UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

EDUARDO YOSHIO DA ROCHA

ANÁLISE DA EVOLUÇÃO DO DISCURSO COM USO DE TÉCNICAS DE MODELAGEM DE TÓPICOS

CURITIBA

2022

EDUARDO YOSHIO DA ROCHA

ANÁLISE DA EVOLUÇÃO DO DISCURSO COM USO DE TÉCNICAS DE MODELAGEM DE TÓPICOS

Analysis of the evolution of discourse using topic modeling techniques

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Luiz Celso Gomes Junior

CURITIBA

2022

EDUARDO YOSHIO DA ROCHA

ANÁLISE DA EVOLUÇÃO DO DISCURSO COM USO DE TÉCNICAS DE MODELAGEM DE TÓPICOS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 24/junho/2022

Luiz Celso Gomes Junior
Doutorado
Universidade Tecnológica Federal do Paraná

Bogdan Tomoyuki Nassu
Doutorado
Universidade Tecnológica Federal do Paraná

André Santanchè
Doutorado
Universidade Estadual de Campinas

CURITIBA
2022

Dedico este trabalho à minha família, amigos, e
a todos que buscam mudar o mundo e ajudar
as pessoas através da tecnologia.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus, por toda a sabedoria e benção a mim concedidas e por sempre estar presente em minha vida e tudo o que faço.

Agradeço também a minha família, pois sem o suporte deles seria muito difícil chegar até aqui e concluir com êxito este trabalho.

Agradeço ao meu orientador Prof. Dr. Luiz Celso Gomes Junior, pela sabedoria transmitida a mim e por todo apoio e assistência concedidos a mim durante todo o trabalho, sempre me incentivando a fazer o meu melhor.

Aos meus colegas de turma e amigos que fiz durante esta trajetória. Sem eles seria muito mais difícil chegar até aqui.

A todo o pessoal do curso e da universidade, professores, coordenadores, incentivadores. A todos aqueles que fizeram parte disso de alguma forma.

RESUMO

Com o crescente uso das redes sociais, cresce também o impacto que o conteúdo nessas redes tem na sociedade atual, influenciando bilhões de pessoas no mundo todo. Uma vez que nesse mundo virtual a diversidade de assuntos discutidos e ideias expressas é muito grande, a análise do conteúdo disposto nas redes sociais torna-se de grande importância para o entendimento da influência que essas redes têm no comportamento da sociedade como um todo. Assim, o presente trabalho tem o objetivo de analisar a evolução dos discursos em postagens nas redes sociais com o intuito de identificar padrões que permitam avaliar mudanças no discurso, eventos incomuns ou outros fatores relevantes nos seus conteúdos. No cenário atual da pandemia gerada pelo vírus Sars-cov-2, o estudo dos discursos de prefeituras municipais brasileiras em postagens em redes sociais, mais precisamente no *Facebook*, torna-se relevante para a mencionada análise do comportamento da população durante a pandemia. Este trabalho procurou avaliar a aplicação de duas técnicas de aprendizado de máquina para a análise da evolução do discurso: (i) LDA (*Latent Dirichlet Allocation* ou Alocação Latente de Dirichlet) com discretização em intervalos de tempo, e (ii) TOT (*Topics Over Time* ou Tópicos ao Longo do Tempo), sem discretização temporal. Os modelos foram comparados quantitativamente e qualitativamente. A principal contribuição deste trabalho é a comparação das vantagens e desvantagens de cada modelo. Não se pôde obter um resultado claro de qual dos dois é o melhor modelo, pois apesar de o modelo do LDA com discretização representar melhor a evolução por tópicos, o modelo do TOT apresenta métricas melhores na regressão linear que foi feita como parte da análise quantitativa.

Palavras-chave: modelagem de tópicos; análise de discurso; covid-19; lda; discreto; tot.

ABSTRACT

With the growing use of social networks, the impact that content on these networks has on today's society also grows, influencing billions of people around the world. Since in this virtual world the diversity of subjects discussed and ideas expressed is very large, the analysis of the content displayed on social networks becomes of great importance for understanding the influence that these networks have on the behavior of society as a whole. Thus, the present work aims to analyze the evolution of discourses in posts on social networks in order to identify patterns that allow us to assess changes in discourse, unusual events or other relevant factors in their content. In the current scenario of the pandemic generated by the Sars-cov-2 virus, the study of the speeches of Brazilian municipal governments in posts on social networks, more precisely on Facebook, becomes relevant for the aforementioned analysis of the behavior of the population during the pandemic. This work sought to evaluate the application of two machine learning techniques for the analysis of speech evolution: (i) LDA (Latent Dirichlet Allocation) with temporal discretization in time intervals, and (ii)) TOT (Topics Over Time), without temporal discretization. The models are compared quantitatively and qualitatively. The main contribution of this work is the comparison of the advantages and disadvantages of each model. It was not possible to obtain a clear result of which of the two is the best model, because although the LDA model with discretization better represents the evolution by topics, the TOT model presents better metrics in the linear regression that was performed as part of the quantitative analysis.

Keywords: topic modeling; discouse analysis; covid-19; lda; discrete; tot.

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Vivemos atualmente no período que pode ser definido como a Terceira Revolução Industrial ou Revolução Digital, no qual é crescente o número de pessoas que estão conectadas à internet, e consequentemente às redes sociais. Estima-se que mais de 4,2 bilhões de pessoas utilizam redes sociais pelo mundo, o que representa 53,6% da população mundial (MERCANTIL, 2021). No Brasil, são mais de 150 milhões de usuários de redes sociais, e a taxa de usuários pelo total de habitantes é de 70,3%, uma das maiores dentre todos os países do mundo. É inegável portanto o impacto que as redes sociais têm no cotidiano das pessoas e a influência que o conteúdo dessas redes tem no comportamento da sociedade como um todo. Por esse motivo, cresce a importância de analisar o que é publicado e compartilhado nessas redes sociais, pois postagens que são feitas influenciam na construção social e cultural de cada indivíduo e nas atitudes tomadas por cada usuário consumidor do conteúdo proveniente dessas redes.

Dessa forma, o presente trabalho buscou realizar uma análise da evolução dos discursos presentes nas redes sociais, os quais são caracterizados pelos conteúdos contidos nas postagens realizadas nessas redes por indivíduos ou entidades. O objetivo foi de identificar padrões que permitiram avaliar mudanças nesses discursos, eventos incomuns ou outros acontecimentos relevantes, como eventos importantes da pandemia que influenciaram nos discursos, ou assuntos discutidos que influenciaram no andamento da pandemia. Para isso, foram aplicadas duas técnicas de aprendizado de máquina para a análise da evolução do discurso. A primeira técnica é baseada no modelo do LDA (*Latent Dirichlet Allocation*), um modelo estatístico e generativo com discretização temporal em intervalos de tempo. A segunda técnica baseia-se no modelo do TOT (*Topics Over Time* ou Tópicos ao Longo do Tempo), o qual é também um modelo estatístico e generativo, porém sem discretização do tempo.

Os modelos foram comparados qualitativamente e quantitativamente, sendo a principal contribuição do trabalho a avaliação das vantagens e desvantagens de cada modelo a partir da análise feita. O modelo do LDA mostrou-se melhor ao representar a evolução dos tópicos temporalmente, enquanto que o modelo do TOT apresentou resultados quantitativamente melhores a partir da regressão linear feita, com métricas superiores as do LDA. Assim, não foi possível afirmar qual dos dois modelos é o melhor.

O estudo foi feito com base nos dados referente à pandemia de Covid-19 em municípios do Brasil, avaliando a relação entre os discursos e o andamento da pandemia ao longo do tempo. Os modelos avaliados podem também ser empregados em outros temas de interesse além da pandemia de Covid-19, podendo servir como referência para trabalhos relacionados e fomentando o conhecimento nessa área de pesquisa.

O restante deste trabalho está organizado da seguinte forma: O capítulo 2 abrange a fundamentação teórica em torno da análise do discurso, da modelagem de tópicos e dos modelos estudados; O capítulo 3 apresenta a metologia empregada no trabalho desde a obtenção dos dados, passando pela interpretação dos tópicos e finalizando com as avaliações quantitativas e

qualitativas dos modelos; O capítulo 4 discorre sobre as implementações feitas e os resultados alcançados, descrevendo todo o processo de modelagem dos tópicos, representação gráfica dos mesmos e análise dos modelos implementados; Por último, o capítulo 5 contém as considerações finais em relação aos modelos avaliados, resultados obtidos e hipóteses levantadas, assim como sobre o que poderia ser feito para a continuação do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Análise do discurso

O discurso é uma parte fundamental da vida de todas as pessoas. Em nosso cotidiano sempre há algum tipo de comunicação, seja através de conversas, telefonemas, mensagens, cartas ou instruções. Aprender a falar e se comunicar é fundamental para se relacionar com outras pessoas, compreender diferentes culturas e também para entender os outros e a si mesmo (HARDY; BRYMAN, 2004, 2009). Nossos locais de trabalho, estudo, ou quaisquer outras áreas estão imbuídas de conversas e textos, sejam eles parte do próprio lugar ou de todas as características auxiliares de vida que os cercam. Por conta disso, a análise do discurso torna-se uma importante tarefa para as ciências sociais.

A análise do discurso é entendida de diversas maneiras no campo das ciências sociais. Um dos motivos é que as abordagens analíticas foram feitas baseadas em diferentes áreas do conhecimento como a sociologia, filosofia, psicologia, psicologia social, comunicação, literatura, estudos culturais e etc. Dessa forma, os métodos de pesquisa nessas diversas áreas são naturalmente diferentes, embora existam fatores comuns entre esses métodos. Em alguns casos a análise do discurso pode ser feita abordando temas como fala, narrativa, conversação e assim por diante. Em outros, o discurso pode ser tratado simplesmente como uma palavra para a linguagem em uso ou como um objeto linguístico que pode ser contado e descrito. A análise do discurso é portanto o estudo da vida social, entendida por meio da análise da linguagem em seu sentido mais amplo (incluindo conversa face a face, interação não verbal, imagens, símbolos e documentos), de forma a investigar o significado, seja na conversação ou na cultura (SHAW; BAILEY, 2009). Neste trabalho, somente os tópicos obtidos a partir do discurso e suas evoluções ao longo do tempo são considerados. Assim, a análise do discurso foi feita por meio da modelagem de tópicos ao longo do tempo.

2.2 Modelagem de tópicos

A modelagem de tópicos é uma técnica de aprendizado de máquina não supervisionado capaz de detectar padrões de palavras e frases dentro de um conjunto de documentos e agrupar automaticamente grupos de palavras e expressões semelhantes que melhor caracterizam estes documentos (PASCUAL, 2019). Diversas são as aplicações em que a modelagem de tópicos pode estar presente. Um exemplo é o de um cientista que precisa encontrar artigos acadêmicos em meio a grandes coleções de literatura acadêmica. O uso eficaz dessas coleções requer uma interação de uma forma mais estruturada: encontrando artigos semelhantes aos de seu interesse e explorando a coleção por meio dos tópicos subjacentes que a percorrem (BLEI; LAFFERTY, 2009). O problema central é que os assuntos dos artigos e os artigos relacionados

nem sempre estão prontamente disponíveis nessas coleções e, devido à quantidade de artigos, realizar a relação entre eles de forma manual é altamente ineficiente e demorado. Se mostram pertinentes portanto, métodos que automatizem as tarefas de organização, gerenciamento e entrega dos conteúdos presentes nos artigos.

Ao descobrir padrões de uso de palavras e conectar documentos que exibem padrões semelhantes, os modelos de tópicos demonstram ser uma técnica poderosa para encontrar estruturas úteis em uma coleção de outra forma não estruturada.

2.3 LDA

O LDA, *Latent Dirichlet Allocation* (Alocação Latente de Dirichlet), é um modelo probabilístico generativo de modelagem de tópicos em um corpus. LDA é amplamente utilizado em diversas áreas como biologia, genética (PRITCHARD; STEPHENS; DONNELLY, 2000), medicina, visão computacional (WANG XIAOGANG; GRIMSON, 2007), aprendizado de máquina e etc. A ideia básica é que os documentos são representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras (BLEI; NG; JORDAN, 2003). Neste modelo, são investigadas as correlações entre palavras presentes em coleções de documentos, analisando a similaridade destas palavras, gerando então tópicos que sintetizam os significados das palavras que são correlatas. Dessa forma, é possível agrupar palavras que possuem semânticas em torno de um determinado assunto, produzindo um tópico que representa o assunto desse grupo de palavras. O LDA é um modelo probabilístico pois, em cada tópico gerado, é possível obter a probabilidade de cada palavra que está contida nos documentos representar o tópico gerado.

Na estrutura da Figura 1, é possível observar um exemplo no qual são apresentados 5 documentos com suas respectivas listas de palavras.

Figura 1 – Documentos como uma coleção de palavras

```
Doc1: word1, word3, word5, word45, word11, word 62, word88 ...
Doc2: word9, word77, word31, word58, word83, word 92, word49 ...
Doc3: word44, word18, word52, word36, word64, word 11, word20 ...
Doc4: word85, word62, word19, word4, word30, word 94, word67 ...
Doc5: word19, word53, word74, word79, word45, word 39, word54 ...
```

Fonte: Towards Data Science, 2019 (KULSHRESTHA, 2019).

Com o LDA, é possível obter então tópicos que mais aparecem nos documentos, juntamente com a probabilidade de cada palavra pertencer aos tópicos gerados, como é mostrado na Figura 2. Nela, é possível observar os tópicos que foram gerados (Topic1, Topic2 e Topic3), e a probabilidade de cada palava (Word1, Word2, Word3, Word4...) de pertencer a esses tópicos. Quanto maior a probabilidade, mais significativa é a palavra para aquele tópico.

Figura 2 – Tópicos com as probabilidades de cada palavra

	Word1	word2	word3	word4	.....
Topic1	0.01	0.23	0.19	0.03	
Topic2	0.21	0.07	0.48	0.02	
Topic3	0.53	0.01	0.17	0.04	

Fonte: Towards Data Science, 2019 (KULSHRESTHA, 2019).

No LDA porém, não são levados em consideração o papel gramatical das palavras ou a ordem delas, sendo considerado cada documento apenas como uma BOW (*bag of words* ou "saco de palavras" numa tradução literal). Os tópicos gerados a partir do LDA não apresentam de forma muito evidente qual o assunto relacionado a eles, cabendo à pessoa que aplica o modelo interpretar os resultados obtidos.

2.3.1 O algoritmo

O LDA faz duas suposições principais: documentos são uma mistura de tópicos e tópicos são uma mistura de palavras. Esses tópicos, utilizando a distribuição de probabilidade, geram as palavras. Em linguagem estatística, os documentos são conhecidos como densidade de probabilidade (ou distribuição) de tópicos e os tópicos são a densidade de probabilidade (ou distribuição) de palavras (SETH, 2021).

Qualquer corpus (coleção de documentos) pode ser representado por uma Matriz de Termos do Documento (ou DTM, *Document Term Matrix*), conforme mostra o exemplo da Figura 3. Neste exemplo, as linhas são os documentos e as colunas são as palavras (ou termos).

Figura 3 – Matriz de Termos do Documento (5 documentos e 8 palavras)

	W1	W2	W3	W4	W5	W6	W7	W8
D1	0	1	1	0	1	1	0	1
D2	1	1	1	1	0	1	1	0
D3	1	0	0	0	1	0	0	1
D4	1	1	0	1	0	0	1	0
D5	0	1	0	1	0	0	1	0

Fonte: Analytics Vidhya, 2021 (SETH, 2021).

O LDA converte a DTM em duas outras matrizes: a Matriz dos Tópicos dos Documentos (exemplo da Figura 4) e a Matriz dos Termos dos Tópicos (exemplo da Figura 5). A Matriz dos Tópicos dos Documentos do exemplo contém os possíveis tópicos (representados por K) que os documentos podem ter. Neste exemplo são 6 tópicos e 5 documentos. Já a Matriz dos Termos dos Tópicos contém as palavras (termos) que os tópicos podem ter, tendo no exemplo 8 palavras únicas e 5 tópicos.

Figura 4 – Exemplo de Matriz dos Tópicos dos Documentos (5 documentos e 6 tópicos)

	K1	K2	K3	K4	K5	K6
D1	1	0	0	0	0	0
D2	0	1	0	0	1	1
D3	1	1	0	0	0	0
D4	1	0	0	1	0	1
D5	0	0	1	1	0	0

Fonte: Analytics Vidhya, 2021 (SETH, 2021).

Figura 5 – Exemplo de Matriz dos Termos dos Tópicos (6 tópicos e 8 palavras)

	W1	W2	W3	W4	W5	W6	W7	W8
K1	0	1	1	0	1	0	1	0
K2	1	1	1	1	0	1	1	1
K3	1	0	0	0	0	1	0	0
K4	1	1	0	1	1	0	0	1
K5	0	0	1	1	0	1	1	1
K6	1	0	1	1	1	0	0	1

Fonte: Analytics Vidhya, 2021 (SETH, 2021).

O objetivo final do LDA é encontrar a melhor representação dessas duas matrizes para então encontrar as distribuições documento-tópico e tópico-palavra mais otimizadas. Como o LDA assume que os documentos são uma mistura de tópicos e os tópicos são uma mistura de palavras, o LDA então retrocede ao nível dos documentos para identificar quais tópicos teriam gerado esses documentos e quais palavras teriam gerado esses tópicos.

A modelagem do LDA é um processo iterativo. Na primeira iteração (também chamada de época), os tópicos são atribuídos aleatoriamente a cada palavra de cada documento. O resultado é uma composição dos documentos com os tópicos e dos tópicos com as palavras. Após a primeira iteração, o treinamento continua buscando otimizar as matrizes obtidas iterando sobre todos os documentos e todas as palavras.

O LDA faz uma outra suposição de que todos os tópicos que foram atribuídos estão corretos, exceto a palavra atual. Assim, com base nessas atribuições de palavras-tópicos já corretas, o LDA tenta corrigir e ajustar a atribuição de tópicos da palavra atual com uma nova atribuição. Então, a cada nova iteração são calculadas duas probabilidades:

- P1: proporção de palavras em um documento D que estão atualmente atribuídas a um tópico K;
- P2: proporção de atribuições a um tópico K sobre todos os documentos que vêm desta palavra (na iteração atual). Em outras palavras, é a proporção dos documentos em que a palavra também é atribuída ao tópico.

Com as probabilidades obtidas, é estimada uma nova probabilidade, a qual é o produto de P1 por P2. Através desta nova probabilidade, um novo tópico é identificado, o qual é o tópico mais relevante para a palavra atual. O algoritmo continua até que um estado estacionário seja obtido ou que se atinja o número máximo de iterações (épocas) definido. O ponto de convergência do LDA é o ponto no qual é fornecida a representação mais otimizada da Matriz dos Tópicos dos Documentos e da Matriz dos Termos dos Tópicos.

2.4 LDA temporal

LDA é uma técnica muito poderosa para a análise qualitativa de grandes corpora por causa de seus tópicos altamente interpretáveis. No entanto, o LDA ignora o aspecto temporal presente em muitas coleções de documentos (LEFEBURE, 2018). Estendendo o modelo do LDA, existem modelos que permitem a modelagem de tópicos não somente a partir das palavras contidas nos documentos, como também levando em consideração o fator temporal dos discursos analisados.

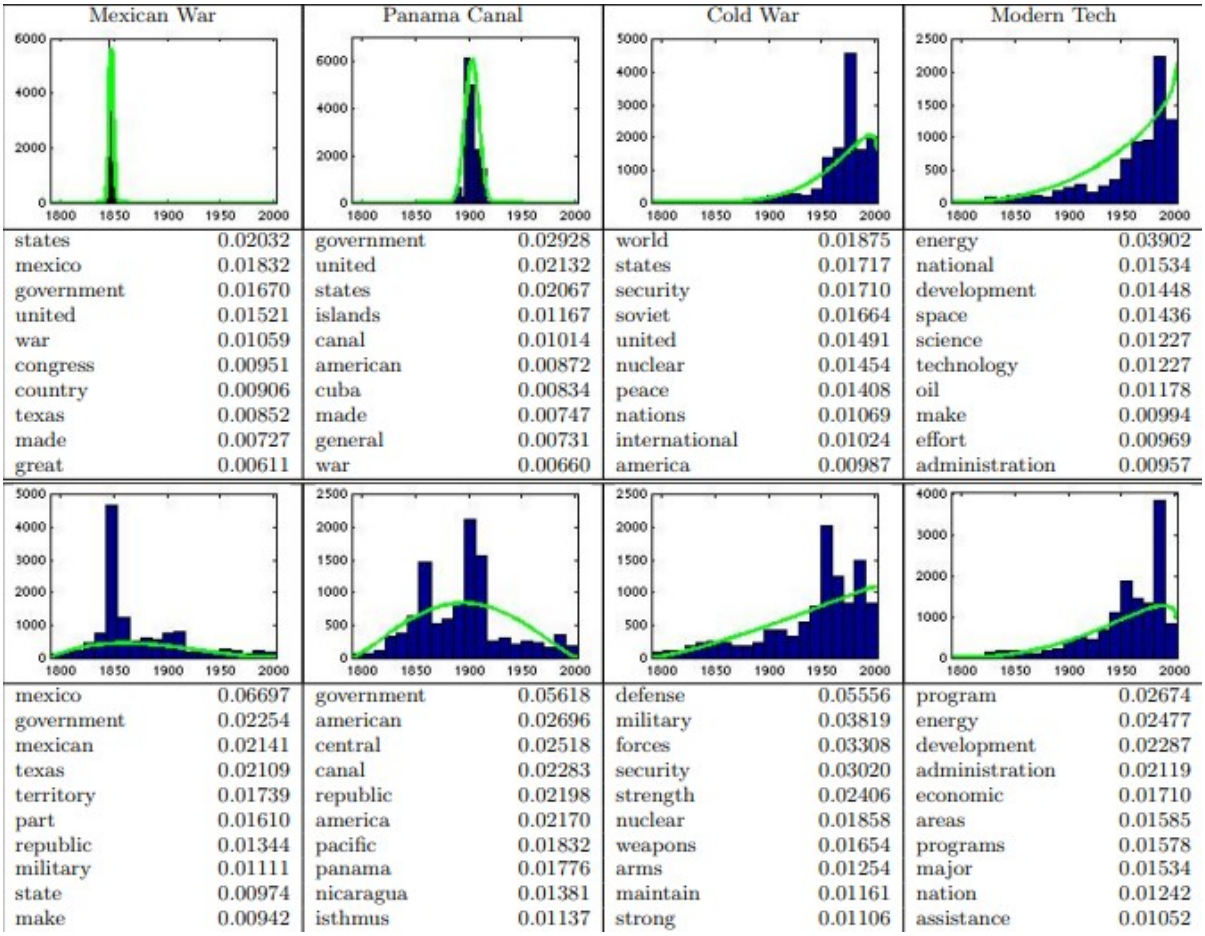
Um dos modelos estudados, mas não implementado neste trabalho, é o *Dynamic Topic Model* ou Modelo de Tópico Dinâmico (também com a sigla de DTM). Ele é um modelo de tópicos temporal que permite capturar a evolução dos tópicos de forma a organizar sequencialmente coleções de documentos (BLEI; LAFFERTY, 2006). Mais especificamente, os documentos em cada intervalo de tempo analisado são modelados com um modelo de tópicos no qual cada tópico evolui de um intervalo de tempo passado ao intervalo de tempo seguinte. O processo de geração de tópicos baseado no LDA não muda, com exceção de que agora a distribuição de tópicos e de termos relacionados a esses tópicos é diferente para cada intervalo de tempo. Nesse caso, os parâmetros para essas distribuições evolui de um intervalo de tempo antecessor ao intervalo de tempo sucessor. Como resultado, tem-se uma série de modelos de tópicos baseados no LDA que são sequencialmente relacionados. Um tópico aprendido por um DTM é, portanto, uma sequência de distribuições relacionadas de termos.

A outra abordagem estudada, e desta vez implementada, é o TOT (*Topics Over Time*, ou Tópicos ao Longo do Tempo), na qual também é feita uma modelagem temporal dos tópicos. A vantagem deste modelo em relação ao DTM é que nele o tempo não é discretizado (dividido em intervalos), mas sim tratado de forma contínua (WANG; MCCALLUM, 2006). Isso elimina o risco de dividir um tópico em dois quando há um breve intervalo de tempo em que o mesmo não aparece. Além disso, outros modelos de tópicos temporais (inclusive o DTM) consideram que o significado (ou associações de palavras) de um tópico muda com o tempo. Em vez disso, no TOT pode-se considerar os próprios tópicos como constantes, enquanto que os padrões de coocorrência de tópicos são os que mudam com o tempo. Assim, diferentemente do DTM, no TOT é feita apenas uma modelagem de tópicos, porém este processo de modelagem envolve várias iterações de treinamento (épocas), bem como é feito no LDA e em outros modelos. No

TOT entretanto, por haver o fator temporal, em cada uma dessas iterações são levados em conta os períodos de tempo de cada palavra para o cálculo das probabilidades.

Uma vantagem do TOT em relação ao LDA, no qual é baseado, é que o TOT apresenta tópicos mais bem localizados no tempo, justamente por considerar o fator temporal, tendo maior precisão ao determinar em qual período um dado tópico ou evento ocorre. Na Figura 6 é possível observar alguns histogramas comparando resultados do TOT (parte superior) com resultados do LDA (parte inferior). Nota-se que no TOT há uma precisão maior para indicar a ocorrência dos eventos ao longo do tempo, juntamente com uma lista de termos mais coerente com o tópico observado. No primeiro tópico por exemplo, o TOT mostra de forma mais clara que a guerra entre México e Estados Unidos (1846-1848) ocorreu poucos anos antes do ano de 1850. No LDA o tema se espalha ao longo dos anos, apresentando outros picos relativos a outros momentos da história como a primeira guerra mundial (1914-1918), quando palavras relacionadas ao assunto "guerra" (*"war"*) também foram utilizadas nos documentos e relações entre México e Estados Unidos desempenharam um pequeno papel. Não fica claro portanto, qual evento está sendo capturado no LDA.

Figura 6 – Comparação do modelo do TOT (parte superior) com o modelo do LDA (parte inferior)



Fonte: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, 2006 (WANG; MCCALLUM, 2006).

Neste trabalho portanto, optou-se por fazer uso do modelo do TOT, comparando-o com o LDA (ou LDA discreto, como também será chamado neste trabalho). O motivo da escolha se deu justamente pelo fato do TOT em tese apresentar uma precisão maior que o LDA, e por não discretizar o tempo como é feito no modelo do DTM por exemplo.

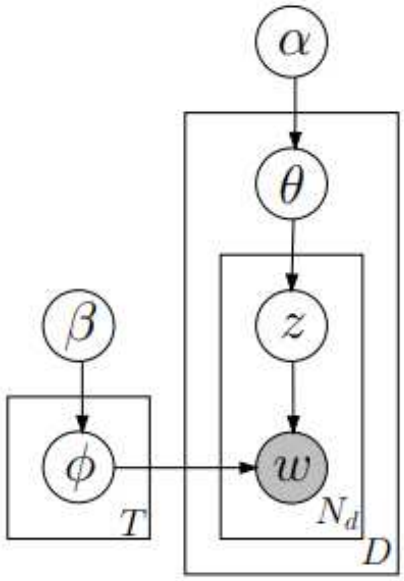
2.5 *Plate Notation* dos modelos

Na inferência Bayesiana, *Plate Notation* (ou Notação de Placas) é um método de representação de variáveis que se repetem em um modelo gráfico. Em vez de desenhar cada variável repetida individualmente, uma placa ou retângulo é usado para agrupar variáveis em um sub-gráfico que se repetem juntas, e um número é desenhado na placa para representar o número de repetições do sub-gráfico na placa (WIKIPéDIA, 2020). Esta notação auxilia no entendimento dos modelos, demonstrando graficamente o funcionamento dos mesmos.

No LDA, a representação do modelo é apresentada na Figura 7. Para cada um dos símbolos, tem-se:

- T $\rightarrow$ Quantidade de tópicos;
- D $\rightarrow$ Quantidade de documentos;
- $\alpha \rightarrow$ Parâmetro do Dirichlet na distribuição dos tópicos por documento;
- $\beta \rightarrow$ Parâmetro do Dirichlet na distribuição das palavras por tópico;
- $N_d \rightarrow$ Quantidade de palavras em um documento d;
- $\theta_d \rightarrow$ Distribuição de tópicos para um documento d;
- $\phi_z \rightarrow$ Distribuição de palavras para um tópico z;
- $z_{di} \rightarrow$ O tópico associado à i-ésima palavra no documento d;
- $w_{di} \rightarrow$ A i-ésima palavra do documento d.

Figura 7 – *Plate Notation* do modelo do LDA



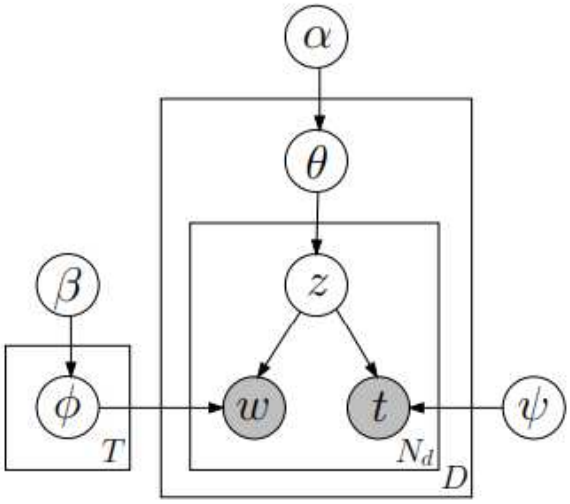
Fonte: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, 2006 (WANG; MCCALLUM, 2006).

Para a construção do modelo probabilístico, a ideia é iterar sobre cada um dos D documentos, os quais possuem uma lista de N_d palavras. Iterando sobre esta lista de palavras, obtém-se uma palavra w, a qual é dependente do tópico z associado a essa palavra, que por sua vez é obtido através da distribuição θ de tópicos por documento. Esta distribuição possui um parâmetro α do Dirichlet, o qual é uma espécie de peso associado a cada tópico da distribuição. Esta mesma palavra w é dependente também da distribuição ϕ de palavras para cada tópico. Esta distribuição por sua vez tem associada a ela um parâmetro β, que analogamente ao α, é uma espécie de peso associado a cada palavra da distribuição. Tanto o parâmetro α quanto o β costumam ter valores menores que 1. Os dois parâmetros costumam ter individualmente o mesmo valor para cada palavra a eles associada, como por exemplo 0.1 para o α e 0.001 para o β, favorecendo assim distribuições de tópicos e de palavras mais esparsas (isto é, menos tópicos por documentos e menos palavras por tópicos) em relação aos parâmetros α e β, respectivamente.

Para o modelo do TOT, a representação muda, pois neste modelo leva-se em consideração o fator temporal dos documentos. A representação é exibida na Figura 8. Além dos elementos presentes no modelo do LDA, existem também elementos referentes ao fator temporal, os quais são:

- $\psi_z \rightarrow$ A distribuição beta de um período de tempo específico para um tópico z;
- $t_{di} \rightarrow$ O *timestamp* associado à i-ésima palavra do documento d.

Figura 8 – *Plate Notation* do modelo do TOT



Fonte: Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, 2006 (WANG; MCCALLUM, 2006).

Desta forma, para cada palavra w tem-se associado também um *timestamp* do documento que contém a palavra. Associado a este *timestamp* existe um parâmetro ψ relativo à distribuição beta de cada *timestamp* para um determinado tópico z.

2.6 Trabalhos relacionados

Vários estudos foram feitos nas áreas de modelagem de tópicos temporal e do LDA. Dentre eles, há o trabalho feito por Menuzzo et al., no qual o tema também é sobre modelagem de tópicos para a análise do discurso (MENUZZO; SANTANCHè; JR, 2021). Nele, a análise é feita buscando avaliar a diversidade, desvio e coesão de discursos de prefeituras municipais do Brasil em postagens direcionadas à pandemia de Covid-19 feitas na rede social *Facebook*. Em abordagem semelhante à proposta no presente trabalho, os autores aplicam o modelo do LDA para avaliar os impactos que os discursos das prefeituras têm em relação à situação da pandemia de Covid-19 ao longo dos meses. Assim, o trabalho apresenta resultados que sugerem que os discursos tendem a se tornar menos diversos e mais coesos conforme o agravamento da pandemia. Ademais, ele indica que em cidades onde o discurso tem maior distância do discurso central empregado, esse discurso tende a se aproximar dos demais discursos conforme a pandemia se agrava.

Embora o presente trabalho se mostre semelhante ao trabalho de Menuzzo et al., aqui buscou-se aplicar, avaliar e comparar os modelos do LDA e do TOT, fazendo uma análise qualitativa e quantitativa em relação aos dois modelos como foi mencionado anteriormente.

3 METODOLOGIA

3.1 Origem e processamento de dados

Como base de dados foram utilizados os conjuntos de dados (*datasets*) obtidos no trabalho sobre a Covid-19 mencionado na seção de trabalhos relacionados (MENUZZO; SANTANCHè; JR, 2021). Neles há todas as informações referentes às postagens das prefeituras municipais no *Facebook* sobre o tema da Covid-19 nos anos de 2020 e 2021, somando mais de 70 mil postagens. Ao todo são 4 conjuntos de dados, os quais foram unidos em um conjunto de dados único para que o processamento pudesse ser feito. O conteúdo principal está naturalmente no corpo dessas postagens, que contém as informações para a análise que foi feita.

Após concatenar todas essas informações, foi feita a filtragem das postagens selecionando apenas as postagens relacionadas à pandemia, removendo publicações referentes a outros assuntos. Isso foi feito utilizando uma lista de palavras contendo os termos mais utilizados em textos sobre a pandemia, de acordo com o artigo de Tiago de Melo e Carlos Figueiredo (MELO; FIGUEIREDO, 2020) no qual é feita uma análise de postagens e notícias sobre a Covid-19. Foram selecionadas as postagens que continham pelo menos uma das palavras dessa lista, descartando as demais postagens. Em seguida foram removidos termos que não tinham importância para análise como links, hashtags, entre outros.

3.2 Definição dos modelos

Com os dados unidos e filtrados, foi feito o processamento referente aos modelos analisados. Foram eliminados das mensagens quaisquer tipos de caracteres ou palavras que não seriam utilizados na análise. Desta forma, o conjunto de dados obtido foi utilizado para a modelagem tanto do LDA discreto quanto do TOT.

Com o LDA discreto, foi dada continuidade à modelagem realizando a chamada tokenização (*tokenization*) das mensagens, na qual as palavras presentes nos discursos foram separadas, gerando uma lista com todas essas palavras. Na fase de lematização, foram retiradas palavras diferentes que na verdade possuem o mesmo significado, como “recuperar” e “recuperou” por exemplo. Após estas etapas, obteve-se uma lista de palavras (*bag of words*), utilizada para a geração de tópicos propriamente dita. Depois, foram retirados caracteres de pontuação e palavras irrelevantes (*stopwords*). Todos esses processos de tokenização, lematização e remoção das *stopwords* foram feitos com o auxílio de bibliotecas da linguagem de programação utilizada, conforme será detalhado na seção 4.1.

A partir da lista de palavras gerada, foi possível executar a modelagem de tópicos utilizando o LDA para produzir tópicos a partir dos documentos que continham as palavras, com a exclusão de caracteres de pontuação e palavras irrelevantes (*stopwords*).

Da mesma forma, no TOT também foi gerada a lista com as *stopwords*, porém o restante da modelagem ocorreu de forma diferente. A partir do conjunto de dados com as informações já filtradas, foi criada uma coluna no conjunto de dados com os *timestamps* referentes à data de cada uma das mensagens. A partir disso, foi feita a inicialização das propriedades utilizadas no TOT, seguida da Amostragem de Gibbs que é feita a partir dos dados fornecidos (WIKIPéDIA, 2021). Ao final da amostragem, obteve-se as probabilidades dos tópicos referentes ao documentos analisados.

Para cada um dos modelos foram gerados 5 tópicos. Esta quantidade de tópicos foi determinada com base na métrica de valor de coerência descrita no trabalho de Menuzzo et al. (MENUZZO; SANTANCHè; JR, 2021), com a qual é possível medir o grau de similaridade semântica entre as palavras presentes nos tópicos.

3.3 Interpretação dos tópicos

Com a aplicação dos modelos do LDA e do TOT, foi feita a interpretação dos tópicos de cada um deles. Esta se deu pela geração de gráficos que permitiram uma melhor visualização dos tópicos, assim como pela representação dos termos resultantes de cada tópico. Em um dos gráficos, referente à evolução dos tópicos ao longo do tempo no LDA, foi adicionada uma variável de número de casos de Covid-19 durante o período avaliado, com o intuito de fazer uma comparação entre os tópicos gerados e o andamento da pandemia. Para o TOT, foram gerados dois gráficos, um deles contendo a distribuição dos tópicos para cada termo e o outro apresentando a evolução temporal dos tópicos.

3.4 Avaliação quantitativa dos modelos

Para a avaliação quantitativa dos modelos, foram feitas regressões lineares com os tópicos de cada um dos modelos com o objetivo de avaliar qual dos dois modelos apresenta as melhores métricas (valores de R^2 e dos coeficientes obtidos, conforme será apresentado na seção 4.3) a partir dos dados fornecidos. Para fazer a regressão, foi agregada uma variável que representa o grau de mobilidade nas cidades em cada um dos dias de medição, obtida a partir de um *dataset* com várias informações a respeito da pandemia. O propósito de agregação dessa variável é avaliar se o discurso das prefeituras tem influência no grau de mobilidade nas cidades brasileiras ou se tanto os discursos quanto o grau de mobilidade são influenciados pelo avanço da pandemia nas cidades.

3.5 Avaliação qualitativa dos modelos

A avaliação qualitativa dos modelos deu-se pela análise dos gráficos e representações criadas, assim como pela análise dos modelos propriamente ditos. Foram avaliados os gráficos e as representações procurando determinar se os mesmos apresentaram informações em conformidade com os dados da pandemia. Buscou-se também avaliar os modelos verificando a coerência nos resultados apresentados e determinar se os mesmos possuem recursos adequados e suficientes para as suas implementações.

4 IMPLEMENTAÇÃO E RESULTADOS

Neste capítulo estão descritas as etapas de desenvolvimento do trabalho e implementação dos modelos, bem como os resultados obtidos a partir desta implementação. A primeira etapa consiste na modelagem dos tópicos feita com cada um dos dois modelos utilizados. Em seguida, são apresentadas as análises quantitativas dos modelos, as quais são formadas pelas regressões lineares feitas a partir dos tópicos gerados pelo LDA discreto e o TOT, bem como de uma variável que representa o grau de mobilidade das cidades em questão. Na sequência, estão descritas as análises qualitativas de cada modelo, avaliando os resultados obtidos com a implementação. Por fim, tem-se a discussão dos resultados com a comparação do desempenho do LDA discreto e do TOT.

4.1 Modelagem dos tópicos com o LDA discreto

A partir do conjunto de dados obtido na etapa de processamento dos dados, foi possível realizar a modelagem dos tópicos contidos nos corpus dos documentos. Foram realizados os processos de tokenização e lematização dos documentos (conforme descrito na seção de Metodologia), seguidos pela criação da lista de palavras consideradas *stopwords*. Todos esses procedimentos foram feitos utilizando a biblioteca NLTK [1] (*Natural Language Toolkit*) da linguagem *Python*, que possui métodos que auxiliam no processamento e análise da linguagem natural. A própria biblioteca disponibiliza uma lista de *stopwords* em português, lista essa que foi utilizada na modelagem com a inclusão de palavras que foram identificadas como *stopwords* nos testes que foram realizados.

Para a modelagem dos tópicos utilizando o LDA discreto (ou somente LDA), foi utilizada a biblioteca Scikit-learn (sklearn)[2] do *Python*, na qual são disponibilizados métodos para a modelagem com o LDA. Com o conteúdo das postagens já lematizado, criou-se a chamada Matriz de Termos do Documento através do método *CountVectorizer*. Este método de vetorização de tokens transpõe todas as palavras/tokens em recursos e fornece uma contagem de ocorrência de cada palavra. Como parâmetros do método, foram passadas as seguintes especificações:

- min_df = 10 $\rightarrow$ Remove palavras com frequência menor do que 10;
- analyzer = 'word' $\rightarrow$ Indica que a análise será feita a partir de palavras e não de caracteres;
- stop_words = stop_words $\rightarrow$ Passa a lista palavras que devem ser desconsideradas na análise;
- lowercase = True $\rightarrow$ Realiza a conversão das palavras paras letras minúsculas;

[1] Documentação disponível em: https://www.nltk.org/
[2] Documentação disponível em: https://scikit-learn.org/stable/

- token_pattern = `r'[(?u)\b\w\w+\b]{3,}'` → Seleciona apenas palavras com 3 caracteres ou mais.

Como resultado da chamada do método, tem-se um Objeto Vetorizador. Este objeto é usado no método *fit_transform*, passando como parâmetro o texto lematizado. Este método é responsável por aprender o dicionário de vocabulário e retornar como resultado uma matriz esparsa. Esta matriz esparsa é utilizada então no método *LatentDirichletAllocation* responsável por gerar o modelo do LDA propriamente dito. Como parâmetros para esse método, tem-se:

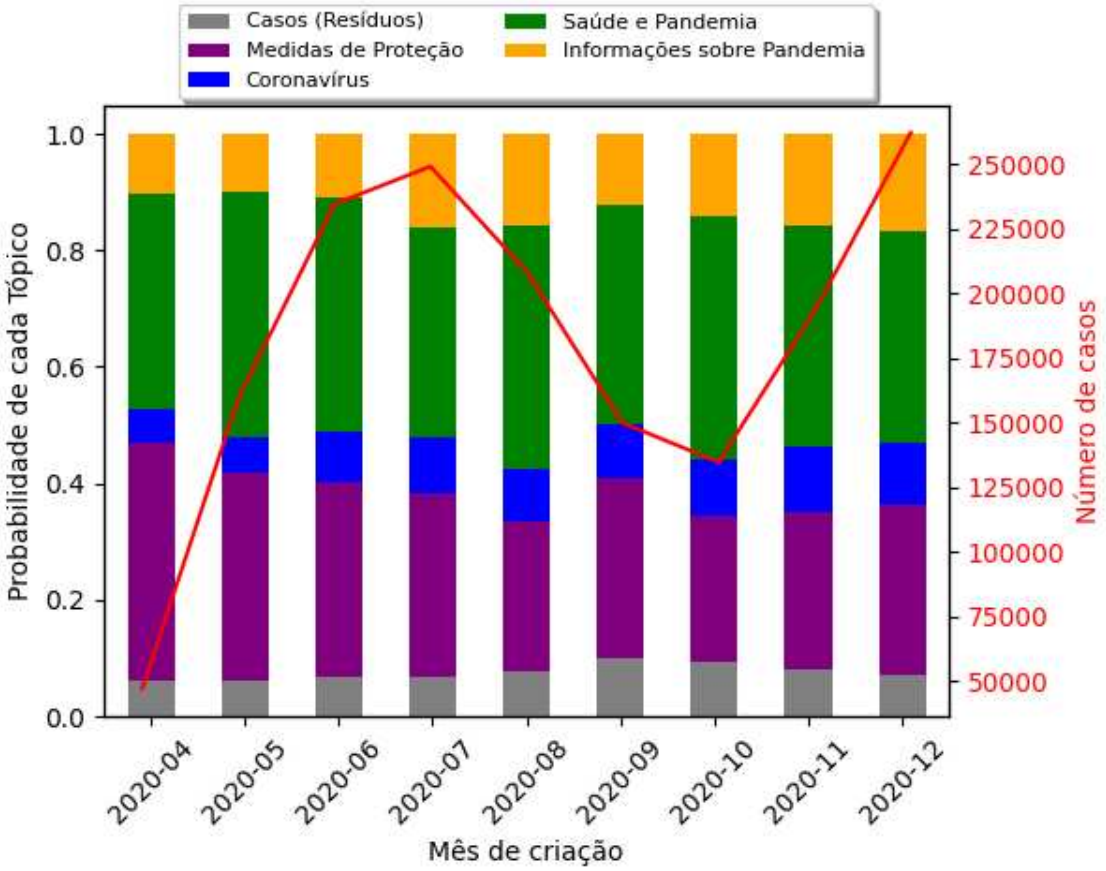
- n_components = 5 → Número de tópicos a serem gerados;
- learning_method = 'online' → Para grandes tamanhos de dados, o método *online* é bem mais rápido que a outra opção de *batch*;
- max_iter = 10 → O número máximo de iterações sobre os dados de treinamento (também conhecido como épocas);
- random_state = 0 →; Parâmetro utilizado na randomização, sendo 0 a semente para o gerador de números;
- n_jobs = -1 → O valor -1 permite o paralelismo das CPUs (todas as CPUs são utilizadas).

Interpretação dos tópicos do LDA discreto

Com a modelagem de tópicos realizada, foi possível então criar um *dataframe* a partir do modelo com cada um dos 5 tópicos e com suas probabilidades de ocorrência para cada documento. Com esse *dataframe*, foi plotado um gráfico que mostra a evolução dos tópicos ao longo do tempo, com a probabilidade de ocorrência de cada tópico em cada mês no período entre abril de 2020 a dezembro de 2020, conforme mostra a Figura 9. Complementando a plotagem, para efeitos de comparação foi adicionada uma representação do número de casos de Covid-19 ao longo do tempo no período acima mencionado. Estes dados foram obtidos através dos conjuntos de dados do site Brasil.io[3], o qual contém dados públicos brasileiros disponibilizados de uma maneira acessível a todos.

[3]Disponível em: https://brasil.io/

Figura 9 – Probabilidade dos tópicos ao longo do tempo - LDA discreto



Fonte: Autoria Própria, 2022.

Com o mesmo modelo do LDA discreto, foi possível também gerar a visualização dos principais termos dos tópicos através da ferramenta do pyLDAvis[4], biblioteca em *Python* que permite uma visualização interativa dos tópicos gerados. Esta visualização possibilitou a interpretação dos tópicos a partir dos principais termos, assim como a construção da tabela da Figura 10 com base nos resultados obtidos. Esta tabela exibe o assunto principal (definido subjetivamente pela análise dos resultados do pyLDAvis), a porcentagem de ocorrência e a lista dos principais termos de cada um dos 5 tópicos gerados.

[4]Documentação disponível em: https://pyldavis.readthedocs.io/en/latest/readme.html/

Figura 10 – Principais termos dos tópicos - LDA discreto

LDA Discreto					
Assunto	Saúde e Pandemia	Medidas de Proteção	Informações sobre Pandemia	Coronavírus	Casos (Resíduos)
Porcentagem	40,90%	31,40%	11,80%	8%	7,80%
Lista de Termos	saúde	coronavírus	covid19	coronavírus	casos
	pandemia	casa	coronavírus	logo	ano
	covid19	todos	casos	covid19	dia
	prefeitura	social	saúde	informações	óbito
	municipal	medidas	leitos	sobre	confirmados
	dia	pessoas	capital	saúde	óbitos
	coronavírus	máscara	municipal	boletim	município
	hospital	covid19	novo	casos	coronavirus
	cidade	contra	novos	link	coronavírus
	hoje	máscaras	sobre	gov	covid19

Fonte: Autoria Própria, 2022.

4.2 Modelagem dos tópicos com o TOT

Este modelo que estende o LDA discreto leva em consideração o fator temporal dos discursos analisados (de acordo com o descrito na seção de LDA temporal). A modelagem feita aqui baseou-se no modelo do TOT (*Topics Over Time*), no qual o tempo é considerado contínuo e não discreto, diferentemente do modelo do LDA discreto.

Conforme o que foi dito na seção 3.2, iniciou-se então a modelagem dos tópicos com o TOT. Para a implementação da modelagem foi utilizada uma biblioteca em *Python* chamada *Topics Over Time*[5], a qual é uma implementação de código aberto da abordagem TOT apresentada por Xuerui Wang e Andrew McCallum em seu artigo (WANG; MCCALLUM, 2006). Esta abordagem também utiliza uma lista de *stopwords*, sendo utilizada aqui a mesma lista criada anteriormente com o LDA discreto. A partir dessa lista, juntamente com os documentos (conteúdo das postagens) e o *timestamp* de cada documento foi possível obter o dicionário de palavras contendo as diferentes palavras dos corpus e outros atributos como a lista de frequência de palavras.

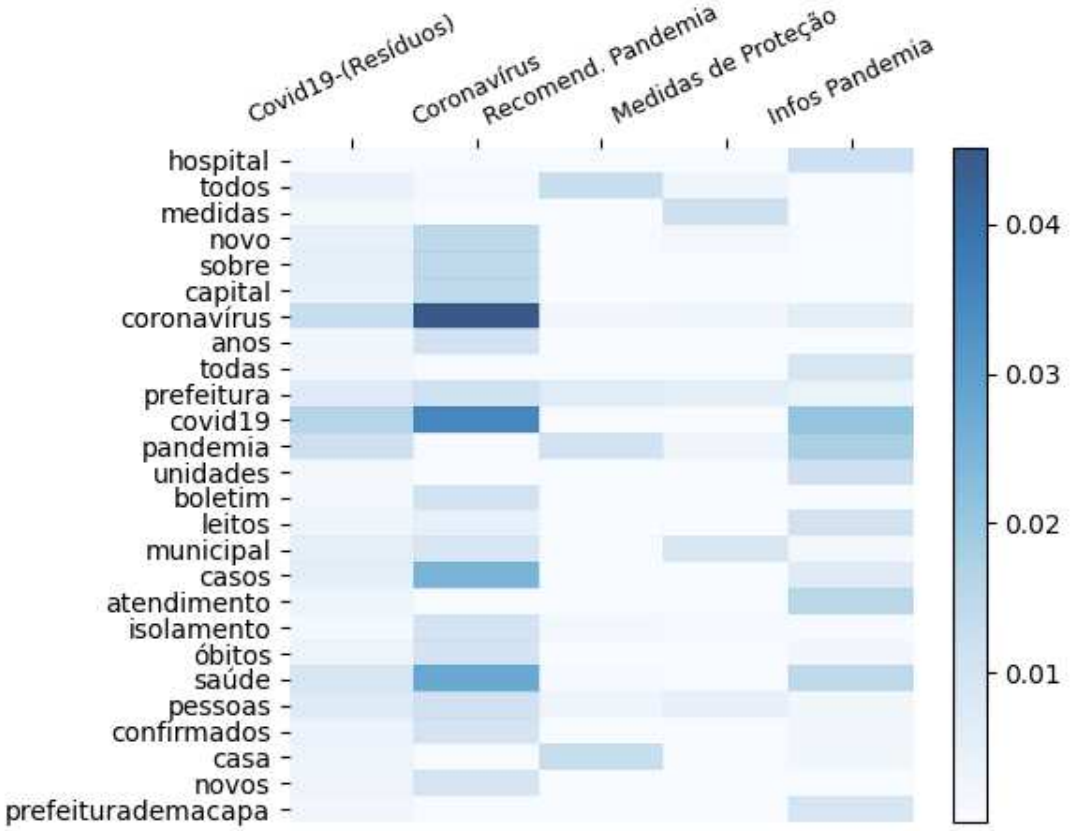
Realizou-se então a inicialização dos parâmetros do TOT (20 parâmetros ao todo), como por exemplo os parâmetros alpha, phi e psi que são utilizados mais adiante. Estes parâmetros foram configurados a partir da manipulação das informações obtidas anteriormente (documentos, *timestamps* e dicionário de palavras), sendo utilizados na etapa seguinte denominada Amostragem de Gibbs (WIKIPéDIA, 2021). O número de iterações utilizado na amostragem foi de 100 vezes, com o intuito de se ter uma boa precisão nos tópicos gerados. Ao final do processo, tem-se as probabilidades de cada um dos 5 tópicos (número especificado nos parâmetros fornecidos) em relação a cada documento.

[5]Documentação disponível em: https://github.com/ahmaurya/topics_over_time

Interpretação dos tópicos do TOT

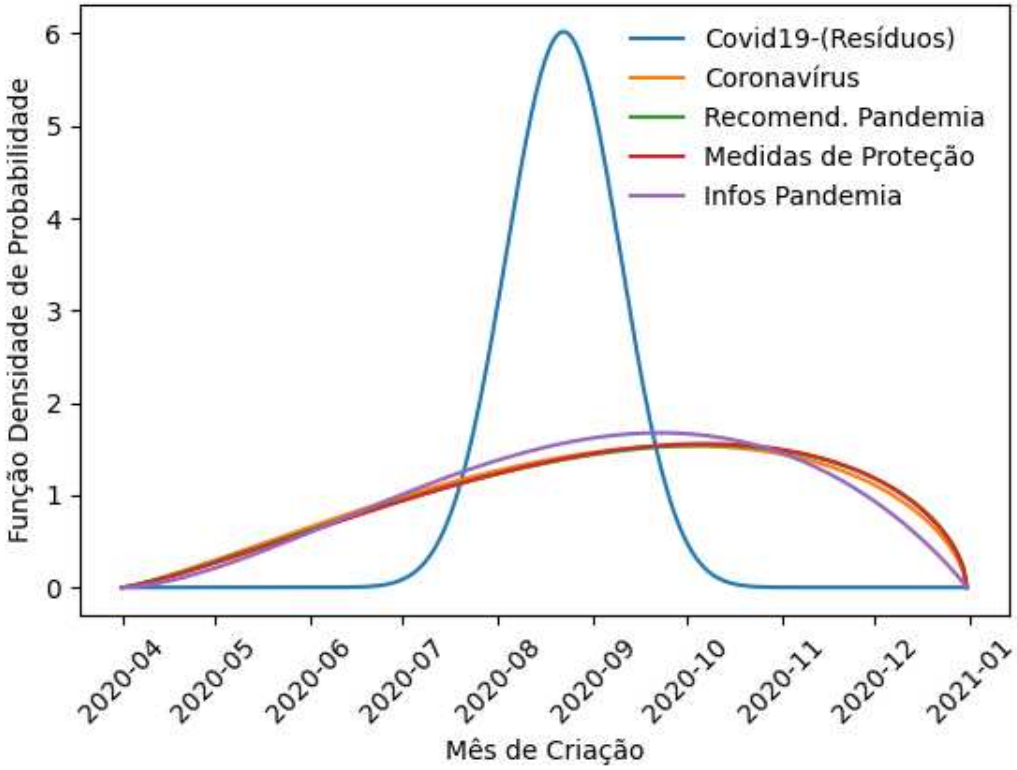
A partir dos dados gerados, foi possível realizar a plotagem de dois gráficos. O primeiro deles apresenta a distribuição dos tópicos em relação aos principais termos que aparecem nos documentos com o valor das probabilidades para cada agrupamento de termos e tópicos, conforme mostra a Figura 11. O segundo gráfico mostra a evolução dos tópicos ao longo do tempo, no período de abril de 2020 a dezembro de 2020 (Figura 12). Este período foi escolhido por conta de ter-se uma maior coerência dos dados em relação à pandemia durante este intervalo de tempo. Por conta disso, o período escolhido para análise é o mesmo tanto no LDA discreto quanto no TOT.

Figura 11 – Distribuição dos tópicos no TOT



Fonte: Autoria Própria, 2022.

Figura 12 – Evolução dos tópicos ao longo do tempo no TOT



Fonte: Autoria Própria, 2022.

Ainda, os dados gerados pela amostragem foram utilizados para realizar a visualização dos tópicos com o pyLDAvis, de forma análoga à visualização com o LDA discreto. Realizou-se também a interpretação dos tópicos e a obtenção dos principais termos conforme tabela da Figura 13.

Figura 13 – Principais termos dos tópicos - TOT

LDA Topics Over Time					
Assunto	Recomendações Pandemia	Coronavírus	Medidas de Proteção	Infos Pandemia	Covid19 (Resíduos)
Porcentagem	24,80%	24,40%	24,30%	17,20%	9,30%
Lista de Termos	casa	coronavírus	medidas	covid19	covid19
	todos	covid19	municipal	pandemia	coronavírus
	pandemia	saúde	social	atendimento	pandemia
	social	casos	atividades	saúde	saúde
	prefeitura	novo	máscaras	hospital	pessoas
	combate	capital	máscara	unidades	prefeitura
	vamos	sobre	ações	leitos	casos
	fundo	prefeitura	prefeitura	todas	cidade
	azul	pessoas	evitar	prefeiturademacapa	municipal
	neste	anos	decreto	pacientes	sobre

Fonte: Autoria Própria, 2022.

4.3 Análise quantitativa dos modelos

Para quantificar a adequação dos modelos gerados, foram implementadas regressões lineares associando os tópicos discutidos em cada cidade a uma variável de mobilidade local. A hipótese é que o discurso das prefeituras pode influenciar a mobilidade dos cidadãos (e.g. com campanhas a favor do distanciamento social). Outra possibilidade é que tanto o discurso quanto a mobilidade sejam influenciados pelo avanço da epidemia na cidade, o que também seria capturado pelo modelo proposto. Portanto, uma modelagem de tópicos que se mostra mais associada à realidade em termos de mobilidade seria potencialmente a melhor modelagem.

4.3.1 Análise de regressão com LDA discreto

Com os resultados obtidos a partir da modelagem de tópicos com o LDA discreto, buscou-se realizar uma regressão linear com o objetivo de analisar o nível de correspondência entre a evolução dos tópicos e o andamento da pandemia. Para fazer a análise, a variável de mobilidade foi utilizada a partir do *dataset* com várias informações sobre a pandemia como por exemplo o número de casos e número de mortes pela Covid-19, novos casos e novas mortes, a mobilidade da população, entre outros. Por haver informações sobre 53 cidades brasileiras neste *dataset* foi realizada uma filtragem obtendo somente informações das capitais brasileiras, no período de abril a dezembro de 2020, em conformidade com as análises feitas anteriormente. Após isso foi feita a regressão linear tendo como variável dependente o grau de mobilidade nas cidades (variável *avg_mobility_7rolling* do *dataset*) e como variáveis independentes as probabilidades de cada tópico. Pelo quinto tópico ser formado majoritariamente por resíduos da modelagem, foram incluídos na análise de regressão apenas os 4 primeiros tópicos, a fim de evitar problemas de multicolinearidade (variáveis independentes correlacionadas). O resultado da regressão pode ser visualizado através da Tabela 1.

Tabela 1 – Regressão Linear - LDA Discreto

Dep. Variable:	avg_mobility_7rolling	**R-squared:**	0.395
Model:	OLS	**Adj. R-squared:**	0.386
Method:	Least Squares	**F-statistic:**	44.07
Date:	Tue, 14 Jun 2022	**Prob (F-statistic):**	1.87e-28
Time:	00:13:30	**Log-Likelihood:**	-1094.9
No. Observations:	275	**AIC:**	2200.
Df Residuals:	270	**BIC:**	2218.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P$>\|$t$\|$**	**[0.025**	**0.975]**
Intercept	22.4495	24.964	0.899	0.369	-26.699	71.598
medidas_de_protecao	-108.6435	26.973	-4.028	0.000	-161.747	-55.540
coronavirus	59.8655	45.458	1.317	0.189	-29.631	149.362
saude_e_pandemia	-56.6861	25.897	-2.189	0.029	-107.672	-5.700
informacoes_sobre_pandemia	24.9557	28.521	0.875	0.382	-31.197	81.108

Omnibus:	9.781	**Durbin-Watson:**	0.673
Prob(Omnibus):	0.008	**Jarque-Bera (JB):**	10.102
Skew:	0.446	**Prob(JB):**	0.00640
Kurtosis:	2.708	**Cond. No.**	94.5

Fonte: Autoria Própria, 2022.

É possível notar que o R^2, variável que representa o quão bem o modelo consegue capturar a variância dos fenômenos, possui um valor de 0.395 (o maior valor possível para o R^2 é de 1). O valor do R^2 ajustado é de 0.386, suficientemente próximo do valor do R^2.

Dois fatores possuem valores positivos de coeficiente, os quais estão associados aos tópicos de Coronavírus e Informações Sobre a Pandemia. Estes porém, possuem valores de P=0.189 e P=0.382 respectivamente, maiores que o valor máximo de P=0.05 usualmente utilizado para determinar se o fator é significante na regressão linear. Os demais fatores têm coeficientes positivos e são considerados significativos. O fator com menor coeficiente (mais negativo) é o associado às Medidas de Proteção, indicando que quanto maior o número de postagens relacionadas a esse tópico, menor é o grau de mobilidade das cidades. Assim ocorre com o fator de Saúde e Pandemia, o qual influencia negativamente no valor da variável dependente de mobilidade.

4.3.2 Análise de regressão com o TOT

De forma análoga à análise de regressão linear utilizando o LDA discreto, a regressão linear com o TOT foi realizada. A diferença se deu apenas pelas probabilidades dos tópicos que neste caso vieram dos resultados obtidos pela modelagem utilizando o TOT. Assim, a variável dependente continua sendo a de mobilidade das cidades, e as variáveis independentes são as probabilidades de cada um dos 4 primeiros tópicos do modelo do TOT. O resultado da regressão encontra-se na Tabela 2.

Tabela 2 – Regressão Linear - TOT

Dep. Variable:	avg_mobility_7rolling	**R-squared:**	0.965
Model:	OLS	**Adj. R-squared:**	0.965
Method:	Least Squares	**F-statistic:**	1876.
Date:	Mon, 13 Jun 2022	**Prob (F-statistic):**	1.32e-195
Time:	23:35:49	**Log-Likelihood:**	-701.95
No. Observations:	275	**AIC:**	1414.
Df Residuals:	270	**BIC:**	1432.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P$>\|$t$\|$**	**[0.025**	**0.975]**
Intercept	-53.6600	0.804	-66.776	0.000	-55.242	-52.078
coronavirus	-838.9460	33.409	-25.111	0.000	-904.721	-773.171
recomendacoes_pandemia	2565.6119	119.333	21.500	0.000	2330.670	2800.553
medidas_de_protecao	-1879.3339	97.501	-19.275	0.000	-2071.293	-1687.375
infos_pandemia	182.1202	10.327	17.635	0.000	161.789	202.452

Omnibus:	390.146	**Durbin-Watson:**	0.606
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	71011.430
Skew:	6.516	**Prob(JB):**	0.00
Kurtosis:	80.637	**Cond. No.**	2.00e+03

Fonte: Autoria Própria, 2022.

Nesta regressão pode-se observar o valor do R^2 igual a 0.965, valor muito próximo ao valor máximo de 1, sendo este muito maior do que o valor obtido com o LDA discreto. Todos os fatores associados aos tópicos se mostraram significativos, com valores de P muito próximos de zero (na tabela os valores aparecem iguais a zero por questões de arredondamento). Dos quatro fatores exibidos, dois apresentaram valores positivos e outros dois tiveram valores negativos. Valores negativos indicam que com o aumento do número de postagens relacionadas a esses tópicos, o grau de mobilidade associado à variável dependente diminui. Para valores positivos o efeito é contrário.

O fator relacionado ao tópico com maior valor do coeficiente é o de Recomendações sobre a pandemia, enquanto que o tópico com menor coeficiente é o de Medidas de Proteção.

A partir desta regressão linear feita com o TOT é possível dizer que o modelo do TOT é surpreendentemente muito bom, pois foram obtidos resultados que dificilmente são alcançados quando são feitas regressões lineares como esta. O valor obtido do R^2 é bem alto, próximo do valor máximo de 1, revelando uma adaptação excelente do modelo aos dados fornecidos. É difícil encontrar uma justificativa para um valor tão alto como esse. Foi visto durante os testes porém, que aumentando o número de iterações na Amostragem de Gibbs aumenta-se também o R^2 na regressão linear, indicando que o modelo passa a ter uma maior precisão na geração dos tópicos conforme o aumento no número de iterações. O valor do R^2 ajustado é exatamente igual ao valor de R^2, sinalizando assim como no LDA discreto, que o número de fatores independentes da regressão está coerente.

Em relação aos fatores associados aos tópicos, a princípio esperava-se que todos os coeficientes fossem negativos. Entretanto, dentre os 4 fatores, 2 apresentaram valores positivos sendo eles o de Recomendações Sobre a Pandemia e Informações sobre a Pandemia, con-

tribuindo positivamente na mobilidade das cidades. Isso pode ter relação com uma tendência maior de mobilidade de pessoas quando fala-se mais sobre recomendações de proteção contra o coronavírus e são divulgadas informações sobre a pandemia, aumentando a segurança e a confiança das pessoas para transitar nas cidades brasileiras tal como conscientizando a população sobre a gravidade da pandemia. O contrário também pode ter ocorrido, numa situação na qual diante do aumento da mobilidade nas cidades, as prefeituras passaram a fazer mais postagens sobre recomendações de proteção contra o coronavírus e sobre informações da pandemia.

4.4 Análise qualitativa dos modelos

4.4.1 Análise do LDA discreto

Com o LDA discreto, é possível observar através do gráfico da Figura 9 que há coerência na evolução dos tópicos ao longo do tempo e uma considerável correspondência entre as probabilidades dos tópicos e o número de casos ao longo do período analisado.

É possível observar por exemplo que o tópico do Coronavírus teve um crescimento em sua probabilidade ao longo dos 8 meses, indicando que o assunto foi sendo mais comentado à medida que a pandemia foi acontecendo. Um comportamento parecido pode ser observado com o tópico de Informações Sobre a Pandemia, tendo também aumentado ao longo do tempo.

Pode-se também estabelecer uma relação inversamente proporcional entre o tópico de Medidas de Proteção e o número de casos da doença. Uma possibilidade de interpretação deste fato é a de que quando as prefeituras diminuem o volume de postagens sobre medidas de proteção contra a Covid-19, a tendência é de que a população diminua os cuidados sanitários contra a doença e o número de casos aumente. Como reflexo do aumento de casos, as prefeituras voltam a realizar postagens sobre medidas de proteção. Ainda, há a possibilidade de tanto a diminuição do número de postagens quanto o aumento no número de casos estarem ligados a um terceiro fator não conhecido, o qual teria influência sobre os outros dois fatores.

Além disso, o tópico de Saúde e Pandemia parece permanecer estável ao longo do tempo, com a possibilidade deste fato ser justificado pelos termos inclusos neste tópico que indicam assuntos não só sobre a pandemia do Coronavírus como também de outros assuntos relacionados à saúde que podem estar inclusos nas postagens.

Por fim, em relação ao modelo propriamente dito, pode-se dizer que os resultados obtidos a partir dele são coerentes. Os coeficientes significativos da regressão de fato contribuem para a diminuição da mobilidade nas cidades, efeito esperado inicialmente. A distribuição e a evolução dos tópicos também estão coerentes, acompanhando o andamento da pandemia de acordo com o previsto. Os recursos para a implementação do modelo foram suficientes e não representaram uma dificuldade no desenvolvimento do projeto, tendo boas opções de bibliotecas

para o modelo, as quais possuem documentações bem escritas que facilitaram o entendimento e o uso das mesmas.

4.4.2 Análise do TOT

Analisando os resultados apresentados na seção de modelagem com o TOT, é possível observar na Figura 11 a distribuição dos tópicos para cada termo, na qual as cores mais fortes indicam maiores ocorrências dos termos nos tópicos. Os principais termos (com cores mais fortes) são “coronavírus”, “covid-19”, “casos”, “saúde” e etc., indicando que a distribuição está de acordo com o esperado, uma vez que o assunto principal das postagens é a pandemia do coronavírus.

No gráfico da Figura 12 observa-se a evolução dos tópicos ao longo do tempo. Nota-se que exceto pelo tópico de Covid-19 (Resíduos), todos os outros tópicos apresentaram um comportamento semelhante, com a probabilidade aumentando gradativamente até um período entre setembro e novembro, indicando o auge do volume de postagens, e depois diminuindo novamente. O tópico de Covid-19 (Resíduos) apresentou tal comportamento irregular justamente por ser um tópico de resíduos, com a menor porcentagem de ocorrência entre os tópicos. Assim, qualquer alteração em relação ao termos ou o volume de postagens desse tópico gera uma grande mudança que pode ser notada visualmente. Outro detalhe importante é que este gráfico foi construído de maneira diferente do gráfico do LDA discreto pois aqui o tempo é considerado contínuo e não há portanto a separação do período em intervalos de tempo, gerando assim um gráfico de variáveis contínuas.

Em relação ao modelo do TOT, o mesmo apresentou resultados coerentes como foi visto na distribuição dos tópicos em relação aos termos. Não era previsto porém, que no gráfico da evolução dos tópicos as respostas dos tópicos fossem tão parecidas, mas este pode ser um comportamento que faz sentido do ponto de vista do número de postagens como foi dito anteriormente. Em comparação com o modelo do LDA, os recursos para a implementação do modelo do TOT são mais escassos. Não há tantas bibliotecas disponíveis e as poucas que existem não têm documentações tão completas, exigindo mais esforço para o entendimento dos métodos fornecidos.

4.5 Discussão dos resultados

A partir dos resultados obtidos, foi visto que os dois modelos apresentaram resultados coerentes ao se analisar a evolução dos tópicos gerados ao longo do tempo, uma vez que no LDA discreto as probabilidades dos tópicos indicam estar em conformidade com o número de casos confirmados de covid-19 nas capitais do Brasil e por consequência, em conformidade com o andamento da pandemia ao longo do período analisado. Com o modelo do LDA temporal

utilizando o TOT (*Topics Over Time*), observou-se que os 4 primeiros tópicos têm comportamentos semelhantes entre si, atingindo o ápice de ocorrência praticamente no mesmo período, o qual foi importante dentro da pandemia como um todo.

Desta forma, o modelo do LDA discreto em princípio demonstra ter um resultado melhor que o modelo do TOT para a evolução dos tópicos temporal justamente pelo fato de que no TOT os tópicos apresentam respostas semelhantes graficamente, enquanto que no LDA discreto há uma maior variação da probabilidade entre os tópicos ao longo do tempo.

Em relação à regressão linear, o modelo do TOT obteve um resultado melhor do que o LDA discreto, com o R^2 alcançando um valor próximo ao máximo possível, contrariando as expectativas de resultado esperadas inicialmente. Apesar do excelente resultado obtido, não há evidências claras que indiquem o motivo do modelo ter apresentado um resultado tão positivo, diferentemente do modelo do LDA discreto que teve um resultado mais de acordo com o esperado. É certo porém, que os modelos se baseiam em dados obtidos exatamente de fatores intrínsecos à pandemia, fato que é possivelmente uma justificativa para um desempenho tão bom principalmente do modelo do TOT.

5 CONCLUSÕES

Diante do que foi apresentado, é possível fazer algumas constatações a respeito do trabalho desenvolvido e dos modelos utilizados. Considerando os objetivos do trabalho estabelecidos inicialmente, é apropriado afirmar que os mesmos foram concluídos com êxito. Através do que foi executado, foi possível pesquisar e aplicar técnicas de modelagem de tópicos analisando os discursos das prefeituras brasileiras nas redes sociais.

Os dois modelos trabalhados apresentaram resultados coerentes e se mostraram adequados às análises feitas. Comparando-os, não é possível afirmar com propriedade qual dos modelos é melhor. O modelo do LDA discreto apresenta resultados que visualmente demonstram ser melhores, como foi possível observar através dos gráficos exibidos, uma vez que o modelo parece captar melhor o andamento da pandemia ao longo do tempo. Entretanto, na regressão linear o TOT apresentou resultados melhores do que o LDA discreto, com um valor de R^2 bem superior, indicando que o modelo ajusta-se melhor aos dados fornecidos.

Com base no que foi apresentado, pode-se dizer que as pesquisas feitas em relação ao tema e aos modelos estudados, assim como a aplicação das técnicas de modelagem foram de grande valor e contribuíram para enriquecer as pesquisas e análises relacionadas ao tema da covid-19, fomentando também o rol de trabalhos existentes a respeito da modelagem de tópicos.

Como continuação do trabalho, é possível estendê-lo aplicando as técnicas de modelagem de tópicos vistas aqui em outros conjuntos de dados, com o objetivo de aprimorar as técnicas de modelagem e os algoritmos implementados, bem como confirmar os resultados apresentados neste trabalho em relação aos modelos analisados. Uma outra alternativa é a de seguir com as pesquisas e análises sobre o tema da covid-19, que demonstra ter muito a ser explorado e demonstra ser de grande interesse e importância para a sociedade brasileira e global.

REFERÊNCIAS

BLEI, D.; LAFFERTY, J. Dynamic topic models. **Proceedings of the ICML**, p. 1, 2006. Citado na página 16.

BLEI, D.; LAFFERTY, J. Topic models. **Computer Science, Columbia University**, p. 1, 2009. Citado na página 12.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of Machine Learning Research**, p. 1–2, 2003. ISSN 0016-6731. Citado na página 13.

HARDY, M.; BRYMAN, A. **Handbook of Data Analysis**. Paperback edition. [*S.l.*]: Sage Publications, 2004, 2009. ISBN 9781848601161. Citado na página 12.

KULSHRESTHA, R. **A Beginner's Guide to Latent Dirichlet Allocation(LDA)**. Towards Data Science, 2019. Disponível em: https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2. Acesso em: 10 nov. 2021. Citado 2 vezes nas páginas 13 e 14.

LEFEBURE, L. **Exploring the UN General Debates with Dynamic Topic Models**. 2018. Disponível em: https://towardsdatascience.com/exploring-the-un-general-debates-with-dynamic-topic-models-72dc0e307696. Acesso em: 11 nov. 2021. Citado na página 16.

MELO, T. d.; FIGUEIREDO, C. M. A first public dataset from brazilian twitter and news on covid-19 in portuguese. **Universidade do Estado do Amazonas, Brazil**, Aug 2020. Citado na página 21.

MENUZZO, V. A.; SANTANCHè, A.; JR, L. C. G. Topic modeling applied to the analysis of diversity and cohesion in discourses. **Universidade de Campinas - UNICAMP**, Sep 2021. Citado 3 vezes nas páginas 20, 21 e 22.

MERCANTIL, M. **Brasil é o terceiro país que mais usa redes sociais no mundo**. Monitor Mercantil, 2021. Disponível em: https://monitormercantil.com.br/brasil-e-o-terceiro-pais-que-mais-usa-redes-sociais-no-mundo/. Acesso em: 11 nov. 2021. Citado na página 10.

PASCUAL, F. **Topic Modeling: An Introduction**. Monkey Learn, 2019. Disponível em: https://monkeylearn.com/blog/introduction-to-topic-modeling/. Acesso em: 08 nov. 2021. Citado na página 12.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945–959, 2000. ISSN 0016-6731. Disponível em: https://www.genetics.org/content/155/2/945. Citado na página 13.

SETH, N. Part 2: Topic modeling and latent dirichlet allocation (LDA) using gensim and sklearn. **Analytics Vidhya**, Jun 2021. Disponível em: https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/. Acesso em: 10 nov. 2021. Citado 2 vezes nas páginas 14 e 15.

SHAW, S. E.; BAILEY, J. Discourse analysis: What is it and why is it relevant to family practice? **PubMed Central**, p. 413–419, 6 2009. Citado na página 12.

WANG, X.; MCCALLUM, A. Topics over time: A non-markov continuous-time model of topical trends. **Department of Computer Science, University of Massachusetts**, Aug 2006. Citado 5 vezes nas páginas 16, 17, 19, 20 e 27.

WANG XIAOGANG; GRIMSON, E. Spatial latent dirichlet allocation. **Computer Science and Artificial Intelligence Lab MIT**, p. 1–2, 2007. Citado na página 13.

WIKIPéDIA. **Plate Notation**. Wikipédia, 2020. Disponível em: https://en.wikipedia.org/wiki/Plate_notation. Acesso em: 15 jun. 2022. Citado na página 18.

WIKIPéDIA. **Amostragem de Gibbs**. Wikipédia, 2021. Disponível em: https://pt.wikipedia.org/wiki/Amostragem_de_Gibbs. Acesso em: 07 jun. 2022. Citado 2 vezes nas páginas 22 e 27.