

Ciência de Dados aplicada à Política

*Identificação de Robôs e Análise do
Comportamento dos Usuários do Twitter diante
da Política Nacional*

Aron Barreira Bordin



Sumário

1 Introdução

2 Fundamentação Teórica

- Sociologia da Comunicação

- Computação

3 Ferramentas e Tecnologias

4 Métodos

5 Análises e Resultados

6 Conclusão

“A crítica arrancou as flores imaginárias que enfeitavam as cadeias, não para que o homem use as cadeias sem qualquer fantasia ou consolação, mas para que se liberte das cadeias e apanhe a flor viva.”

1

Introdução




1 Introdução - Sobre

- Redes sociais revolucionaram a comunicação;
- Surgimento do *prosumidor*: agente que produz e consome informação em tempo real;
- Política 2.0:
 - Redes sociais se tornaram palcos de disputa política;
 - Eleições de Barack Obama, 2008;
 - Ações de propaganda e comunicação através das redes;
 - Uso de *social bots*: robôs automatizados que simulam humanos na rede;
 - Propagação de *fake news* e amadurecimento da pós-verdade.



1 Introdução - Sobre

- Devido a relevância das redes sociais como meio de comunicação, tornaram-se também ameaças ao exercício da democracia;
- Estudos realizados pela FGV-DAPP identificaram robôs presentes nas eleições brasileiras e em diversos momentos políticos;
- Poderiam as redes sociais manipular ou interferir nas eleições?
- Quais seriam os temas e o teor das discussões *online* sobre política?



1 Introdução - Justificativa

- Devido a já comprovada presença de robôs e *fake news* nas discussões políticas, buscou-se estudar o uso do *Twitter* como meio de comunicação;
- Diversos autores desenvolveram métodos de identificação dos robôs e *fake news*, enquanto este trabalho busca realizar análises do teor das mensagens dos usuários;
- Com uso de técnicas de *Ciência de Dados* o sistema desenvolvido é capaz de capturar dados do *Twitter*, realizar análises e produzir relatórios, tudo em tempo real;
- O sistema é de fácil utilização, sendo indicado para cientistas de dados, cientistas sociais e outros pesquisadores estudarem mais a fundo os impactos do *Twitter* na política nacional.

2

Fundamentação Teórica

Parte 1 - Sociologia da Comunicação:
Para a Análise Crítica

Parte 2 - Computação:
Para o Desenvolvimento



2 Fundamentação Teórica - Comunicação

- A comunicação e a informação são fundamentais para a sociedade atual;
- Comunicação de massas:
 - quando o número de pessoas que expressam suas opiniões é muito maior do que as que recebem;
 - objetiva orientar ou influenciar as massas através da comunicação.

“a mídia desempenha um papel crucial na sociedade contemporânea com base em seu fundamento mercantil e em seu caráter capitalista: *ela seleciona, organiza, sistematiza e difunde informações.*” (FERREIRA, 2000)



2 Fundamentação Teórica - Propaganda Política e Manipulação

- **Influência significativa no governo Wilson (EUA, 2016):**
 - criou uma comissão para propaganda estatal;
 - tinha como objetivo tornar a opinião pública favorável a entrada na guerra;

“em seis meses, [a comissão conseguiu] transformar uma população pacifista numa população histórica e belicosa que queria destruir tudo o que fosse alemão, partir os alemães em pedaços, entrar na guerra e salvar o mundo. [...] e após a guerra, foram utilizadas essas mesmas táticas para insuflar o histórico Pânico Vermelho.” (CHOMSKY, 2015, p. 11)



2 Fundamentação Teórica - Propaganda Política e Manipulação

- A manipulação ocorria através de:
 - notícias falsas;
 - convencimento de membros intelectuais e influentes na sociedade, pois uma vez que a propaganda é aceita pela classe erudita, fica mais fácil de manipular o restante da população;
- Essa forma de manipulação foi aceita e desenvolvida por diversos intelectuais da época, sendo chamada de “Teoria Progressista do Pensamento Liberal Democrático”.



2 Fundamentação Teórica - Propaganda Política e Manipulação

- Teoria Progressista do Pensamento Liberal Democrático:
 - defende que a massa pode ser manipulada e “guiada” através da comunicação visando um consenso;
 - defende uma visão supremacista da sociedade, onde existiria uma seleta “classe especializada” que seria responsável por criar e difundir as opiniões para as massas;
 - segundo essa teoria, em oposição ao grupo intelectual, existe um “rebanho desorientado” que tem como papel ser “espectador” da democracia.

“A lógica é cristalina. A propaganda política está para uma democracia assim como o porrete está para um Estado totalitário” (CHOMSKY, 2015, p. 21)



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

- Via de mão-dupla entre a mídia, a opinião pública e as instituições políticas;
- Santos (1992, p. 105-106) apresenta três definições para opinião pública:
 1. “a opinião pública é constituída pelo conjunto das opiniões expressas pelos meios de comunicação de massas, uma vez que é apenas através deles que uma opinião se torna pública”;
 2. seria construída “pelas opiniões do público em geral, independentemente do seu acesso à comunicação social para as expressar”;
 3. “a opinião pública não existe, é um conceito demasiado vasto e amplo, incapaz de traduzir os pensamentos de um público fragmentado onde, na verdade, prolifera um grande número de opiniões diferentes e contraditórias”.



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

- O sociólogo/filósofo Jürgen Habermas trabalha com a terceira definição;
- Para ele, além da opinião pública ser uma farsa, esse conceito surge do que ele define como “esfera pública burguesa”;
- São as pessoas privadas que quando reunidas num espaço público que definem o que passa a ser chamado de “público”;
- Como os representantes das esferas públicas são enviesados, as opiniões por eles produzidas deixam de serem públicas pois representam apenas uma pequena fatia da sociedade;
- O autor defende que conforme as sociedades se desenvolveram, esses espaços foram dominados e conquistados pelo mercado.



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

Devido o mau funcionamento da esfera pública burguesa devemos nos preocupar, pois:

“costumamos pensar que os meios de comunicação são essenciais para a democracia, mas, atualmente, eles geram problemas ao próprio sistema democrático, pois não funcionam de maneira democrática para os cidadãos. [Isso ocorre pois] se põem a serviço dos interesses dos grupos que os controlam” (MORAES; RAMONET; SERRANO, 2013, p. 53).



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

- Além da disfuncionalidade da esfera pública, há o problema de quem a grande mídia representa;
- Primeiro, é importante diferenciar “comunicação” de “informação”:
 - **comunicação:** tem como objetivo elogiar, representar ou defender a instituição a que representa;
 - **informação:** funciona de maneira distinta, é como um contrapeso ao discurso institucional dominante. A informação é um fato que ocorreu.
- Entretanto, hoje as grandes mídias se tornaram grande dependentes do mercado financeiro e das grandes empresas, pois dependem de anúncios, financiamentos, são parceiros de grupos de investimento, dentre outros.



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

- Como consequência, as grandes mídias acabaram se tornando grande porta-vozes do mercado, trabalhando mais como um agente de comunicação do mercado do que um agente informacional.

“Hoje, megaempresas, fundos de investimentos, magnatas das finanças e do petróleo e bancos [...] têm participações acionárias e propriedades cruzadas em mídia. [...] eleva-se a dependência de grupos de mídias a entidades de crédito [...]. Com a expansão de seus negócios e o lastro financeiro assegurado por bancos e fundos de investimentos, conglomerados de mídia se convertem em atores econômicos de primeira linha” (MORAES; RAMONET; SERRANO, 2013, p. 24-25)



2 Fundamentação Teórica - Mídia, Opinião Pública e Política

- Chomsky exemplifica essa interdependência entre mercado, mídia e instituições públicas com um dos primeiros casos estudados;
- Segundo o autor, após a primeira guerra e a crise de 29, com o objetivo de cortar direitos trabalhistas, buscaram formas de desmobilizar a classe trabalhadora;
- Foi desenvolvida a “Fórmula do Vale Mohawk”, método que utiliza a mídia com o objetivo de tornar a opinião pública em favorável aos empresários e contra os trabalhadores, sendo utilizada até hoje.

“O plano era imaginar formas de colocar a população contra os grevistas, apresentando-os como desordeiros, nocivos à população e contrários ao interesse geral” (CHOMSKY, 2015, p. 22-25)



2 Fundamentação Teórica - *Graph Database*

- BD em que são utilizados grafos como estrutura de dados;
- As consultas, armazenamento, e relacionamento fazem uso da Teoria dos Grafos;
- Cada grafo é constituído por um conjunto de vértices (nós) e arestas (relacionamentos entre os vértices).



2 Fundamentação Teórica - *Graph Database*

- Segundo Robinson, Webber e Eifrem (2013, p. 8-10), as vantagens são:
 - **performance:**
 - relacionamento é algo nativo, melhorando o tempo das operações;
 - não precisa de *joins* ou *múltiplas consultas*;
 - **flexibilidade:**
 - é possível representar os dados como eles são, sem trabalho extra para os relacionamentos;
 - devido a estrutura em grafo, o formato dos dados pode ser alterado de forma dinâmica;
 - **agilidade:**
 - permite dados não-estruturados de forma nativa.



2 Fundamentação Teórica - *Big Data*

- McAfee et al. (2012) diferencia *big data* dos outros tipos de dados através de três pontos:
 - **volume:** produção de alto volume de dados;
 - **velocidade:** operações frequentemente necessitam de alta velocidade no tempo de resposta, mesmo dispondo de alto volume de dados; além disso, comumente a produção dos dados ocorre com alta velocidade;
 - **variedade:** uma das principais características é a alta variedade dos dados, geralmente em diferentes formatos, mesclando dados estruturados com não-estruturados.
- Devido essas características dos *big data*, são necessárias ferramentas e tecnologias específicas para sua análise.



2 Fundamentação Teórica - Ciência de Dados

- Área que se popularizou nos últimos anos, principalmente com a propagação dos *big data*;
- Interdisciplinar;
- Composta por um conjunto de ferramentas, modelos e algoritmos para extrair informações a partir de conjunto de dados;
- Um cientista de dados geralmente precisa ter um bom conhecimento em programação, estatística/probabilidade, e na produção de relatórios;

“um cientista de dados é uma pessoa que é melhor em estatística do que qualquer programador e melhor programador do que qualquer estatístico.” Josh Wills

3

Ferramentas e Tecnologias



3 Ferramentas e Tecnologias - *Scala*

- Linguagem de programação baseada em *Java*;
- Desenvolvida desde 2001;
- Combina orientação a objetos com programação funcional de forma concisa e de alto nível;
- Código fonte pode ser compilado em:
 - *Java Bytecode*: para ser executado em uma JVM;
 - *JavaScript*: para ser executado em um navegador de internet;
 - **Nativo**: para ser executado de forma nativa no SO.
- Quando compilado em *Java Bytecode*, é possível: chamar, criar, importar e estender métodos e classes Java de modo natural.


```

38 import java.time.ZoneId
37 import java.time.temporal.ChronoField.EPOCH_DAY
36 import util.control.Breaks._
35
34
33 class DataLoader(val limit: Integer) {
32   implicit val g: ScalaGraph = Janus.g
31   private val startDate = LocalDateTime.of(2018, 4, 19, 0, 0, 0)
30   private val endDate = LocalDateTime.now
29
28   private val duration = (endDate.getLong(EPOCH_DAY) - startDate.getLong(EPOCH_DAY)).toInt
27
26   def iterDate: Iterator[Date] = {
25     for (days <- Iterator.range(0, duration)) yield {
24       val date = startDate.plusDays(days)
23
22       val dateA = Date.from(date.atZone(ZoneId.systemDefault).toInstant)
21       val dateB = Date.from(date.plusDays(1).atZone(ZoneId.systemDefault).toInstant)
20
19       val result = g.V.hasLabel("tweet")
18         .has(Key[Boolean]("tweet_full"), true)
17         .has(Key[String]("by_user_screen_name"))
16         .has(Key[Date]("date"), P.gte(dateA))
15         .has(Key[Date]("date"), P.lt(dateB))
14         .group(By(Key[String]("by_user_screen_name")))
13         .toList
12
11       result.foreach(groups => {
10         groups.foreach(group => {
9           breakable {
8             val size: Integer = group._2.size()
7             if (size < limit) break
6
5             println(group._1 + " -> " + size)
4           }
3         })
2         println("#####")
1       })
47 }
1     dateA
2   }
3 }
4 }

```



3 Ferramentas e Tecnologias - *JanusGraph*

- Implementação de *graph database* que “é otimizado para armazenamento e consulta em grafos contendo centenas de bilhões de vértices e arestas distribuídos em um *cluster*”;
- Escalabilidade linear e elástica;
- Distribuição e replicação dos dados: melhor performance e tolerância de erros;
- Suporta ACID;
- Código aberto, sob licença Apache 2;
- Utiliza a linguagem Gremlin para manipulação e consulta dos dados.



3 Ferramentas e Tecnologias - *ElasticSearch*

- Motor de buscas e de análises em documentos/texto;
- RESTful;
- Distribuído e escalável;
- Desenvolvido em *Java*, código fechado e gratuito;
- Contém clientes oficiais em: Java, C#, PHP, Python, Groovy, Ruby, dentre outros;
- Pode ser utilizado para realizar buscas em todos os tipos de documentos;
- Resultados praticamente em tempo-real.



3 Ferramentas e Tecnologias - *Kibana*

- Interface gráfica para o *ElasticSearch*;
- Com ele é possível:
 - realizar buscas;
 - criar visualizações (gráficos e *dashboards*);
 - executar *queries* e configurar o *ElasticSearch*;
- Gratuito e *open source*.



3 Ferramentas e Tecnologias - *Twitter Streaming API*

- *Twitter* fornece uma série de *APIs* para interação com a rede social;
- Com o *Twitter Streaming* é possível capturar *tweets* em tempo real através de filtros por palavra-chave e por nome de usuário;
- Desse modo, a *API* envia uma cópia de cada publicação e seus metadados que ocorre enquanto o sistema está em execução.



Métodos



4 Métodos - Captura dos Dados

- O sistema foi desenvolvido em *Scala* e configurado para capturar todos os *tweets* de acordo com um determinado filtro;
- O filtro foi composto por 388 termos e 186 perfis de interesses, sendo estes:
 - portais de notícias;
 - perfis de candidatos;
 - perfis de celebridades (jornalistas, políticos, influenciadores digitais, etc);
 - termos sobre temas políticos (aborto, desemprego, dívida, eleição, etc);
 - dentre outros.

4 Métodos

- Os dados capturados são convertidos para a estrutura de grafo e inseridos no *JanusGraph*;
- A estrutura de dados contém:
 - vértices: *tweet, user, hashtag, url*;
 - arestas: *hashtaged, mentioned, linked, quoted, replied, retweeted, tweeted*.



Andrew Pierce @toryboypierce · Apr 18

Humiliation of @jeremycorbyn at PMQs by @theresa_may will feature in @papers @SkyNews at 10.30 with me @Kevin_Maguire

95

43

252



Barrie 🌹

@BarrieJenks

Follow

Replying to @toryboypierce @jeremycorbyn and 3 others



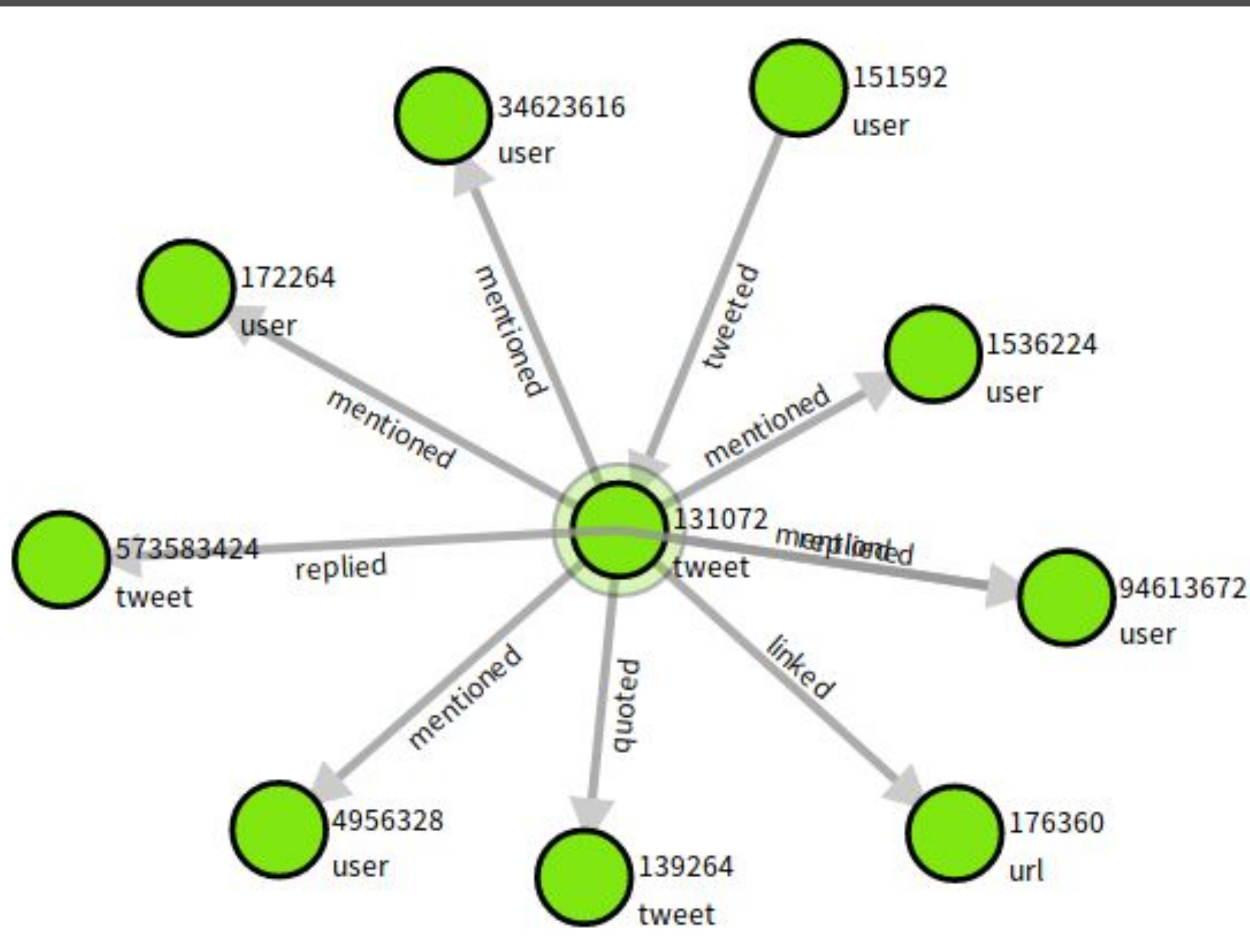
WeThePeople #GTTO @gaurangmorjaria

More 'Truth Bombs'. Has PM been caught lying again? So not only were the cards destroyed under a Tory govt, but also that the decision to do so was taken by a Tory govt! Seems warned by civil servants, Tory MPs & by Corbyn, Diane, Lammy amongst...

Show this thread

2:52 AM - 19 Apr 2018





54 MILHÕES

De *Tweets* foram capturados e analisados
entre 17 de abril e 8 de setembro.

350 GB

Big Data com aproximadamente 350 GB
de dados para processamento e consulta
em tempo real.

350k/dia

Capturando aproximadamente 350 mil
tweets por dia, processando e gerando
relatórios em tempo real.



4 Métodos

- Após serem inseridos no banco, uma *trigger* automaticamente atualiza os dados no *ElasticSearch*;
- A partir do *ElasticSearch* é possível realizar buscas e análises nos conteúdos armazenados;
- O *Kibana* foi configurado para visualização dos dados em tempo real, conforme são capturados e para exploração de dados históricos.

5

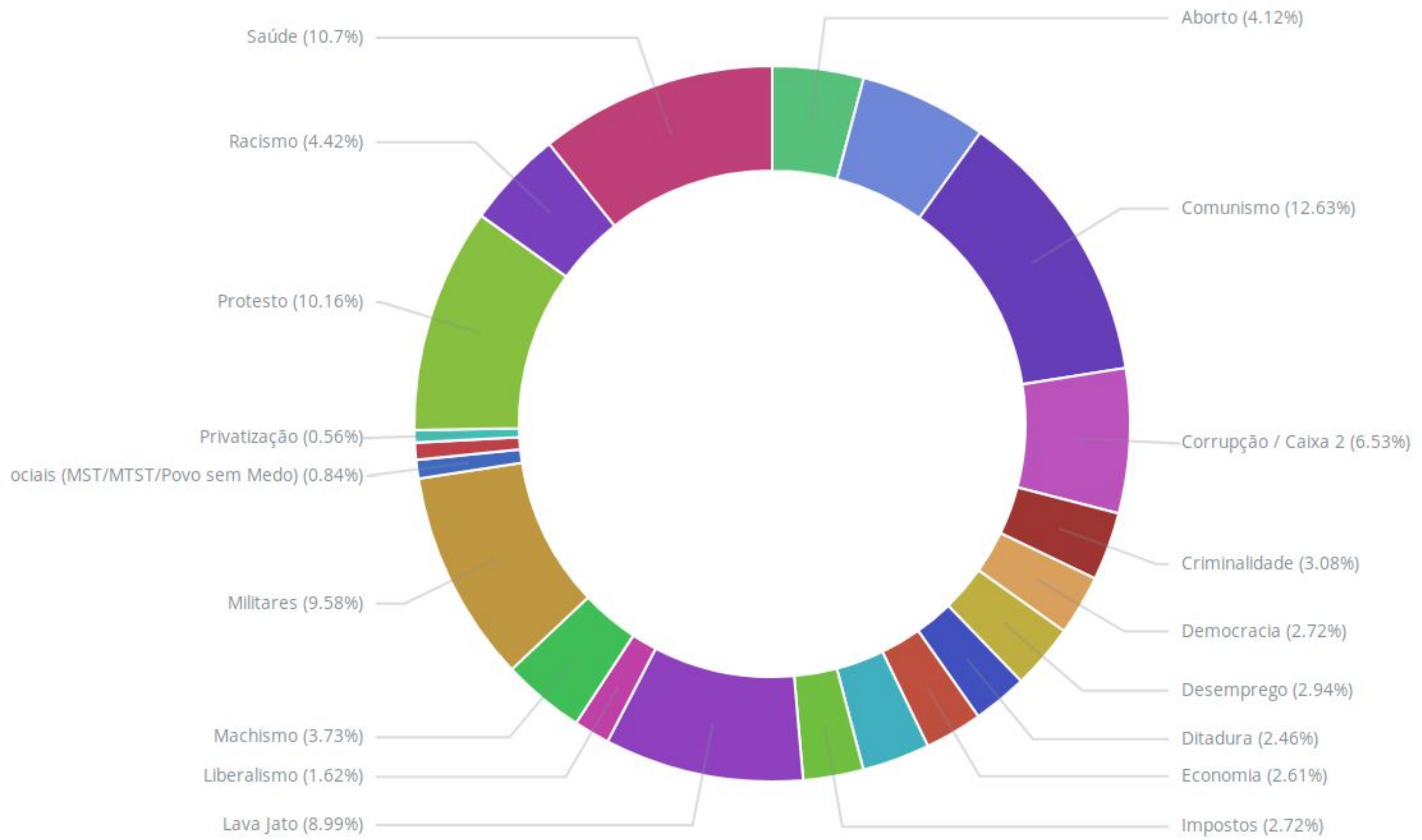
Análises e Resultados



5 Análises e Resultados - Popularidade dos Assuntos

- Foram mapeados 21 assuntos políticos de interesse dos internautas;
- Dentro do período analisado, os assuntos que mais se destacaram foram:
 - comunismo (12,63%);
 - protestos (10,16%);
 - saúde (10,7%);
 - militares (9,58%).

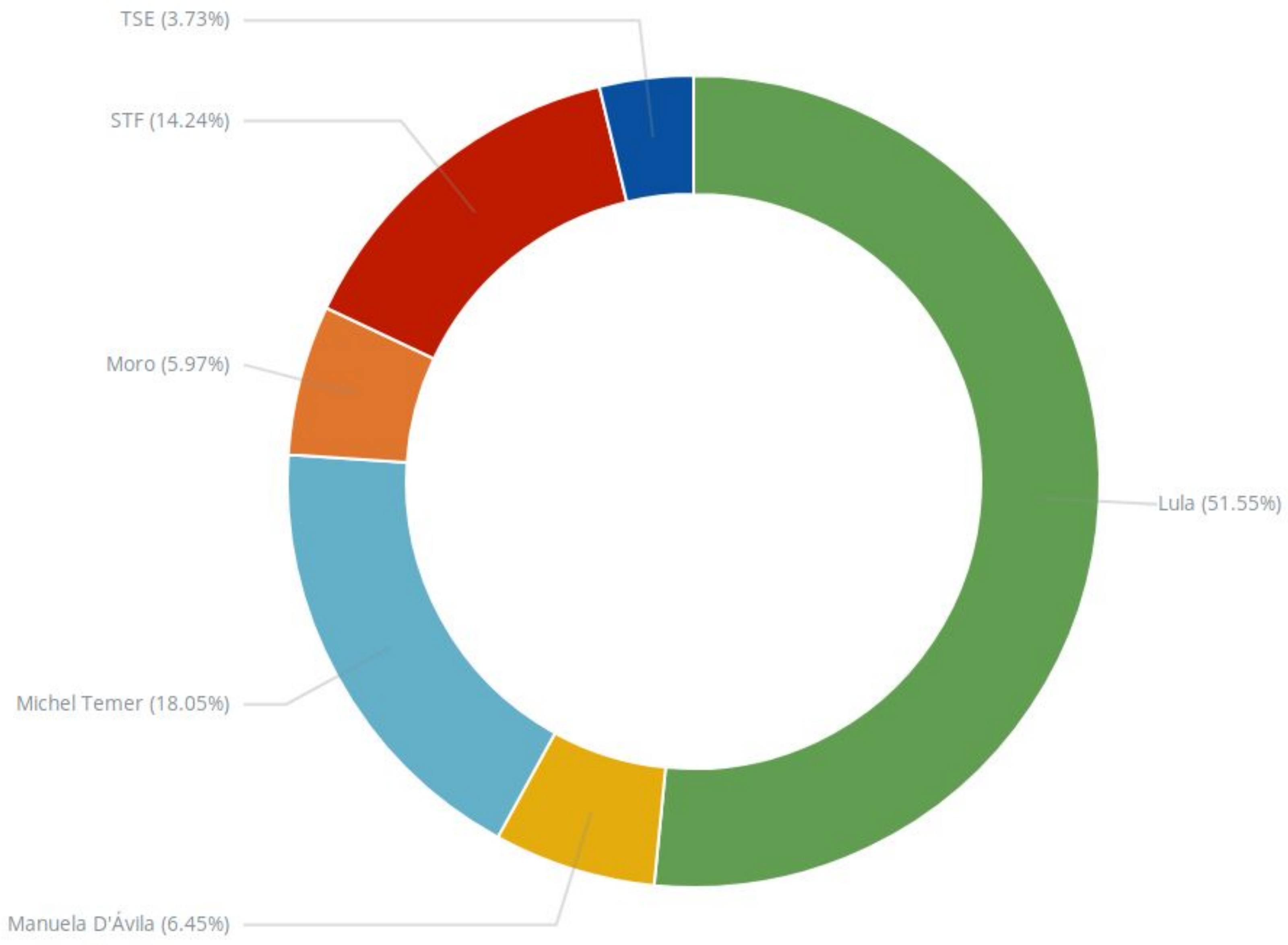
- ▶ Aborto
- ▶ Armamento
- ▶ Comunismo
- ▶ Corrupção / Caixa 2
- ▶ Criminalidade
- ▶ Democracia
- ▶ Desemprego
- ▶ Ditadura
- ▶ Economia
- ▶ Feminismo
- ▶ Impostos
- ▶ Lava Jato
- ▶ Liberalismo
- ▶ Machismo
- ▶ Militares
- ▶ Movimentos Sociais (
- ▶ Previdência
- ▶ Privatização
- ▶ Protesto
- ▶ Racismo
- ▶ Saúde



5 Análises e Resultados - Popularidade das Celebidades Políticas

- Foram mapeados 6 celebridades políticas de interesse dos internautas;
- Consideramos celebridade como uma entidade política que poderia de alguma forma influenciar o debate das eleições, que são eles:
 - Lula, Manuela D'Ávila (que durante parte do período era candidata antes de se tornar vice), Temer, Sérgio Moro, STF e TSE.
- Dentro do período analisado, as celebridades que mais se destacaram foram:
 - Lula (51,55%);
 - Temer (18,05%);
 - STF (14,24%).

- Michel Temer
- Moro
- STF
- TSE

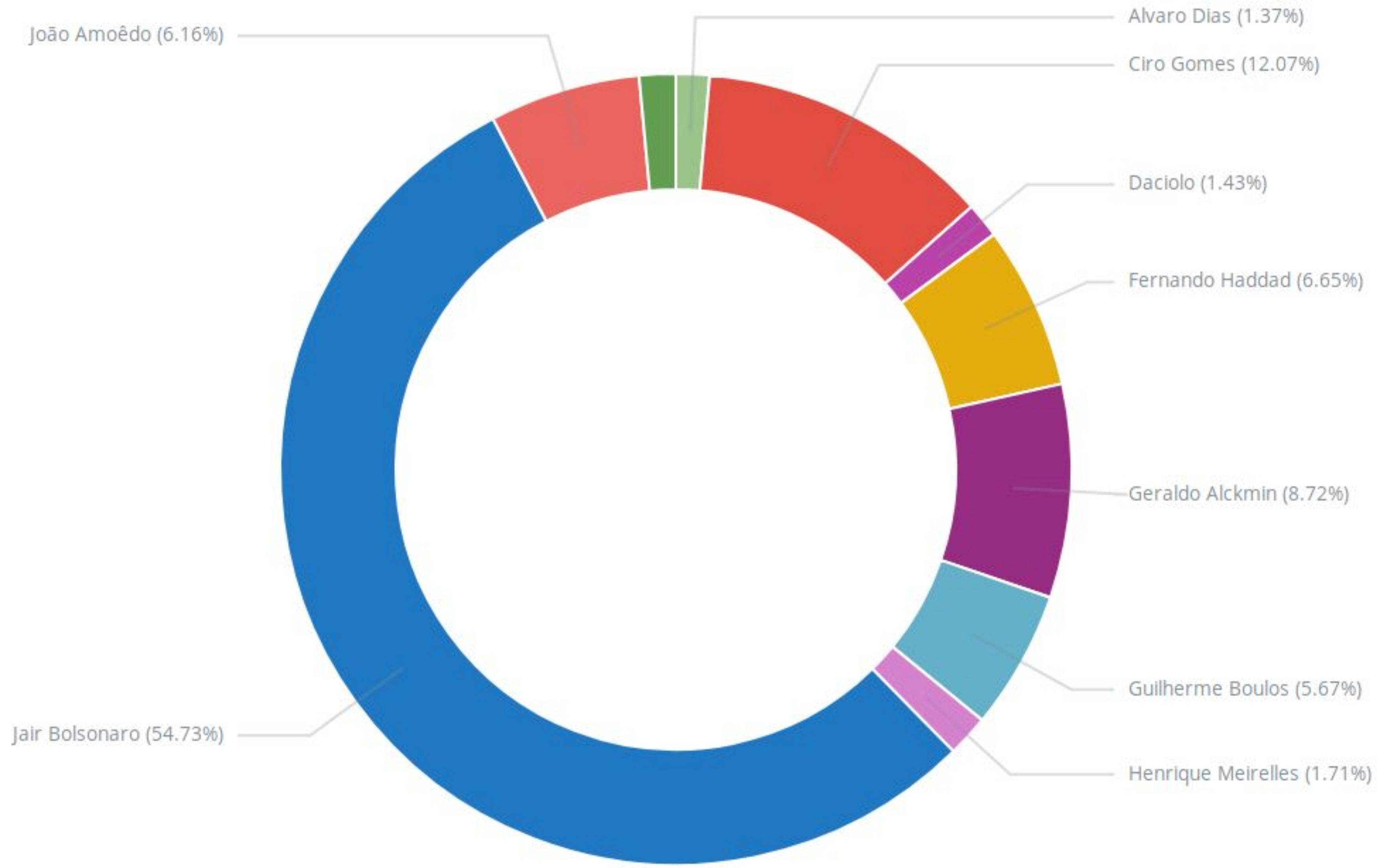




5 Análises e Resultados - Popularidade dos Candidatos

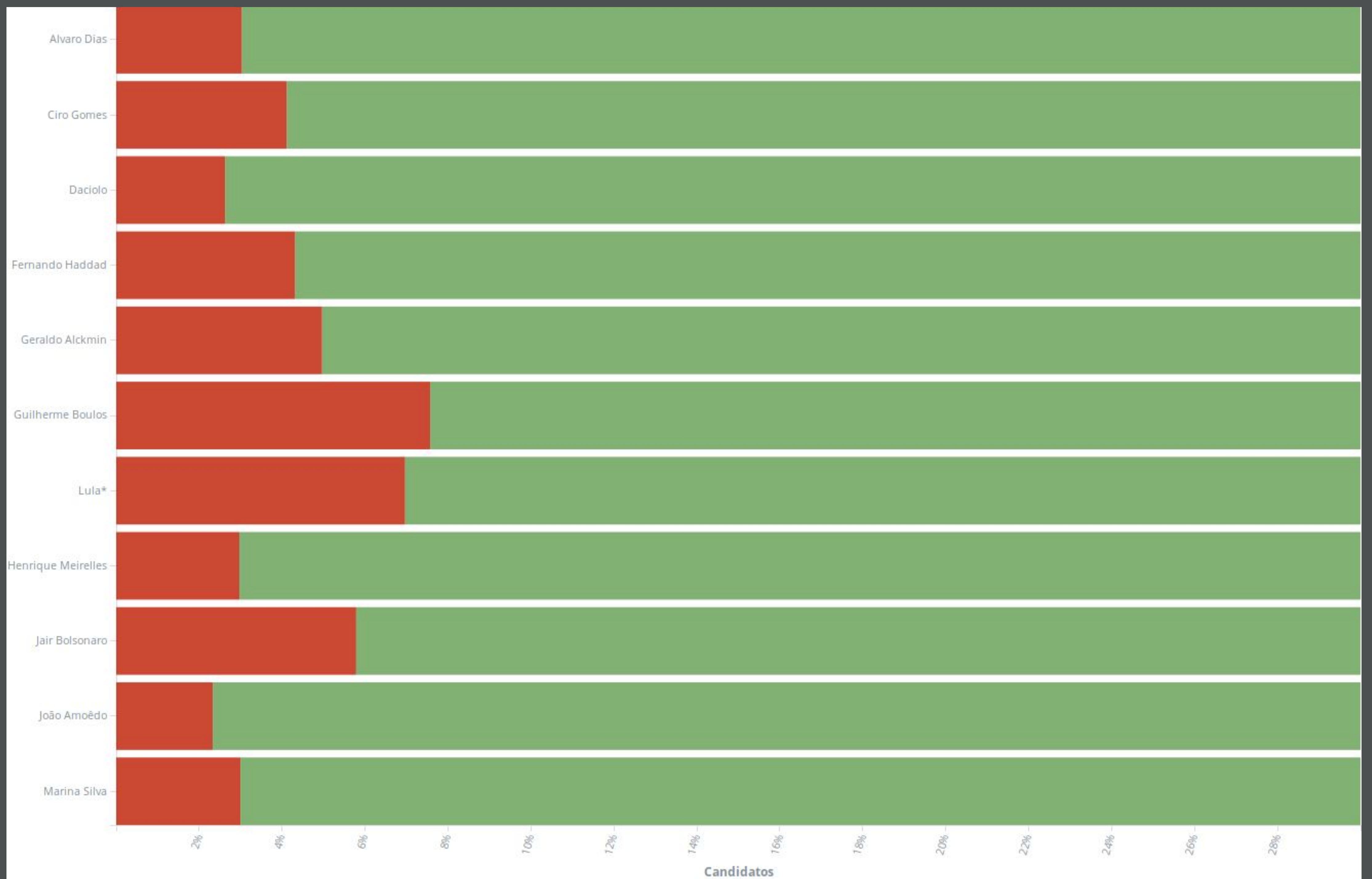
- No caso dos candidatos a popularidade é medida através da quantidade de respostas a *tweets* dos candidatos, citações por nome e citações por nome de usuário;
- Dentre os candidatos à presidência no período analisado, os que mais se destacaram foram:
 - Bolsonaro (54,73%);
 - Ciro Gomes (12,07%);
 - Alckmin (8,72%).

- Daciolo
- Fernando Haddad
- Geraldo Alckmin
- Guilherme Boulos
- Henrique Meirelles
- Jair Bolsonaro
- João Amoêdo
- Marina Silva



5 Análises e Resultados - Discurso de Ódio

- Foi desenvolvido um modelo para verificar se o conteúdo do tweet contém discurso de ódio ou não;
- Dentre as citações de cada candidato, verificamos a proporção das menções na rede que continham discurso de ódio;
- Dentre os candidatos analisados, os com maior quantidade de discurso de ódio em suas menções são:
 - Guilherme Boulos (7,6%);
 - Lula (7,0%);
 - Jair Bolsonaro (5,8%);
 - Alckmin (5,0%).





Conclusão

Obrigado!

Aron Barreira Bordin