Nota Técnica

AVALIAÇÃO DO CARNAVAL POR MEIO DAS MÍDIAS SOCIAIS

Brasília, junho de 2019

codeplan
COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL

Secretaria de Economia do Distrito Federal

Governo do Distrito Federal

GOVERNO DO DISTRITO FEDERAL
Ibaneis Rocha
Governador

Paco Brito
Vice-Governador

SECRETARIA DE ECONOMIA DO DISTRITO FEDERAL
André Clemente Lara de Oliveira
Secretário

COMPANHIA DE PLANEJAMENTO DO DISTRITO FEDERAL - CODEPLAN
Jeansley Lima
Presidente

Juliana Dias Guerra Nelson Ferreira Cruz
Diretora Administrativa e Financeira

Bruno de Oliveira Cruz
Diretor de Estudos e Pesquisas Socioeconômicas

Daienne Amaral Machado
Diretora de Estudos e Políticas Sociais

Erika Winge
Diretora de Estudos Urbanos e Ambientais

EQUIPE RESPONSÁVEL

Diretoria de Estudos e Pesquisas Socioeconômicas - DIEPS/Codeplan
Gerência de Estudos Regionais e Metropolitanos- GEREM/DIEPS/Codeplan

- Henrique de Mello de Assunção - Assistente

Revisão e copidesque
Heloísa Faria Herdy

Editoração Eletrônica
GEREM/DIEPS

RESUMO

O Carnaval é uma importante festa brasiliense com uma participação significativa do Governo do Distrito Federal em sua execução. Esta Nota Técnica faz uma avaliação do Carnaval no Distrito Federal por meio da análise de tweets públicos coletados no período de 2010 a 2019. Realizou-se uma avaliação textual dos tweets por meio de nuvens de palavras e avaliação de polaridade dos tweets usando a ferramenta *Amazon Comprehend*. Todas as análises foram feitas usando o programa R (R Core Team, 2019). Observa-se que houve uma redução do número de comentários neutros nos últimos anos sobre o Carnaval brasiliense e que existem muitos comentários relacionados a outras cidades quando se fala sobre o Carnaval do Distrito Federal. Alguns blocos brasilienses recebem mais comentários positivos do que o conjunto de blocos. De modo geral há mais tweets positivos do que negativos em relação ao Distrito Federal. Este trabalho mostra o potencial de pesquisas que utilizam redes sociais.

PALAVRAS-CHAVE: Carnaval; Twitter; Aprendizado de máquinas.

SUMÁRIO

1. INTRODUÇÃO

O Carnaval é uma festa indiscutivelmente importante no contexto brasileiro. Apesar de suas dimensões serem particulares às tradições de cada cidade que individualizam as festividades, é possível identificar contemporaneamente um traço comum que consiste em práticas que as caracterizam com grandes negócios, responsáveis por uma movimentação significativa e complexa da economia (MIGUEZ; LOIOLA, 2011). Milhões de brasileiros participam de diversos tipos de comemorações, como festas de ruas, escolas de samba, festas privadas e retiros religiosos. Essas comemorações geram conteúdo nas redes sociais que pode trazer percepções e informações relevantes à organização da festa pelo poder público, de forma a aprimorar as políticas públicas promovidas no período.

As redes sociais têm exercido forte influência sobre a política pública recente. Apesar de evidências mais marcantes serem do período eleitoral nacional e internacional, a finalidade dessa nota diz respeito ao uso do conteúdo presente nas redes sociais em termos de dados para avaliação do Carnaval no Distrito Federal. No contexto científico, por exemplo, a utilização de dados de redes sociais é utilizada em diversos estudos. Como exemplo, temos as propostas de medida do impacto social de artigos científicos baseados em tweets que são entendidas como complemento às métricas científicas tradicionais, tendo em vista que a análise dos tweets pode prever quais artigos científicos serão altamente citados com base nos três primeiros dias da sua publicação (EYSENBACH, 2011). Outras pesquisas também utilizaram dados de redes sociais para avaliações em saúde pública (REECE; DANFORTH, 2017), mercados de ações (BOLLEN; MAO; ZENG, 2011) e de feriados nacionais (HU, 2013).

Nesse sentido, a utilização das redes sociais como fonte de dados permite a coleta de grande quantidade de menções espontâneas de um segmento da população referente ao tema de pesquisa. As redes sociais são utilizadas por milhões, e, às vezes, por bilhões de pessoas no mundo que compartilham informações sobre diferentes tópicos. Este volume de dados possibilita diferentes análises por pesquisadores. Nesta pesquisa, por meio de mineração de dados das manifestações no Twitter referentes ao Carnaval do Distrito Federal, foi possível adquirir um amplo número de tweets e com eles realizar uma análise quanto à percepção do Carnaval do DF pelos usuários dessa rede social, exercício pioneiro no âmbito do Carnaval no Brasil.

Esta Nota Técnica busca fornecer informações sobre o Carnaval do Distrito Federal de modo a subsidiar o planejamento desse evento pelo governo distrital. Este trabalho integra um projeto piloto de quatro instrumentos para a análise do Carnaval no Distrito Federal. Além da presente nota, estão *156 do Carnaval*, uma pesquisa que utiliza a central telefônica do GDF (156), que investiga o perfil dos participantes dos blocos de Carnaval e sua percepção dos serviços públicos no decorrer da festa; *Entrevista com os blocos*, um levantamento nos blocos de Carnaval sobre a cadeia produtiva mobilizada na sua execução, o envolvimento da comunidade e suas maiores dificuldades; e *Análise da arrecadação*, uma análise do efeito do Carnaval sobre a arrecadação de ICMS e ISS do Distrito Federal.

2. Análise das redes sociais

A avaliação das redes sociais foi feita a partir de dados coletados no Twitter. Foram realizadas duas coletas: uma com um robô e outra com a API oficial e o programa R (R Core Team, 2019) conjuntamente com pacote *rtweet* (KEARNEY, 2018). As coletas feitas com o robô são referentes ao período de 2010 a 2019, enquanto as feitas com a API oficial são referentes a 11 de fevereiro até 7 de março de 2019.

O Twitter é um rede social em que os usuários se comunicam por tweets, pequenas mensagens de textos com um número limitado de caracteres (o limite era de 140 caracteres até novembro de 2017 e 280 após essa data); essas mensagens podem conter vídeos, imagens, links, entre outros conteúdos. O Twitter é uma grande rede social com 335 milhões de usuários ativos[1].

É possível coletar tweets do Twitter por meio de ferramentas pagas, da utilização direta da *Application Programming Interfaces* (API) do Twitter ou por *web scrapping*. Dois métodos de coletas de dados foram utilizados neste trabalho: um robô feito em linguagem Python[2] que coletou tweets públicos referentes ao Carnaval e à API do Twitter, acessada por meio do programa R (R Core Team, 2019) e do pacote *rtweet*, para obtenção de tweets que continham as hashtags disponibilizadas pela Secretaria de Estado de Cultura do Distrito Federal para acompanhamento. Todos os textos coletados foram transformados de modo a terem apenas letras minúsculas; essa alteração foi feita para o programa analisar corretamente os textos. Os tweets apresentados nesta Nota Técnica estão como escritos pelo autor, porém com todas as letras minúsculas.

2.1. Coleta por robô

O robô coletou tweets que contivessem os termos ”carnaval”e ”brasilia”de modo a obter tweets referentes ao Carnaval brasiliense. Esta regra de seleção permite obter tweets de interesse e reduz a coleta de tweets não relacionados ao tema. Uma desvantagem desta abordagem é a redução do número de tweets coletados. Usando esta regra, foram coletados 45.396 tweets, e esses tweets foram publicados entre primeiro de janeiro de 2010 até 3 de maio de 2019. A base foi reduzida para os dois meses com maior atividade em cada ano, para focar em mensagens referentes ao período de Carnaval. Com estes cortes, a base final possui 37.735 tweets; destes, 3.895 são referentes ao 2019. A quantidade de tweets coletados por ano está na Tabela 1.

Tabela 1 – Número de tweets por ano

Ano	Número de tweets
2010	2.386
2011	4.399
2012	4.640
2013	4.615
2014	3.020
2015	4.205
2016	3.537
2017	3.443
2018	3.595
2019	3.895

Fonte: Elaboração própria a partir da extração de dados do Twitter, período de 2010-2019

2.2. Análise textual

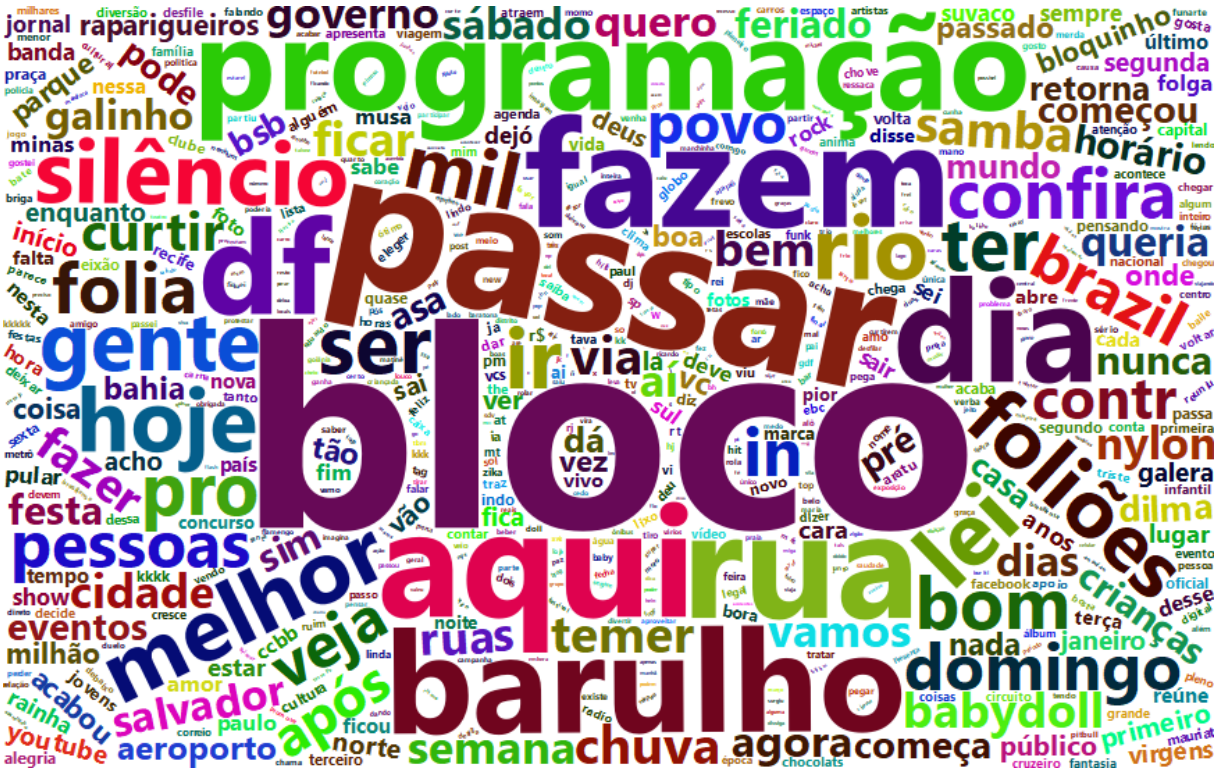
Realizou-se inicialmente uma análise das características textuais dos tweets por meio de nuvens de palavras. Foram removidas palavras que não adicionam significado na frase tais como ”que”, ”de”, ”a”, entre outras, além das palavras ”brasilia”e ”Carnaval”por estarem na restrição da busca. Os resultados podem ser visualizados por meio de duas nuvens de palavras: a primeira referente aos anos de 2016 a 2018 (Gráfico 1) e a segunda referente o ano de 2019 (Gráfico 2).

O principal assunto em ambos os anos são os referentes à festa em si. As palavras ”bloco”e ”blocos”são as palavras com maior destaque em ambos os anos. Outras referentes à festa em si e a demandas em relação

[1] https://en.wikipedia.org/wiki/List_of_virtual_communities_with_more_than_100_million_active_users - acessado em 29/03/2019

[2] O robô esta disponível em: https://github.com/Jefferson-Henrique/GetOldTweets-python

Gráfico 1 – Nuvem de palavras dos anos de 2016 a 2018



Fonte: Elaboração própria.

à festa, tais como ”silêncio”, ”barulho”, ”samba”, aparecem com alguma proeminência. Um ponto interessante é que a palavra ”chuva”varia de acordo com o ano, indicando que o efeito das chuvas no Carnaval varia entre os anos apurados. Há alguma conversa sobre política durante a festa nas duas nuvens de palavra, com palavras como ”Dilma”, ”Temer”, ”Lei”, aparecendo na nuvem de palavras, em 2019 e devido à alteração do quadro político as palavras são alteradas, como o termo ”bolsonaro”e ”presidente”ganhando mais destaque.

As palavras ”bloco”e ”blocos”são mencionadas recorrentemente nos tweets coletados. Esta característica pode ser observada nas nuvens de palavras já apresentadas. Uma característica interessante dos comentários é que a palavra ”bloco”tem maior proporção de comentários positivos do que a palavra ”blocos”em todos os anos em que estas aparecem com alguma proeminência, com no mínimo dez tweets diferentes usando as palavras, isto ocorreu nos anos de 2011, 2014, 2015, 2016, 2018 e 2019. Constata-se que existe um maior número de críticas ao conjunto de blocos enquanto as pessoas tendem a demonstrar seus sentimentos positivos a blocos específicos.

O Carnaval de Brasília é mencionado em relação a outras cidades, com menções a localidades como ”rio”, ”salvador”e ”bahia”. Existe uma comparação feita entre o Carnaval brasiliense e de outras regiões brasileiras, especialmente as que possuem o Carnaval mais consolidado como a Bahia e o Rio de Janeiro. Isso ocorre porque as redes sociais facilitam a comparação entre o Carnaval de diferentes regiões. Logo, é necessário levar em consideração a realidade de outros carnavais para a realização do Carnaval brasiliense, e em especial os mais famosos, como o do Rio e da Bahia.

2.3. Análise de sentimentos

A análise de sentimentos compreende o estudo do sentimento atribuído a uma frase, podendo ser feita por diferentes métodos. Nas últimas décadas, técnicas computacionais têm sido utilizadas para lidar com o volume crescente de dados que necessitam ser analisados. Entre essas técnicas existem utilização de léxicos, algoritmos de *machine learning* e *deep learning*. As diferentes abordagens observam padrões nos textos de modo a avaliar que sentimentos podem ser atribuídos ao texto.

A análise de sentimentos referentes aos tweets foi conduzida fazendo uso do serviço de avaliação de sentimentos providos pela Amazon[3]. A ferramenta classificou os tweets em neutro, positivo, negativo e

[3] *Amazon Compreheend* disponível em https://aws.amazon.com/pt/

Gráfico 2 – Nuvem de palavras do ano de 2019

Fonte: Elaboração própria.

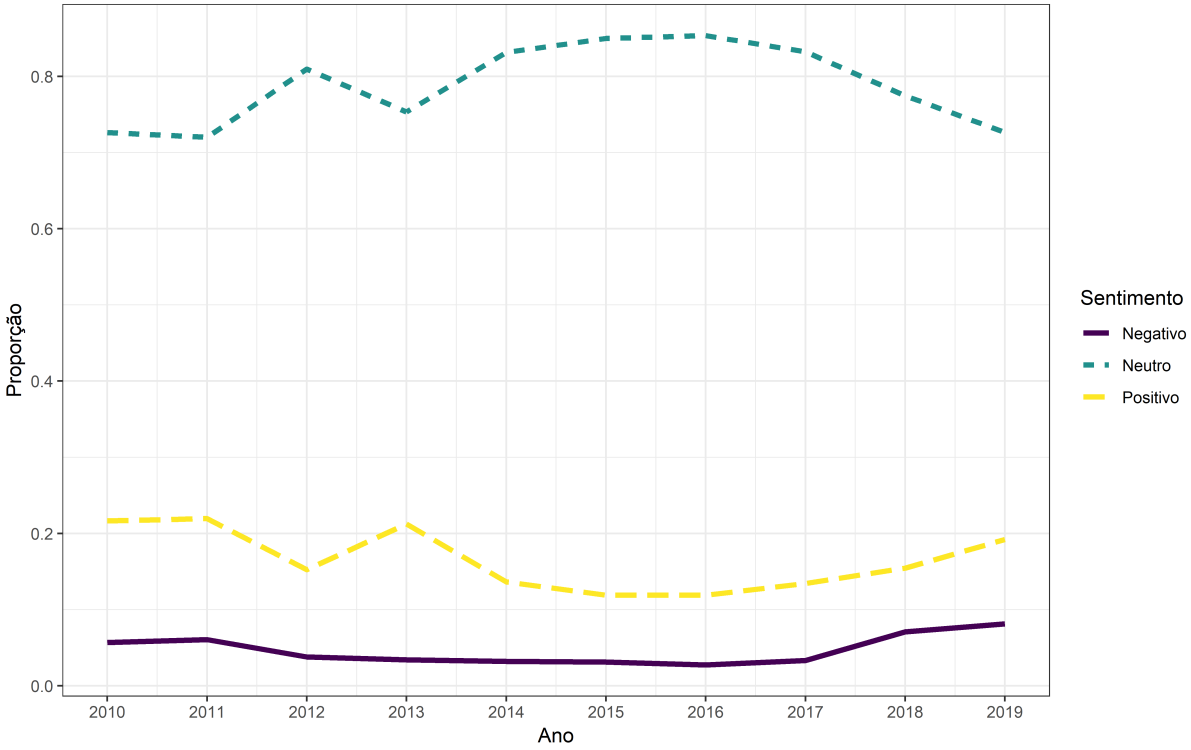
misto. Devido aos poucos tweets classificados como mistos, estes foram adicionados à categoria neutro. Então, trabalhou-se com três categorias: positivo, neutro e negativo. Como exemplos de tweets temos:

1. Tweets classificados como positivos:
   (a) ”Esse Carnaval de Brasília vai da bom”
   (b) ”primeira vez na vida que eu vejo espirito de Carnaval em Brasília, todo lugar tocando axé ou samba, mto good”
2. Tweets classificados com neutros:
   (a) ”Indo para a praça dos prazeres, Carnaval em Brasília [...]”
   (b) ”Carnaval? ta tendo sim! # brasilia # babydolldenylon # galinho babydoll de nylon [...]”
3. Tweets classificados como negativos:
   (a) ”Senão há dinheiro para educação, saúde e segurança, como pode haver dinheiro para os blocos de Carnaval? Pobre Brasília!”
   (b) ”não é que eu odeio bloco de Carnaval em Brasília! eu só odeio o lugar e as pessoas. apenas isso”

Os tweets referentes ao Carnaval foram avaliados principalmente como neutros (29.716, 78,7% dos tweets), seguidos de positivos (6.269, 16,6%) e negativos (1.750, 4,6%). Esta classificação depende do limiar utilizado pela ferramenta para classificação de textos como positivos ou negativos, um alto limiar mantém a maioria dos textos como neutros enquanto limiares mais baixos classificam maior proporção dos textos como negativos e positivos.

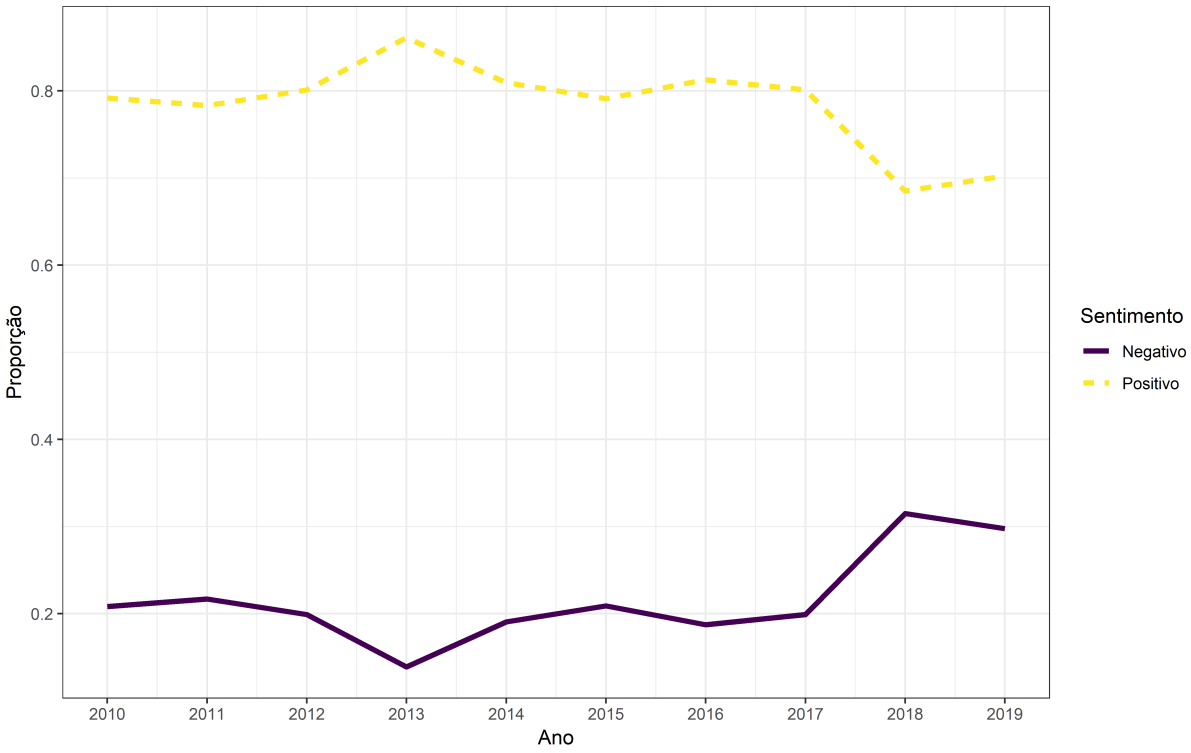
Houve um aumento da proporção dos comentários neutros até o ano de 2016 com uma queda dos comentários positivos e negativos no período. A partir do ano de 2017, observa-se uma mudança na tendência, um aumento da proporção de comentários não neutros. Esta redução é causada por um aumento da proporção de comentários positivos e negativos durante o período. A evolução da classificação dos comentários pode ser

Gráfico 3 – Gráfico de Análise de sentimento usando AWS



Fonte: Elaboração própria a partir da extração de dados do Twitter, período de 2010-2019.

Gráfico 4 – Gráfico de Análise de sentimentos positivos/negativos usando AWS

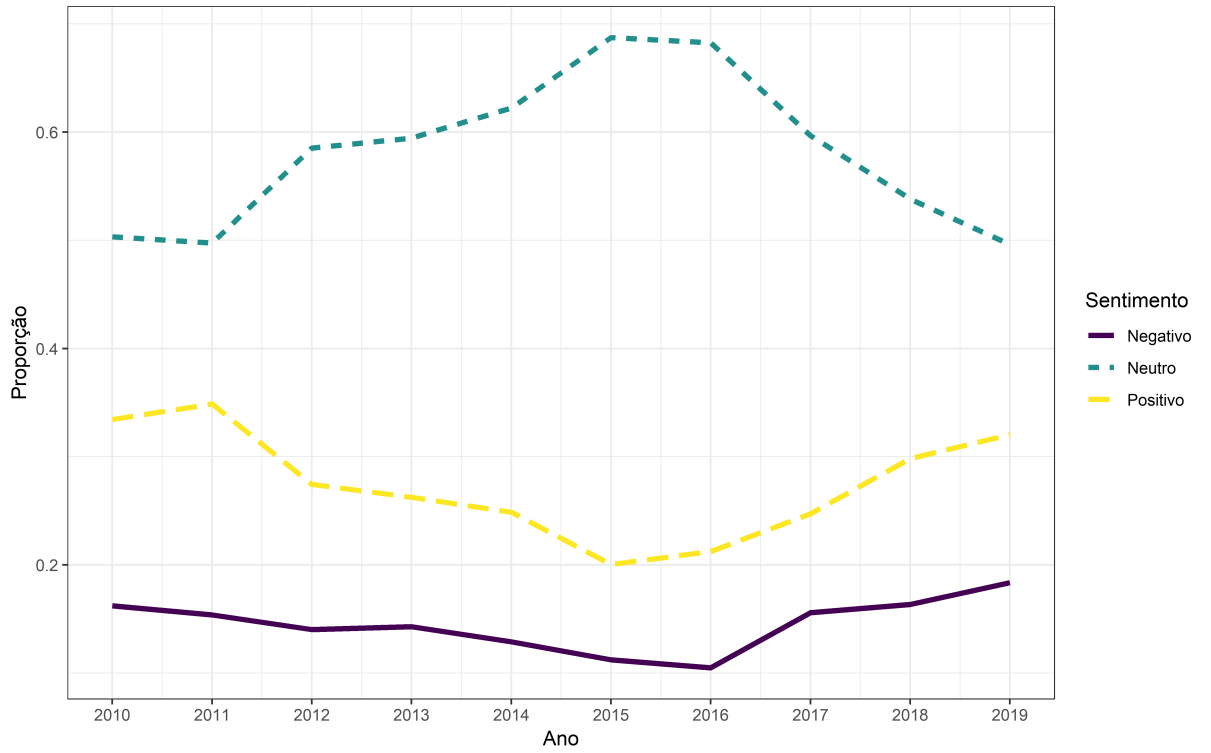


Fonte: Elaboração própria a partir da extração de dados do Twitter, período de 2010-2019.

observada no Gráfico 3. Considerando apenas os tweets não neutros, a proporção de comentários positivos e negativos segue uma tendência estável até o ano de 2017, com um choque positivo em 2013 e, em 2018, um choque negativo, com uma lenta recuperação em 2019, conforme mostra o Gráfico 4.

Foi realizada uma análise complementar utilizando o léxico OpLexicon V3.0 (SOUZA; VIEIRA, 2012), disponível no pacote (GONZAGA, 2017), para classificar os tweets seguindo uma metodologia *bag of words*. Nessa metodologia, considera-se que algumas palavras são negativas, tais como ”doença”, enquanto outras são positivas, como ”alegria”; quando uma frase tem maior número de palavras positivas do que negativas ela é considerada positiva, caso haja maior número de palavras negativas, a frase é considerada negativa; por fim, caso o número seja igual, a frase é considerada neutra. A evolução da polaridade dos tweets referentes ao Carnaval segundo esta metodologia pode ser vista no Gráfico 5, enquanto a evolução da proporção de tweets não neutros, no Gráfico 6.

Gráfico 5 – Gráfico de Análise de Sentimentos usando Léxico



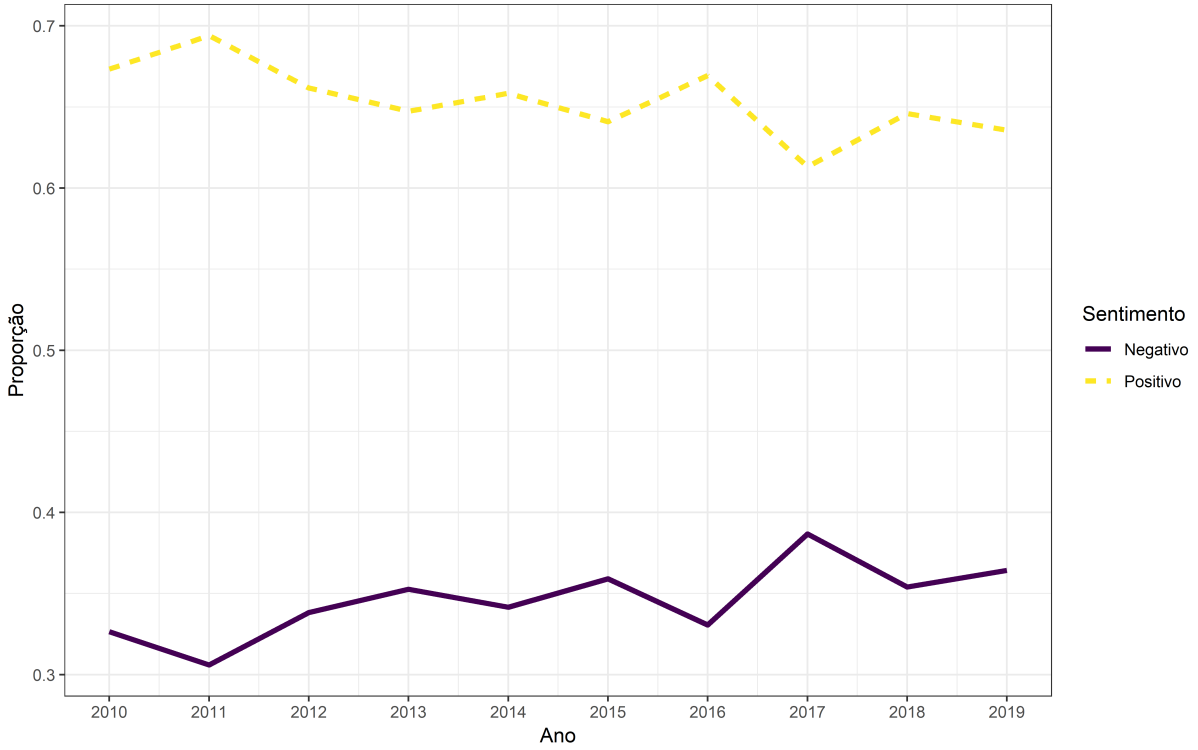
Fonte: Elaboração própria a partir da extração de dados do Twitter, período de 2010-2019.

Segundo esta metodologia, o crescimento da proporção de tweets neutros foi maior entre 2012 e 2016 e sua queda em 2017 e, em anos subsequentes, também foi maior. Registrou-se ainda um leve aumento da proporção de tweets negativos em relação aos positivos ao longo do tempo, sem a recuperação observada em 2019. De modo geral, ambas as metodologias concordam com o aumento recente da polarização enquanto houve um aumento do número de comentários neutros na metade da década, entretanto a avaliação usando léxicos aponta movimentações mais bruscas.

2.4. *Hashtags* passadas pela Secretária de Estado de Cultura do Distrito Federal

A Secretaria de Estado de Cultura disponibilizou uma série de *hashtags* que utilizaria durante o Carnaval para a avaliação do seu uso pela Companhia de Planejamento do Distrito Federal. Para isso, foram feitas três coletas de tweets com as *hashtags* passadas pela Secretaria usando a API oficial do Twitter. As coletas foram realizadas nos dias 20/02/2019, 27/02/2019 e 07/03/2019 utilizando o programa R e o pacote *rtweets*. No total, foram obtidos 208 tweets que continham as *hashtags* passadas pela Secretaria de Cultura. A maioria dos tweets foi feita por órgãos do governo; dos 208 tweets coletados, 108 partiram das contas no Twitter da Secretaria de Cultura, do Governo do Distrito Federal e do secretário de Cultura do Distrito Federal. Portanto, o uso dessas *hashtags* ficou restrito a ações governamentais. Não foi observado o uso das *hashtags* disponibilizadas pela Secretaria na base de tweets coletados por robô sobre o Carnaval de Brasília.

Gráfico 6 – Gráfico de Análise de Sentimentos positivos/negativos usando Léxico



Fonte: Elaboração própria a partir da extração de dados do Twitter, período de 2010-2019.

2.5. Facebook e Instagram

A coleta de dados do Facebook e do Instagram enriqueceriam a pesquisa, pois uma porcentagem significativa da população brasileira utiliza estas duas plataformas. Logo, a coleta de menções públicas dessas plataformas permitiria obter dados de manifestações espontâneas de uma parcela significativa da população brasileira a respeito do evento de interesse, gerando assim uma grande quantidade de dados sobre o Carnaval. Entretanto, não foi possível obter os dados do Facebook e do Instagram para esta pesquisa.

O Facebook, companhia que detém as plataformas Facebook (rede social) e Instagram, alterou suas políticas de mineração de dados por terceiros em 2018 devido ao escândalo da *Cambridge Analytic*. Portanto, as APIs de ambas as plataformas sofreram grandes alterações no ano de 2018. Sendo assim, não se encontrou no âmbito deste trabalho uma ferramenta que atendesse as necessidades e limitações da pesquisa e possibilitasse a obtenção de dados dessas plataformas.

3. Conclusão

O presente estudo é uma análise inovadora dos comentários do Twitter no âmbito do Carnaval e teve como objetivo obter informações sobre a sua caracterização no Distrito Federal. Para isto, foram observadas menções públicas obtidas do Twitter. Foram analisados os temas discutidos sobre o Carnaval durante o período do evento, bem como uma avaliação dos sentimentos expressos pelos usuários nestes tweets.

A avaliação da polaridade mostra que houve um aumento do percentual de comentários positivos e negativos em detrimento do número de comentários neutros em anos recentes. As duas metodologias utilizadas na avaliação de polaridade dos comentários corroboram com a tendência geral de comentários. De forma geral, tem havido uma piora do teor dos comentários, com queda dos comentários positivos e aumento dos negativos. O ano de 2019 é uma exceção à concordância dos dois dicionários, dado que o primeiro indica uma melhora com relação a 2018 e o segundo, uma piora relativa.

Os resultados também mostram um debate mais acalorado sobre o Carnaval de Brasília, com um menor número de pessoas neutras à festa. E que o aumento de comentários não neutros não necessariamente seja algo bom; é necessário observar se houve um aumento da proporção de comentários positivos, pois os resultados das duas metodologias usadas divergem quanto a este ponto.

Além disso, o Carnaval de Brasília encontra-se em um contexto nacional, em que é comparado com diferentes carnavais do Brasil, como os que ocorrem no Rio de Janeiro e em Salvador. As redes sociais permitem que comparações sejam feitas de modo muito mais rápido, portanto constata-se que o Carnaval brasiliense não ocorre em um vácuo mas em relação aos outros que ocorrem no Brasil. Sendo assim, o gerenciamento do Carnaval, bem como suas avaliações devem levar em conta como o Carnaval brasiliense compara-se com as outras festas brasileiras, em especial as mais estabelecidas: a do Rio de Janeiro e Salvador.

Por fim, vale dizer que os sentimentos de tweets referentes a blocos específicos são mais positivos comparados com os dos blocos em geral. Os usuários tendem a focar suas percepções positivas em um bloco específico enquanto as suas percepções negativas são voltadas a ”blocos”em geral. Isso limita o alcance de blocos de sucesso e blocos bem avaliados têm de influenciar a percepção do público aos blocos brasilienses em geral e ao Carnaval como um todo.

Uma limitação relevante da pesquisa foi ser restrita ao Twitter e a uma pequena base de dados. Não foi possível coletar informações de outras redes sociais com maior número de usuários brasilienses como WhatsApp, Instagram e Facebook. Ainda assim, entende-se que cada rede social fornece dados distintos, em decorrência do formato das suas manifestações, e o conteúdo presente no Twitter auxilia no entendimento do Carnaval brasiliense. Como a pesquisa está limitada a usuários do Twitter, não é possível ampliar as conclusões deste estudo à população do DF devido ao viés de seleção. Trata-se de um estudo com grande potencial de ampliação em pesquisas futuras, tendo em vista as informações e ferramentas emergentes a serem exploradas por meio das redes sociais.

REFERÊNCIAS

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of computational science**, Elsevier, v. 2, n. 1, p. 1–8, 2011.

EYSENBACH, G. Can tweets predict citations? metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. **Journal of Medical Internet Research**, JMIR Publications, v. 13, n. 4, 2011.

GONZAGA, S. **lexiconPT: Lexicons for Portuguese Text Analysis**. [S.l.], 2017. R package version 0.1.0. Disponível em: <https://CRAN.R-project.org/package=lexiconPT>.

HU, W. Real-time twitter sentiment toward thanksgiving and christmas holidays. **Social Networking**, Scientific Research Publishing, v. 2, n. 02, p. 77, 2013.

KEARNEY, M. W. **rtweet: Collecting Twitter Data**. [S.l.], 2018. R package version 0.6.7. Disponível em: <https://cran.r-project.org/package=rtweet>.

MIGUEZ, P.; LOIOLA, E. A economia do carnaval da bahia. **Bahia Análise e Dados**, UFBA, v. 21, n. 2, p. 285–299, 2011.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>.

REECE, A. G.; DANFORTH, C. M. Instagram photos reveal predictive markers of depression. **EPJ Data Science**, SpringerOpen, v. 6, n. 1, p. 15, 2017.

SOUZA, M.; VIEIRA, R. Sentiment analysis on twitter data for portuguese language. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2012. p. 241–247.

Companhia de Planejamento do Distrito Federal - Codeplan
Setor de Administração Municipal
SAM, Bloco H, Setores Complementares
Ed. Sede Codeplan
CEP: 70620-080 - Brasília-DF
Fone: (0xx61) 3342-2222
www.codeplan.df.gov.br
codeplan@codeplan.df.gov.br