CERP USP
Centro de Pesquisa da Religião e Políticas Públicas

CERP – Relatórios Metodológicos

Metodologia para Coleta e Análise de Dados de Redes Sociais para o Boletins CERP - *2ª Edição (Versão 1.2)*

Jun, 2022

Índice

Coletor CERP
Versão 1.2

Introdução

O objetivo deste relatório é apresentar a metodologia da versão 1.2 do *Coletor CERP*.

A principal inovação desta versão em relação à anterior é a ampliação do conjunto de contas acompanhadas, de 667 para 739.

O *Coletor CERP* coleta, diariamente, dados dos *tweets* feitos por pessoas e instituições que são relevantes para a agenda de políticas públicas no Brasil, segundo a vinculação desses agentes a entidades ou organizações religiosas. O vínculo pode ser formal ou informal, e o mesmo usuário pode ser vinculado a mais de um grupo, conforme o caso.

Ao todo, o Monitor distribui cada um dos *tweets* coletados em pelo menos um de 13 grupos:

- Igreja Católica
- Igrejas Evangélicas
- Lideranças religiosas católicas
- Lideranças religiosas evangélicas
- Mídia
- Parlamentares
- Parlamentares que pertencem à Frente Parlamentar em Defesa da Liberdade Religiosa e da Cultura de Paz
- Parlamentares que pertencem à Frente Parlamentar Evangélica do Congresso Nacional
- Parlamentares que pertencem à Frente Parlamentar Mista Católica Apostólica Romana
- Parlamentares que pertencem à Frente Parlamentar Mista da Liberdade Religiosa, Refugiados e Ajuda Humanitária
- Parlamentares que pertencem à Frente Parlamentar para a Liberdade Religiosa do Congresso
- Parlamentares que não pertencem a nenhuma das frentes supracitadas
- Parlamentares que não pertencem à Frente Parlamentar Evangélica do Congresso Nacional

A coleta dos dados é feita por meio de API registrada junto ao *Twitter*, implementada com o auxílio do software R, versão 4.1.1 (R Core Team, 2021), com o pacote *rtweet* (Kearney, 2019), o pacote *readxl* (Wickham e Bryan, 2019), o pacote *tm* (Feinerer e Hornik, 2020) e o pacote *tidyverse* (Wickham et al., 2019).

O processo de coleta de dados pode ser dividido em três grandes etapas: identificação das contas a serem acompanhadas, coleta e agregação dos *tweets* e coleta de dados

secundários. Cada uma das próximas três seções descreverá em detalhes como cada uma dessas etapas é executada.

Identificação das contas a serem acompanhadas

A identificação das contas associadas a cada indivíduo ocorre em duas etapas. Na primeira etapa, é construída uma lista com os nomes das pessoas e instituições que são relevantes em cada um dos 13 grupos supracitados. Os nomes podem ser acrescidos de cargos e pronomes de tratamento, conforme o caso. Cada elemento dessa lista está associado a um termo de busca, usado para buscar as contas associadas a ele.

Ao todo, a lista completa tem 707 indivíduos, distribuídos em grupos conforme a contagem abaixo.

- Igrejas Evangélicas: 50
- Igreja Católica: 4
- Lideranças religiosas evangélicas: 8
- Lideranças religiosas católicas: 3
- Mídia: 15
- Parlamentares: 627
- FP Evangélica do Congresso Nacional: 204
- FP Mista Católica Apostólica Romana: 215
- FP para a Liberdade Religiosa do Congresso: 203
- FP Mista da Liberdade Religiosa, Refugiados e Ajuda Humanitária: 217
- FP em Defesa da Liberdade Religiosa e da Cultura de Paz: 209
- Parlamentares que não pertencem a nenhuma das frentes supracitadas: 185
- Parlamentares que não pertencem à FP Evangélica do Congresso Nacional: 423

Os termos de busca associados às igrejas são, basicamente, seus nomes (em mais de uma variação, quando cabível) ou o nome de associações, congregações, convenções ou confederações vinculadas a elas. Os líderes religiosos, por sua vez, são os nomes das pessoas que exercem papéis oficiais nas igrejas listadas ou que têm forte expressão popular e associam suas imagens públicas a alguma religião ou experiência religiosa (os nomes podem ser acrescidos de termos como “Padre”, “Bispo”, “Pastor”). Para a mídia, foram listados telejornais, jornais impressos e eletrônicos de grande circulação e relevância nacional.

A vinculação dos parlamentares às foi construída com base nas listagens de deputados signatários em cada uma destas frentes, na 56ª legislatura, segundo a listagem publicada no site https://www.camara.leg.br/internet/deputado/frentes.asp. A formalização da composição de cada uma das frentes ocorreu em publicações no Diário da Câmara dos Deputados nas datas abaixo indicadas:

- FP Evangélica do Congresso Nacional: 17/04/2019
- FP Mista da Liberdade Religiosa, Refugiados e Ajuda Humanitária: 25/04/2019
- FP em Defesa da Liberdade Religiosa e da Cultura de Paz: 17/05/2019
- FP Mista Católica Apostólica Romana: 31/05/2019

- FP para a Liberdade Religiosa do Congresso: 17/06/2019

Portanto, a listagem de nomes associados a cada uma dessas frentes reflete a composição do Congresso no momento em que elas foram instauradas. Em outras palavras, deputados e senadores que eventualmente não estejam mais no exercício de seus mandatos não foram retirados das listas.

O termo de busca foi construído iniciando com a palavra Deputado, Deputada, Senador ou Senadora (conforme o caso), seguida pelo nome do parlamentar (como registrado na frente).

Por sua vez, a listagem com os deputados e senadores que não pertencem a nenhuma das frentes foi construída a partir da relação de todos os deputados e senadores da 56ª legislatura que não constam nas listagens anteriores. A listagem com os parlamentares que não pertencem à FP Evangélica do Congresso Nacional foi construída de forma análoga.

Na segunda etapa, usando o pacote *rtweet* e a função *search_users*, são buscadas as contas do *Twitter* mais próximas de cada um dos termos de busca da lista. O máximo de contas associadas a cada termo de busca é três.

Permite-se que um termo de busca seja associado a mais de uma conta porque há diversos casos de pessoas que mantêm contas institucionais e pessoais. Há também o caso de pessoas que mantêm mais de uma conta institucional, como políticos que têm contas separadas para as atividades de campanha e para a divulgação de resultados dos seus mandatos. No caso das igrejas, por exemplo, é comum que uma mesma igreja tenha contas separadas por região. Optou-se por coletar os dados de todas as contas vinculadas a cada pessoa (mesmo quando administrada por terceiros) ou instituição, de forma a obter a expressão mais completa possível do seu pensamento e ideologia.

Ao todo, foram encontradas 739 contas únicas, distribuídas em cada um dos grupos:

- Igrejas Evangélicas: 70
- Igreja Católica: 8
- Lideranças religiosas evangélicas: 17
- Lideranças religiosas católicas: 3
- Mídia: 36
- Parlamento: 605
- FP Evangélica do Congresso Nacional: 186
- FP Mista Católica Apostólica Romana: 206
- FP para a Liberdade Religiosa do Congresso: 188
- FP Mista da Liberdade Religiosa, Refugiados e Ajuda Humanitária: 203
- FP em Defesa da Liberdade Religiosa e da Cultura de Paz: 199
- Parlamentares que não pertencem a nenhuma das frentes supracitadas: 191
- Parlamentares que não pertencem à FP Evangélica do Congresso Nacional: 419

Somando as contas associadas a cada um dos grupos (2.331), tem-se que, em média, cada conta apareceu em 3,15 grupos (2.331 / 739).

Vale destacar que, para eliminar contas voltadas a públicos fora do Brasil, caso alguma das contas resultantes da consulta ao *search_users* tenha apresentado, no conjunto dos seus 40 últimos *tweets* (no momento da busca), mais palavras vazias em inglês, espanhol ou italiano do que em português, ela não terá sido incluída.[1] Além disso, os resultados do algoritmo são conferidos manualmente, para garantir que as contas encontradas sejam das pessoas e instituições que se pretende acompanhar.

A lista de usuários de cada grupo é estática, de modo que o algoritmo de identificação das contas associadas a cada termo de busca foi executado somente uma vez. A versão da listagem de contas acompanhadas foi elaborada em 06/06/2022.

Coleta e Agregação dos *tweets*

A coleta dos *tweets* é feita usando a função *get_timeline*, do pacote *rtweet*. Na primeira coleta de dados de um usuário, são extraídos os seus últimos 3.500 *tweets*. Depois da primeira coleta, somente são coletados os *tweets* que ocorreram após o último *tweet* registrado no banco de dados. Não são coletadas respostas a *tweets*, somente *tweets* originais.

Durante a coleta de dados, os *tweets* são armazenados em um *data frame*. Antes de salvar o *data frame* em disco, ocorrem algumas operações de limpeza da base de dados:

- São extraídos todos os links http, https e ftp
- São removidos os caracteres especiais “!”, “#”, “$”, “%”, “(“, “)”, “*”, “,”, “.”, “:”, “;”, “<”, “=”, “>”, “@”, “^”, “_”, “|”, “~”, “{”, “}”, “[“, “]”, “-“, “+”, “/”, “?”, “‘” e “”””
- São removidas as mudanças de linha e de parágrafo

Uma vez finalizadas essas operações, o *data frame* com os *tweets* coletados é salvo, agregando os resultados da coleta corrente com os resultados das coletas anteriores.

Coleta de Dados Secundários

Além dos dados de *tweets*, são diariamente coletadas as informações sobre o número de seguidores de cada uma das contas acompanhadas, por meio da função *lookup_users*. Com isso, é possível acompanhar a evolução das contas analisadas.

Adicionalmente, são coletados dados gerais de identificação pessoal, filiação partidária e condição eleitoral de parlamentares, usando as APIs da Câmara dos Deputados e do Senado Federal.

[1] Para a definição de palavra vazia, foi usado o conjunto de termos da função *stopwords*, do pacote *tm*.

Monitor das Lideranças Religiosas, Políticas e da Mídia
Versão 1.1

Introdução

O Monitor das Lideranças Relgiosas, Políticas e da Mídia tem por objetivo avaliar como o discurso de pessoas influentes para políticas públicas no Brasil se divide segundo a sua participação em instituições religiosas.

Para atingir este objetivo, ele utiliza a base de dados *Coletor CERP* e o software R, versão 4.1.1 (R Core Team, 2021), com os pacotes *tm* (Feinerer e Hornik, 2020), *readxl* (Wickham e Bryan, 2019), *tidytext* (Silge e Robinson, 2016), *SnowballC* (Bouchet-Valat, 2020), *dplyr* (Wickham et al., 2021), *tidyverse* (Wickham et al., 2019), *wordcloud* (Fellows, 2018), *RColorBrewer* (Neuwirth, 2014) e *lexiconPT* (Gonzaga, 2017).

Coleta e agrupamento de dados

O algoritmo parte da extração dos dados de *tweets* da base *Coletor CERP*, carregando e distribuindo os dados disponíveis no *Coletor CERP* segundo os grupos a que os indivíduos estão vinculados. Quando os dados de um grupo são carregados, o algoritmo realiza uma série de limpezas da base de dados. A primeira delas se refere à filtragem dos dados para que permaneçam somente os *tweets* realizados no intervalo de tempo que se escolheu analisar. No caso, uma semana.

Então, utilizam-se as funções do pacote *tm* para fazer com que todas as palavras sejam representadas em letras minúsculas, excluindo os números e qualquer forma residual de pontuação.

Em seguida, ocorre a remoção de expressões comuns sem significado (preposições, pronomes, advérbios) usando o pacote *tm*, com sua a listagem de termos vazios disponibilizada na função *stopwords*. A listagem de palavras do pacote foi calibrada para o português, e toda a forma de acentuação das palavras listadas foi removida, de modo a compatibilizar sua estrutura com a estrutura das palavras do banco de dados *Coletor CERP*. Vale destacar que os termos da função *stopwords* são complementados por outros: palavras em inglês (eventualmente, algumas contas brasileiras que geralmente escrevem seus *tweets* em português usam expressões em inglês, como *the*, *there* e *their*), numerais, pronomes, preposições e advérbios não contemplados na lista original da função *stopwords*, assim como substantivos ou adjetivos que indicam tempo (*dia*, *noite*, *manhã*) ou têm significado genérico (*coisa*).

A lista completa de termos removidos pode ser encontrada no Anexo I, ao final desta seção.

Em seguida, é construída uma matriz contendo todas as palavras nas colunas e todos os *tweets* nas linhas, usando a função *TermDocumentMatrix*, do pacote *tm*. A partir dessa matriz, são computadas as frequências com que cada palavra aparece no discurso de cada grupo.

O procedimento descrito acima é repetido para cada um dos grupos do *Coletor CERP*.

Para os dados secundários (como número de usuários e filiação partidária), que mudam ao longo do tempo, são usados os resultados da última extração de dados dentro do período de análise considerado.

Identificação de radicais

Uma vez finalizado o tratamento inicial de dados, as palavras são agregadas conforme seus radicais comuns, usando a função *Stem* do pacote *SnowballC*. Quando são identificadas duas palavras com o mesmo radical, elas são agrupadas. A frequência da palavra agrupada é igual à soma das frequências das palavras de mesmo radical, em cada grupo, e a palavra agrupada é igual à palavra, entre aquelas com o mesmo radical, com maior frequência em todos os grupos (de forma que as estatísticas de frequência de palavras dos diferentes grupos sejam comparáveis).

Por exemplo, suponha que a frequência com que apareçam os termos “mulher” e “mulheres” seja a indicada na Tabela 1. Nesse caso, após o procedimento descrito anteriormente, serão atribuídas 6 ocorrências da palavra “mulher” para o Grupo A, 10 ocorrências da palavra “mulher” para o Grupo B e a palavra “mulheres” será removida da base de dados.

Tabela 1: exemplo de agregação de radicais

Palavra	Frequência Grupo A	Frequência Grupo B
“mulher”	4	7
“mulheres”	2	3

Geração das nuvens e dos arquivos de frequência de palavras

Uma vez finalizado este processo, as nuvens de palavras são construídas usando a função *wordcloud*, do pacote *Wordcloud*, e o pacote *RcolorBrewer*. Na nuvem de palavras só entram termos que tenham frequência mínima de 3, e o máximo de palavras exibidas é 200.

Já as frequências de palavras são calculadas em três formatos distintos:

1. Ocorrências absolutas de cada termo
2. Frequência relativa de cada termo, para cada indivíduo de cada grupo
3. Frequência relativa de cada termo, para cada indivíduo de cada grupo, subtraído da média da frequência relativa deste termo em todos os outros indivíduos

Enquanto a primeira forma de cálculo da frequência leva em consideração os termos mais e menos usados, a segunda controla para o fato de que as contas postam em diferentes frequências (algumas contas muito mais ativas do que outras), em *tweets* de tamanhos variados (algumas contas com postagens tipicamente mais longas que outras). Por fim, a terceira forma de cálculo da frequência desconta os termos tipicamente utilizados.

Análise de sentimento

Para a análise de sentimento, é utilizado o banco de dados *sentiLex_lem_PT02,* do pacote *lexiconPT*.

Primeiramente, os *tweets* de cada usuário, em cada grupo, são classificados em uma categoria, conforme a tabela abaixo. O *tweet* entra em uma categoria quando contém algum dos termos daquela categoria. O mesmo *tweet* pode ser classificado em múltiplas categorias, dependendo do seu conteúdo.

Categoria	**Termos**
Gasolina	gasolina
Diesel	diesel
Petróleo	petroleo
Petrobras	petrobras
Inflação	inflacao, igpm, ipca
IPTU	iptu
ICMS	icms
IPVA	ipva
IRPF	irpf, imposto renda pessoa fisica
IRPJ	irpj, imposto renda pessoa juridica
CSLL	csll, contribuicao social lucro liquido
Educação Básica	escola publica

Educação Superior	faculdade publica, faculdades publicas, universidade publica, universidades publicas, instituto federal, institutos federais
IDEB	ideb
FUNDEB	fundeb
ENEM	enem
ENADE	enade
FIES	fies, fgeduc
SISU	sisu
PROUNI	prouni
PRONATEC	pronatec
Bolsa Família	bolsa familia
Auxílio Emergencial	auxilio emergencial
Auxílio Brasil	auxilio brasil
Inclusão produtiva	inclusao produtiva
BPC	bpc, beneficio prestacao continuada
Microcrédito	microcredito
Fomento urbano	fomento urbano
Fomento rural	fomento rural
CRAS	cras
CREAS	creas

Desmatamento	desmatamento
SUS	sistema unico saude
Aborto	aborto
Vacina	vacina
Máscara	mascara
COVID	covid
Pandemia	pandemia
Endemia	endemia
UBS	unidade basica saude
FGTS	fgts, fundo garantia
PASEP	pasep
Abono Salarial	abono salarial
Salário Família	salario familia
INSS	inss
Aposentadoria	aposentadoria
SINE	sistema nacional emprego
Jair Bolsonaro	bolsonaro
Lula	lula

Quando um *tweet* é classificado em uma categoria, sua polarização é calculada. A polarização é calculada somando o valor da polarização de todas as palavras do *tweet*, de acordo com o dicionário supramencionado. O cálculo ignora a polarização associada aos termos que definem cada uma das categorias.

O indicador de polarização é calculado para cada um dos grupos do *Coletor CERP*, de modo que os dados permitem a comparação da polarização dos diferentes grupos.

Anexo I: listagem de palavras excluídas

dia, dias, semana, semanas, mes, meses, ano, anos, hora, horas, minuto, minutos, segundo, segundos, noite, manha, madrugada, hoje, amanha, agora, sempre, atual, alem, nunca, enquanto, durante, pre, pos, primeiro, primeira, ultimo, ultima, ultimos, ultimas, proximo, proxima, proximos, proximas, inicio, fim, comeco, antes, depois, vezes, demarco, segunda, terca, quarta, quinta, sexta, feira, sabado, domingo, seg, ter, qua, qui, sex, sab, dom, janeiro, fevereiro, marco, abril, maio, junho, julho, agosto, setembro, outubro, novembro, dezembro, jan, fev, mar, abr, mai, jun, jul, ago, set, out, nov, dez, the, there, their, new, york, mil, milhao, bilhao, milhoes, bilhoes, trilhao, trilhoes, um, uma, dois, duas, tres, quatro, cinco, seis, sete, oito, nove, dez, todos, todas, mim, vosso, vossa, ambos, alta, baixa, atras, todo, toda, tudo, nada, desse, deste, desses, destes, dessa, dessas, desta, destas, nessa, nessas, nesta, nestas, nesse, neste, nesses, nestes, outro, outra, outros, outras, algum, alguma, alguns, algumas, alguem, aqui, ali, acima, abaixo, qualquer, quaisquer, inves, apos, por, porque, pois, que, qual, quais, onde, cada, quanto, tanto, demais, perto, longe, melhor, pior, pra, pro, muito, muita, muitos, muitas, pouco, pouca, poucos, poucas, apenas, assim, atraves, entao, portanto, porem, mas, entanto, todavia, contudo, logo, ainda, maior, menor, algo, maiores, menores, mais, menos, mesmo, mesma, mesmos, mesmas, tao, tal, apos, sobre, grande, grandes, pequeno, pequenos, pequena, pequenas, sim, nao, bom, ruim, unico, unica, extremamente, bla, coisa, coisas, sendo, total, diversas, aaaa

Referências

Bouchet-Valat, M. (2020). *SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library*. R package version 0.7.0. https://CRAN.R-project.org/package=SnowballC

Feinerer, I.; Hornik, K. (2020). *tm: Text Mining Package*. R package version 0.7-8. https://CRAN.R-project.org/package=tm

Fellows, I. (2018). *wordcloud: Word Clouds*. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

Gonzaga, S. (2017). lexiconPT: Lexicons for Portuguese Text Analysis. R package version 0.1.0. https://CRAN.R-project.org/package=lexiconPT

Kearney, M. (2019). *rtweet: Collecting and analyzing Twitter data, Journal of Open Source Software*, 4, 42.1829. doi:10.21105/joss.01829 (R package version 0.7.0)

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer

R Core Team (2021). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Silge, J.; Robinson, D. (2016). *tidytext: Text Mining and Analysis Using Tidy Data Principles in R*. _JOSS_, *1*(3). doi: 10.21105/joss.00037 (http://dx.doi.org/10.21105/joss.00037)

Wickham, H.; Bryan, J. (2019). *readxl: Read Excel Files*. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham, H. et al., (2019). *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, H.; François, R.; Henry, L.; Müller, K. (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr