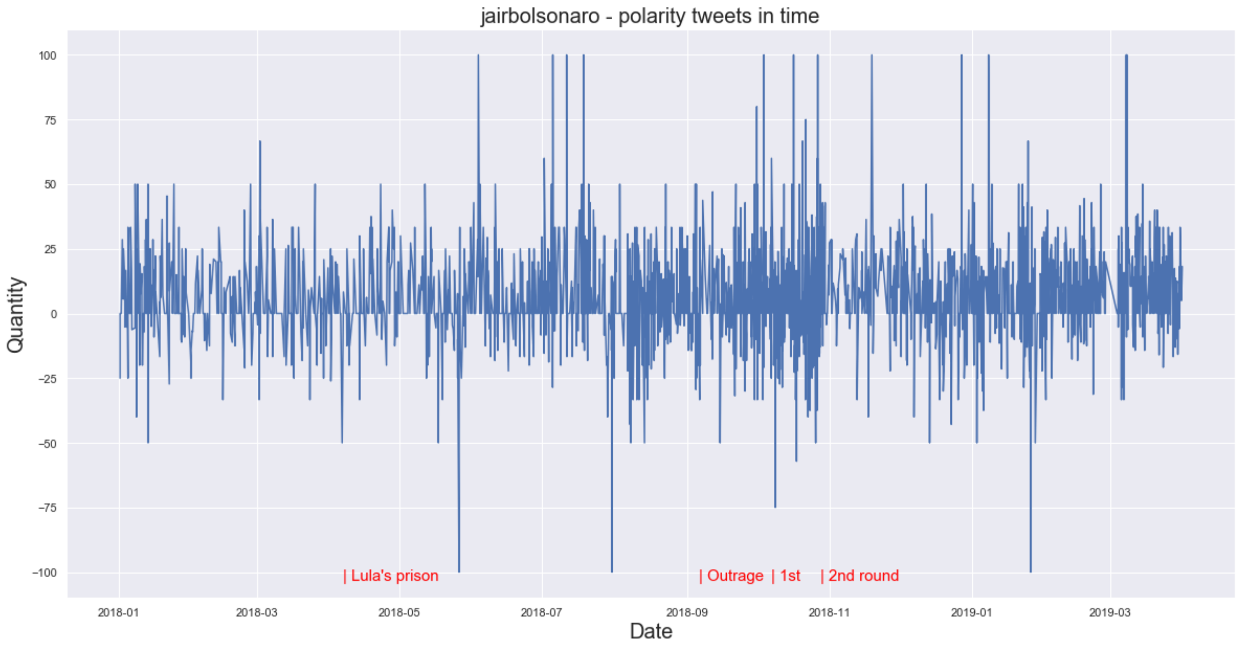
executed in 16ms, finished 18:56:23 2019-04-09

```
Recebemos há pouco ligação do Presidente dos EUA, @realDonaldTrump nos parabenizando por esta eleição histórica! Manif
estamos o desejo de aproximar ainda mais estas duas grande nações e avançarmos no caminho da liberdade e da prosperida
de!

 [TEXT ANNOTATION]

recebemos
haver
ligação  ==>  {'negative'}
presidente  ==>  {'positive', 'trust'}
eua
parabenizar
eleição
histórico
manifestamos
desejar
aproximar
nação  ==>  {'trust'}
avançar  ==>  {'anticipation', 'joy', 'positive', 'fear', 'surprise'}
caminhar
liberdade  ==>  {'anticipation', 'joy', 'trust', 'positive', 'surprise'}
prosperidade  ==>  {'joy', 'positive'}
```

## Creating New features

Additional 2 features were created:

- Polarity ➔ (positive_score – negative_score)*100
- Polarity_pond ➔ Polarity*favorites



From each one of the screen names, the number of positives and negatives tweets were calculated based on the polarity:

screen_name = 'jairbolsonaro'

```
Number of Positive tweets:  1188
Number of Negative tweets:  492
```

screen_name = 'CarlosBolsonaro'

```
Number of Positive tweets:  1165
Number of Negative tweets:  766
```

screen_name = 'FlavioBolsonaro'

```
Number of Positive tweets:  441
Number of Negative tweets:  259
```

screen_name = 'BolsonaroSP'

```
Number of Positive tweets:  1446
Number of Negative tweets:  680
```

screen_name = 'LulaOficial'

```
Number of Positive tweets:  1548
Number of Negative tweets:  637
```

screen_name = 'gleisi'

```
Number of Positive tweets:  1048
Number of Negative tweets:  487
```

screen_name = 'Haddad_Fernando'

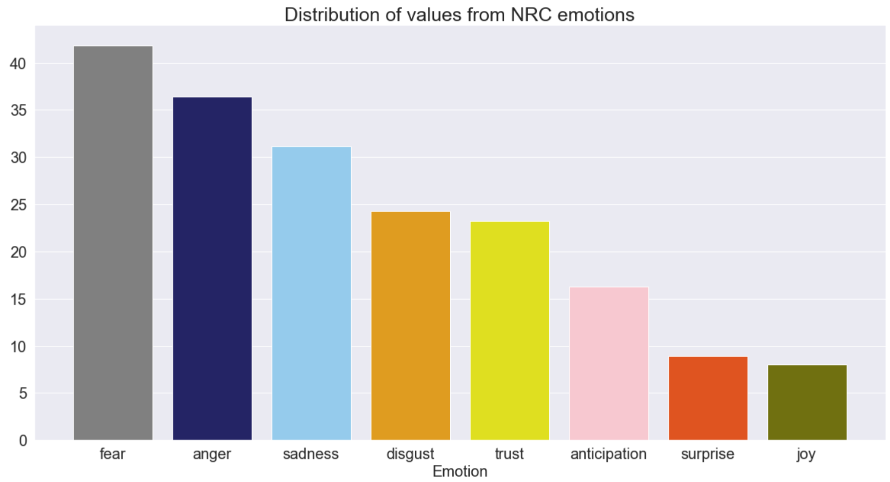
```
Number of Positive tweets:  1060
Number of Negative tweets:  463
```

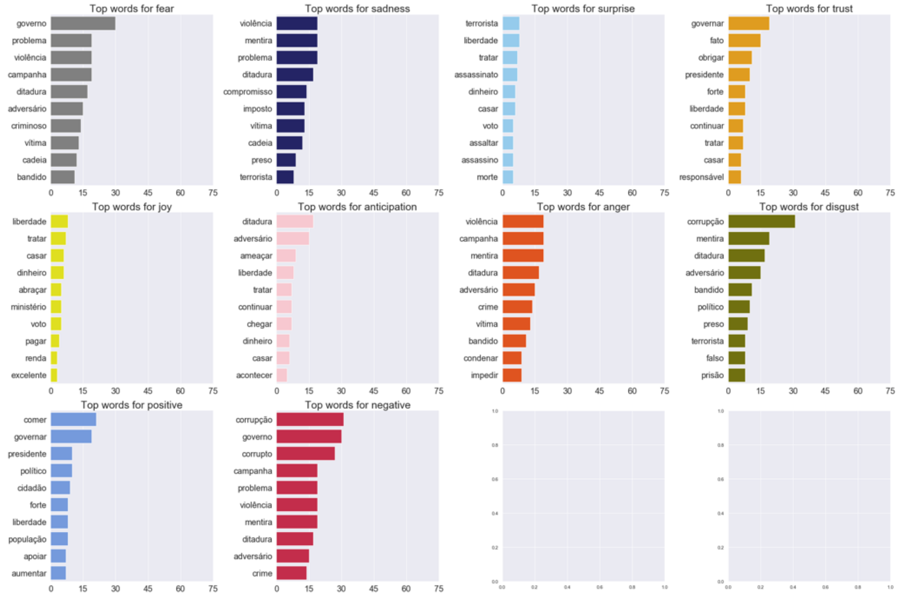
screen_name = 'BlogDoPim'

```
Number of Positive tweets:  1136
Number of Negative tweets:  852
```

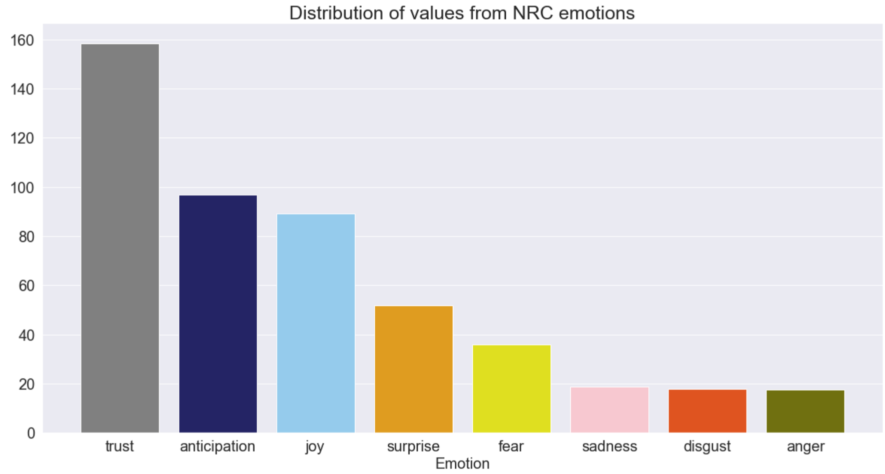
Note: It is interesting to note that in general 2/3 of tweets are “positive” and 1/3 are negative. Could be coincidence but must be deeper analyzed.

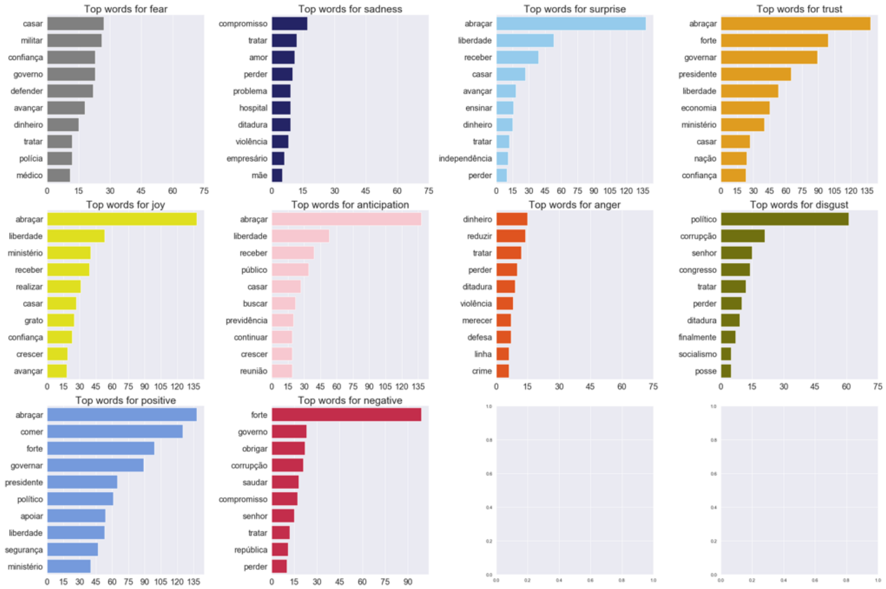
For the weighted polarity score, the emotions were analyzed (for each one) separated for negative tweets:  Jair Bolsonaro



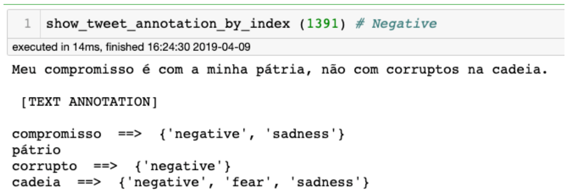


And for positive tweets:





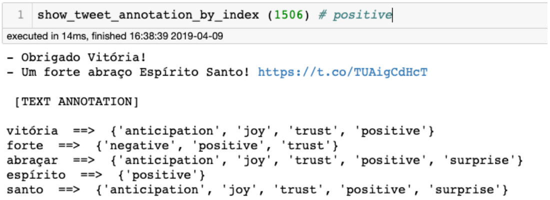
The most negative and most positive tweets were deeper analyzed.



```
show_tweet_annotation_by_index (1391) # Negative
```

executed in 14ms, finished 16:24:30 2019-04-09

```
Meu compromisso é com a minha pátria, não com corruptos na cadeia.

 [TEXT ANNOTATION]

compromisso  ==>  {'negative', 'sadness'}
pátrio
corrupto  ==>  {'negative'}
cadeia  ==>  {'negative', 'fear', 'sadness'}
```



```
show_tweet_annotation_by_index (1506) # positive
```

executed in 14ms, finished 16:38:39 2019-04-09

```
- Obrigado Vitória!
- Um forte abraço Espírito Santo! https://t.co/TUAigCdHcT

 [TEXT ANNOTATION]

vitória  ==>  {'anticipation', 'joy', 'trust', 'positive'}
forte  ==>  {'negative', 'positive', 'trust'}
abraçar  ==>  {'anticipation', 'joy', 'trust', 'positive', 'surprise'}
espírito  ==>  {'positive'}
santo  ==>  {'anticipation', 'joy', 'trust', 'positive', 'surprise'}
```

## Next steps

1. Review Pre-Processing cleaning, mainly POS and use of N-grans
2. Use of # Favorites and/or re-tweets to weight the score of each tweet's emotions (TBC)
3. Analyze temporal relationship among different tweet users
4. Study the possibility of finding subject detection on tweets
5. Compare the results with tweet sentiment/emotion labeled datasets
6. Explore other approaches than Lexicon, as ML, Lexicon+ML, others?

## Bibliography

1. Crowdsourcing a Word–Emotion Association Lexicon, Saif et al. 2013
2. A holistic lexicon-based approach to opinion mining, Ding et al. 2008
3. Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination, Kolchyna 2015
4. O sentimento político em redes sociais: big data, algoritmos e as emoções nos tweets sobre o impeachment de Dilma Rousseff , Malina et al. 2017
5. Lexicon-Based Methods for Sentiment Analysis Taboada et al. 2014
6. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods Ribeiro et al. 2016
7. Sentiment Analysis for 2018 Presidential Election De Nadai et al. (https://github.com/rdenadai/sentiment-analysis-2018-president-election)
8. Data, data Alex Ingberg (https://towardsdatascience.com/data-data-1fedfac91c79)