MINERAÇÃO DE TEXTO: UMA ANÁLISE DOS TWEETS DE POLÍTICOS DO RIO DE JANEIRO

Introdução

A comunicação foi revolucionada a partir da disseminação do uso de redes sociais virtuais. Diante dessa nova realidade, políticos adotaram perfis nas redes e passaram a ter uma forma de interação mais direta com seus seguidores. O *Twitter*, de forma especial, tornou-se quase um meio oficial de comunicação. A rede social criada em 2006 consiste em compartilhar mensagens curtas, de até 280 caracteres. Rapidamente, ela ganhou popularidade e, em 2012, atingiu mais de 100 milhões de usuários e cerca de 340 milhões de mensagens por dia.

Nesse sentido, a análise dos textos postados permite uma análise do posicionamento político e de como é sua interação com a sociedade. Esse estudo, portanto, apresenta uma aplicação de mineração de texto aos *tweets* postados por três políticos do Rio de Janeiro. Foram escolhidos políticos de campos distintos. São eles: Carlos Bolsonaro (PSL), Marcelo Freixo (PSOL) e Rodrigo Maia (DEM). A escolha deveu-se por serem três políticos do mesmo estado e que possuem relevância no cenário nacional. Rodrigo Maia é o atual presidente da Câmara dos Deputados e Marcelo Freixo é um importante nome da oposição ao Governo Federal, tendo disputado a última eleição para a presidência da Câmara. Apesar de Carlos Bolsonaro ser vereador, ele é filho do atual Presidente da República. Ressalta-se, também, que ele é um dos principais responsáveis pela comunicação do grupo de apoio ao Governo nas redes sociais.

Objetivos

O estudo teve como objetivo analisar, utilizando a técnica de mineração de texto, o posicionamento de políticos a partir de publicações no *Twitter*. Esse objetivo compreende verificar a frequência dos termos, bem como testar um modelo que possa identificar o autor de um *tweet* específico, sem saber *a priori*.

Material e Método

Este estudo foi realizado a partir da extração dos textos postados no *Twitter* pelos políticos Carlos Bolsonaro (PSL-RJ), Marcelo Freixo (PSOL-RJ) e Rodrigo Maia (DEM-RJ). Para tanto, utilizou-se o pacote *twitteR* (Gentry, 2015). Foram consideradas somente as mensagens originadas pelos próprios políticos, sendo excluídas respostas ou compartilhamentos. A API do *Twitter* limita a quantidade de postagens a serem recuperadas. A quantidade de mensagens de cada político é distinta, pois a proporção entre *tweets*,

retweets e repostas é diferente. Optou-se por buscar o maior volume de texto permitido para aumentar a quantidade de dados para análise em detrimento da padronização, seja pelo período ou quantidade de mensagens. A análise consistiu, primeiramente, em realizar uma nuvem de palavras para cada um, a fim de avaliar os termos utilizados mais frequentemente. O próximo passo foi realizar a análise de sentimento, para avaliar a frequência e a proporção entre termos positivos e negativos nos *tweets* de cada político. Para tanto utilizou-se o dicionário *SentiLex-PT02*, disponível no pacote *lexicon-PT* (Gonzaga, 2017). Em seguida, foi aplicada a técnica de TF-IDF para avaliar os termos mais utilizados por cada político que não foram utilizados ou foram pouco utilizados pelos demais. Por fim, utilizando o pacote *topicmodels* (Grün B, 2011), foi feita uma modelagem utilizando a técnica de *topic modelling* e o algoritmo LDA (*Latent Dirichlet Allocation*). As análises foram realizadas utilizando o software R, versão 3.5.

Resultados e Discussão

Analisando a nuvem de palavras de cada político (Figura 1), nota-se uma diferença uma clara diferença nas expressões mais usadas por cada político.

Carlos Bolsonaro **Marcelo Freixo** **Rodrigo Maia**

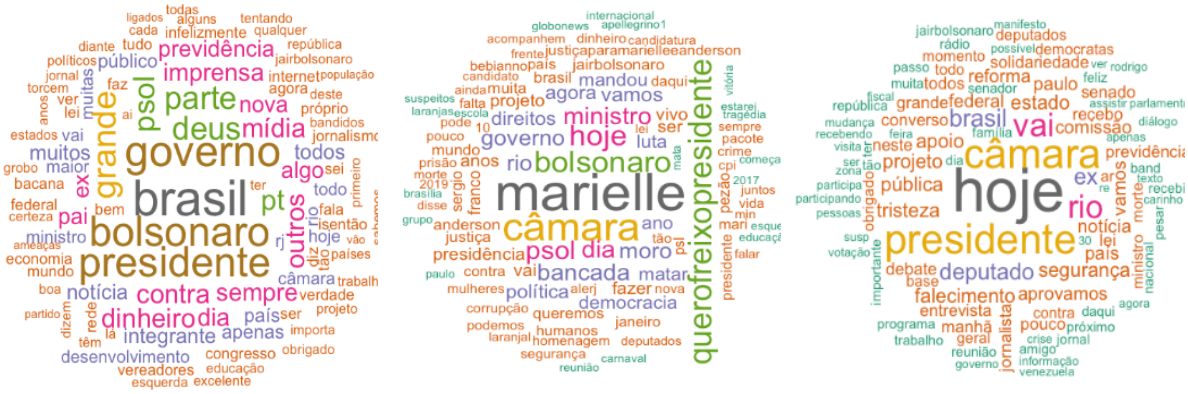

Figura 1 – Nuvem de palavras dos políticos baseada na frequência de cada termo.

Fonte: Elaboração própria.

A análise de sentimentos (Figura 2) das publicações ao longo do tempo revela uma maior presença de termos negativos nos textos dos três políticos (barras em vermelho), em comparação com os positivos (barras em azul). É observada uma posição mais neutra do Deputado Rodrigo Maia, enquanto Carlos Bolsonaro e Marcelo Freixo, possuem muitos textos com sentimento negativo, este último especialmente após fevereiro, mês em que foi iniciado o novo período legislativo com novo governo. Explica-se os diferentes intervalos de tempo entre os *tweets* dos três políticos pela diferença na quantidade de postagens que fazem. Por

exemplo, como Rodrigo Maia é o que posta com menor frequência, nota-se um período maior entre uma postagem e outra.

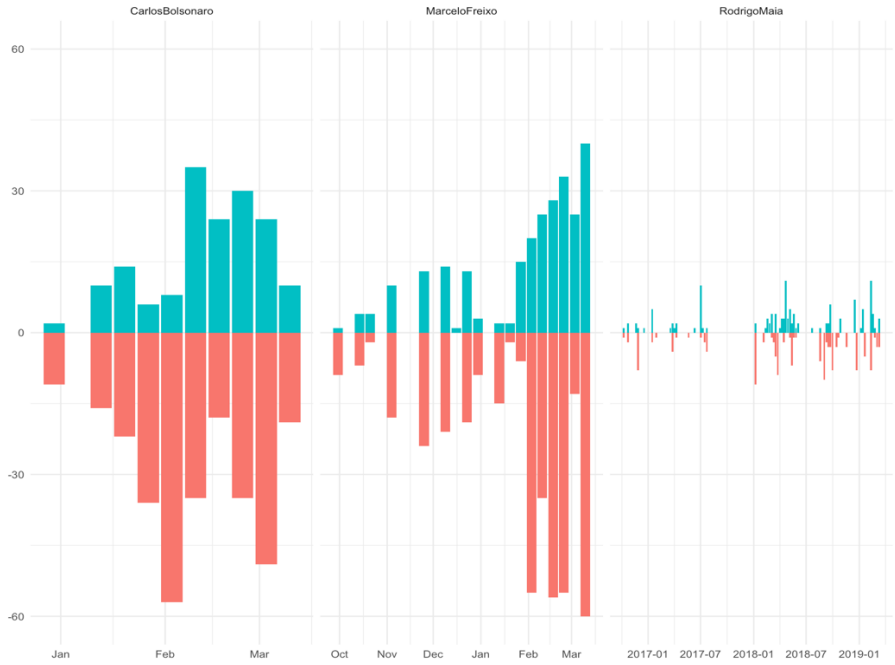

Figura 2 – Análise de sentimento agrupado por semana com base nas publicações dos políticos.

Fonte: Elaboração própria.

A fim de verificar a importância das palavras por políticos, foi utilizada a técnica TF-IDF (*term frequency–inverse document frequency*). Dessa forma, pode-se inferir quais palavras estão associadas a qual político, baseado no valor TF-IDF (Figura 3).

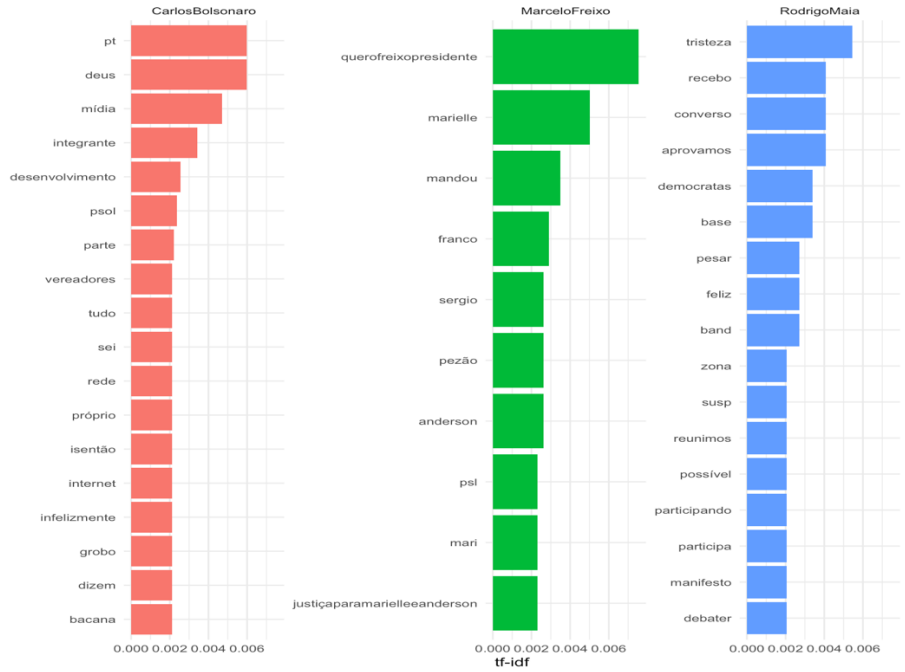

Figura 3 – Termos mapeados pela técnica TF-IDF

Fonte: Elaboração própria.

O *topic modelling* é um método de classificação, no qual para cada palavra é atribuída uma probabilidade (beta) de pertencer a um grupo. Assim, é possível, a partir de um conjunto de palavras, inferir a probabilidade de pertencer a determinado político, mesmo sem conhecer *a priori* a autoria. Nesse sentido, o gráfico abaixo (Figura 4) apresenta o resultado do modelo utilizando o algoritmo LDA. Foram escolhidos três grupos, a fim de discriminar os textos de cada político. Nota-se que o primeiro grupo (1) é referente a palavras utilizadas pelo vereador Carlos Bolsonaro. Já o segundo (2) é referente aos textos do deputado Marcelo Freixo. Por fim, o terceiro grupo (3) é relativo a palavras utilizadas pelo deputado Rodrigo Maia.

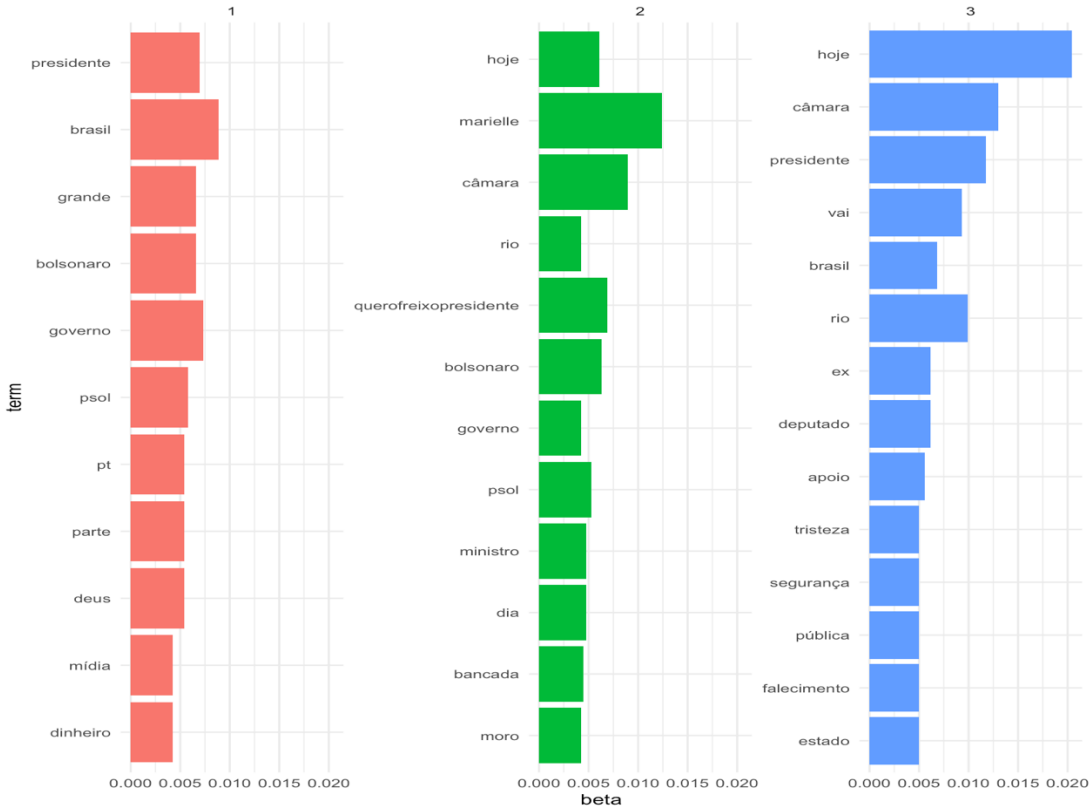

Figura 4 – Topic Modeling

Fonte: Elaboração própria.

Conclusão

As análises realizadas reforçam a percepção de que as redes sociais são utilizadas predominantemente para ataques a opositores e suas ideias, e menos para a defesa de ideias próprias.

O vereador Carlos Bolsonaro é aquele que mais utiliza a rede social e seus textos possuem muitos termos negativos. Destaca-se também a ligação das postagens à religião e o confronto com a mídia e oposição, PT e PSOL. Já o deputado Marcelo Freixo mudou seu comportamento após janeiro desse ano. Nota-se um aumento dos ataques ao PSL e também a outros políticos do Rio de Janeiro como Sérgio Cabral e Pezão. Como esperado, há muitas referências ao caso Marielle. O deputado Rodrigo Maia prefere usar o *Twitter* para compartilhar eventos de sua agenda ou projetos aprovados pela Câmara. Há destaque para lamentos a eventos trágicos, como alagamentos e o incêndio no Centro de Treinamento do Flamengo. Por fim, nota-se que foi possível discriminar os políticos a partir dos textos utilizados, aplicando o modelo de *topic modelling*.

Referências

Daniel Falbel (2019). ptstem: Stemming Algorithms for the Portuguese Language. R package version 0.0.4. https://CRAN.R-project.org/package=ptstem.

Grün B, Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models.” _Journal of Statistical Software_, *40*(13), 1-30. doi: 10.18637/jss.v040.i13 (URL: http://doi.org/10.18637/jss.v040.i13).

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

Ingo Feinerer and Kurt Hornik (2018). tm: Text Mining Package. R package version 0.7-6. https://CRAN.R-project.org/package=tm.

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

Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR.

Julio Trecenti and Fernando Correa (2014). abjutils: Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association. R package version 0.0.1. https://github.com/abjur/abjutils.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

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