

Análise exploratória de *tweets* do Governo de Santa Catarina utilizando Modelagem de Tópicos

Leonardo H. Rocha¹, Denio Duarte¹, Geomar A. Schreiner¹, Guilherme Dal Bianco¹

¹Universidade Federal da Fronteira Sul (UFFS)
Campus Chapecó
Chapecó – SC – Brazil

leoheiro@hotmail.com, {duarte, guilherme.dalbianco}@uffrs.edu.br

geomarschreiner@gmail.com

Abstract. *Topic modeling approaches have been widely used to discover latent topics from document collections. Twitter is one of the most used microblogs to spread news to people. Based on tweets extracted from the Santa Catarina official account from 2019 to 2021, we propose a set of experiments to discover the main subjects discussed in that period. As a result, we could identify the most important subjects that were discussed and presented to the Santa Catarina citizens.*

Resumo. *Modelagem de tópicos tem sido amplamente utilizada para descobrir tópicos e agrupar documentos de uma coleção de documentos de entrada. O Twitter é uma das plataformas mais utilizadas para divulgação de notícias, sendo uma fonte valiosa de informação para os cidadãos. Utilizando tweets extraídos da conta oficial do Governo de Santa Catarina e LDA para extração de tópicos, este trabalho visa identificar os assuntos recorrentes divulgados nos anos de 2019, 2020 e 2021. Como resultado, é possível identificar os assuntos mais importantes publicados pela conta oficial no período selecionado.*

1. Introdução

A *Web* tem se tornado uma fonte de informação muito usada nos últimos anos por indivíduos, especialmente redes sociais com acesso rápido a fontes de informações. Consequentemente, a sociedade é bombardeada com informações sobre diversos acontecimentos, sejam sobre segurança, saúde, educação, meio ambiente, entre outros assuntos do cotidiano. A quantidade massiva de conteúdo faz com que acontecimentos passem despercebidos diante de tanta informação.

Dentre as redes sociais, o *Twitter* proporciona acesso rápido a perfis de informação do governo, aproximando a população do que está acontecendo em seu Estado ou município. Este tipo de tarefa pode ser automatizado utilizando ferramentas que processam e ajudam a extrair informações de grandes volumes de dados. Entretanto, essa tarefa inclui vários desafios para extrair informações, incluindo erros de tipografia, ambiguidade do texto e tratamento caracteres especiais, entre outros. Como consequência, vários estudos têm sido realizados na extração de tópicos em documentos. Uma abordagem promissora para descobrir informações latentes em coleções de documentos é a modelagem de tópicos

[Blei et al. 2003], onde um tópicos representa um conjunto de palavras que descreve um assunto e os documentos são uma mistura de tópicos. Os tópicos são descobertos com base na co-ocorrência de palavras de conjunto de documentos.

Neste trabalho, é utilizada a abordagem de modelagem de tópicos *Latent Dirichlet Allocation* (LDA) [Blei 2012], para extrair tópicos mensais de um conjunto de *tweets* da conta oficial do governo de Santa Catarina no período de janeiro de 2019 até dezembro de 2021. O objetivo principal deste trabalho, é entender o comportamento dos assuntos das publicações no período selecionado da conta oficial do Governo de Santa Catarina e como os assuntos mudaram em decorrência da pandemia de *Covid-19* que atingiu o país em 2020.

Os *tweets* são divididos em grupos mensais e aplicados ao modelo LDA usando um número K de tópicos fixo em cinco e uma combinação de hiperparâmetros. Desta forma, o modelo é avaliado usando a métrica C_v , que de uma forma geral, calcula a qualidade dos tópicos gerados. Os resultados apontam que o uso do LDA para extrair tópicos a partir de uma coleção formada por *tweets* auxiliam a sumarizar como o Governo de Santa Catarina interage com os cidadãos por meio do Twitter.

2. Modelagem de Tópicos

A abordagem de modelagem de tópicos Latent Dirichlet Allocation (LDA) [Blei et al. 2003] é um dos modelos para modelagem probabilística de tópicos mais usados para extração de tópicos em documentos [Chehal et al. 2021]. Caracteriza-se por atribuir inicialmente probabilidades às palavras do dicionário encontradas na coleção. A distribuição é feita usando a família de distribuição discreta multivariada de Dirichlet. Assim, os tópicos são derivados de distribuições probabilísticas de palavras de uma coleção de documentos de entrada. Um conjunto de palavras que, pela relação de ordem, frequência e semântica, representam determinados assuntos (temas). Assim, por meio dessas relações, é possível definir um tema como um tópico, uma distribuição probabilística de palavras com frequência e semântica que faz sentido ao contexto do tópico.

A Tabela 1 apresenta um exemplo com três tópicos e suas cinco principais palavras ao lado das respectivas probabilidades de ocorrência no tópico (coluna $P(w)$). Observe que, como não há rótulos, o domínio deve ser definido a partir da semântica de cada tópico. Por exemplo, o primeiro tópico deve referir-se a pandemia no estado de Santa Catarina.

Tabela 1. Três tópicos com as top-5 palavras e suas probabilidades.

word	P(w)	word	P(w)	word	P(w)
covid	0.007	região	0.01	saber	0.007
hospital	0.005	obra	0.009	processo	0.005
caso	0.005	milhão	0.007	governosc	0.004
população	0.005	retomar	0.006	ação	0.004
realizar	0.005	saúde	0.006	pandemia_coronavirus	0.004

A modelagem de tópicos é baseada na ideia de que os documentos são uma mistura de tópicos, ou seja, os documentos exibem vários tópicos [Steyvers and Griffiths 2007]. Assim, documentos podem ser gerados a partir de diferentes distribuições de tópicos.

Um documento pode ser definido como uma sequência de palavras $\mathbf{w}=w_1, w_2, \dots, w_n$, onde n é o número de palavras em \mathbf{w} . Desse modo, uma coleção é um conjunto de m documentos $D=\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$. Documento pode ser qualquer conteúdo baseado em texto, por exemplo, um artigo ou comentário em uma rede social.

Em modelagem de tópicos, a maioria das abordagens considera um documento como um conjunto de palavras, ou seja, a ordem das palavras no documento não importa. O pré-processamento deve ser realizado sobre a coleta de documentos para prepará-la para a extração dos tópicos. O pré-processamento pode ser composto pelas seguintes etapas [Steyvers and Griffiths 2007]: (i) remoção de *stop words*, removendo palavras espúrias da coleção, (ii) *tokenização*, transformando a coleção em uma lista de palavras, (iii) *stemming*, que reduz as palavras à sua forma raiz, e (iv) lematizar, isto é, agrupar as formas flexionadas de uma palavra.

A Figura 1 representa o modelo LDA [Blei 2012] graficamente. Os retângulos representam iterações: o mais externo representa os documentos e o mais interno representa a escolha repetida de tópicos e palavras dentro de um documento. Além disso, assumindo a LDA como processo generativo, a Figura 1 pode ser explicada da seguinte forma:

1. Para cada documento w em um *corpus* D :
 - (a) Escolha $N \sim \text{Poisson}(\xi)$
 - (b) Escolha $\Theta \sim \text{Dir}(\alpha)$
 - (c) Para cada N nas palavras w_n :
 - i. Escolha um tópico $z_n \sim \text{Multinomial}(\Theta)$
 - ii. Escolha uma palavra w_n de $p(w_n | z_n\beta)$, probabilidade multinomial condicionada no tópico z_n

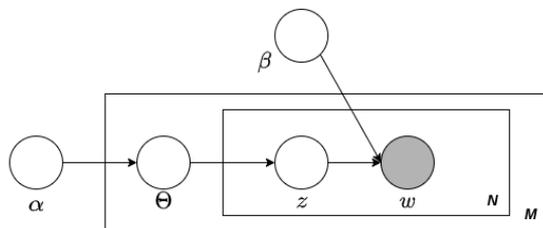


Figura 1. Representação gráfica do modelo LDA.

O hiperparâmetro β é a contagem de observação anteriores no número de vezes que as palavras aparecem antes de um tópico antes que qualquer outra palavra do corpus seja observada; um β mais alto significa que mais palavras estão associadas a um determinado tópico. O hiperparâmetro α desempenha o mesmo papel, mas em relação aos documentos. Note-se também que a LDA considera que os documentos podem apresentar vários tópicos por documento, por exemplo, sobre política, pode-se discutir economia e corrupção. No entanto, cada tópico associado a documentos tem uma probabilidade diferente. A soma da probabilidade de todos os tópicos associados a um determinado documento é igual a um.

Outra questão importante ao lidar com modelagem de tópicos é encontrar o número certo de tópicos para uma determinada coleção. Como qualquer método não supervisionado, temos que contar com uma métrica para verificar a melhor combinação dos hiperparâmetros e o número de tópicos [Duarte and Ståhl 2019].

Na modelagem de tópicos, avaliar modelos é um desafio, como em qualquer método não supervisionado, pois os conjuntos de dados não possuem rótulos para verificar a consistência dos resultados. A avaliação pode ser feita por humanos; no entanto, é uma tarefa onerosa. [Röder et al. 2015] apresentam um estudo comparando várias métricas de coerência para modelagem de tópicos, cujo objetivo foi descobrir qual métrica é a mais próxima da avaliação humana dos tópicos. No mesmo estudo, foi identificado que a métrica mais correlacionada com a percepção humana foi a c_v .

Neste trabalho, usamos c_v para experimentos para encontrar a melhor combinação de α , β e o número de tópicos (K).

3. Trabalhos Relacionados

Os autores em [Fukuyama and Wakabayashi 2018] propõe um método para extrair tópicos do *Twitter* para examinar se os tópicos extraídos correspondem aos eventos reais ocorridos no mesmo período. Os *tweets* extraídos correspondem a 12 dias em agosto de 2012. Foi utilizada a abordagem *Biterm Topic Modeling* (BTM) [Yan et al. 2013] para extrair cinco tópicos que foram comparados com eventos ocorridos. Através de análise manual, os autores identificaram que os tópicos extraídos correspondiam aos eventos do mesmo período. Da mesma forma, Pereira [Pereira 2019] utilizou BTM e o *tweets* para identificar os assuntos recorrentes durante as Olimpíadas Rio 2016. Os *tweets* foram separados por período e os tópicos extraídos confrontados com as notícias referentes aos jogos olímpicos. Os resultados confirmaram (com em [Fukuyama and Wakabayashi 2018]) que os eventos mais importantes (*e.g.*, conquista de medalhas e cerimônias) eram amplamente discutidos via *tweets*.

Um outro trabalho [Hidayatullah et al. 2018], utilizou LDA e uma coleção de documentos formada por *tweets* sobre cinco ligas de futebol: Indonésia, *Premier League* (Inglaterra), *Bundesliga* (Alemanha), *La Liga* (Espanha) e Serie A (Itália). Os contas oficiais @VIVAbola, @panditfootball, @detiksport, @Bolanet e @GOAL_ID foram utilizadas para extração e o intervalo entre o dia 01 de janeiro de 2017 a 24 de dezembro de 2017. Foram extraídos 10 tópicos e os mesmos foram sobre os pré-jogos da *Premier League*, rivalidade entre *Manchester United* e *Liverpool*, jogos da *La Liga*, sobre o clássico Real Madrid e Barcelona, Série A, Copa do Mundo, clubes da Europa e 3 tópicos sobre o campeonato da Indonésia. Os autores afirmam que os resultados foram satisfatórios, pois resumem bem as discussões sobre futebol nas cinco ligas.

O trabalho aqui apresentado se assemelha aos citados no uso de coleções de documentos extraídas do *Twitter* e aplicação de modelagem de tópicos para descobrir os assuntos mais relevantes. As maiores diferenças foi o uso de uma conta diferente e a aplicação do LDA em três anos de *tweets*, criando subcoleções mensais.

4. Experimentos

Para a realização das análises, foram coletados *tweets* da conta oficial do Governo de Santa Catarina (GovSC), de janeiro de 2019 a dezembro de 2021, totalizando 11.122 *tweets*. A coleção foi dividida em meses, totalizando 36 partes. A análise foi realizada utilizando a linguagem Python e suas bibliotecas de desenvolvimento, usando a implementação do modelo LDA da biblioteca *Tomotopy*.

4.1. Projeto

A etapa de pré-processamento foi dividida em três passos. No primeiro passo, foram retirados links, filtrando e removendo palavras que possuíam “www”, “http” e “https” em sua cadeia de caracteres e também foram retiradas menções a perfis removendo palavras que possuem “@”. No segundo passo, foram aplicados métodos da biblioteca do *gensim* para retirar caracteres especiais e palavras com menos de três caracteres e, além disso, foram removidas as *stop words*. No terceiro passo, o conjunto de notícias passou pelo processo de lematização, utilizando os recursos da biblioteca *NLPyPort*. E por último, no quarto passo, foram gerados bigramas utilizando os recursos da classe *Phrases* do *gensim*. Palavras que aparecem juntas com uma determinada frequência são concatenadas por `_`, por exemplo, as palavras “abastecimento” e “água” são juntadas formando o bigrama “abastecimento_água”, dado que ambas aparecem lado a lado em uma frequência considerável. A Tabela 2 mostra a quantidade de notícias e a média de palavras pós e pré pré-processamento, e na Tabela 3 é possível ver um exemplo do antes e depois do pré-processamento.

Ano	Quantidade de notícias	Média de palavras pré pré-processamento	Média de palavras pós pré-processamento
2019	2336	29.94	11.70
2020	4103	31.01	11.78
2021	4683	28.90	11.07

Tabela 2. Quantidade de notícias por ano e média de palavras pós e pré pré-processamento.

Original
Já no 1º dia será possível conferir exposição no Museu da Imagem e do Som de SC e no Espaço das Oficinas de Arte, frequentar a Biblioteca de Arte e Cultura, tomar um café no Bário e até mesmo assistir a um espetáculo de sapateado no Teatro Ademir Rosa. https://t.co/o5IBIhQFtH .
Pre-processado
dia conferir exposição museu_imagem som espaço oficina arte frequentar biblioteca arte_cultura tomar café bário assistir espetáculo sapateado teatro ademir rosa

Tabela 3. Comparação de pré-processamento do texto original e pré-processado.

Para escolher os melhores hiperparâmetros do modelo LDA, foram geradas combinações para cada fatia da coleção seguindo a documentação do modelo, com número fixo de tópicos em cinco (devido ao resultado de experimentos anteriores e a pouca quantidade de *tweets* mensais). A melhor combinação foi escolhida pelo melhor resultado da métrica C_v . A Tabela 4 apresenta os melhores hiperparâmetros por coleção de *tweets*, de janeiro de 2019 a dezembro de 2021.

Os assuntos discutidos mensalmente são agrupados por ano e são apresentados nas próximas seções. As Tabelas 5, 6 e 7 apresentam os tópicos, com suas *top-10* palavras, de maior probabilidade por mês e por ano. Perceba que devido ao formato dos *tweets* (textos curtos), existem várias palavras com pouca semântica e outras no formato *hashtag*, *e.g.*, receber, dia, verdadeiro, governosc e vivaaponte. A identificação dos assuntos que os tópicos extraídos tratavam foi feita observando os *tweets* com maior probabilidade de estarem associados aos tópicos.

Fatia	α	β	Fatia	α	β	Fatia	α	β
7/2019	0.25	0.25	11/2021 9/2019	0.1	0.2	11/2019 8/2019	0.15	0.25
1/2020 5/2019	0.25	0.2	9/2021	0.1	0.15	7/2020 4/2019	0.15	0.2
5/2021 12/2020	0.25	0.1	7/2021 10/2020 8/2020	0.1	0.1	6/2019	0.15	0.15
4/2021 2/2019	0.2	0.25	1/2022 3/2020	0.1	0.05	8/2021 1/2021 11/2020 3/2019	0.15	0.1
6/2020	0.2	0.2	3/2021 9/2020	0.1	0.05	12/2021 10/2019 1/2019	0.05	0.1
2/2020 12/2019	0.2	0.1	4/2020	0.05	0.2	5/2020	0.05	0.05
2/2021	0.2	0.05	10/2021	0.05	0.15	6/2021	0.1	0.25

Tabela 4. Melhores hiperparâmetros por coleção de tweets.

4.2. 2019

O ano de 2019 começou com *tweets* sobre a posse do novo governador, dicas para turismo, matrícula nas escolas estaduais que são assuntos típicos de início de ano. A Tabela 5 apresenta os tópicos, com suas top-10 palavras, de maior probabilidade por mês.

Mês	Tópicos (top-10 words)
Jan	feira, oficial, posse, confiro_previsão, governador_eleger, possegovsc, instante, santa_catarina, governador, passagem_veículo
Fev	receber, arroz, ano_letivo, escola, programa, estadual, acontecer, santa_catarina, empresa, gestor
Mar	hoje, ganhar, informação, escola, curso, acontecer, projeto, site, integração, segurança_pública
Abr	ação, parceria, doença, integrar, precisar, ajudar, lutar, casan, saber, resposta
Mai	governosc, bom, produção, santacatarina, mundo, família, planta_ornamental, saber, notícia_integra, rolou_semana
Jun	obra, região, educação, gestão, evento, plano, rede, recurso, estadual, santa_catarina
Jul	escola, educação, investimento, estadual, atleta, governosc, arte, competição, ensino, dia_julho
Ago	nota, natureza_ano, curricular_ciência, evolução_componente, temático_vida, gênero_unidade, termo_identidade, menção_única, base_nacional, base_estreito
Set	economia, gestão, milhão, governosc, santacatarina, trânsito, aeroporto, obra, processo, eficiência
Out	digital, evento, cidade, quase, edição_jogo, santacatarina, abrir, algum, catarinense, modalidade
Nov	santa_catarina, país, dia, pessoa, ano, processo_seletivo, dado_síntese, percentual, ibge, atuação
Dez	dia, luz, vivaaponte, ponte_hercílio, santa_catarina, histórico, hercílio_luz, reabertura_ponte, vi-vaaponte_santacatarina, gente

Tabela 5. Relação dos hot-topics de cada mês de 2019.

Os meses de janeiro, fevereiro março são marcados por notícias sobre o agro-negócio e combate a dengue. A posse no novo governador em janeiro foi amplamente discutida, bem como a pesca e a estiagem que afetava o estado. Como esperado, fevereiro e março também foram marcados com *tweets* sobre o carnaval. Em março, foi pedido cuidado com a febre amarela e incentivo à doação de sangue.

Abril e maio focam sobre a educação no estado, assuntos como eventos nas escolas, combate ao bullying são amplamente discutidos. Abril é marcado com *tweets* sobre a Páscoa e a chegada do frio. Maio também foca na violência contra a mulher e o agronegócio.

As postagens de junho, julho e agosto ainda focam nas ações para a educação e a chegada do inverno e do frio. Investimento em obras públicas e infraestruturas são divulgados. Também, inovação e tecnologia e o início da safra de tainha são apresentados. Em agosto, são divulgados a campanha de vacinação contra a gripe e conscientização sobre a violência contra a mulher.

Obras de infraestrutura, investimento em saúde, vacinação contra sarampo e o Jogos Abertos Paradesportivos (Parajasc) são discutidos em setembro e outubro. Eventos culturais, dicas de turismo e informações sobre vagas de empregos também são recorrentes nesses meses. As postagens dos últimos meses do ano focam na reabertura da Ponte Hercílio Luz, adesão do estado à CNH digital e festas de fim de ano.

4.3. 2020

O ano de 2020 continuou a tratar sobre a abertura da Ponte Hercílio Luz para pedestres, dicas para turismo, matrícula nas escolas estaduais entre outros e, obviamente, os meses seguintes focaram na pandemia. A Tabela 6 apresenta os tópicos, com suas top-10 palavras, de maior probabilidade por mês. Perceba que devido ao formato dos *tweets* (textos curtos), existem várias palavras com pouca semântica e outras no formato *hashtag*, e.g, receber, dia, verdadeiro e vivaaponte.

Mês	Tópicos (top-10 words)
Jan	florianópolis, pra, semana, vivaaponte, escola, luz, resultado, receber, passarela, conferir
Fev	serviço, pra, trilha, usina, leito, dia, inauguração_pavimento, ala_hospital, governador, ampliação
Mar	coronavírus_tomar, diariamente_medida, sério_necessário, informação_oficial, compartilhar_informação, verdadeiro, ligar_ajude, sintoma_atendimento, site_dúvida, mobilizado_combate
Abr	casa, acompanhar_via, ação_combater, responsável, poder_ficar, família, governosc, aovivo_atualização, pandemia_coronavíru, combate_coronavíru
Mai	município, leito_uti, receber, aovivo_ação, respirador, acompanhar_via, governosc, garantir, região, santa_catarina
Jun	região, obra, milhão, retomar, saúde, governosc, santacatarina, importante, saber, brasil
Jul	catarinense, governosc, região, nacional, investir, obra, milhão, saber, realizar, santacatarina
Ago	investimento, governosc, milhão, obra, região, investir, infraestrutura, acesso, educação, catarinense
Set	governosc, scontracovid_coronavíru, covid_santacatarina, covid, santacatarina_governosc, atendimento, continuar_acompanhamento, coronavírus_covid, acompanhamento_scontracovid, paciente_acompanhamento
Out	coronavírus, covid, continuar_acompanhamento, covid_santacatarina, governosc, região, santacatarina, scontracovid_coronavíru, considerar_recuperar, caso_confirmar
Nov	programa, estiagem, ação, santacatarina, rural, projeto, inovação, governosc, secretaria_agricultura, água
Dez	governosc, covid_santacatarina, governador, região, município, scontracovid_coronavíru, obra, estadual, santacatarina_governosc, acompanhamento_scontracovid

Tabela 6. Relação dos *hot-topics* de cada mês de 2020.

Março marca a divulgação do enfrentamento da pandemia, e.g., isolamento, primeira morte e distribuição de *kits*. A divulgação de investimento em infraestrutura

também está presente. Já em abril, assuntos relativos à pandemia dominam os *tweets*: fabricação de álcool em gel, aquisição de testes rápidos, autorização de serviços essenciais e uso de máscara são recorrentes.

Maio e junho também são marcados, como era de se esperar, por assuntos sobre a pandemia. Aumento de leitos de UTI, divulgação dos casos e compra de respirados são exemplos de assuntos discutidos. Outros assuntos como alerta sobre a estiagem no estado e obras de infraestrutura aparecem também, mas como menos frequência. Dois eventos isolados, mas de grande impacto foram também recorrentes em junho: o ciclone bomba que atingiu o Oeste do Estado e a ameaça do ataque dos gafanhotos vindo da Argentina.

Julho apresenta uma discussão mais recorrente que maio e junho sobre a pandemia. O inverno fez os casos aumentarem, assim, medidas restritivas, aumento de leitos de UTI voltam a ser discutidos. Agosto segue o mesmo fluxo. Porém, em julho, um ciclone extratropical que ocorreu na região litorânea e editais de fomento à pesquisa da Fapesc foram igualmente discutidos. Agosto também foi marcado pela onda de frio e investimento em infraestrutura.

Os meses seguintes, setembro, outubro, novembro e dezembro, apresentam menos *tweets* sobre a pandemia. Setembro divulga o provável retorno presencial às aulas, resumo das ações do enfrentamento da pandemia, focando também no problema de estiagem e editais de fomento da Fapesc. Outubro foca, como era de se esperar, na conscientização do câncer de mama (Outubro Rosa). A estiagem e a campanha Novembro Azul (câncer de próstata) também são assuntos em novembro. Dezembro já começa a discutir o plano de vacinação contra Covid, mas foca ainda no problema da estiagem e também o mega assalto a um banco ocorrido em Criciúma.

Como era de se esperar, o ano de 2020 foi marcado pela divulgação, principalmente, relativa à pandemia. Mesmo assim, o Governo do Estado de Santa Catarina divulgou e faz campanhas para assuntos importantes para os cidadãos.

4.4. 2021

O ano 2021 continuou sendo marcado pela pandemia causada pelo coronavírus. Sendo a vacinação o foco principal, porém o aumento de casos em fevereiro e março também foi amplamente divulgado. A Tabela 7 apresenta os tópicos mensais com maior probabilidade.

Os meses de janeiro, fevereiro e março foram marcados por postagens com ampla divulgação de chegada de lotes de vacinas, bem como o plano de vacinação. O Porto de Imbituba, a exportação de carne suína e as chuvas fortes na Grande Florianópolis também foram destaques em janeiro. Em fevereiro e março o foco foi no enfrentamento da segunda onda de Covid, com o aumento de leitos de UTI, os casos no Oeste Catarinense e volta do isolamento. O investimento em empresas para minimizar os efeitos da pandemia também foi divulgado. Em março, a vice-governadora assume, pois o governador estava em processo de *impeachment*.

Os meses de abril, maio e junho ainda apresentam uma preocupação com a segunda onda, mas existe uma melhora na matriz de risco. A vacinação é o assunto mais recorrente, mas a volta às aulas presenciais e atividades esportivas também são postados. No mês de maio houve uma quebra na safra de milho e em junho foram divulgados in-

Mês	Tópicos (top-10 words)
Jan	santa.catarina, vacina_covid, catarinense, dose, feira, começar, continuar, vacina, hoje, santacatarina
Fev	feira, covid, região, dose, saber, paciente, região_oeste, ser, governosc, receber
Mar	feira, governosc, scontracovid_coronavírus, vacinação, decreto, covid_santacatarina, covid, pessoa_acompanhamento, município, realizar
Abr	dose, covid, vacinação, vacina_covid, governosc, total_dose, feira, hoje, santacatarina, scontracovid_coronavírus
Mai	scontracovid_coronavírus vacinação, covid_santacatarina, governosc, dose, santacatarina, covid, município, chegar, vacina
Jun	saber, rural, projeto, desenvolvimento, investimento, governo, santacatarina, investir_milhão, gestão, governosc
Jul	vacinação, covid, coronavírus_covid, dose, santacatarina_governosc, vacina, acompanhamento_scontracovid, chegar, país, milhão
Ago	milhão, confira, governosc, região, programa, feira, obra, lançar, infraestrutura, saúde
Set	governador, ação, município, confira, anunciar, investimento, anunciar_investimento, dia, obra, laguna
Out	milhão, investimento, programa, governador, repasse, município, indaial, educação_especial, ato, educação
Nov	governador, oeste, anunciar, escola, município, cop, investimento, milhão, obra, construção
Dez	milhão, investimento, município, projeto, recurso, história, governosc, programa, governo, plano

Tabela 7. Relação dos *hot-topics* de cada mês de 2021.

vestimentos na agricultura e combate a estiagem. Junho também foi marcado por fortes chuvas no estado e a instalação de acesso à internet na zona rural.

Os meses de julho, agosto e setembro ainda focam no enfrentamento a segunda onda de Covid. Cuidados com higiene, uso de máscara, recursos destinados a hospitais são exemplos de postagens. A vacinação também é bem divulgada, além dos cuidados com a chegada do frio. O mês de setembro segue com divulgações semelhantes, mas foca também em investimentos em educação e infraestrutura.

Os últimos meses do ano continuam focando na campanha de vacinação contra a Covid. Outubro divulga a campanha Outubro Rosa e novembro a Novembro Azul. Agronegócio e investimentos na infraestrutura também fazem parte das postagens. Em todos os meses, os boletins da situação da pandemia também são divulgados. Dezembro também foca nas festas de fim de ano e na corrida realizada na ponte Hercílio Luz.

A extração de tópicos da conta oficial do Governo de Santa Catarina no *Twitter*, permitiu identificar os assuntos mais recorrentes discutidos no triênio 2019-2020-2021. Apesar de alguns tópicos não serem muito informativos em relação aos assuntos abordados, a associação dos tópicos com os *tweets* auxiliou no processo de identificação da semântica das dez palavras mais recorrentes dos tópicos. Por exemplo, mostrou que o governo enfatizou o enfrentamento à pandemia de forma sistemática e constante a partir de fevereiro de 2020. O arquivo em bit.ly/3xSZk7m apresenta todos os tópicos extraídos.

5. Conclusão

Este trabalho apresentou uma análise exploratória dos *tweets* postados pela conta oficial do Governo do Estado de Santa Catarina (@GovSC). A análise baseou-se na extração de

tópicos das coleções de documentos criadas a partir dos *tweets*. Foram extraídas postagens de 2019 a 2021. A coleção foi fatiada por períodos mensais, gerando 36 novas coleções. Cada coleção foi pré-processada e aplicada a abordagem LDA para extrair tópicos.

A análise mostrou que o governo enfatizou o enfrentamento à pandemia de forma sistemática e constante a partir de fevereiro de 2020. Também focou na divulgação de investimento públicos, educação, campanhas Outubro Rosa e Novembro Azul, além de atendimentos a catástrofes climáticas. O resultado mostrou que o uso de modelagem de tópicos e consulta aos *tweets* associados aos tópicos descobertos podem auxiliar o entendimento de como o governo catarinense divulga suas ações.

Como trabalhos futuros se pretende: entender a evoluções dos tópicos (*i.e.*, como as ações divulgadas pelo governo catarinense se alteram no decorrer do tempo) e considerar os *links* presentes nos *tweets* para enriquecer as coleções geradas.

Agradecimentos: Leonardo H. Rocha é parcialmente financiado pela Universidade Federal da Fronteira Sul. Projeto PES-2021-0458.

Referências

- Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Chehal, D., Gupta, P., and Gulati, P. (2021). Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations. *Journal of Ambient Intelligence and Humanized Computing*, 12(5):5055–5070.
- Duarte, D. and Ståhl, N. (2019). Machine learning: a concise overview. In Said, A. and Torra, V., editors, *Data Science in Practice*, pages 27–58. Springer.
- Fukuyama, S. and Wakabayashi, K. (2018). Extracting time series variation of topic popularity in microblogs. In *Proceedings of iiWAS*, pages 365–369, New York, New York, USA. ACM.
- Hidayatullah, A. F., Pembrani, E. C., Kurniawan, W., Akbar, G., and Pranata, R. (2018). Twitter topic modeling on football news. In *2018 3rd ICCCS*, pages 467–471. IEEE.
- Pereira, M. (2019). Análise exploratória de tweets utilizando modelagem de tópicos para textos curtos: caso olimpíadas rio 2016. <https://rd.uffrs.edu.br/handle/prefix/3371>. Monografia, UFFS, Chapecó, Brasil.
- Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth WSDM*, pages 399–408, USA. ACM.
- Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In Landauer, T. K., McNamara, D. S., Dennis, S., and Kintsch, W., editors, *Handbook of latent semantic analysis*, chapter 21, pages 424–440. Laurence Erlbaum Associates.
- Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd WWW*. ACM.