UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Giordanno Brunno Bergamini Gomes

Contribuições à Análise Histórica e Social em Rede Social Baseada em Processamento de Linguagem Natural

Campinas

2023

Giordanno Brunno Bergamini Gomes

Contribuições à Análise Histórica e Social em Rede Social Baseada em Processamento de Linguagem Natural

Dissertação apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica, na Área de Engenharia de Computação.

Orientador: Prof. Dr. Romis Ribeiro de Faissol Attux

Este trabalho corresponde à versão final da dissertação/tese defendida pelo aluno Giordanno Brunno Bergamini Gomes, e orientada pelo Prof. Dr. Romis Ribeiro de Faissol Attux.

Campinas

2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

G585c

Gomes, Giordanno Brunno Bergamini, 1993-
Contribuições à análise histórica e social em rede social baseada em processamento de linguagem natural / Giordanno Brunno Bergamini Gomes. – Campinas, SP : [s.n.], 2023.

Orientador: Romis Ribeiro de Faissol Attux.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Processamento de linguagem natural (Computação). 2. Humanidades digitais. 3. Mídia social. I. Attux, Romis Ribeiro de Faissol, 1978-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Informações Complementares

Título em outro idioma: Contributions to historical and social analysis in a social network based on natural language processing
Palavras-chave em inglês:
Natural language processing (Computing)
Digital humanities
Social media
Área de concentração: Engenharia de Computação
Titulação: Mestre em Engenharia Elétrica
Banca examinadora:
Romis Ribeiro de Faissol Attux [Orientador]
Leonardo Tomazeli Duarte
Diego Jair Vicentin
Data de defesa: 13-07-2023
Programa de Pós-Graduação: Engenharia Elétrica

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0005-1098-8395
- Currículo Lattes do autor: http://lattes.cnpq.br/8744722427127148

COMISSÃO JULGADORA - DISSERTAÇÃO DE MESTRADO

Candidato: Giordanno Brunno Bergamini Gomes RA: 146244
Data de defesa: 13 de julho de 2023
Título da Tese: “Contribuições à Análise Histórica e Social em Rede Social Baseada em Processamento de Linguagem Natural”

Prof. Dr. Romis Ribeiro de Faissol Attux (Presidente)
Prof. Dr. Leonardo Tomazeli Duarte
Prof. Dr. Diego Jair Vicentin

A Ata de Defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no SIGA (Sistema de Fluxo de Dissertação/Tese) e na Secretaria de Pós-Graduação da Faculdade de Engenharia Elétrica e de Computação.

Agradecimentos

É com profunda gratidão que expresso meus sinceros agradecimentos a todos que contribuíram para a realização desta dissertação de mestrado. Este momento representa não apenas uma etapa acadêmica, mas também uma jornada repleta de aprendizado e colaboração.

Primeiramente, desejo agradecer ao meu orientador, Romis Ribeiro de Faissol Attux, pela sua orientação valiosa, paciência e constante incentivo ao longo deste processo. Suas ideias e expertise foram fundamentais para o desenvolvimento deste trabalho e para o meu crescimento como pesquisador.

Minha profunda gratidão também se estende aos membros do meu comitê avaliador, Leonardo Tomazeli Duarte e Diego Jair Vicentin, por dedicarem seu tempo e expertise na avaliação e aprimoramento deste estudo. Suas contribuições críticas foram essenciais para elevar a qualidade desta dissertação.

Não posso deixar de reconhecer o apoio generoso dos meus colegas de mestrado. Nossas discussões e trocas de ideias enriqueceram minha perspectiva e contribuíram para a evolução deste trabalho. Também sou grato aos amigos e familiares que me apoiaram ao longo dessa jornada, oferecendo palavras de encorajamento nos momentos mais desafiadores.

Por fim, dedico um agradecimento especial à minha família. Minha mãe, Regina. Meu pai, Luiz. Meu irmão, Neill. Seu amor incondicional e constante apoio foram a força motriz que me impulsionou a superar obstáculos e a persistir na busca pelo conhecimento. Esse agradecimento também inclui minha namorada, Tuanny, meus amigos, Gustavo Novara, Leonardo, Victor, Gustavo Gomes, Gileno e Daniel.

Este trabalho não teria sido possível sem a contribuição de cada uma dessas pessoas e instituições. Sou profundamente grato por fazerem parte da minha jornada acadêmica e por terem deixado uma marca indelével em minha trajetória.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

O diálogo entre as ciências humanas e a engenharia da computação ainda é tímido, principalmente diante dos desafios impostos pela nossa moderna realidade tecnológica. No entanto, tal diálogo tem o potencial de gerar ideias teóricas e práticas de grande impacto. Visando ajudar a preencher essa lacuna, este trabalho propõe, primeiramente, realizar uma revisão bibliográfica sobre pesquisas com contribuições do Processamento de Linguagem Natural (PLN) às ciências humanas. Depois, realizar uma abordagem computacional para uma tarefa profundamente relacionada às ciências humanas, a de empregar o PLN na análise de texto de mídias sociais. Os pesquisadores que trabalham nesse campo frequentemente se deparam com a necessidade de extrair informações de grandes massas de dados textuais. Uma dessas aplicações é a modelagem de tópicos, tarefa que requer a descoberta dos tópicos discutidos em textos - para lidar com isso, existem várias técnicas disponíveis, como Alocação de Dirichlet Latente (LDA), Modelo de Tópico Bitermo (BTM), Versão de Tópicos de Representações de Codificador Bidirecional de Transformadores (BERTopic) e Fatoração de Matriz Não-Negativa (NMF). Neste trabalho, desenhamos uma configuração metodológica e realizamos uma análise comparativa das técnicas acima mencionadas sobre dados recuperados do Twitter. Por meio dessa mídia social, buscamos contribuir metodologicamente para o estudo de questões políticas, econômicas e sociais, bem como avaliar os méritos relativos das técnicas de modelagem de tópicos. Os resultados indicam um desempenho de coerência de tópico superior para BERTopic, em segundo lugar para NMF, seguido por BTM e, por último, por LDA. Ao fim, realizamos uma discussão sobre limitações dos modelos de tópicos, seus desafios, boas práticas e perspectivas futuras.

Palavras-chaves: humanidades digitais; processamento de linguagem natural (computação); mídia social.

Abstract

The dialogue between human sciences and computer engineering is still timid, especially in view of the challenges imposed by our modern technological reality. However, such a dialogue has the potential to generate theoretical and practical ideas of great impact. Aiming to help fill this gap, this work proposes, firstly, to carry out a bibliographic review on research with contributions from Natural Language Processing (NLP) to the human sciences. Then, the idea is to perform a computational approach to a task deeply related to human sciences, to employ NLP in text analysis of social media. Researchers working in this field are often faced with the need to extract information from large masses of textual data. One of these applications is topic modeling, a task that requires the discovery of topics discussed in texts - to deal with this, there are several techniques available, such as Latent Dirichlet Allocation (LDA), Biterm Topic Model (BTM), Topic Bidirectional Encoder Representations of Transformers (BERTopic) and Non-Negative Matrix Factoring (NMF). In this work, we design a methodological setup and perform a comparative analysis of the aforementioned techniques on data retrieved from Twitter. Through this social media, we seek to methodologically contribute to the study of political, economic and social issues, as well as to assess the relative merits of topic modeling techniques. The results indicate a superior topic coherence performance for BERTopic, in second place for NMF, followed by BTM and lastly by LDA. At the end, we held a discussion about limitations of topic models, their challenges, good practices and future perspectives.

Keywords: digital humanities; natural language processing (computing); social media.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A ascensão das Humanidades Digitais representa uma tentativa de superar a divisão tradicional entre as ciências humanas e a computação, buscando explorar as vastas possibilidades proporcionadas pela tecnologia da informação. Como apontado por Sancassani (2020), um marco da emergência das Humanidades Digitais como campo de estudo e um termo a ser utilizado foi a publicação do livro Schreibman, Siemens e Unsworth (2004). Embora esse campo esteja emergindo como uma ponte entre esses dois domínios aparentemente separados, ainda há, entre eles, uma notável distância (LAZER et al., 2009) (LAZER et al., 2020).

Um potencial de aproximação das ciências humanas com a computação é, sem dúvida, o processamento de linguagem natural (PLN), dado que textos dos mais variados tipos são amplamente utilizados nas humanidades. Há vários exemplos interessantes dessa confluência (ROBILA; ROBILA, 2020). Logo, um primeiro objetivo do trabalho é realizar uma revisão bibliográfica sobre pesquisas com contribuições de PLN às ciências humanas.

Dentro do âmbito do PLN, a *modelagem de tópicos* surge, organicamente, como um caminho para determinar os assuntos discutidos em grandes conjuntos de textos ou em textos de grande porte. Vários exemplos de modelos da literatura serão apresentados na seção 2.4 do capítulo 2.

A modelagem de tópicos, como uma ferramenta de mineração de texto, torna-se um recurso de grande valia, visto que aproximadamente 80% dos dados no mundo não são estruturados,[1] em que grande parte são textos. Essa quantidade, cabe dizer, aumenta a cada dia.

As mídias sociais são fontes de grande importância para as humanidades, e uma rede social que se destaca pela grande disponibilidade de textos é o Twitter, que possuía cerca de 368 milhões de usuários ativos em 2022.[2] Por meio dela, é possível coletar uma grande quantidade de dados para estudar o que está sendo dito por muitas pessoas sobre uma enorme variedade de assuntos. Também é possível estudar os dados por país e outras divisões geográficas.

Neste trabalho, apresentamos uma nova metodologia para análise de mídias sociais, buscando, assim, contribuir para o uso de PLN no contexto das ciências humanas. Também temos como objetivo comparar modelos de diferentes abordagens (probabilística, matricial, especialista em textos curtos, neural de aprendizado profundo) para a modelagem de tópicos nessas mídias e para avaliar aquele que possui maior desempenho para essa

[1] https://www.analyticsinsight.net/the-future-of-data-revolution-will-be-unstructured-data/
[2] https://www.statista.com/statistics/303681/twitter-users-worldwide/

tarefa. Como abordagem probabilística, usamos o LDA; como abordagem matricial, temos a estratégia de NMF; como metodologia especializada em textos curtos, trabalhamos com o BTM, e, como abordagem neural, utilizamos o BERTopic.

O trabalho está estruturado da seguinte forma: o capítulo 2 traz uma revisão sobre contribuições de PLN às ciências humanas, mineração de texto, mineração de texto no Twitter, modelagem de tópicos, métricas e modelos utilizados no trabalho; o capítulo 3 apresenta a metodologia proposta; os resultados estão no capítulo 4 e as conclusões, juntamente com perspectivas de trabalhos futuros, são expostas no capítulo 5.

2 REVISÃO DA LITERATURA

Neste capítulo, partindo de discussões acerca das contribuições de Processamento de Linguagem Natural (PLN) às humanidades e da tarefa da mineração de texto, apresentaremos a tarefa específica abordada nesta dissertação: a tarefa de modelagem de tópicos (ZONG; XIA; ZHANG, 2021). Após uma discussão inicial, passaremos a uma fundamentação técnica mais aprofundada, inclusive de métricas utilizadas para avaliar os modelos.

2.1 Exemplos de Contribuições do Processamento de Linguagem Natural às Humanidades

Desde o surgimento da pesquisa em inteligência artificial (IA), houve grande interesse em construir sistemas computacionais capazes de lidar com a linguagem natural (HAENLEIN; KAPLAN, 2019). As primeiras aplicações de IA em processamento da linguagem natural (PLN) incluíram esforços em aplicações como tradução automática de texto (HIRSCHBERG; MANNING, 2015) e chatbots a partir da inspiração do teste de Turing (HAENLEIN; KAPLAN, 2019). Com o advento das modernas redes neurais e da explosão de disponibilidade de dados provocada pela internet, a área de PLN atingiu níveis de desempenho impressionantes com os já bem conhecidos grandes modelos de linguagem (LLMs, *Large Language Models*), como exemplificado pelo fenômeno do ChatGPT (THORP, 2023).

De acordo com Chowdhary (2020), o Processamento de Linguagem Natural consiste em um conjunto de técnicas computacionais que têm como objetivo a análise e representação automática de linguagens humanas, motivadas pela teoria. Enquanto Joseph et al. (2016), por meio de vários pesquisadores, define como uma área de pesquisa e aplicação que investiga a maneira pela qual os computadores podem ser empregados para compreender e manipular textos ou fala em linguagem natural, com o intuito de realizar tarefas úteis.

O Processamento de Linguagem Natural abrange diversas áreas. Também Joseph et al. (2016), por meio de Church e Rau (1995), categoriza o PLN nas seguintes cinco áreas:

- Compreensão da Linguagem Natural;
- Geração de Linguagem Natural;

- Reconhecimento de fala ou voz;
- Tradução Automática;
- Correção ortográfica e verificação gramatical.

Na aurora da Linguística Computacional, os cientistas se empenharam em escrever os vocabulários e as regras das línguas humanas para que pudessem ser compreendidos por computadores. Contudo, essa tarefa mostrou-se difícil devido à variedade, ambiguidade e dependência de contexto presentes nessas línguas (HIRSCHBERG; MANNING, 2015).

Na década de 1980, mas, principalmente, na década de 1990, o PLN foi transformado por pesquisadores que passaram a construir modelos baseados em grandes quantidades de dados empíricos da linguagem. O PLN baseado em estatística ou em *corpus* foi um dos primeiros casos de sucesso notável do uso de *Big Data*, muito antes de se reconhecer amplamente o poder da aprendizado de máquina ou de se introduzir o termo *Big Data* (HIRSCHBERG; MANNING, 2015).

Na última década, o PNL avançou rapidamente. Esse avanço, como mencionado acima, pode ser creditado à internet e ao nível de paralelismo atingido pelo hardware atual (JOSEPH et al., 2016). De fato, o aprendizado profundo (LECUN; BENGIO; HINTON, 2015) trouxe a possibilidade de lidar organicamente com tarefas que são difíceis de serem realizadas apenas com base em regras e critérios fixos. Antes desse desenvolvimento, não era possível representar todo o significado gerado pela ambiguidade das linguagens por meio da escrita de regras ou algoritmos de aprendizado de máquina como árvores de decisão e modelos probabilísticos. O aprendizado profundo é capaz de resolver eficientemente esse problema (JOHRI et al., 2021).

Visto que os textos estão entre os principais dados para estudo das humanidades, a área foi influenciada de diversas maneiras, o que é demonstrado nas subseções seguintes sobre história, sociologia, ciência política, antropologia, demografia e economia. Entretanto, não temos a pretensão de realizar uma revisão extensiva: apresentaremos, a seguir, uma breve revisão do uso de PLN nessas áreas.

2.1.1 História

Nesta subseção, apresentamos um panorama de trabalhos que trazem contribuições de PLN à área de História.

Primeiramente, temos o trabalho de Duong, Pivovarova e Zosa (2021) que apresenta um novo método de detecção automática de mudança de discurso em coleções de textos e, assim, de rastreamento de dinâmicas de discurso em *corpora* históricos, podendo ser relevante para a pesquisa na área. Esse método consistiu em utilizar uma técnica de

agrupamento de documentos textuais e, depois, uma combinação de uma rede neural recorrente com uma rede neural convolucional.

O estudo de Manjavacas e Fonteyn (2022) avaliou a adequação de grandes modelos de linguagem, como o BERT, a aplicações em textos históricos. Em seus resultados, há a conclusão de que um modelo BERT pré-treinado do zero com *corpora* históricos é mais adequado a essas aplicações.

Em Sumikawa, Jatowt e Düring (2018), foi feita a extração de referências temporais de textos da rede social Twitter ao longo de 11 meses, nos anos de 2016 e 2017. Essa extração foi feita por meio de uma ferramenta computacional baseada em regras chamada *HeidelTime*. Com essas referências temporais, foi possível analisar memórias coletivas (HALBWACHS, 1950) no Twitter.

Aguilar, Chastang e Tannier (2022) apresentam um modelo Bi-LSTM combinado com uma camada final de CRF para detectar automaticamente seções em cartas medievais em latim, o que pode acelerar a recuperação de evidências para auxiliar em hipóteses históricas. Conclui-se que o modelo é robusto para conjuntos de cartas externas, o que confirma que pode ser generalizado para cartas de períodos e origens diferentes do que foi utilizado no estudo.

Em Walter et al. (2021), foram analisados vieses políticos e raciais em *corpora* históricos por meio representações semânticas de palavras em espaços vetoriais obtidos por meio de uma rede neural artificial. Os resultados das medidas de vieses se alinharam com tendências históricas comumente percebidas, indicando uma viabilidade para analisar tendências de vieses históricos por meio desse método.

2.1.2 Sociologia

Na área de sociologia, também há trabalhos que possuem contribuições de PLN - discutiremos alguns deles nesta seção.

A tese de Nelson (2014) analisou os tópicos obtidos por meio de *Latent Dirichlet Allocation* (LDA) para comparar textos produzidos de quatro organizações feministas nas cidades de Chicago e Nova Iorque nos Estados Unidos das América, uma de cada cidade nos períodos 1900-1917 e 1960-1970. Com essas comparações e mais outros estudos, o trabalho demonstrou que o movimento das mulheres não deve ser concebido como duas ondas separadas, mas sim como um movimento contínuo que oscila e evolui ao longo do tempo.

Flores (2017) emprega dados textuais do Twitter para demonstrar como as leis anti-imigração no Arizona intensificaram a opinião pública contrária à imigração. O estudo foi feito por meio da análise de sentimento e de outra abordagem para avaliar mais de 250 mil *tweets*. O autor constata que a lei teve um impacto negativo na média de sentimentos

expressos em *tweets* relacionados aos imigrantes, mexicanos e hispânicos, mas não teve impacto significativo nos sentimentos expressos sobre asiáticos ou negros.

Por fim, citamos AlMaghlouth et al. (2015), que analisa a produção acadêmica sobre os levantes árabes de 2010 a 2012 por meio de técnicas de PLN – como medida semântica estatística – além de outras vertentes. O trabalho demonstra que há uma predominância de língua inglesa e de países não-árabes – principalmente Estados Unidos da América – no discurso científico. Conclui-se que há uma hegemonia de legitimidade no conhecimento por conta em parte ao local onde os artigos são produzidos.

2.1.3 Ciência Política

Nesta subseção, apresentamos uma breve revisão de alguns trabalhos que possuem contribuições de PLN à ciência política.

Brown (2018) realizou classificação de opiniões políticas de textos do Twitter por meio de análise de sentimento executada por um modelo de regressão logística. Demonstra-se, portanto, o uso de ferramentas de PLN para identificar opiniões e extrair tendências dos dados textuais. Para atingir tal objetivo, desenvolveu-se um estudo de caso que examina uma iniciativa para levar um projeto de lei anti-trabalho de Ohio (EUA) a um referendo estadual.

Em Kulkarni et al. (2018), foi proposto um novo modelo que teve melhor desempenho que o estado da arte para revelar a ideologia política de artigos de notícia. Inspirado em recentes avanços em inferência neural, o modelo foi construído com a ideia de multivisualização baseado em atenção.

Roberts et al. (2014) demonstra a utilidade do modelo de tópico estrutural (STM, do inglês *structural topic model*) em tornar as análises de respostas abertas de pesquisas por entrevistas mais fáceis e reveladora. Realiza-se essa demonstração com experimentos e uma análise de dados abertos no Estudo Eleitoral Nacional Americano (ANES, do inglês *American National Election Study*). Uma contribuição crucial do método é que ele incorpora informações sobre o documento, como sexo do autor, afiliação política e atribuição de tratamento.

2.1.4 Antropologia

Nesta subseção, apresentamos uma revisão de trabalhos em que se aplica PLN à antropologia.

O estudo em Krieg, Berning e Hardon (2017) analisou mais de 20 mil relatos de experiências com drogas de um portal virtual e gerou importantes percepções sobre os emaranhados sociais e políticos de consumo, fenomenologia das drogas e redução de danos.

Um dos métodos foi a análise de frequência de pares de palavras e redes de co-ocorrência de palavras-chave.

Em Munk, Olesen e Jacomy (2022), construiu-se um modelo de redes neurais que busca associar comentários do Facebook a reações de emojis. Comparando-se com os resultados de pessoas realizando a mesma tarefa, foram apresentadas três conclusões. A primeira consiste em entender que quando a máquina falha em uma situação há um real potencial para uma descrição densa, logo, representado tipicamente uma situação ambígua. A segunda constitui-se no compreendimento da delimitação onde o trabalho da máquina termina e o da etnografia começa, justamente onde há essa identificação de uma necessidade por uma descrição densa. A terceira conclusão é que os experimentos demonstraram ser possível utilizar redes neurais de maneira que a interpretabilidade pós-análise esteja em concordância com as práticas etnográficas já estabelecidas.

Abramson et al. (2018) realizam uma discussão que ferramentas computacionais, como mineração de dados, podem ajudar a escalar a etnografia, melhorar a transparência, permitir replicação básica e enfrentar desafios fundamentais em torno da validade interna e externa. Além disso pontua que a etnografia computacional não substitui as formas existentes de investigação social, mas as complementa e as estende. Também apresenta exemplos de visualizações simples de mineração de texto de discussões com menções de laços sociais entre pacientes com câncer avançado.

2.1.5 Demografia

Na área de demografia, destacamos o trabalho de Xu et al. (2022), que buscou aplicar e avaliar modelos de tópicos em dados de pesquisas demográficas. Nesse estudo, foram utilizados modelos de tópicos neurais baseados em modelos de linguagem para analisar as respostas de mulheres holandesas sobre seus planos de fertilidade. Em seguida, comparou-se os resultados obtidos com os julgamentos de especialistas, levando em consideração o cálculo de coerência dos tópicos. Os resultados indicaram que os modelos neurais produzem tópicos mais próximos da interpretação humana em comparação com o LDA. Essa pesquisa demonstra que a demografia pode se beneficiar significativamente com a adoção de novos métodos de PLN.

Em Roy et al. (2022), houve um estudo demográfico com o objetivo de analisar os dados de alunos matriculados em duas instâncias de Cursos on-line abertos massivos (MOOCs, do inglês *Massive Open Online Courses*) introdutórios de Python, oferecidos antes e durante a pandemia do COVID-19, a fim de verificar se houve alguma alteração no comportamento dos alunos e nos resultados gerais de aprendizado durante esse período. Além disso, os autores procederam à análise da interação dos alunos no fórum de discussão do curso, utilizando o processamento de linguagem natural para realizar uma análise de sentimento dos comentários. Os resultados revelaram tendências comportamentais que

apresentaram diferenças significativas durante a pandemia. O presente estudo constatou que alunos de diversas demografias apresentaram resultados e comportamentos de aprendizagem notavelmente distintos durante a pandemia, em relação a um conjunto de MOOCs de Python.

Tutubalina e Nikolenko (2017) realizam uma pesquisa acerca da mineração automática de informações demográficas em textos gerados pelos usuários de medicamentos, comparando técnicas de processamento de linguagem natural, incluindo extensões de modelos de tópicos e redes neurais profundas. O objetivo foi abordar esse problema em conjuntos de dados extraídos de sites relacionados à saúde. Esse estudo apresenta os primeiros resultados acerca do problema prático e relevante de aprendizagem automática das características demográficas do usuário, a partir de suas avaliações sobre produtos ou serviços médicos.

2.1.6 Economia

Nesta seção, apresentamos uma revisão de trabalhos que trazem contribuições do PLN à economia.

Em Ahrens e McMahon (2021), há o objetivo de fornecer à comunidade de pesquisa uma nova série de choques de política monetária baseada em discursos de bancos centrais. Para atingir esse propósito, usou-se uma modelagem de tópicos supervisionada com a capacidade de lidar com texto e variáveis numéricas para estimar um índice de dispersão de sinal de política monetária.

No estudo de Dierckx, Davis e Schoutens (2021), realizou-se modelagem de tópicos para extrair temas de notícias financeiras para investigar como suas narrativas se relacionam com os movimentos no Índice de Volatilidade CBOE. Comparou-se LDA com uma técnica combinada de modelos doc2vec e mistura gaussiana. Ao final, após utilização de outras técnicas, mostrou-se que os recursos de notícias obtidos conseguem prever os movimentos do índice.

Na pesquisa Kim e Yoon (2021), analisou-se a seção de gerenciamento, discussão e análise das divulgações corporativas e extraiu-se seu sentimento específico do contexto para prever falências iminentes. Foi empregado o modelo BERT para realizar essa análise de sentimento e demonstrou-se que ele supera as previsões de outros modelos, melhorando a acurácia de predição de falências iminentes. Além de que foi descoberto que o sentimento textual tem capacidade preditiva adicional para variáveis financeiras bem conhecidas.

2.1.7 Um Panorama sobre os Exemplos Apresentados

Até este ponto, exploramos exemplos de contribuições do Processamento de Linguagem Natural (PLN) em diversas áreas das humanidades, incluindo História, Socio-

logia, Ciência Política, Antropologia, Demografia e Economia. Embora essas áreas possam parecer distintas em seus objetivos e métodos, a aplicação do PLN oferece uma série de pontos de convergência e interconexões que enriquecem nosso entendimento e abordagem desses campos.

O PLN atua como uma ponte entre a linguagem humana e o mundo digital, permitindo a extração de informações significativas a partir de vastas quantidades de texto. Suas técnicas abrem possibilidades para analisar padrões, sentimentos, opiniões e relações que, muitas vezes, não seriam facilmente acessíveis por métodos tradicionais. O uso de algoritmos e modelos de aprendizado de máquina, como redes neurais e modelos de tópicos, permite aos pesquisadores identificar tendências, construir taxonomias, analisar sentimentos e revelar insights que contribuem para uma compreensão mais profunda das áreas estudadas.

Nesse contexto interdisciplinar, é notável como o PLN fornece ferramentas para a análise automatizada de textos que, de outra forma, seria humanamente inviável processar. Ao explorar a história, por exemplo, o PLN permite detectar mudanças de discurso, analisar vieses políticos e raciais, e até mesmo acompanhar a evolução do movimento das mulheres ao longo do tempo. Da mesma forma, na sociologia, o PLN é usado para compreender a opinião pública sobre questões sociais, analisar padrões de sentimentos em relação à imigração e investigar a influência de diferentes grupos de pesquisa no discurso científico.

Na ciência política, o PLN capacita a análise de opiniões políticas nas redes sociais, a identificação de ideologias em artigos de notícias e a análise de discursos em respostas abertas em pesquisas de opinião. Na antropologia, a mineração de textos revela insights sobre o consumo de drogas, a interação nas redes sociais e a aplicação de ferramentas computacionais para melhorar a transparência e escalabilidade da etnografia.

Além disso, o PLN encontra espaço na demografia, onde modelos de tópicos e análises de sentimentos auxiliam na compreensão dos planos de fertilidade e comportamentos de aprendizagem dos alunos. Na economia, as análises de sentimentos extraídos de discursos de bancos centrais e de notícias financeiras oferecem informações valiosas para previsões econômicas e de mercado.

O fio condutor que une esses exemplos é a capacidade do PLN de revelar padrões, informações e nuances escondidas nos dados textuais, independentemente do campo de aplicação. Essa abordagem transcende barreiras disciplinares, permitindo que os pesquisadores abordem questões complexas de maneira inovadora. À medida que o PLN continua a evoluir, sua interseção com as humanidades continua a enriquecer nossa compreensão das dinâmicas sociais, culturais, políticas e econômicas que moldam nosso mundo.

2.2 Mineração de Texto

Nesta seção, faremos uma discussão sobre mineração de texto, um processo que utiliza Processamento de Linguagem Natural e engloba a tarefa realizada neste trabalho – a modelagem de tópicos. Também faremos uma discussão das principais metodologias para lidar com esse processo.

A mineração de dados textuais se caracteriza como uma tecnologia integrada de processamento de linguagem natural, classificação de padrões e aprendizado de máquina. O termo mineração se refere à descoberta, busca, indução e refinamento, pois os resultados alvo estão geralmente escondidos e ocultos (ZONG; XIA; ZHANG, 2021).

Para Kumar, Kar e Ilavarasan (2021, p. 2), "mineração de texto, também chamada de análise de texto, é uma técnica de inteligência artificial que converte dados não estruturados em dados estruturados usando processamento de linguagem natural para aprimorar a análise usando algoritmos de aprendizado de máquina."

Também segundo Zong, Xia e Zhang (2021), algumas tecnologias de mineração de texto são:

- Classificação de texto: tarefa que busca dividir um dado texto em tipos textuais predefinidos;
- Agrupamento de texto: tarefa que se propõe em dividir um dado conjunto de textos em categorias diferentes e, ao invés da classificação, não predefinindo um número de categorias;
- Modelo de tópico: tarefa que busca tópicos referentes a um texto, sendo que cada tópico pode ser expressado por um grupo de palavras presentes no texto que possuem forte correlação e compartilham os mesmos conceitos e semânticas;
- Análise de sentimento de texto e mineração de opinião: um tipo especial de classificação de texto em que ela é feita baseada em informação subjetiva como visões e atitudes expressadas no texto ou julgamentos positivos ou negativos;
- Detecção e rastreamento de tópico: refere-se à mineração e triagem de tópicos de texto de inúmeras notícias e comentários;
- Extração de informação: tarefa associada à extração de informação factual como entidades, atributos de entidades, relação entre entidades, e eventos de linguagem natural não estruturada e semiestruturada para saída de dados estruturados;
- Sumarização automática de texto: tecnologia referente à geração automática de resumos usando métodos de processamento de linguagem natural.

O estudo de Subalalitha (2019) usa extração de informação para coletar conhecimentos históricos de um clássico poema tâmil chamado Kurunthogai. A estrutura de extração de informação proposta extrai informações de textos bilíngues, alcançou uma precisão de 88,8% e pode ser aplicado para qualquer tipo de literatura.

Um exemplo de aplicação de sumarização de texto automática é o trabalho de Mahajan e Purohit (2021), que utiliza essa tecnologia para obter informações sobre monumentos históricos famosos da Índia na forma de um resumo, sendo uma contribuição à área patrimonial histórica.

2.3 Mineração de Texto no Twitter

Dado que o objetivo deste estudo é trabalhar com a mineração de textos sobre temáticas ligadas às humanidades no contexto da maior amplitude possível de pessoas, ou seja, abarcando uma maior representatividade factível e acessível da sociedade, a escolha principal para a fonte de dados se torna as mídias sociais. Porém, é importante pontuar que essa representatividade possui limitações, abarcando um recorte específico da sociedade, composto por pessoas que possuem acesso à internet e são alfabetizadas.

Escolhido o tipo de fonte, foi necessário decidir qual mídia social seria mais adequada para o estudo. Dentre as opções, estavam as redes sociais (e.g., Facebook, LinkedIn), os blogs (e.g., Huffington Post), os microblogs (e.g., Twitter, Tumblr), os fóruns, marcadores sociais (e.g., Delicious, Pinterest), wikis (e.g., Wikipedia, Wikihow), notícias sociais (e.g., Reddit) e as de compartilhamento de mídia (e.g., Youtube, Instagram) (FARZINDAR; INKPEN, 2020, p. 2).

Dentre as mídias sociais com mais usuários, que possuem dados textuais e APIs (do inglês *Application Programming Interfaces*) funcionais e consolidadas, estão o Facebook, LinkedIn, Twitter, Reddit e Quora. Também por esse motivo, essas mídias sociais são as mais estudadas na literatura. A classificação por número de usuários pode ser observada na figura 2.1.

Uma API é uma interface através da qual desenvolvedores terceiros podem conectar novos complementos a um serviço existente. Ela é, basicamente, um meio que permite a um programa de computador se comunicar com outro programa (LOMBORG; BECHMANN, 2014, p. 256). É também por meio de APIs que se torna possível coletar sistematicamente os dados das mídias sociais.

De acordo com Russell e Klassen (2019, p. 24), enquanto algumas redes sociais como Facebook e LinkedIn requerem uma aceitação mútua da conexão entre usuários, o modelo de relacionamento do Twitter permite que você acompanhe os últimos acontecimentos de qualquer outro usuário, mesmo que outro usuário escolha não lhe seguir de

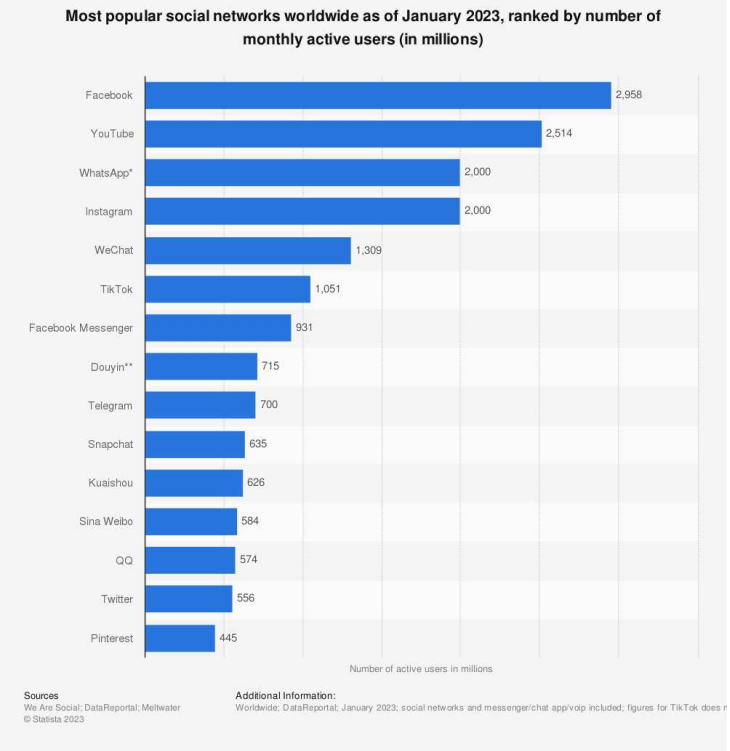


Figura 2.1 – Classificação de mídias sociais em relação ao número de usuários ativos mensais em janeiro de 2023. Segundo a fonte, os números de usuários podem não representar indivíduos únicos. É possível, ainda, que alguns desse usuários sejam *bots*.

Fonte: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

volta ou mesmo saiba da sua existência.

Segundo Farzindar e Inkpen (2020, p. 2), o Twitter se caracteriza como um microblog, ou seja, similar a um blog mas com conteúdo limitado (número de caracteres). O blog pode ser compreendido como um diário online no qual o blogueiro pode criar o conteúdo e exibi-lo em ordem cronológica reversa. Blogs são, geralmente, mantidos por uma pessoa ou uma comunidade. Os comentários dos blogs são postagens realizadas por usuários que se vinculam a blogs ou postagens de jornais online.

O Twitter foi criado em março de 2006 e lançado em julho do mesmo ano. Em 2012, declarou possuir mais de 140 milhões de usuários ativos, com 340 milhões de *tweets* por dia.[1] Em 2023, o número de usuários ativos mensais declarado em relatórios da empresa chegou a 556 milhões, como pode ser observado na figura 2.1. Entretanto, deve-se frisar que uma parte desses usuários são *bots*:[2] o trabalho de Rodríguez-Ruiz et al. (2020) aponta uma proporção de 15 % do total.

Além de possuir grande quantidade de usuários, o Twitter sempre teve seus

[1] https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html

[2] Contas automatizadas e controladas por software, programadas para criar novas publicações ou interagir com outros usuários, normalmente com objetivos específicos em mente.

dados disponíveis gratuitamente, públicos por padrão e principalmente textuais por meio de sua API gratuita e pública, resultando em um grande volume de pesquisa científica (GIGLIETTO; ROSSI; BENNATO, 2012, p. 148).

Portanto, como a mídia social escolhida foi o Twitter, apresentaremos uma breve revisão bibliográfica de trabalhos com mineração de texto utilizando essa fonte de dados. Priorizaremos, por ora, trabalhos que não usam modelagem de tópicos, i.e. que utilizam outras técnicas de mineração, já que modelos de tópicos serão tratados posteriormente.

Como exemplo de mineração de texto no Twitter, temos o trabalho de Karami, Bennett e He (2018), que realizou mineração de opinião pública para explorar a discussão de questões econômicas no Twitter durante a eleição presidencial de 2012 nos Estados Unidos da América. Eles utilizaram dois métodos de mineração de texto, modelagem de tópicos com LDA e análise de sentimento.

Outro exemplo é o artigo Shamoi et al. (2022) que executou uma mineração de opinião pública sobre a dieta baseada somente em vegetais (vegana) no Twitter utilizando análise de sentimento.

2.4 Modelagem de Tópicos

Como já apresentada na seção 2.2, a modelagem de tópicos é uma técnica de mineração de texto usada para descobrir automaticamente os tópicos discutidos em um conjunto de documentos (também conhecido como *corpus*), sendo que cada tópico é representado por um conjunto de palavras que possuem forte correlação entre si e partilham os mesmos conceitos e semânticas.

A modelagem de tópicos tem aplicações no auxílio do entendimento e resumo de grandes coleções de documentos, identificação de tópicos emergentes em um campo de pesquisa ou em um determinado período de tempo, e também pode na utilização em recomendação de conteúdo personalizado e análise de tendências.

A modelagem de tópicos foi criada (e continua sendo utilizada) para fins de recuperação de informação e classificação de texto em aplicações práticas. No entanto, recentemente, tem ganhado destaque também na área de humanidades digitais e ciências sociais no contexto de leitura distante,[3] em que é cada vez mais empregada para abordar questões de pesquisa específicas relacionadas às distribuições de conteúdo em textos literários, decisões de judiciais, debates políticos e legais, cobertura da mídia e publicações acadêmicas (SHADROVA, 2021).

[3] A leitura distante é um tipo de análise, em oposição à leitura próxima que consiste na análise detalhada e minuciosa de um texto. Na leitura distante, utilizam-se técnicas computacionais de mineração de texto para extrair e analisar padrões de superfície presentes em grandes *corpora* (SHADROVA, 2021).

Churchill e Singh (2022, p. 1) definem um modelo de tópico como um modelo matemático não-supervisionado que recebe como entrada um conjunto de documentos e retorna um conjunto de tópicos que representa o conteúdo dos documentos em uma maneira coerente e precisa. Ou seja, encontrar os temas centrais associados a uma coleção de documentos é o objetivo da modelagem de tópicos, ao poder realizar a compressão de um corpus de milhares de documentos em um pequeno resumo que captura os assuntos mais prevalentes presentes (CHURCHILL; SINGH, 2022, p. 1).

De acordo com Churchill e Singh (2022), os precursores dos modelos de tópicos têm sua origem, em 1990, com a análise semântica latente (mais comumente conhecida como LSA, do inglês *latent semantic analysis*), também chamada por indexação semântica latente (mais conhecida como LSI, do inglês *latent semantic indexing*). O LSA, que foi proposto por Deerwester et al. (1990), realiza a decomposição de valor singular da matriz de palavras por documento, reduzindo assim sua dimensionalidade e obtendo os tópicos.

Em 1999, Hofmann (1999) propôs a Indexação Semântica Latente Probabilística (mais conhecida como pLSI, do inglês *Probabilistic Latent Semantic Indexing*), introduzindo os tópicos como uma mistura probabilística de palavras com base em suas probabilidades conjuntas com documentos (CHURCHILL; SINGH, 2022).

Segundo Churchill e Singh (2022), o termo "modelo de tópico" é cunhado em Blei, Ng e Jordan (2001) ao propor um dos modelos mais conhecidos e utilizados da área, a Alocação de Dirichlet Latente (mais conhecido como LDA, do inglês *Latent Dirichlet Allocation*.

Após esse modelo, ocorreram muitas variações dele direcionadas a problemas específicos como para fontes de dados diferentes (e.g. Twitter), para uma captura da evolução de tópicos em relação ao tempo (e.g. Modelos de Tópicos Dinâmicos), para maior rapidez de inferência e para maior capacidade em tratar de bases de dados maiores (e.g. LDA Online); entretanto também surgiram novos modelos não tão marcadamente baseados no LDA (CHURCHILL; SINGH, 2022).

Os modelos de tópicos "modernos", i.e. aqueles que surgiram de 2011 a 2022, são os baseados em: fatoração de matriz não-negativa; grafos; misturas de métodos tradicionais e modernos; aprendizagem aumentada de metadados; aprendizado supervisionado; aprendizado por reforço; redes neurais rasas e profundas (CHURCHILL; SINGH, 2022).

Exemplos de modelos de tópicos na literatura são:

- *Lantent Semantic Analysis* (LSA) (DEERWESTER et al., 1990);
- *Non-negative Matrix Factorization* (NMF) (PAATERO; TAPPER, 1994);
- *Probabilistic Latent Semantic Analysis* (pLSA) (HOFMANN, 1999);

- *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003);
- *Hierarchical Dirichlet Processes* (HDP) (TEH et al., 2004);
- *Structural Topic Model* (STM) (ROBERTS et al., 2013);
- *Biterm Topic Model* (BTM) (YAN et al., 2013);
- *Word2vec Gaussian Mixture Model* (W2V-GMM) (SRIDHAR, 2015);
- *Correlation Explanation* (CorEx) (GALLAGHER et al., 2017);
- *Topic to vectors* (top2vec) (ANGELOV, 2020)
- *Topic Bidirectional Encoder Representations from Transformers* (BERTopic) (GROOTENDORST, 2022b).

Um exemplo de modelagem de tópicos foi o trabalho de Valdez, Pickett e Goodson (2018), no qual analisaram transcrições de debates presidenciais dos Estados Unidos da América em 2016 usando análise semântica latente. Os tópicos resultantes eram paralelos às pesquisas mais frequentes na Internet relacionadas a políticas na época e, quando divididas por candidato, as mudanças em tópicos emergentes refletiram posições políticas individuais.

2.4.1 Modelagem de Tópicos em Textos Curtos

Como as plataformas de mídias sociais estão cada vez mais sendo adotadas como fontes valiosas de informação, e elas são normalmente caracterizadas como texto curto, esparso e de baixa densidade, a pesquisa em modelagem de tópicos em textos curtos (STTM, do inglês *short text topic modeling*) ganhou recentemente um grande interesse (MURSHED et al., 2022).

As ocorrências de palavras em documentos curtos desempenham um papel menos discriminativo em comparação com documentos longos, caso em que o modelo tem contagens de palavras suficientes para saber como estas estão relacionadas (HONG; DAVISON, 2010). Ademais, os contextos limitados tornam mais difícil para os modelos de tópicos identificar os sentidos de palavras ambíguas em documentos curtos (YAN et al., 2013).

A falta de co-ocorrência de palavras em textos curtos torna improvável que modelos de tópicos de textos longos tenham um bom desempenho, fazendo com que sua aplicação para textos curtos seja inexpressiva em termos de desempenho (ABDEL-HAFEZ; XU, 2013). Esse é um dos principais motivos pelos quais o desenvolvimento de modelos próprios para textos curtos se tornou necessário, já existindo, portanto, uma literatura ampla, como se pode se perceber em Murshed et al. (2022).

Os algoritmos tradicionais de modelagem de tópicos para textos longos (e.g., PLSA e LDA), que se baseiam na co-ocorrência de palavras, não conseguem lidar com textos curtos de forma eficaz, já que as informações de co-ocorrência de palavras são muito limitadas neles (QIANG et al., 2019, p. 1).

Uma forma simples que busca aliviar o problema de esparsidade em textos curtos é agregá-los em pseudodocumentos longos antes de treinar um modelo de tópico padrão (YAN et al., 2013). Esse é o caso de Weng et al. (2010) e Hong e Davison (2010). Outra maneira simples de lidar com o problema é fazer suposições mais fortes sobre os dados como, por exemplo, assumir que um documento curto abrange apenas um único tópico (YAN et al., 2013). Mas também existem meios mais complexos de enfrentar o problema de esparsidade, esses são os modelos desenvolvidos especificamente para a modelagem de tópicos em textos curtos.

O trabalho Li et al. (2021) apresentou um novo modelo de tópicos específico para textos curtos, nomeado Mistura Multinomial Dirichlet Laplaceana (LapDMM, do inglês *Laplacian Dirichlet Multinomial Mixture*). Para avaliação, eles utilizaram bancos de dados de textos curtos como o BaiduQA coletado por Cheng et al. (2014), que consiste em amostras de perguntas de um popular site chinês de perguntas e respostas.

2.4.2 Alocação de Dirichlet Latente

O modelo de Alocação de Dirichlet Latente (LDA, do inglês *Latent Dirichlet Allocation*) foi proposto por Blei, Ng e Jordan (2001) e se tornou, ao longo dos anos, um dos modelos mais conhecidos e utilizados em modelagem de tópicos como já comentado na seção 2.4.

Ele corresponde a um modelo probabilístico generativo de um *corpus*, em que os documentos são representados como misturas de tópicos latentes, sendo cada tópico caracterizado por uma distribuição de palavras. As proporções de mistura de valor contínuo são distribuídas como uma variável aleatória latente de Dirichlet (BLEI; NG; JORDAN, 2001).

No LDA, assume-se que existem k tópicos latentes subjacentes de acordo com os quais os documentos são gerados e que cada tópico é representado como uma distribuição multinomial sobre as palavras $|V|$ no vocabulário (BLEI; NG; JORDAN, 2001).

Conforme Blei, Ng e Jordan (2001), um documento de N palavras e $\mathbf{w} = \langle w_1, ..., w_N \rangle$ é gerado pelo seguinte processo: primeiro, θ é amostrado de um distribuição Dirichlet$(\alpha_1, ..., \alpha_k)$; segundo, para cada uma das N palavras, um tópico $z_n \in \{1, ..., k\}$ é amostrado de uma distribuição Multinomial(θ): $p(z_n = i|\theta) = \theta_i$; finalmente, cada palavra w_n é amostrada, condicionada no z_n-ésimo tópico, de uma distribuição multinomial $p(w|z_n)$. Intuitivamente, θ_i pode ser pensado como um grau ao qual o tópico i é referido

no documento. A probabilidade de um documento é então descrito como:

$$p(\mathbf{w}|\alpha,\beta) = \int_{\theta} p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n,\beta)p(z_n|\theta) \right) d\theta, \tag{2.1}$$

onde $p(\theta|\alpha)$ é Dirichlet, $p(z_n|\theta)$ é uma multinomial parametrizada por θ, e $p(w_n|z_n,\beta)$ é uma multinomial sobre as palavras. Este modelo é parametrizado por variáveis Dirichlet k-dimensionais $\alpha = \langle \alpha_1, ..., \alpha_k \rangle$ e uma β matriz $k \times |V|$, que controlam as k distribuições multinomiais sobre as palavras. A representação gráfica do modelo LDA é apresentada na figura 2.2.

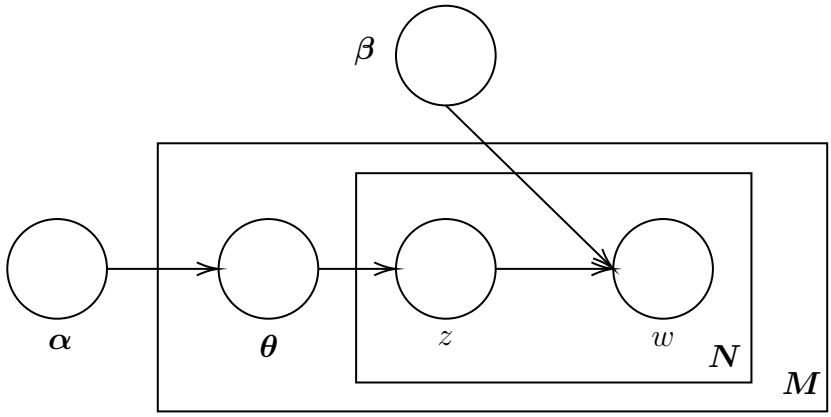


Figura 2.2 – Representação gráfica do modelo LDA.

Fonte: (BLEI; NG; JORDAN, 2003).

Apesar da equação 2.1 ser intratável para o cálculo exato, há algoritmos de inferência aproximada que podem ser considerados para realizar a LDA, como aproximação Laplace, aproximação variacional e Cadeia de Markov Monte Carlo (BLEI; NG; JORDAN, 2003). Utilizou-se, por exemplo, um algoritmo variacional baseado em convexidade simples para inferência em Blei, Ng e Jordan (2003).

Apesar de seu amplo e variado uso, o LDA possui limitações: segundo Jónsson (2016), o modelo não é um modelo com desempenho expressivo para modelagem de tópicos em textos curtos.

A aplicação de LDA é ilustrada pelo trabalho de H, Zainuddin e Wabula (2022), que realizou a modelagem e visualização de tópicos que são discutidos no Twitter pela comunidade Makassar da Indonésia por meio do modelo LDA. Também por meio de LDA, o estudo Lyu e Luli (2021) identificou os tópicos emergentes da discussão pública no Twitter relacionada ao COVID-19 sobre os Centros de Controle e Prevenção de Doenças nos Estados Unidos da América. Por fim, Xue et al. (2020) também examinaram discussões relacionadas ao COVID-19 postadas por usuários do Twitter através de identificações de tópicos e temas com LDA.

2.4.3 Modelo de Tópico Bitermo

Uma proposta para desvendar tópicos em textos curtos é o modelo de tópico bitermo (BTM, do inglês biterm topic model), que foi apresentado por Yan et al. (2013). Sua razão é contornar a esparsidade severa de dados em documentos curtos que os modelos de tópicos convencionais (e.g. LDA e PLSA) sofrem.

Enquanto esses modelos convencionais capturam padrões de coocorrência de palavras em nível de documento para revelar tópicos, o BTM modela diretamente a geração de padrões de coocorrência de palavras em todo o corpus (YAN et al., 2013).

Um bitermo é, portanto, um par de palavras desordenado coocorrido em um contexto curto e cada bitermo é extraído de um tópico específico. Um contexto curto é uma janela de texto apropriada que contém coocorrências significativas de palavras. Como os documentos são pequenos, eles são tratados como unidades de contexto individuais. Assim, todas as duplas de palavras distintas em um documento curto são extraídas como um bitermo. Os bitermos extraídos de todos os documentos da coleção compõem os dados de treinamento do BTM (YAN et al., 2013).

Suponha que α e β são os parâmetros de Dirichlet. O processo gerador específico do corpus em BTM pode ser descrito da seguinte forma:

1. *Para cada tópico z*

   a) *extraia uma distribuição de palavras específica do tópico $\phi_z \sim Dir(\beta)$*

2. *Extraia uma distribuição de tópicos $\theta \sim Dir(\alpha)$ para toda a coleção*

3. *Para cada bitermo b no conjunto de bitermos B*

   a) *extraia uma atribuição de tópico $z \sim Multi(\theta)$*

   b) *extraia duas palavras: $w_i, w_j \sim Multi(\phi_z)$*

A probabilidade conjunta de um bitermo $b = (w_i, w_j)$ pode ser escrita como:

$$p(b) = \sum_z p(z)p(w_i|z)p(w_j|z) = \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{2.2}$$

Conseguinte, a função de verossimilhança de todo o corpus é:

$$p(B) = \prod_{(i,j)} \sum_z \theta_z \phi_{i|z} \phi_{j|z} \tag{2.3}$$

O modelo BTM, ilustrado na figura 2.3, resolve o problema de esparsidade de dados do LDA ao utilizar a distribuição de tópicos em nível de corpus θ para atribuir tópicos

z, com base em uma mistura de unigramas. Além disso, o BTM supera a desvantagem da mistura de unigramas, dividindo os documentos em bitermos. Essa abordagem permite ao modelo manter a correlação entre as palavras e capturar múltiplos gradientes de tópicos em um único documento, já que as atribuições de tópicos de diferentes bitermos são independentes (YAN et al., 2013).

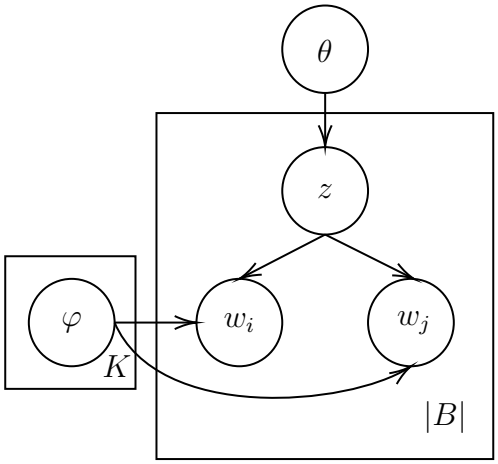


Figura 2.3 – Representação gráfica do modelo BTM.

Fonte: (YAN et al., 2013).

Todavia, o BTM não modela o processo de geração de documentos, diferentemente dos modelos de tópicos convencionais. Consequentemente, as proporções de tópicos de um documento não podem ser obtidas diretamente durante o processo de aprendizado de tópicos. Para inferir os tópicos em um documento, é necessário assumir que as proporções de tópicos de um documento são equivalentes à expectativa das proporções de tópicos de bitermos gerados a partir do documento (YAN et al., 2013), que pode ser descrito como:

$$p(z|d) = \sum_{b} p(z|b)p(b|d). \tag{2.4}$$

Na equação 2.4, $p(z|b)$ pode ser calculado pela fórmula de Bayes baseada nos parâmetros estimados no BTM:

$$p(z|b) = \frac{p(z)p(w_i|z)p(w_j|z)}{\sum_z p(z)p(w_i|z)p(w_j|z)}, \tag{2.5}$$

onde $p(z) = \theta_z$, e $p(w_i|z) = \phi_{i|z}$. Por fim, obtém-se $p(b|d)$ por uma distribuição empírica de bitermos no documento como a estimação:

$$p(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}, \tag{2.6}$$

onde $n_d(b)$ é a frequência do bitermo b no documento d. No entanto, assim como no LDA, a inferência no BTM não pode ser exata. Para solucionar essa questão, existem diversos

algoritmos que podem ser utilizados para fazer a inferência de forma aproximada, como amostragem Gibbs, inferência variacional e estimação posterior máxima. No caso de Yan et al. (2013), foi utilizada a amostragem Gibbs para inferir os parâmetros $\{\phi, \theta\}$.

2.4.4 Fatoração de Matriz Não-Negativa

A Fatoração de Matriz Não-Negativa (NMF, do inglês *Non-Negative Matrix Factorization*) foi introduzida por Paatero e Tapper (1994) sob o conceito de Fatoração de Matriz Positiva, de acordo com Wang e Zhang (2013). Entretanto, a abordagem não era, então, utilizada para modelagem de tópicos, o que veio a ocorrer posteriormente.

A NMF é um procedimento matemático no qual uma matriz formada por valores não-negativos é decomposta em duas novas matrizes de forma que o produto dessas duas novas matrizes seja igual à matriz original (CHURCHILL; SINGH, 2022).

A matriz que é decomposta é a matriz documento-palavra, que consiste em um conjunto de documentos onde cada documento é representado por um vetor de palavras. As duas matrizes menores resultantes são a matriz tópico-palavra, que pode ser interpretada como a distribuição de tópicos em relação às palavras, e a matriz tópico-documento, que pode ser interpretada como a distribuição de tópicos em relação aos documentos (CHURCHILL; SINGH, 2022).

O número de tópicos desejado pelo usuário, representado por k, determina a dimensionalidade final das matrizes menores. Se houver m documentos e n palavras no vocabulário, a matriz maior terá um tamanho de $m \times n$, enquanto as matrizes menores terão tamanhos de $m \times k$ e $k \times n$ (MURSHED et al., 2022).

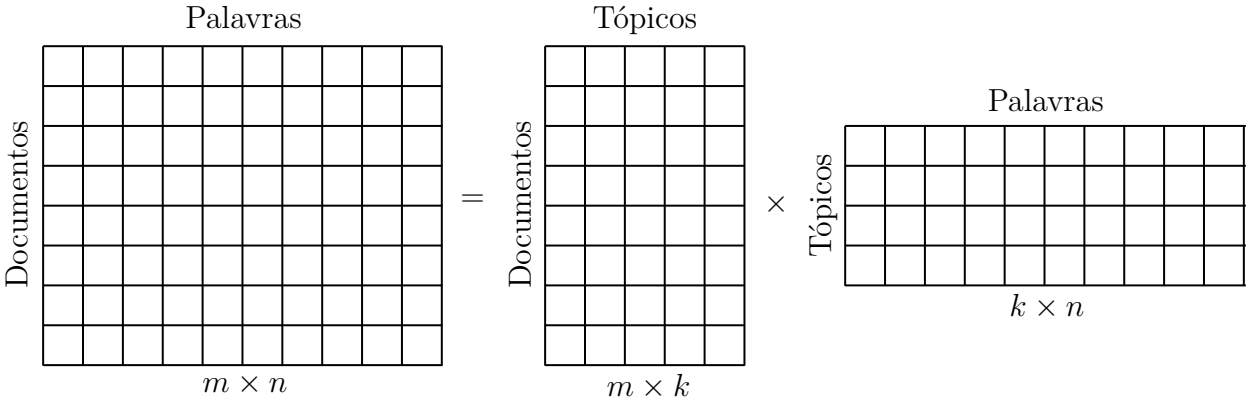


Figura 2.4 – Processo da NMF.

Fonte: (MURSHED et al., 2022).

Um exemplo de utilização de NMF foi o trabalho de Sharaff e Nagwani (2016), que o aplicaram em e-mails para identificar tópicos. A partir dos resultados é mostrado que o algoritmo NMF tem melhor desempenho que outros dois modelos também utilizados.

Essas três técnicas foram avaliadas em termos de todos os parâmetros de desempenho para identificação de tópicos usando a técnica de similaridade baseada em assunto.

2.4.5 BERTopic

Como último modelo para experimentação, comparação e análise em textos curtos do Twitter, lançamos mão de um modelo mais recente que utiliza métodos estado da arte para tarefas de processamento de linguagem natural. Esse é o BERTopic, uma variação de Representações de Codificador Bidirecional de Transformadores (BERT, do inglês Bidirectional Encoder Representations from Transformers)(DEVLIN et al., 2018) e proposto por Grootendorst (2022b).

Modelos convencionais, como LDA e NMF, e também o BTM, representam um documento como um *bag of words*[4] e modelam cada documento como uma mistura de temas latentes. Essa representação possui limitação ao desconsiderar as relações semânticas entre as palavras por não levar em conta o contexto delas numa sentença (GROOTENDORST, 2022b).

Uma maneira de lidar com esse desafio emergiu com o surgimento de técnicas de representação de texto, que se tornaram rapidamente populares no campo do Processamento de Linguagem Natural. O BERT e suas variações, em particular, têm sido bastante eficazes na criação de representações de vetores de palavras e sentenças contextualmente relevantes. As características semânticas dessas representações vetoriais permitem codificar o significado do texto de tal forma que textos semelhantes estejam próximos no espaço vetorial (GROOTENDORST, 2022b).

Como resposta a esta questão, as técnicas de incorporação de texto tornaram-se rapidamente populares no campo do Processamento de Linguagem Natural. Mais especificamente, as representações de codificador bidirecional de transformadores (BERT) e suas variações, mostraram grande resultado na geração de representações contextuais de vetores de palavras e sentenças.

BERTopic é um modelo de linguagem baseado em transformer pré-treinado que gera representações de documentos, as agrupa e, ao fim, gera representações de tópicos com o procedimento TF-IDF (do inglês *Term Frequency - Inverse Document Frequency*) (AIZAWA, 2003) baseado em classe.

Descrevendo de forma mais detalhada: empregam-se três etapas independentes para obter representações de tópicos flexíveis que possam ser utilizadas em diferentes casos de uso. Primeiro, são geradas incorporações de documentos utilizando um modelo de

[4] A abordagem *bag of words* (BoW) tem como representação a transformação de um documento em um vetor de tamanho constante, composto por todos os termos (palavras ou n-gramas) encontrados no conjunto de textos. Cada termo recebe uma ponderação numérica, que pode ser determinada pela sua frequência, por exemplo (LI et al., 2020).

linguagem pré-treinado, visando obter informações em nível de documento. Em seguida, é reduzida a dimensionalidade das incorporações de documentos e agrupados semanticamente documentos similares em tópicos distintos. Por fim, desenvolve-se uma versão baseada em classe do TF-IDF para extrair a representação de cada tópico, superando assim a abordagem baseada em centroide. Com essas etapas, pode-se obter um modelo de tópico flexível, que pode ser utilizado em diversos casos de uso, incluindo modelagem de tópico dinâmico (GROOTENDORST, 2022b). Na figura 2.5 podem ser observadas as etapas realizadas pelo modelo mais outras opcionais.

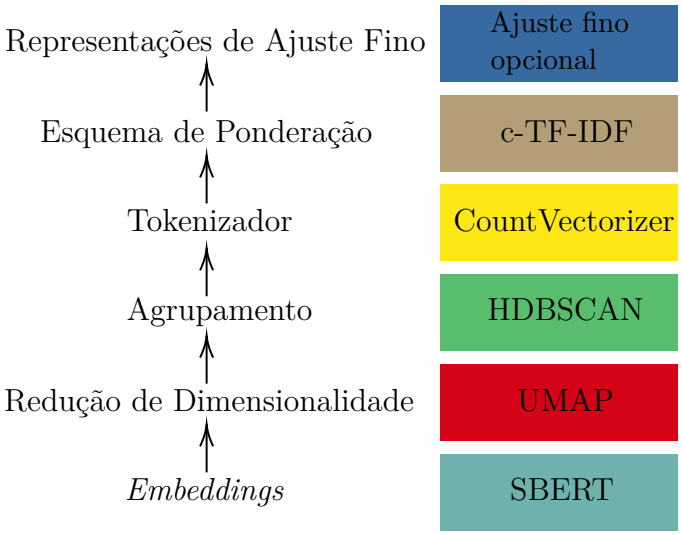


Figura 2.5 – Algoritmo do BERTopic.

Fonte: https://maartengr.github.io/BERTopic/algorithm/algorithm.html

2.5 Métricas de Avaliação

Apesar da dificuldade de avaliar modelos não-supervisionados, particularmente em aplicações nas línguas naturais humanas, existem métricas com esse objetivo disponíveis. Duas delas serão discutidas a seguir.

2.5.1 Coerência de Tópicos

Coerência de tópicos é uma métrica estado da arte para LDA no contexto de corpora textuais genéricos (PANICHELLA, 2021). Ela foi proposta por Mimno et al. (2011), e permite que se mensure a qualidade dos tópicos gerados por modelos.

A informação mútua pontual (PMI, do inglês *pointwise mutual information*) é a métrica de coerência mais amplamente utilizada. Embora existam muitas variantes do PMI, em sua essência, essa métrica visa medir a proximidade das palavras em cada tópico com base em suas coocorrências relativas entre si. Sua popularidade cresceu porque

o cálculo do PMI não requer um conjunto de tópicos corretos como base de comparação (CHURCHILL; SINGH, 2022). PMI é calculada de acordo com a seguinte equação:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_i)}\right), \tag{2.7}$$

Considerando a equação, $P(w_i, w_j)$ representa a frequência ou probabilidade de observar as palavras w_i e w_j juntas em uma mesma janela.[5] $P(w_i)$ e $P(w_j)$ representam, respectivamente, a frequência ou probabilidade de observar as palavras w_i e w_j separadamente. Quanto mais próxima a frequência de co-ocorrência de duas palavras estiver da frequência de ocorrência individual das duas palavras, melhor será a pontuação atribuída ao par de palavras fornecido. A constante ϵ pode ser utilizada para evitar a ocorrência de um logaritmo com argumento igual a zero (CAMPAGNOLO; DUARTE; BIANCO, 2022).

A coerência de tópico C_V é uma versão de PMI com correlação mais forte com avaliações humanas (RöDER; BOTH; HINNEBURG, 2015), o que justifica sua escolha para avaliar os modelos deste trabalho. De acordo com a definição de Röder, Both e Hinneburg (2015), C_V é uma medida que combina a similaridade do cosseno e o PMI normalizado (NPMI) com uma janela deslizante booleana com tamanho de 110 palavras. Essa medida busca capturar tanto a proximidade entre as palavras quanto a informação mútua e a similaridade vetorial (CHURCHILL; SINGH, 2022). A equação de NPMI é dada por:

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(P(w_i, w_j))}, \tag{2.8}$$

De acordo com Campagnolo, Duarte e Bianco (2022), uma janela deslizante é um subconjunto de um determinado número de palavras consecutivas e que pode ser movido palavras por palavra para qualquer um dos lados, i.e., uma janela deslizante é uma estratégia de segmentar palavras de um documento, ajudando a identificar como elas aparecem próximas ou distantes umas das outras em relação a segmentação.

C_V utiliza uma variação de NPMI para calcular a coerência sobre uma janela deslizante, computando a co-ocorrência de uma palavra de um dado tópico contra todas as palavras de um mesmo tópico (CAMPAGNOLO; DUARTE; BIANCO, 2022). Para obter C_V, primeiro se calcula vetores entre as palavras de acordo com a seguinte fórmula:

$$v_{w_i,w_j} = NPMI(w_i, w_j)^{\gamma} = \left(\frac{PMI(w_i, w_j)}{-\log(P(w_i, w_j)) + \epsilon}\right)^{\gamma}, \tag{2.9}$$

em que γ é o peso usado para dar mais força para palavras mais associativas. Depois, é calculado um vetor para cada palavra:

[5] O termo "janela" corresponde a um subconjunto de palavras obtido por uma estratégia de segmentação de um documento (CAMPAGNOLO; DUARTE; BIANCO, 2022).

$$\vec{v}_{w_i} = \{NPMI(w_i, w_0)^\gamma, ..., NPMI(w_i, w_n)^\gamma\} \tag{2.10}$$

em que w_n é a última palavra e n é número de palavras no tópico. Por fim, as distâncias dos vetores são medidas usando similaridade de cosseno:

$$\vec{v}_c = \vec{v}_{w_0} + ... + \vec{v}_{w_n} \tag{2.11}$$

$$C_V = \frac{1}{n} \cdot (cos(\vec{v}_{w_0}, \vec{v}_c) + ... + cos(\vec{v}_{w_n}, \vec{v}_c)) \tag{2.12}$$

Portanto, essa é a definição de C_V, em que quanto maior seu valor, melhor o desempenho.

Um exemplo de uso de métrica C_V para avaliar o desempenho de modelos é o trabalho de Gui et al. (2019). Nele há a proposta de um modelo neural de tópicos com aprendizado por reforço que supera - nessa métrica e em outra - modelos neurais de tópicos estado da arte.

A ferramenta utilizada para cálculo de coerência de tópico C_V foi a classe *CoherenceModel* da biblioteca Gensim (ŘEHůřEK; SOJKA, 2010). Nela é possível introduzir os tópicos gerados pelos modelos e seus textos correspondentes para obter a métrica.

2.5.2 Distância de Jaccard

Essa métrica, que mede a dissimilaridade entre conjuntos, é complementar ao coeficiente de Jaccard, desenvolvido por Grove Karl Gilbert em 1884 (MURPHY, 1996) e depois, independentemente, por Paul Jaccard em 1912 (KOSUB, 2016). O coeficiente de Jaccard pode ser calculado como:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.13}$$

E a distância de Jaccard (d_J) como:

$$d_J = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{2.14}$$

Essa métrica indica que quanto maior o valor dela, mais diferentes são os conjuntos. No caso deste trabalho, usamos para comparar os tópicos gerados por diferentes modelos LDA nos experimentos em tamanho de amostra. Para a comparação, é computada, então, uma média das distâncias entre os tópicos dos modelos para medir o quão diferentes eles são. A distância de Jaccard foi calculada por meio da implementação presente no

próprio modelo LDA da classe LdaMulticore da biblioteca Gensim (ŘEHůřEK; SOJKA, 2010).

3 METODOLOGIA

Este capítulo apresenta a metodologia empregada neste trabalho do ponto de vista de: avaliação da quantidade dos dados; coleta, pré-processamento dos dados e inserção dos dados nos modelos para análise de sensibilidade aos parâmetros e análise de desempenho. As seções estabelecem uma sequência e detalham cada uma das etapas.

3.1 Delimitação Baseada no Idioma e Contagem de Tweets

O primeiro passo da metodologia consistiu em avaliar de que maneira seria possível concentrar nossa análise no conteúdo produzido no contexto brasileiro, uma vez que esse enfoque era uma motivação central da pesquisa. Embora o Twitter possua um mecanismo de busca que inclui a localização apontada por quem posta uma mensagem, percebemos que havia um número significativo de mensagens marcadas como tendo sido postadas no Brasil que não eram relevantes para nós, como mensagens postadas por visitantes atraídos por eventos como a Copa do Mundo de 2014 ou a Olimpíada de 2016.

Ponderamos então que poderíamos utilizar como crivo o idioma português, uma vez que tínhamos a percepção de que o número de usuários brasileiros seria amplamente majoritário nesse recorte lusófono. Essa ponderação foi confirmada pelo ranking trazido na figura 3.1 e também pela consideração de que o número de usuários brasileiros (19,05 milhões) é mais do que 13 vezes o número de usuários de Portugal (1,40 milhões),[1] terceiro país com maior quantidade de pessoas lusófonas no mundo.[2] Angola como segundo país com mais falantes de português, possui apenas 71,4 mil usuários no Twitter em 2022.[3]

Tendo validado o uso do idioma português como uma hipótese de trabalho razoável, fizemos as contagens de *tweets* em português sobre assuntos políticos, econômicos e sociais em português desde a criação do Twitter, em 26 de março de 2006, a 31 de dezembro de 2021. O filtro de busca para cada assunto consistia em o *tweet* conter, no seu texto, a "palavra-chave" e/ou "#palavra-chave" referente a ele. Por exemplo, o texto conter "desemprego" e/ou "#desemprego". As palavras-chaves do filtro de busca foram: desemprego, inflação, fome, capitalismo, socialismo, comunismo, liberalismo e neoliberalismo. Essas palavras foram pensadas como exemplos de assuntos políticos, econômicos e sociais, sendo, assim, prováveis objetos de estudo para uma pesquisa na área.

É importante mencionar que o filtro de busca da API do Twitter é indiferente

[1] https://datareportal.com/reports/digital-2022-portugal#: :text=Numbers%20published%20in%20Twitter's%20advertising,total%20population%20at%20the%20time.
[2] https://www.worlddata.info/languages/portuguese.php
[3] https://datareportal.com/reports/digital-2022-angola

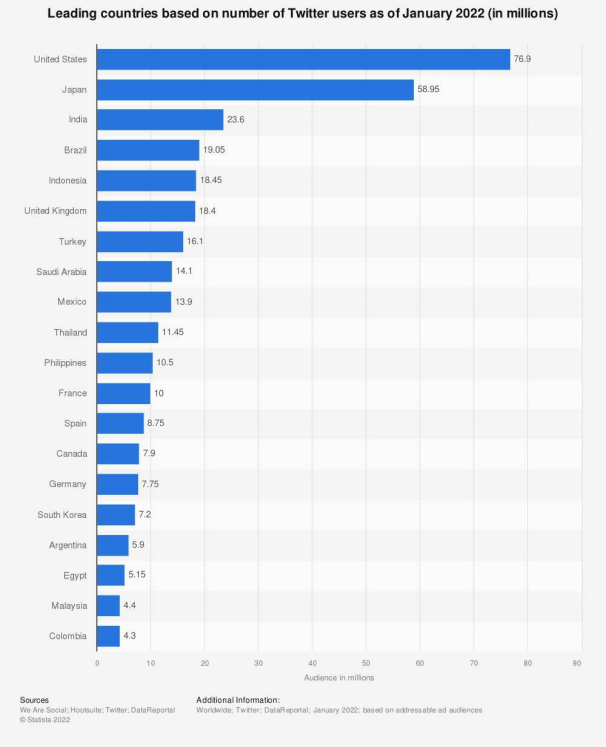


Figura 3.1 – Países com o maior quantidade de usuários em 2022.

Fonte: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

em relação a letras maiúsculas e minúsculas, abrangendo todas as variações possíveis para as palavras escolhidas para a contagem.

Essas contagens totais de palavras-chaves serão apresentadas no capítulo 4 por meio da tabela 4.1. Além disso foram realizadas contagens de *tweets* ao longo dos anos. Elas podem ser observadas no capítulo 4 por meio dos gráficos das figuras 4.1, 4.2, 4.3, 4.4 e 4.5. Após terem sido realizadas as contagens, passamos às coletas dos textos dos *tweets*.

3.2 Coleta de Dados do Twitter

Nas coletas, o tamanho de amostra é uma questão fundamental: é necessário investigar a quantidade de *tweets* que podem ser representativos de determinado assunto que se quer estudar. Entretanto, a distribuição estatística dos dados deste trabalho era-nos desconhecida *a priori*. Não havia, assim, um modelo que determinasse um tamanho de amostra que pudesse garantir uma representatividade de uma população para uma definida confiança e margem de erro. Há, portanto, uma dificuldade em se definir a quantidade de *tweets* para que se possa trabalhar.

Uma estratégia apresentada por Krippendorff (2018) é realizar experimentos de amostragem para descobrir um tamanho do conjunto. A primeira coleta adotada neste trabalho consistiu em obter três amostras de 50 mil tweets no ano de 2021 e realizar

comparações entre elas para analisar a adequação do tamanho. As três amostras coletadas foram compostas pela palavra-chave "comunismo", isto é, continham "comunismo" e/ou "#comunismo". A opção por essa palavra se dá por sua centralidade no discurso da extrema-direita, que ocupou grande espaço no debate político brasileiro nos últimos dez anos (MACHADO; COLEVATI, 2021) e também desempenha função de ser um exemplo de aplicação da metodologia desenvolvida neste trabalho de mestrado em um assunto social, político e econômico. Por conta dessa opção, as outras palavras-chaves ficam como oportunidade para estudos futuros.

Antes da coleta, também foi feita uma visualização da quantidade diária de *tweets* sobre esse assunto em 2021. Esse período foi selecionado por ser o ano completo mais recente em relação à data da coleta, que ocorreu em 2022, e, assim, representar a perspectiva mais atual possível. No capítulo 4, a figura 4.10 apresenta esse comportamento.

Para realizar a coleta, utilizamos a API do Twitter.[4] Como ela retorna no máximo 500 *tweets* por resposta, a cada conjunto com esse máximo retornado, eram fornecidos à API uma data e um horário pseudoaleatórios no ano de 2021 para seguir uma amostragem aleatória simples de acordo com Krippendorff (2018), visto que todos *tweets* possuem a mesma probabilidade de serem incluídos na amostra. Todavia, nem sempre o máximo de 500 foi retornado por haver *tweets* indisponíveis.

Outra questão é que o Twitter possui diferentes tipos de *tweets*,[5] eles são: *tweets* gerais, *tweets* de status, *retweets*, *replies*, menções e comentários. A coleta abrangeu todos esses tipos.

Coletados os *tweets* para as 3 amostras, foram realizadas inspeções das distribuições em relação à quantidade diária ao decorrer do ano de 2021 para verificar se estavam bem espalhadas ao longo desse período e pudesse ser representativo para esse ano. No capítulo 4, as figuras 4.11, 4.12 e 4.13 exibem essas distribuições para as amostras.

3.3 Pré-Processamento

Inspecionadas as distribuições das amostras, realizamos nos textos dos *tweets* um pré-processamento de acordo com o indicado em Allahyari et al. (2017). Ele consistiu nas seguintes etapas: 1) tratar os caracteres e outros processos; 2) quebrar o texto em unidades individuais (processo conhecido como *tokenizing* em inglês); 3) remoção de palavras de parada (conhecidas como *stopwords* em inglês); 4) lematização; por fim, 5) outra remoção de palavras de parada geradas pela lematização.

No tratamento dos caracteres, houve a remoção dos itens não-alfanuméricos[6] e

[4] https://developer.twitter.com/en/docs/twitter-api
[5] https://help.twitter.com/en/using-twitter/types-of-tweets
[6] Caracteres que não são letras latinas.

algarismos indo-arábicos. Retiramos usuários mencionados no texto e URLs. Por meio de uma lista manual, substituímos as abreviações por suas palavras completas correspondentes.

Juntamente com a quebra dos textos em unidades individuais (*tokenizing*), todas as letras foram transformadas em minúsculas e removemos as pontuações.

Palavras de parada são aquelas muito frequentes em um texto e que não possuem um valor semântico muito grande, como preposições e conjunções. A remoção delas foi feita por meio da lista da biblioteca NLTK (BIRD; KLEIN; LOPER, 2009) e de outras palavras adicionadas manualmente a essa lista.

A lematização consiste em transformar palavras em sua forma mais simples, como colocá-las no singular, masculino ou, no caso de verbo, colocar no infinitivo, i.e., agrupar as várias formas flexionadas de uma palavra para que possam ser analisadas como um único item considerando a análise morfológica (ALLAHYARI et al., 2017). Esse processo foi realizado utilizando um modelo treinado de rede neural convolucional da biblioteca spaCy (HONNIBAL; MONTANI, 2017), sendo a versão pequena desse modelo a usada por conta de seu menor custo computacional e de memória.

Ao fim do pré-processamento, foi realizada mais uma remoção de palavras de parada porque mais delas foram geradas pela lematização.

3.4 Experimentos em Tamanho de Amostra com LDA

Foram realizados três experimentos com o modelo LDA (apresentado no capítulo 2) para avaliar o tamanho de amostra. Eles estão na subseções a seguir, LDA nas Amostras de 50 mil tweets, LDA com Técnica de Divisão pela Metade e LDA com Aumento de Tamanho de Amostra.

Uma razão por trabalhar com LDA, além de ser o modelo mais clássico para modelagem de tópicos, é por sua implementação ser a mais rápida em tempo de execução de código. Ela se chama *LdaMulticore* e pertence à biblioteca Gensim (ŘEHůřEK; SOJKA, 2010).

Todos os experimentos em tamanho de amostra com LDA foram realizados com a configuração de cinco tópicos, alfa e eta simétricos de acordo com a equação 3.1.

3.4.1 LDA nas Amostras de 50 mil tweets

Pré-processadas as três amostras, aplicamos o modelo LDA apresentado na seção 2.4.2 com cinco tópicos para cada uma das três amostras em busca de comparar seus resultados e analisar a representatividade do tamanho de amostra.

Entretanto, a comparação entre tópicos gerados por diferentes modelos de LDA

é complexa devido aos seguintes fatores:

- caráter estocástico do modelo, sendo necessário executar várias vezes cada modelo e fazer uma comparação entre médias;
- caráter não supervisionado do modelo, não existindo, portanto, tópicos gerados esperados e sendo necessária uma métrica adequada para análise do desempenho;
- necessidade de escolher seus hiperparâmetros para se obter tópicos gerados ótimos;
- necessidade de um especialista no tema dos tweets para comparar melhor os tópicos gerados pelos modelos.

Visto essas complexidades, é necessária uma métrica adequada para análise de desempenho do modelo. A utilizada foi a métrica apresentada na subseção 2.5.1 do capítulo 2, a coerência de tópicos C_V, indicada como a versão com correlação mais forte com avaliações humanas (RöDER; BOTH; HINNEBURG, 2015). Sendo essa métrica utilizada, primeiramente, para avaliar o tamanho de amostra e, depois, também para o desempenho dos outros modelos (BTM, BERTopic e NMF) que foram utilizados para assim conseguir ser possível realizar uma comparação entre todos os modelos.

Como comparação entre esses 3 modelos LDA, utilizamos a média de coerência de tópicos C_V de 10 execuções para cada amostra como métrica. Resultando, portanto, em 30 execuções ao todo. Na tabela 4.2 do capítulo 4 se encontram os resultados dos cálculos.

Também utilizamos a distância de Jaccard entre os tópicos gerados pelos diferentes modelos em cada execução e calculamos a média de distância para avaliar o tamanho de amostra. No capítulo 4, os resultados estão presentes na tabela 4.3. Essa métrica foi utilizada somente nas análises de tamanho de amostra.

3.4.2 Técnica de Divisão pela Metade

Essa é uma técnica proposta por Krippendorff (2018), também chamada de *Split-Half Technique*, em que dividem-se as amostras ao meio para analisar se elas apresentam o mesmo comportamento em relação ao conjunto inteiro. Realizamos essa técnica nas três amostras.

Divididas as amostras, executamos 10 vezes o modelo LDA com 5 tópicos em cada metade. Realizamos novamente o cálculo de coerências C_V e da distância de Jaccard entre os tópicos gerados pelas metades. Nas tabelas 4.4 e 4.5 do capítulo 4 se encontram as médias e desvios padrão das execuções.

3.4.3 Experimento de Aumento de Tamanho de Amostra

Como o comportamento das amostras mudaram ao dividir pela metade, realizamos um experimento de amostragem também proposto por Krippendorff (2018), que é realizar outra coleta com tamanho diferente de amostra para comparação.

Fizemos 3 coletas de 100 mil *tweets*, o dobro do tamanho de amostra inicial. O intuito foi avaliar as médias de coerência C_V e distância de Jaccard entre os tópicos para analisar se esse tamanho de amostra altera o comportamento dessas métricas em relação ao tamanho de amostra inicial. Também foram realizadas 10 execuções para cada amostra. Os resultados dos cálculos das médias e desvios padrão para cada amostra estão nas tabelas 4.6 e 4.7 do capítulo 4.

3.5 Escolha dos Hiperparâmetros do LDA

Terminados os experimentos de tamanho de amostra, decidimos trabalhar com 50 mil *tweets* por ter obtido uma coerência C_V maior e ter um custo computacional menor do que 100 mil *tweets*.

Como o modelo LDA e seus métodos de otimização apresentam vários hiperparâmetros, é necessário realizar a afinação, isto é, buscas para encontrar aqueles que apresentam o melhor desempenho.

Antes dessa afinação é necessário definir quais hiperparâmetros são mais relevantes para o desempenho, visto que uma busca de todos se torna inviável computacionalmente. Logo, realizamos a afinação de hiperparâmetros escolhidos e trabalhados na literatura (PANICHELLA, 2021). Esses são: número de tópicos, alfa e eta.

Primeiro realizamos uma busca grossa, ou seja, com passos grandes de valores para ter uma amplitude maior do espaço de busca. Depois realizamos uma busca fina, isto é, com passos pequenos de números de tópicos enquanto os valores de alfa e eta estavam fixos.

Baseado em Panichella (2021), os valores de hiperparâmetros na busca grossa foram:

- número de tópicos (k): 2, 42, 82, 122, 162
- alfa: ‘symmetric’, ‘asymmetric’, 0.01, 0.1, 1, 10
- eta: ’symmetric’, ‘auto’, 0.1, 1, 10

Sendo ‘symmetric’ equivalente ao inverso de k, ‘asymmetric’ equivalente ao inverso da soma do índice do tópico com a raiz quadrada de k e ‘auto’ há um aprendizado assimétrico do corpus. Essas equivalências são apresentadas nas equações 3.1 e 3.2.

$$symmetric = \frac{1}{k} \tag{3.1}$$

$$asymmetric(t) = \frac{1}{t + \sqrt{k}} \tag{3.2}$$

em que t é o índice do tópico, assumindo valores de 0 a $k - 1$.

Os perfis resultantes das buscas em gráficos estão nas figuras 4.14, 4.15 e 4.16.

Os valores de hiperparâmetros na busca fina foram:

- número de tópicos: [2, 4, 6, ..., 158, 160, 162];
- alfa e eta com maiores valores encontrados de coerência C_V na busca grossa como também combinações propostas na literatura.

Nas tabelas 4.8, 4.9 e 4.10 do capítulo 4 estão as combinações de número de tópicos, alfa e eta com maiores valores encontrados de coerência C_V para cada busca fina de cada amostra.

3.6 Aplicação do BTM nas amostras e escolha dos hiperparâmetros

Após os experimentos com o LDA, realizamos a afinação do modelo BTM para as 3 amostras. A justificativa de escolha desse modelo foi por ter obtido melhor desempenho em Jónsson (2016).

Nesse processo foi necessário também definir quais hiperparâmetros buscar. Para essa questão utilizamos como referência Jónsson (2016), em que os hiperparâmetros buscados foram: número de tópicos, alfa e beta (equivalente a eta).

As combinações de valores testadas também foi como Jónsson (2016), i.e., com as seguintes configurações:

- número de tópicos (k): [10 50 100 200];
- alfa: [$1/k$ $50/k$ $100/k$];
- beta: [0,001 0,01 0,5]

Os valores de hiperparâmetros com maior coerência C_V estão na tabela 4.12 do capítulo 4.

Na aplicação do modelo BTM neste trabalho, utilizamos o pacote bitermplus[7] que implementa Yan et al. (2013) em Cython.

[7] https://bitermplus.readthedocs.io/en/stable/index.html

3.7 Aplicação do BERTopic nas amostras e escolha dos hiperparâmetros

Com o BERTopic, aplicamos a modelagem de tópicos para as amostras. Na tabela 4.13 do capítulo 4 estão os números de tópicos encontrados e respectivas coerências para cada amostra.

Em busca de um número de tópicos com maior coerência de tópico C_V, realizamos a redução desse número. No capítulo 4, os gráficos representando essas buscas se encontram na figura 4.17 e os valores máximos de coerência C_V dos gráficos da figura 4.17 estão na tabela 4.14 com seus respectivos números de tópicos.

Foi utilizada a implementação presente no repositório Grootendorst (2022a) para aplicar o modelo BERTopic.

3.8 Aplicação do NMF nas amostras e escolha dos hiperparâmetros

Depois dos experimentos com o BERTopic, realizamos a escolha de hiperparâmetros do último modelo, o NMF. A escolha consistiu em buscar o número de tópicos que obtivessem a maior coerência de tópicos C_V para cada uma das três amostras. Essas buscas estão representadas nos gráficos da figura 4.18 do capítulo 4. Nesse mesmo capítulo, a tabela 4.15 apresenta os números de tópicos com as máximas coerências de tópicos C_V encontradas.

A implementação do NMF utilizada foi a *nmf*, presente na biblioteca Gensim (ŘEHůřEK; SOJKA, 2010), que utiliza o algoritmo incremental eficiente de Zhao e Tan (2017). A configuração de todos os hiperparâmetros (exceto o número de tópicos) seguiram o padrão na implementação.[8]

[8] Os hiperparâmetros do padrão são: corpus = None, id2word = None, chunksize = 2000, passes = 1, kappa = 1.0, minimum_probability = 0.01, w_max_iter = 200, w_stop_condition = 0.0001, h_max_iter = 50, h_stop_condition = 0.001, eval_every = 10, normalize = True, random_state = None.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados obtidos a partir da aplicação da metodologia apresentada no capítulo 3, bem como a sua análise. As seções obedecem a uma sequência de explosição análoga à adotada para a metodologia.

4.1 Contagem de *Tweets*

Seguindo o fio condutor do conteúdo da seção 3.1, apresentamos, na tabela 4.1, os resultados obtidos para as contagens totais de *tweets* contendo o conjunto pré-determinado de palavras-chave.

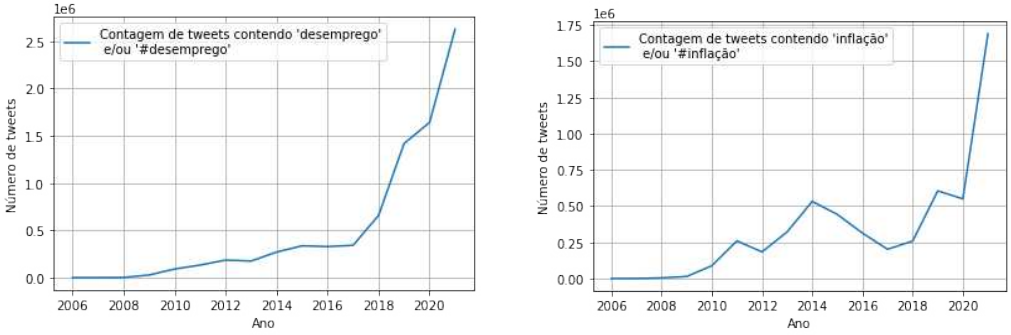
Tabela 4.1 – Contagens totais de *tweets* contendo as "palavras-chaves" e/ou "#palavras-chaves" no período especificado.

"palavra-chave" e/ou "#palavra-chave"	Contagem de *tweets* de 26/03/2006 a 31/12/2021
Desemprego	8.501.164
Inflação	5.875.542
Fome	108.330.899
Capitalismo	6.368.290
Socialismo	4.486.471
Comunismo	7.808.580
Liberalismo	1.352.170
Neoliberalismo	811.017
Total de *tweets* em português	53.998.570.272

Os valores associados a cada palavra-chave possuem uma ordem de grandeza plausivel, uma vez que os números totais do Twitter estavam em cerca de 200 bilhões de postagens anuais em agosto de 2022.[1]

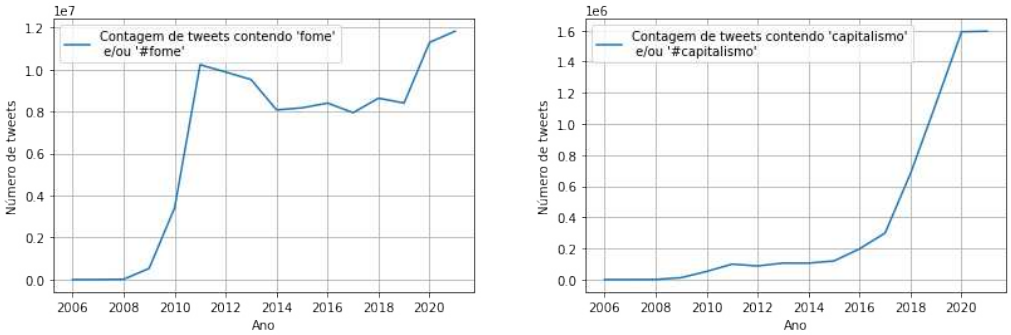
Os resultados das contagens de *tweets* por ano estão presentes nas figuras 4.1, 4.2, 4.3, 4.4 e 4.5. Nelas, é observável um comportamento geral de crescimento do número de *tweets* contendo as palavras-chaves selecionadas, no decurso dos anos, por conta do aumento do número de usuários do Twitter. Há também particularidades em relação a cada palavra-chave, como períodos de rápido crescimento e picos, que, de alguma forma, se relacionam com a situação do debate político brasileiro. Embora, numa primeira análise, tenhamos discutido algumas hipóteses explicativas, como o itinerário que leva das chamadas "Jornadas de Junho" de 2013 ao impeachment da presidenta Dilma Rousseff e, em seguida, à ascensão de Jair Bolsonaro (FARIA, 2021), julgamos que qualquer conclusão deve ser objeto de um trabalho cuidadoso que transcende o escopo desta dissertação. Análises mais aprofundadas podem ser feitas, futuramente, com o apoio de especialistas.

[1] https://www.dsayce.com/social-media/tweets-day/



(a) (b)

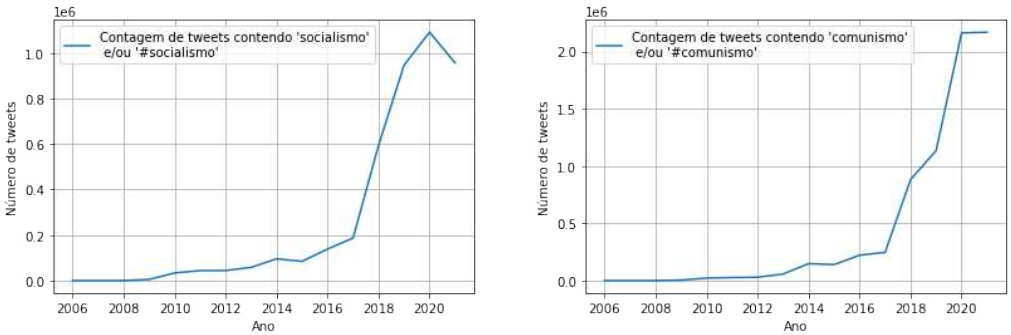
Figura 4.1 – Contagens de *tweets* com 'palavra-chave' e/ou '#palavra-chave': (a) desemprego, (b) inflação



(a) (b)

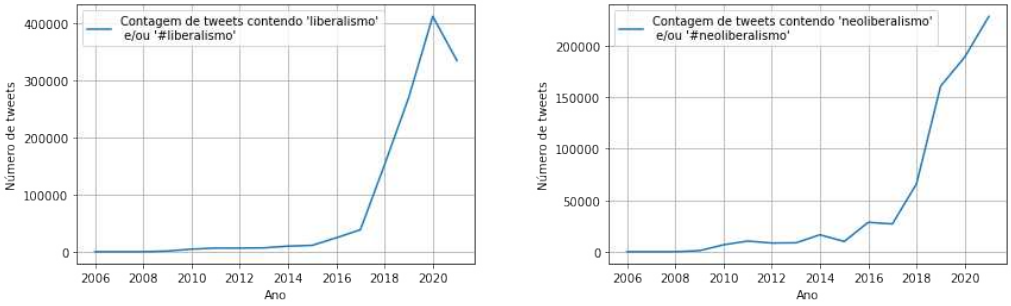
Figura 4.2 – Contagens de *tweets* com 'palavra-chave' e/ou '#palavra-chave': (a) fome, (b) capitalismo

Na figura 4.5, cabe destacar um comportamento curioso: a quantidade de *tweets* em português declinou entre os anos de 2012 e 2015, permaneceu aproximadamente constante entre 2016 e 2017, e somente retomou o aumento a partir do ano de 2018. Esse comportamento, novamente, poderá ser objeto de futuros estudos suportados por



(a) (b)

Figura 4.3 – Contagens de *tweets* com 'palavra-chave' e/ou '#palavra-chave': (a) socialismo, (b) comunismo



(a) (b)

Figura 4.4 – Contagens de *tweets* com 'palavra-chave' e/ou '#palavra-chave': (a) liberalismo, (b) neoliberalismo

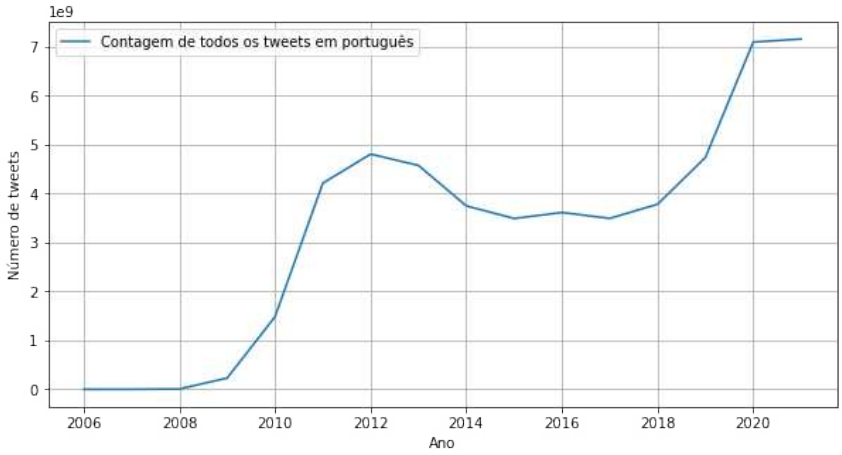
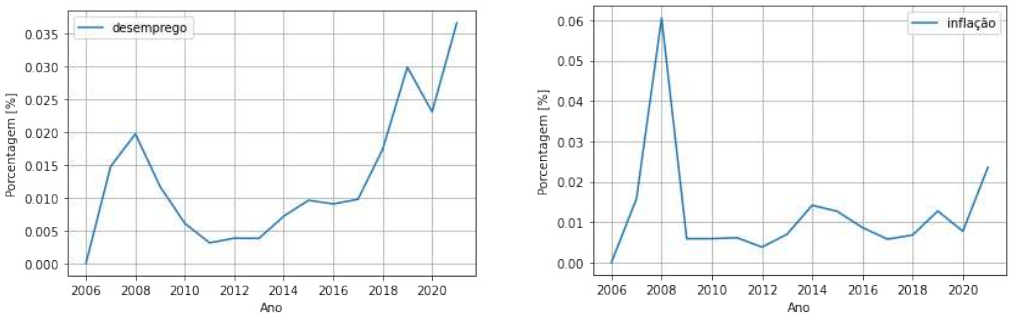


Figura 4.5 – Contagem total de *tweets* em português.

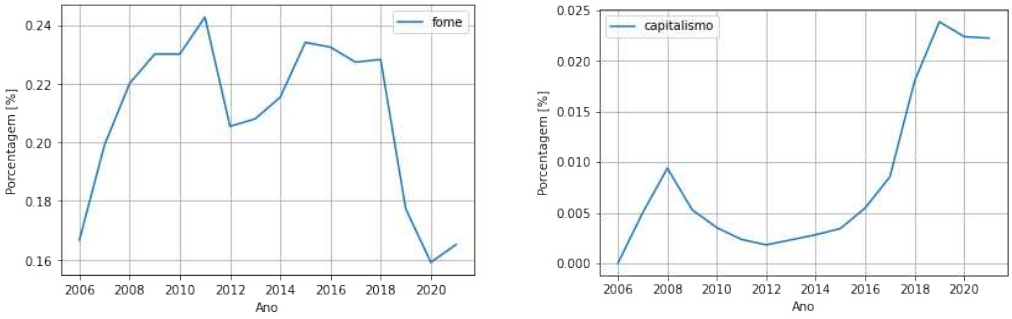


(a) (b)

Figura 4.6 – Porcentagem de tweets contendo 'palavra-chave' e/ou '#palavra-chave' em relação ao total de tweets em português: (a) desemprego, (b) inflação

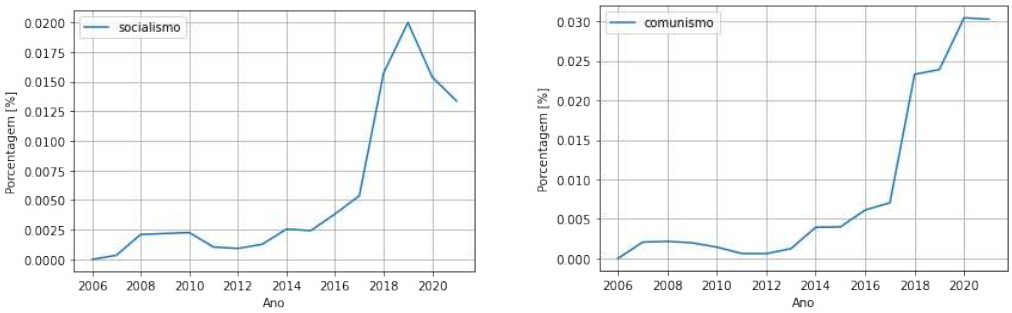
pesquisadores(as) das humanidades.

Nas figuras 4.6, 4.7, 4.8 e 4.9, estão os gráficos apresentando as contagens de *tweets* contendo as palavras-chaves em porcentagem em relação ao total de *tweets* em português. Com esses gráficos, é possível observar o comportamento dessas contagens proporcionalmente ao número de usuários e *tweets* no decorrer dos anos.



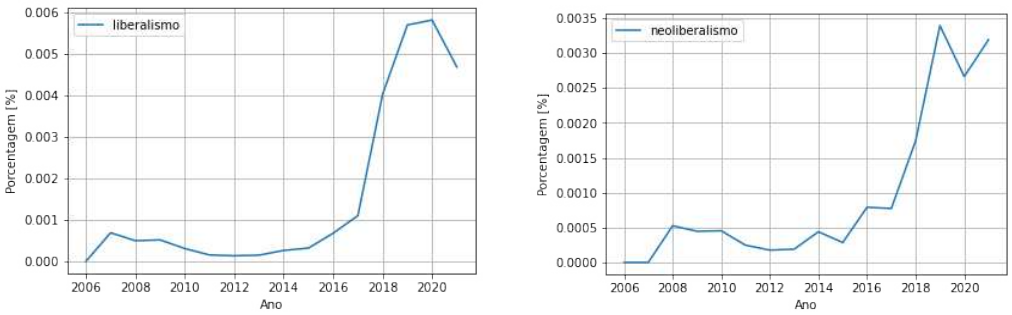
(a) (b)

Figura 4.7 – Porcentagem de tweets contendo 'palavra-chave' e/ou '#palavra-chave' em relação ao total de tweets em português: (a) fome, (b) capitalismo



(a) (b)

Figura 4.8 – Porcentagem de tweets contendo 'palavra-chave' e/ou '#palavra-chave' em relação ao total de tweets em português: (a) socialismo, (b) comunismo



(a) (b)

Figura 4.9 – Porcentagem de tweets contendo 'palavra-chave' e/ou '#palavra-chave' em relação ao total de tweets em português: (a) liberalismo, (b) neoliberalismo

Ao visualizar os gráficos das figuras 4.6, 4.7, 4.8 e 4.9, é possível observar comportamentos curiosos. Alguns continuaram apresentando crescimento como aqueles contendo as palavras-chaves “comunismo” e, aproximadamente, “neoliberalismo”. Outros obteram oscilações maiores como “desemprego”, “capitalismo”, “socialismo”, e “liberalismo”. “fome” obteve um aumento e depois voltou às porcentagens de início. O termo “inflação” obteve um grande aumento nos anos iniciais e depois se estabilizou em porcentagens próximas aos anos iniciais. Todos esse comportamentos podem ser importantes objetos de estudos futuros para especialistas no tema.

Na figura 4.10, mostram-se as quantidades diárias obtidas de *tweets* que contêm a palavra-chave “comunismo” e/ou “#comunismo” no decorrer de 2021. Em termos gerais, a quantidade diária de *tweets* não ultrapassa a marca de 10 mil. Contudo, notam-se flutuações pontuais intensas ao longo do ano (principalmente as que ultrapassam a marca de 20 mil), as quais são provavelmente explicadas por eventos ocorridos nos dias correspondentes. Por exemplo, o primeiro dia que ultrapassou 10 mil *tweets* foi 06/01/2021, dia em que ocorreu a invasão do Capitólio nos Estados Unidos da América (ROCHA; DUWE, 2022).

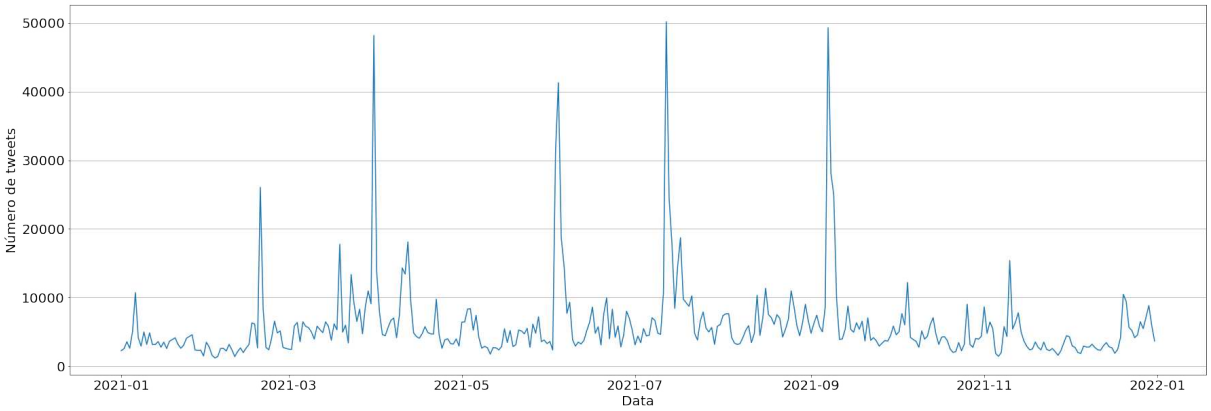


Figura 4.10 – Quantidades diárias de *tweets* que contêm ‘comunismo’ e/ou ‘#comunismo’.

4.2 Coleta de Dados no Twitter

Nas figuras 4.11, 4.12 e 4.13, estão os resultados referentes aos *tweets* coletados das três amostras. Nelas, representam-se as distribuições dos *tweets* coletados ao longo de 2021. A linha vermelha indica a média de todas as contagens: nas três amostras, seu valor foi um pouco inferior a 500 *tweets*, o que significa que houve poucos *tweets* que não foram enviados pela API em cada conjunto de 500. Isso geralmente ocorre por conta de usuários que configuram seu conteúdo como privado. As barras azuis representam as quantidades diárias de *tweets* que foram coletados, sendo que alguns tiveram mais de uma coleta devido ao fornecimento pseudo-aleatório de datas para colher a resposta da API.

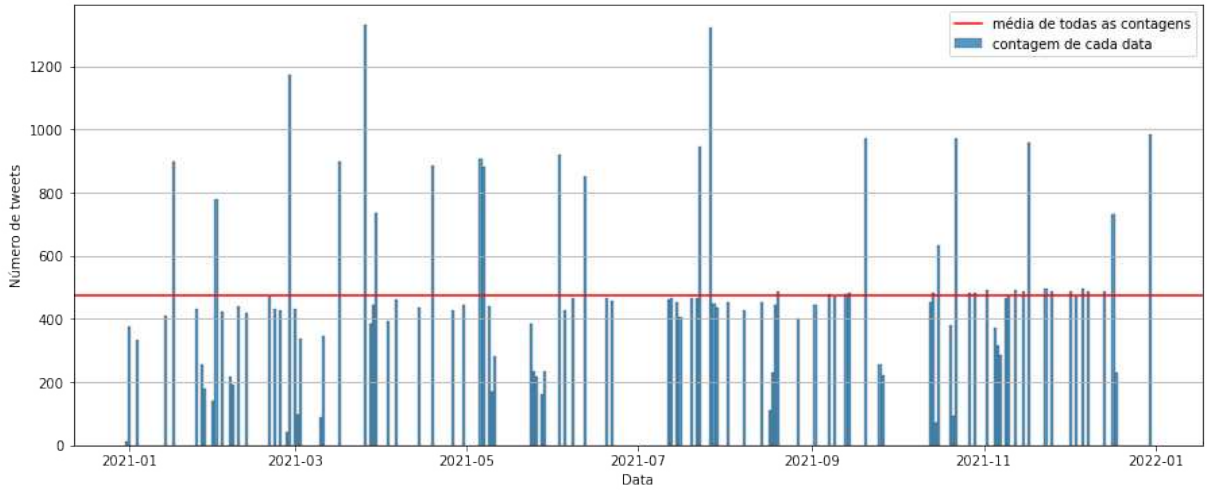


Figura 4.11 – Distribuição de *tweets* coletados ao decorrer de 2021 (amostra 1).

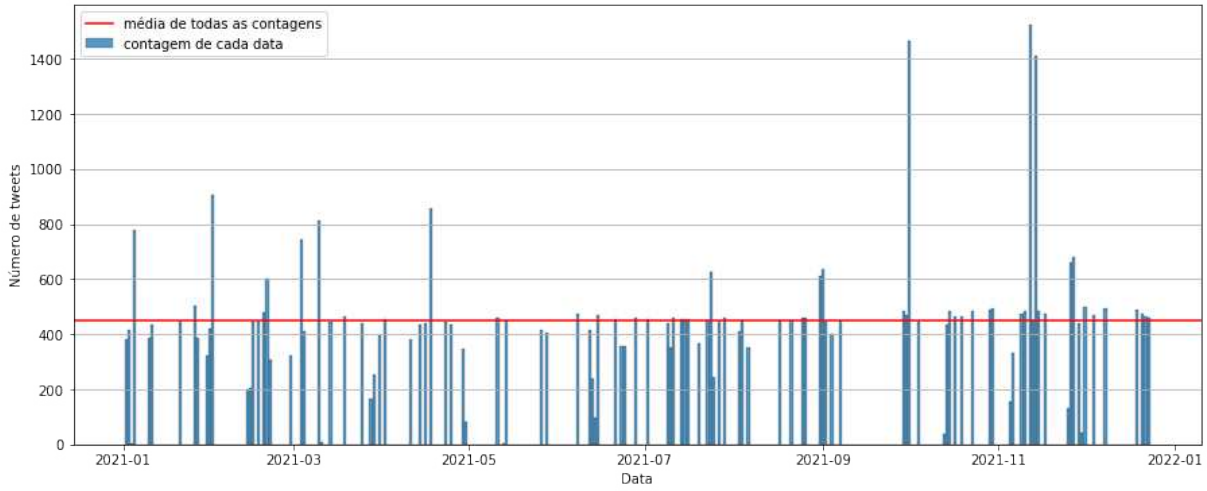


Figura 4.12 – Distribuição de *tweets* coletados ao decorrer de 2021 (amostra 2).

4.3 Experimentos em Tamanho de Amostra com LDA

Nesta seção, encontram-se os resultados dos três experimentos em tamanho de amostra realizados, bem como uma discussão a seu respeito. Os experimentos foram: a) LDA nas amostras de 50 mil tweets, b) a técnica de divisão de amostra pela metade e c) experimento de aumento de tamanho de amostra. Passaremos a uma breve exposição de cada um deles.

4.3.1 LDA nas Amostras de 50 mil tweets

Nas tabelas 4.2 e 4.3, estão os resultados da aplicação de LDA com cinco tópicos junto às três amostras, com tamanho de amostra igual a 50 mil *tweets*.

Na tabela 4.2, observa-se que a média total de coerência de tópico C_V resultada foi de 0,299, o desvio padrão foi igual a 0,050 e coeficiente de variação foi 16,7 %. Compararemos essa média com as outras médias dos outros experimentos em tamanho de

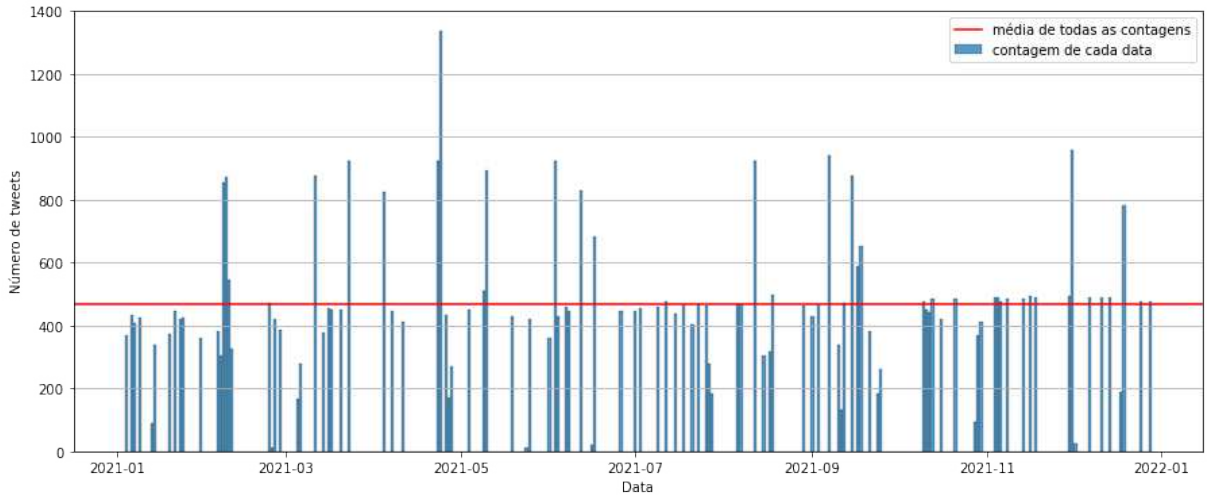


Figura 4.13 – Distribuição de *tweets* coletados ao decorrer de 2021 (amostra 3).

Tabela 4.2 – Médias, desvios padrão e coeficientes de variação das coerências C_V para LDA de cinco tópicos de cada amostra com 50 mil *tweets*.

Amostra	Média de C_V	Desvio padrão de C_V	Coeficiente de Variação de C_V (%)
1	0,300	0,054	18,0
2	0,306	0,050	16,3
3	0,294	0,046	15,6
Total	0,299	0,050	16,7

Tabela 4.3 – Médias, desvios padrão e coeficientes de variação da distância de Jaccard (d_J) entre os tópicos gerados das amostras com 50 mil *tweets*.

Amostras	Média de d_J	Desvio padrão de d_J	Coeficiente de Variação de d_J (%)
1 e 2	0,888	0,014	1,6
1 e 3	0,889	0,014	1,6
2 e 3	0,875	0,022	2,5
Total	0,884	0,018	2,0

amostra nas subseções 4.3.2 e 4.3.3 para avaliar o tamanho mais adequado.

Na tabela 4.3, uma média total de distância de Jaccard foi 0,884, o desvio padrão total foi de 0,018 e o coeficiente de variação foi de 2 %. Esse último valor demonstra que há uma pequena variabilidade entre as distâncias de Jaccard, indicando que há diferenças com intensidades similares entre os tópicos geradas das amostras e que não há uma amostra que destoa muito das outras.

4.3.2 Técnica de Divisão pela Metade

Nas tabelas 4.4 e 4.5, estão os resultados de aplicação de LDA em cada metade de cada amostra de 50 mil *tweets*, i.e., sobre amostras de 25 mil postagens.

Tabela 4.4 – Médias, desvios padrão e coeficiente de variação de coerências C_V para cada metade de cada amostra com 25 mil *tweets*.

Amostra	Metade	Média de C_V	Desvio padrão de C_V	Coeficiente de Variação de C_V (%)
1	1	0,258	0,030	11,6
1	2	0,248	0,023	9,3
2	1	0,337	0,038	11,3
2	2	0,323	0,036	11,1
3	1	0,242	0,013	5,4
3	2	0,242	0,018	7,4
Total		0,275	0,026	9,5

Tabela 4.5 – Médias, desvios e coeficiente de variação de distância Jaccard entre as metades de cada amostra com 25 mil *tweets*.

Amostra	Metade	Média de d_J	Desvio padrão de d_J	Coeficiente de Variação de d_J (%)
1	1 e 2	0,850	0,016	1,9
2	1 e 2	0,838	0,032	3,8
3	1 e 3	0,822	0,025	3,0
Total		0,837	0,024	2,9

Na tabela 4.4, constata-se que a média total de coerência de tópico C_V resultante foi 0,275, o desvio padrão foi de 0,026 e o coeficiente de variação foi de 9,5 %. Houve, portanto, uma diminuição na coerência do modelo ao se dividirem as amostras pela metade, mas a variação também foi reduzida, indicando uma menor diferença geral entre os modelos com as metades das amostras.

Na tabela 4.5, verifica-se que houve uma média total de distância de Jaccard de 0,837, desvio padrão total de 0,024 e coeficiente de variação de 2,9 %. As distâncias diminuíram entre as amostras principalmente devido à redução do tamanho de amostra. Na seção 4.8, tem-se um resumo das comparações entre os experimentos de tamanho de amostra.

4.3.3 Experimento de Aumento de Tamanho de Amostra

Nas tabelas 4.6 e 4.7, estão os resultados do experimento de amostragem que foi duplicar o tamanho de amostra, i.e. coletar 100 mil *tweets* e aplicar LDA com cinco tópicos.

Tabela 4.6 – Médias, desvios padrão e coeficientes de variação de coerências C_V do LDA para cada amostra de 100 mil *tweets*.

Amostra	Média de C_V	Desvio padrão de C_V	Coeficiente de Variação de C_V (%)
4	0,228	0,025	11,0
5	0,231	0,019	8,2
6	0,246	0,016	6,5
Total	0,235	0,022	9,4

Tabela 4.7 – Médias, desvios de distância e coeficientes de variação Jaccard entre as combinações de LDA com diferentes amostras de 100 mil *tweets*.

Amostras	Média de d_J	Desvio padrão de d_J	Coeficiente de Variação de d_J (%)
4 e 5	0,850	0,030	3,5
4 e 6	0,842	0,020	2,4
5 e 6	0,833	0,025	3,0
Total	0,842	0,027	3,2

Na tabela 4.6, constata-se que a média total de coerência de tópico C_V resultante foi 0,235, o desvio padrão foi de 0,022 e o coeficiente de variação foi de 9,4 %. Também houve uma diminuição na coerência do modelo ao se duplicar o tamanho de amostra, porém, a variação também foi reduzida, indicando uma menor diferença geral entre os modelos com as metades das amostras.

Na tabela 4.7, verifica-se uma média total de distância de Jaccard de 0,842, desvio padrão total de 0,027 e coeficiente de variação de 3,2 %. Na seção 4.8, apresenta-se um resumo das comparações entre os experimentos de tamanho de amostra.

4.4 Escolha dos Hiperparâmetros do LDA

Nas figuras 4.14, 4.15 e 4.16 a seguir, estão os perfis das buscas grossas dos três hiperparâmetros do LDA (número de tópicos, alfa e eta) em relação às coerências C_V calculadas.

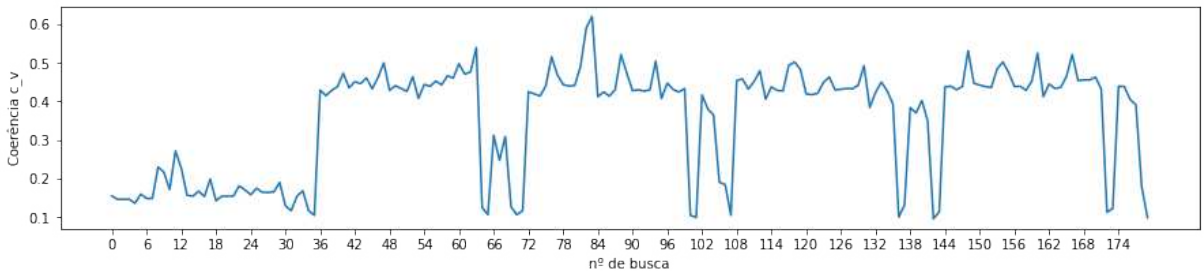


Figura 4.14 – Busca grossa dos três hiperparâmetros do LDA em relação à coerência C_V (amostra 1).

Esses gráficos das figuras 4.14, 4.15 e 4.16 são importantes para observar como o valor de coerência varia em relação aos hiperparâmetros e demonstrar a necessidade da escolha deles. Os valores de coerência C_V variaram de aproximadamente 0,1 a 0,6, apresentando uma amplitude considerável, em torno de um aumento 500% referente à diferença do maior valor ao menor.

Nas tabelas 4.8, 4.9 e 4.10, estão os resultados das buscas finas realizadas para cada uma das três amostras. Além de buscas finas de números de tópicos realizadas com

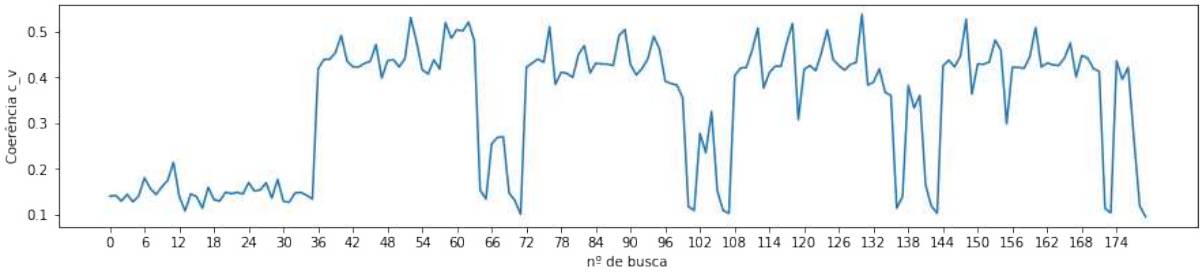


Figura 4.15 – Busca grossa dos três hiperparâmetros do LDA em relação à coerência C_V (amostra 2).

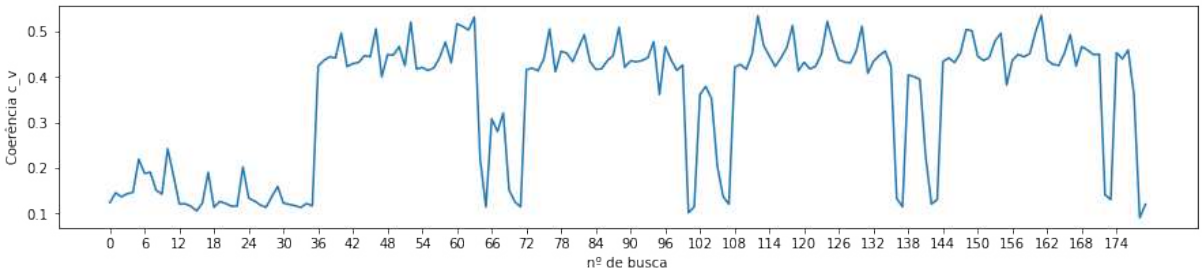


Figura 4.16 – Busca grossa dos três hiperparâmetros do LDA em relação à coerência C_V (amostra 3).

alfa e eta ótimos encontrados na busca grossa, estão presentes buscas finas de números de tópicos adicionais realizadas com base na literatura.

Tabela 4.8 – Hiperparâmetros do LDA com maiores valores de coerência C_V da amostra 1 na busca fina.

Busca fina	Alfa	Eta	Nº de tópicos (k) ótimo	Coerência C_V
1	1	0.1	40	0,555
2	*asymmetric*	*symmetric*	80	0,453
3	*asymmetric*	10	98	0,594
4	10/k	0.01	6	0,504

Busca fina 1: Configuração de alfa e eta com segundo maior valor de coerência na busca grossa da amostra 1.

Busca fina 2: Configuração de alfa e eta "asymmetric-symmetric" proposto por Wallach, Mimno e McCallum (2009).

Busca fina 3: Configuração ótima de alfa e eta encontrada na busca grossa da amostra 1.

Busca fina 4: Configuração de alfa e eta de acordo com Jónsson (2016).

Os hiperparâmetros do LDA encontrados com maior coerência C_V para as respectivas amostras estão na tabela 4.11.

Tabela 4.9 – Hiperparâmetros do LDA com maiores valores de coerência C_V da amostra 2 na busca fina.

Busca fina	Alfa	Eta	Nº de tópicos (k) ótimo	Coerência C_V
1	0,1	1	116	0,540
2	*symmetric*	*symmetric*	22	0,478
3	*asymmetric*	*symmetric*	18	0,464

Busca fina 1: Configuração ótima de alfa e eta encontrada na busca grossa da amostra 2.

Busca fina 2: Configuração padrão do *LdaMulticore* do Gensim.

Busca fina 3: Configuração de alfa e eta "asymmetric-symmetric" proposto por Wallach, Mimno e McCallum (2009).

Tabela 4.10 – Hiperparâmetros do LDA com maiores valores de coerência C_V da amostra 3 na busca fina.

Busca fina	Alfa	Eta	Nº de tópicos (k) ótimo	Coerência C_V
1	0,1	1	34	0,554
2	0,01	10	22	0,561

Busca fina 1: Configuração ótima de alfa e eta encontrada na busca grossa da amostra 2.

Busca fina 2: Configuração ótima de alfa e eta encontrada na busca grossa da amostra 3.

Tabela 4.11 – Hiperparâmetros ótimos encontrados para cada amostra no LDA.

Amostra	Nº de tópicos	Alfa	Eta	Coerência C_V
1	98	*asymmetric*	10	0,594
2	116	0,1	1	0,540
3	22	0,01	10	0,561

4.5 Aplicação do BTM nas amostras e escolha dos hiperparâmetros

Na tabela 4.12, estão os resultados das buscas realizadas para escolha dos hiperparâmetros do BTM. Nela se encontram as configurações que obtiveram maior coerência de tópicos C_V para cada amostra.

Tabela 4.12 – Hiperparâmetros ótimos do BTM encontrados para cada uma das 3 amostras.

Amostra	Nº de tópicos	Alfa	Beta	Coerência C_V
1	100	1	0,5	0,624
2	100	0,01	0,5	0,592
3	100	0,5	0,5	0,610

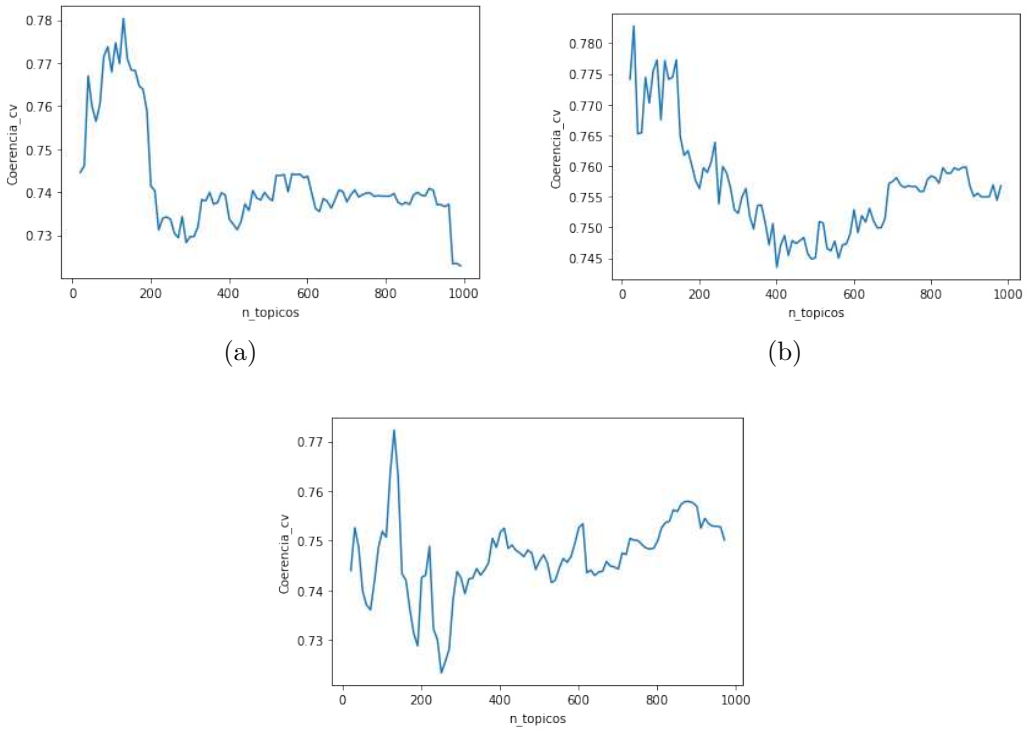
4.6 Aplicação do BERTopic nas amostras e escolha dos hiperparâmetros

Na tabela 4.13, estão os resultados dos número de tópicos encontrados automaticamente pelo BERTopic para cada amostra e suas respectivas coerências C_V.

Tabela 4.13 – Número de tópicos e respectivas coerências encontradas pelo algoritmo do BERTopic.

Amostra	Nº de tópicos	Coerência C_V
1	991	0,723
2	989	0,754
3	977	0,749

Na figura 4.17, estão as buscas pelos números de tópicos com maior coerência C_V ao se reduzir esse mesmo número a partir do encontrado automaticamente pelo BERTopic.



(a) (b)

(c)

Figura 4.17 – BERTopic: Números de tópicos em relação à coerência C_V da amostra 1 (a), amostra 2 (b) e amostra 3 (c).

Na tabela 4.14, apresentam-se os resultados dos números de tópicos com maiores coerências C_V para cada amostra e o respectivo ganho de coerência em relação ao número de tópicos inicial encontrado automaticamente pelo BERTopic.

Tabela 4.14 – Maiores valores de coerências C_V ao reduzir o número de tópicos do BERTopic para cada amostra.

Amostra	Nº de tópicos	Coerência C_V	Ganho de Coerência C_V (%)
1	131	0,780	7,9
2	31	0,783	3,8
3	131	0,772	3,1

Verifica-se, então, que, ao reduzir o número de tópicos até seu valor ótimo, observamos ganhos de coerência C_V em relação ao número de tópicos inicialmente encontrados.

4.7 Aplicação do NMF nas amostras e escolha dos hiperparâmetros

Nos gráficos da figura 4.18, estão os valores de coerência de tópicos C_V calculados para cada número de tópicos, e, na tabela 4.15, estão os números de tópicos encontrados com maiores coerências C_V para cada amostra.

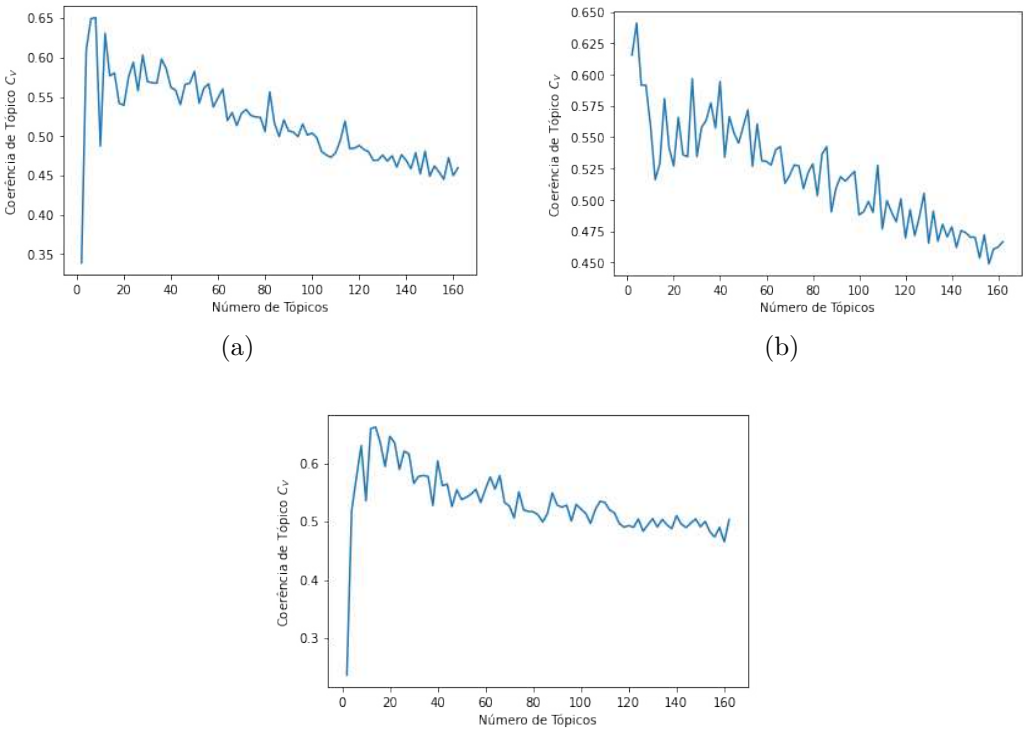


Figura 4.18 – NMF: Números de tópicos em relação à coerência C_V da amostra 1 (a), amostra 2 (b) e amostra 3 (c).

Por meio desses resultados, avalia-se que os maiores valores de coerência C_V foram com números de tópicos menores que 15, ocorrendo um decréscimo aparente desse desempenho à medida que esse número aumenta apesar das oscilações.

Tabela 4.15 – Número de tópicos e respectivas coerências encontradas pelo algoritmo do NMF.

Amostra	Nº de tópicos	Coerência C_V
1	8	0,651
2	4	0,641
3	14	0,662

4.8 Resumo dos Resultados

Na tabela 4.16, observa-se um resumo sobre os experimentos em tamanho de amostra, ou seja, em quantidade de *tweets* coletados.

Tabela 4.16 – Médias totais de coerência C_V e distância de Jaccard para os três tamanhos diferentes de amostra.

Métrica	25 mil tweets	50 mil tweets	100 mil tweets
Coerência C_V	0,275	0,299	0,235
Distância de Jaccard	0,837	0,884	0,842
C_V/d_J	0,329	0,338	0,279

Optamos em trabalhar com 50 mil *tweets* nas amostras por conta do maior desempenho em coerência de tópicos C_V e um médio custo computacional, i.e. optamos por um valor menor que 100 mil tweets e maior que 25 mil tweets. Também a razão entre coerência C_V e distância de Jaccard (d_J) para 50 mil *tweets* foi maior que a de outros tamanhos, o que também indica um maior desempenho.

Na tabela 4.17, é apresentado um resumo dos desempenhos alcançados nos experimentos de afinação dos hiperparâmetros de cada modelo.

Tabela 4.17 – Valores máximos de coerência de tópico C_V para cada amostra em cada modelo.

Amostra	LDA	BTM	BERTopic	NMF
1	0,594	0,624	0,780	0,651
2	0,540	0,592	0,783	0,641
3	0,561	0,610	0,772	0,662
Média	0,565	0,609	0,778	0,651

De acordo com os experimentos, o BERTopic obteve melhor desempenho, seguido do NMF, em terceiro o BTM e por último o LDA. Isso demonstra que esse modelo neural de modelagem de tópicos apresenta vantagem em relação aos modelos mais antigos, LDA e BTM, que se baseiam mais estritamente em ferramentas probabilísticas e NMF, que se baseia em técnicas de álgebra linear.

O BERTopic poderia ter usado os dados sem realizar pré-processamento por se tratar de um modelo que consegue considerar o contexto de um texto. Ele tem a capacidade de ponderar a ordem das palavras, as palavras de parada (*stopwords*) e pontuações. Provavelmente, ele poderia obter um desempenho ainda maior em relação aos outros modelos. Entretanto, optamos em manter a mesma metodologia para todos os modelos para realizar a comparação.

4.9 Limitações, Desafios e Boas Práticas

Uma discussão sobre limitações, desafios e boas práticas sempre é necessária. No artigo de Shadrova (2021), há um importante debate referente aos modelos de tópicos. Também há uma discussão epistemológica e um argumento adicional para uma possível vantagem do BERTopic em relação aos outros modelos, principalmente por conta da etapa que ele realiza de clusterização no espaço vetorial, diferentemente dos outros modelos (LDA, BTM e NMF), que montam os tópicos como uma distribuição de palavras mais prováveis a partir do *bag of words*.

Uma limitação que Shadrova (2021) apresenta da modelagem de tópicos surge do fato de que uma porcentagem muito pequena de palavras ocorre com frequência suficiente para fornecer informações estatisticamente significativas, com muitas palavras ocorrendo apenas uma ou poucas vezes em grandes *corpora*. Consequentemente, o desafio de usar a modelagem de tópicos para fins acadêmicos não é apenas esclarecer o escopo da evidência que ela fornece, mas também desenvolver as melhores práticas para incorporar os resultados da modelagem de tópicos no discurso acadêmico e integrar o método à pesquisa linguística e baseada em texto. Simplesmente relatar resultados exploratórios de modelos de tópicos acarreta no risco de tratá-los como conhecimento confirmado e objetivo, em vez de informações básicas. Isso poderia invalidar os esforços para modelar o conhecimento de forma cuidadosa e explícita usando métodos computacionais nas ciências humanas e sociais.

A leitura multiescala, também conhecida como a consideração simultânea de múltiplas camadas de complexidade, pode superar parcialmente as limitações da modelagem de tópicos enquanto ainda permite a leitura distante (MOA; ROSS, 2019). No entanto, se uma análise quantitativa do texto for necessária em um contexto específico, uma maneira de evitar os problemas com a modelagem de tópicos é considerar algoritmos que não dependem de abordagens de *bag of words* ou independência de tópicos (SHADROVA, 2021). Lamirel et al. (2020) apresentam tal abordagem na forma de agrupamento de maximização de recursos de palavras em artigos relacionados ao campo Ciência da Ciência na China. Na maximização de recursos, as palavras são tratadas como recursos e as combinações de recursos são comparadas em termos de similaridade ou distância. Essa abordagem não

assume independência de tópicos e é capaz de modelar relacionamentos entre *clusters*, que podem ser visualizados como gráficos de contraste. Métodos de agrupamento deste tipo oferecem claras vantagens em termos de alinhamento dos modelos conceituais e matemáticos, podendo produzir melhores resultados em termos de qualidade do conteúdo extraído. Lamirel et al. (2020) demonstram isso comparando os clusters de palavras com tópicos extraídos de LDA do mesmo corpus. No entanto, os agrupamentos de palavras ainda requerem interpretação e não correspondem a tópicos ou campos específicos, mas são construídos como tal por especialistas, que podem ser influenciados por apofenia e viés de confirmação na interpretação *post hoc*.

A leitura à distância é frequentemente proposta como uma alternativa à leitura atenta de grandes coleções de textos, principalmente na era do *big data*. No entanto, confiar apenas em palavras e não em anotações mais profundas que descrevem estruturas de significado de ordem superior ou ontologias relevantes para um determinado campo pode resultar em um alto grau de incerteza epistemológica e linguística. Embora o agrupamento de maximização de recursos seja uma opção em alguns casos, é importante considerar as limitações de tais abordagens e a necessidade de maior desenvolvimento de métodos que integrem técnicas computacionais com as ciências humanas e sociais (SHADROVA, 2021).

5 Conclusão

Neste trabalho, apresentamos uma contribuição à análise automática de tópicos tendo em vista a investigação de elementos históricos e sociais em textos de redes sociais. Também efetuamos uma análise comparativa de metodologias de processamento de linguagem natural (PLN) para a realização dessa análise, buscando, assim, trazer elementos que propiciem um maior uso de tais ferramentas sobre *corpora* de grandes dimensões.

Como primeiro passo, realizamos uma revisão bibliográfica sobre contribuições de PLN às ciências humanas, estabelecendo um panorama de trabalhos e aplicações no contexto de diferentes áreas (história, sociologia, ciência política, antropologia, demografia e economia). Em seguida, foi feita uma revisão sobre mineração de texto, uma tarefa de PLN frequentemente empregada nas ciências humanas e que, em certo sentido, engloba a modelagem de tópicos, sendo, portanto, de grande importância para este esforço.

O passo seguinte foi revisar as fontes de dados, principalmente o Twitter, uma das redes sociais mais populares do mundo moderno, e o meio escolhido como foco deste trabalho. Conhecidas as fontes, pudemos discutir a modelagem de tópicos cobrindo os modelos selecionados para abordar a tarefa. Concluindo a revisão, foram abordadas as métricas de avaliação de desempenho utilizadas neste trabalho.

Após essa primeira etapa, apresentamos uma metodologia para realizar análise de mídias sociais com base em modelagem de tópicos. Primeiramente, definimos uma forma de especificar os dados para o Brasil tendo por base a língua portuguesa. Em segundo lugar, apresentamos um procedimento para a coleta e o pré-processamento das postagens. Em seguida, definiu-se um método de análise de tamanho de amostra para definir uma quantidade de coletas que fosse representativa para o estudo. E, por fim, foram indicados elementos para melhorar o desempenho dos modelos por meio de seus hiperparâmetros.

Com base nos resultados, pudemos realizar uma análise comparativa e conseguimos observar que houve um melhor desempenho por parte modelo mais recente, que é caracterizado como uma estratégia neural de aprendizado profundo na categoria dos *transformers* pré-treinados, o BERTopic. Também se verificou a vantagem de modelos que não são baseados em *bag of words* (SHADROVA, 2021), como é o caso do BERTopic, que cria representações de textos no espaço vetorial multi-dimensional levando em consideração todo o contexto e por meio disso realiza o agrupamento dos textos em tópicos.

Como perspectivas e tópicos para trabalhos futuros, podemos apontar a análise dos tópicos produzidos pelos modelos, principalmente aqueles gerados pelo BERTopic, que obteve o melhor desempenho. Uma análise desse tipo deveria, necessariamente, incluir pesquisadores(as) da área de humanidades, num esforço multidisciplinar.

Os resultados e a discussão de Shadrova (2021) também dão suporte ao potencial do BERTopic, o que nos indica ser promissor usá-lo para analisar outros temas de cunho político / social utilizando a metodologia desse trabalho. Tais análises poderiam incluir elementos ligados à capacidade dinâmica da ferramenta, i.e. de lidar com o decurso do tempo.

Por fim, outra ideia seria introduzir uma técnica de sumarização nos grupos de documentos encontrados pelo BERTopic, em vez de TF-IDF de classe. Isso poderia contornar o emprego de *bag of words* nessa etapa final do modelo.

Referências

ABDEL-HAFEZ, A.; XU, Y. A survey of user modelling in social media websites. *Comput. Inf. Sci.*, Canadian Center of Science and Education, v. 6, n. 4, set. 2013. Citado na página 27.

ABRAMSON, C. M. et al. The promises of computational ethnography: Improving transparency, replicability, and validity for realist approaches to ethnographic analysis. *Ethnography*, v. 19, n. 2, p. 254–284, 2018. Disponível em: <https://doi.org/10.1177/1466138117725340>. Citado na página 19.

AGUILAR, S. T.; CHASTANG, P.; TANNIER, X. Automatic medieval charters structure detection : A Bi-LSTM linear segmentation approach. *Journal of Data Mining and Digital Humanities*, Episciences.org, v. 2022, out. 2022. Disponível em: <https://hal.science/hal-03410057>. Citado na página 17.

AHRENS, M.; MCMAHON, M. Extracting economic signals from central bank speeches. In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 93–114. Disponível em: <https://aclanthology.org/2021.econlp-1.12>. Citado na página 20.

AIZAWA, A. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, v. 39, n. 1, p. 45–65, 2003. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457302000213>. Citado na página 33.

ALLAHYARI, M. et al. *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. arXiv, 2017. Disponível em: <https://arxiv.org/abs/1707.02919>. Citado 2 vezes nas páginas 40 e 41.

ALMAGHLOUTH, N. et al. Who frames the debate on the arab uprisings? analysis of arabic, english, and french academic scholarship. *International Sociology*, v. 30, n. 4, p. 418–441, 2015. Disponível em: <https://doi.org/10.1177/0268580915580157>. Citado na página 18.

ANGELOV, D. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020. Citado na página 27.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with python*. [S.l.]: O'Reilly Media, 2009. Citado na página 41.

BLEI, D.; NG, A.; JORDAN, M. Latent dirichlet allocation. In: DIETTERICH, T.; BECKER, S.; GHAHRAMANI, Z. (Ed.). *Advances in Neural Information Processing Systems*. MIT Press, 2001. v. 14. Disponível em: <https://proceedings.neurips.cc/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf>. Citado 2 vezes nas páginas 26 e 28.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, v. 3, n. Jan, p. 993–1022, 2003. Citado 2 vezes nas páginas 27 e 29.

BROWN, C. H. *Analyzing group behavior from language use with natural language processing and experimental methods: three applications in political science and sociology.* [S.l.]: The University of Texas at Austin, 2018. Citado na página 18.

CAMPAGNOLO, J. M.; DUARTE, D.; BIANCO, G. D. Topic coherence metrics: How sensitive are they? *Journal of Information and Data Management*, v. 13, n. 4, Oct 2022. Disponível em: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/2181>. Citado na página 35.

CHENG, X. et al. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, v. 26, n. 12, p. 2928–2941, 2014. Citado na página 28.

CHOWDHARY, K. R. *Fundamentals of artificial intelligence*. 1. ed. New Delhi, India: Springer, 2020. Citado na página 15.

CHURCH, K. W.; RAU, L. F. Commercial applications of natural language processing. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 38, n. 11, p. 71–79, nov 1995. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/219717.219778>. Citado na página 15.

CHURCHILL, R.; SINGH, L. The evolution of topic modeling. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 10s, nov 2022. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3507900>. Citado 3 vezes nas páginas 26, 32 e 35.

DEERWESTER, S. et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990. Disponível em: <https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9>. Citado na página 26.

DEVLIN, J. et al. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>. Citado na página 33.

DIERCKX, T.; DAVIS, J.; SCHOUTENS, W. Quantifying news narratives to predict movements in market risk. In: _____. *Data Science for Economics and Finance: Methodologies and Applications*. Cham: Springer International Publishing, 2021. p. 265–285. ISBN 978-3-030-66891-4. Disponível em: <https://doi.org/10.1007/978-3-030-66891-4_12>. Citado na página 20.

DUONG, Q.; PIVOVAROVA, L.; ZOSA, E. Benchmarks for unsupervised discourse change detection. In: Sumikawa , Y. et al. (Ed.). *Proceedings of the 6th International Workshop on Computational History (HistoInformatics 2021)*. Germany: CEUR-WS.org, 2021. (CEUR workshop proceedings). International Workshop on Computational History, HistoInformatics2021 ; Conference date: 30-09-2021 Through 01-10-2021. Disponível em: <https://sites.google.com/view/histoinformatics2021workshop/home>. Citado na página 16.

FARIA, M. A. G. No meio do caminho tinha uma estrela: o antipetismo das mobilizações à ascensão de Jair Bolsonaro (2013-2018). mar. 2021. Accepted: 2021-03-18T16:04:04Z Publisher: Universidade Federal de São Paulo. Disponível em: <https://repositorio.unifesp.br/handle/11600/60558>. Citado na página 46.

FARZINDAR, A. A.; INKPEN, D. *Natural language processing for social media, third edition*. 3. ed. Cham, Switzerland: Springer International Publishing, 2020. (Synthesis Lectures on Human Language Technologies). Citado 2 vezes nas páginas 23 e 24.

FLORES, R. D. Do anti-immigrant laws shape public sentiment? a study of arizona's sb 1070 using twitter data. *American Journal of Sociology*, v. 123, n. 2, p. 333–384, 2017. Disponível em: <https://doi.org/10.1086/692983>. Citado na página 17.

GALLAGHER, R. J. et al. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 5, p. 529–542, 2017. Citado na página 27.

GIGLIETTO, F.; ROSSI, L.; BENNATO, D. The open laboratory: Limits and possibilities of using facebook, twitter, and youtube as a research data source. *Journal of Technology in Human Services*, Routledge, v. 30, n. 3-4, p. 145–159, 2012. Disponível em: <https://doi.org/10.1080/15228835.2012.743797>. Citado na página 25.

GROOTENDORST, M. *BERTopic*. [S.l.]: GitHub, 2022. <https://github.com/MaartenGr/BERTopic>. Citado na página 45.

GROOTENDORST, M. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2203.05794>. Citado 3 vezes nas páginas 27, 33 e 34.

GUI, L. et al. Neural topic model with reinforcement learning. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 3478–3483. Disponível em: <https://aclanthology.org/D19-1350>. Citado na página 36.

H, M. K.; ZAINUDDIN, H.; WABULA, Y. Twitter social media conversion topic trending analysis using latent dirichlet allocation algorithm. *Journal of Applied Engineering and Technological Science (JAETS)*, v. 4, n. 1, p. 390–399, Dec. 2022. Disponível em: <https://journal.yrpipku.com/index.php/jaets/article/view/1143>. Citado na página 29.

HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, v. 61, n. 4, p. 5–14, 2019. Disponível em: <https://doi.org/10.1177/0008125619864925>. Citado na página 15.

HALBWACHS, M. La mémoire collective [la memoria colectiva]. *Paris, Francia: Presses Universitaires de France*, 1950. Citado na página 17.

HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. *Science*, v. 349, n. 6245, p. 261–266, 2015. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aaa8685>. Citado 2 vezes nas páginas 15 e 16.

HOFMANN, T. Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 1999. (SIGIR '99), p. 50–57. ISBN 1581130961. Disponível em: <https://doi.org/10.1145/312624.312649>. Citado na página 26.

HONG, L.; DAVISON, B. D. Empirical study of topic modeling in twitter. In: *Proceedings of the First Workshop on Social Media Analytics*. New York, NY, USA: Association for Computing Machinery, 2010. (SOMA '10), p. 80–88. ISBN 9781450302173. Disponível em: <https://doi.org/10.1145/1964858.1964870>. Citado 2 vezes nas páginas 27 e 28.

HONNIBAL, M.; MONTANI, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 2017. Citado na página 41.

JOHRI, P. et al. Natural language processing: History, evolution, application, and future work. In: ABRAHAM, A.; CASTILLO, O.; VIRMANI, D. (Ed.). *Proceedings of 3rd International Conference on Computing Informatics and Networks*. Singapore: Springer Singapore, 2021. p. 365–375. ISBN 978-981-15-9712-1. Citado na página 16.

JÓNSSON, E. An evaluation of topic modelling techniques for twitter. In: . [S.l.: s.n.], 2016. Citado 3 vezes nas páginas 29, 44 e 55.

JOSEPH, S. R. et al. Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, v. 6, n. 3, p. 207–210, 2016. Citado 2 vezes nas páginas 15 e 16.

KARAMI, A.; BENNETT, L. S.; HE, X. Mining public opinion about economic issues. *Int. J. Strat. Decis. Sci.*, IGI Global, v. 9, n. 1, p. 18–28, jan. 2018. Citado na página 25.

KIM, A. G.; YOON, S. Corporate bankruptcy prediction with domain-adapted BERT. In: *Proceedings of the Third Workshop on Economics and Natural Language Processing*. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. p. 26–36. Disponível em: <https://aclanthology.org/2021.econlp-1.4>. Citado na página 20.

KOSUB, S. *A note on the triangle inequality for the Jaccard distance*. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1612.02696>. Citado na página 36.

KRIEG, L. J.; BERNING, M.; HARDON, A. Anthropology with algorithms? *Medicine Anthropology Theory*, v. 4, n. 3, Sep. 2017. Disponível em: <http://www.medanthrotheory.org/article/view/4783>. Citado na página 18.

KRIPPENDORFF, K. *Content analysis*. 4. ed. Thousand Oaks, CA: SAGE Publications, 2018. Citado 4 vezes nas páginas 39, 40, 42 e 43.

KULKARNI, V. et al. Multi-view models for political ideology detection of news articles. *CoRR*, abs/1809.03485, 2018. Disponível em: <http://arxiv.org/abs/1809.03485>. Citado na página 18.

KUMAR, S.; KAR, A. K.; ILAVARASAN, P. V. Applications of text mining in services management: A systematic literature review. *International Journal of Information Management Data Insights*, v. 1, n. 1, p. 100008, 2021. ISSN 2667-0968. Disponível em: <https://www.sciencedirect.com/science/article/pii/S266709682100001X>. Citado na página 22.

LAMIREL, J.-C. et al. An overview of the history of science of science in china based on the use of bibliographic and citation data: a new method of analysis based on clustering with feature maximization and contrast graphs. *Scientometrics*, v. 125, 05 2020. Citado 2 vezes nas páginas 60 e 61.

LAZER, D. et al. Computational social science. *Science*, v. 323, n. 5915, p. 721–723, 2009. Disponível em: <https://www.science.org/doi/abs/10.1126/science.1167742>. Citado na página 13.

LAZER, D. M. J. et al. Computational social science: Obstacles and opportunities. *Science*, v. 369, n. 6507, p. 1060–1062, 2020. Disponível em: <https://www.science.org/doi/abs/10.1126/science.aaz8170>. Citado na página 13.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, n. 7553, p. 436–444, maio 2015. Citado na página 16.

LI, P. et al. Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Know.-Based Syst.*, Elsevier Science Publishers B. V., NLD, v. 193, n. C, apr 2020. ISSN 0950-7051. Disponível em: <https://doi.org/10.1016/j.knosys.2019.105436>. Citado na página 33.

LI, X. et al. Topic extraction from extremely short texts with variational manifold regularization. *Mach. Learn.*, Springer Science and Business Media LLC, v. 110, n. 5, p. 1029–1066, maio 2021. Citado na página 28.

LOMBORG, S.; BECHMANN, A. Using apis for data collection on social media. *The Information Society*, Routledge, v. 30, n. 4, p. 256–265, 2014. Disponível em: <https://doi.org/10.1080/01972243.2014.915276>. Citado na página 23.

LYU, J. C.; LULI, G. K. Understanding the public discussion about the centers for disease control and prevention during the covid-19 pandemic using twitter data: Text mining analysis study. *J Med Internet Res*, v. 23, n. 2, p. e25108, Feb 2021. ISSN 1438-8871. Disponível em: <http://www.jmir.org/2021/2/e25108/>. Citado na página 29.

MACHADO, M. G.; COLEVATI, J. Anticomunismo e Gramscismo Cultural no Brasil. *Revista Aurora*, v. 14, n. Edição Especial, p. 23–34, jul. 2021. ISSN 1982-8004. Number: Edição Especial. Disponível em: <https://revistas.marilia.unesp.br/index.php/aurora/article/view/12690>. Citado na página 40.

MAHAJAN, R.; PUROHIT, A. Text summarization for information of famous indian historical monuments. In: TIWARI, A. et al. (Ed.). *Soft Computing for Problem Solving*. Singapore: Springer Singapore, 2021. p. 499–509. ISBN 978-981-16-2709-5. Citado na página 23.

MANJAVACAS, E.; FONTEYN, L. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining and Digital Humanities*, Episciences.org, NLP4DH, jun. 2022. Disponível em: <https://hal.inria.fr/hal-03592137>. Citado na página 17.

MIMNO, D. et al. Optimizing semantic coherence in topic models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, 2011. p. 262–272. Disponível em: <https://aclanthology.org/D11-1024>. Citado na página 34.

MOA, B.; ROSS, S. Big data analytics for multiscale reading. In: *Doing More Digital Humanities*. London ; New York, NY : Routledge/Taylor & Francis Group, 2020.: Routledge, 2019. p. 199–236. Citado na página 60.

MUNK, A. K.; OLESEN, A. G.; JACOMY, M. The thick machine: Anthropological ai between explanation and explication. *Big Data & Society*, v. 9, n. 1, p. 20539517211069891, 2022. Disponível em: <https://doi.org/10.1177/20539517211069891>. Citado na página 19.

MURPHY, A. H. The finley affair: A signal event in the history of forecast verification. *Weather and Forecasting*, American Meteorological Society, Boston MA, USA, v. 11, n. 1, p. 3 – 20, 1996. Disponível em: <https://journals.ametsoc.org/view/journals/wefo/11/1/1520-0434_1996_011_0003_tfaase_2_0_co_2.xml>. Citado na página 36.

MURSHED, B. A. H. et al. Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artif. Intell. Rev.*, Springer Science and Business Media LLC, p. 1–128, out. 2022. Citado 2 vezes nas páginas 27 e 32.

NELSON, L. K. *The Power of Place: Structure, Culture, and Continuities in U.S. Women's Movements*. Tese (Doutorado) — UC Berkeley, 2014. Disponível em: <https://escholarship.org/uc/item/8794361r>. Citado na página 17.

PAATERO, P.; TAPPER, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, v. 5, n. 2, p. 111–126, 1994. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/env.3170050203>. Citado 2 vezes nas páginas 26 e 32.

PANICHELLA, A. A systematic comparison of search-based approaches for lda hyperparameter tuning. *Information and Software Technology*, v. 130, p. 106411, 2021. ISSN 0950-5849. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584920300069>. Citado 2 vezes nas páginas 34 e 43.

QIANG, J. et al. Short text topic modeling techniques, applications, and performance: A survey. *CoRR*, abs/1904.07695, 2019. Disponível em: <http://arxiv.org/abs/1904.07695>. Citado na página 28.

ROBERTS, M. E. et al. The structural topic model and applied social science. In: HARRAHS AND HARVEYS, LAKE TAHOE. *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. [S.l.], 2013. v. 4, p. 1–20. Citado na página 27.

ROBERTS, M. E. et al. Structural topic models for open-ended survey responses. *American Journal of Political Science*, v. 58, n. 4, p. 1064–1082, 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12103>. Citado na página 18.

ROBILA, M.; ROBILA, S. A. Applications of artificial intelligence methodologies to behavioral and social sciences. *Journal of Child and Family Studies*, v. 29, n. 10, p. 2954–2966, out. 2020. Citado na página 13.

ROCHA, L. D. d.; DUWE, R. Feridas abertas e processos inconclusos: o fantasma da guerra civil e os eventos de janeiro de 2021 nos estados unidos. *Revista Tempo e Argumento*, v. 14, n. 36, p. e0107, set. 2022. Disponível em: <https://www.revistas.udesc.br/index.php/tempo/article/view/2175180314362022e0107>. Citado na página 50.

RöDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2015. (WSDM '15), p. 399–408. ISBN 9781450333177. Disponível em: <https://doi.org/10.1145/2684822.2685324>. Citado 2 vezes nas páginas 35 e 42.

RODRíGUEZ-RUIZ, J. et al. A one-class classification approach for bot detection on twitter. *Computers & Security*, v. 91, p. 101715, 2020. ISSN 0167-4048. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167404820300031>. Citado na página 24.

ROY, A. et al. How covid-19 affected computer science mooc learner behavior and achievements: A demographic study. In: *Proceedings of the Ninth ACM Conference on Learning @ Scale*. New York, NY, USA: Association for Computing Machinery, 2022. (L@S '22), p. 345–349. ISBN 9781450391580. Disponível em: <https://doi.org/10.1145/3491140.3528328>. Citado na página 19.

RUSSELL, M. A.; KLASSEN, M. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More*. 3rd edition. ed. [S.l.]: O'Reilly Media, 2019. ISBN 978-1-4919-8504-5. Citado na página 23.

SANCASSANI, V. Resenha do livro "a companion to digital humanities", de susan schreibman, ray siemens e john unsworth (eds.). *TECCOGS: Revista Digital de Tecnologias Cognitivas*, Pontifical Catholic University of Sao Paulo (PUC-SP), n. 21, dez. 2020. Citado na página 13.

SCHREIBMAN, S.; SIEMENS, R.; UNSWORTH, J. (Ed.). *A companion to digital humanities*. London, England: Blackwell Publishing, 2004. (Blackwell Companions to Literature and Culture). Citado na página 13.

SHADROVA, A. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021, out. 2021. Disponível em: <https://jdmdh.episciences.org/8608>. Citado 5 vezes nas páginas 25, 60, 61, 62 e 63.

SHAMOI, E. et al. Sentiment analysis of vegan related tweets using mutual information for feature selection. *PeerJ Comput. Sci.*, PeerJ, v. 8, n. e1149, p. e1149, dez. 2022. Citado na página 25.

SHARAFF, A.; NAGWANI, N. K. Email thread identification using latent dirichlet allocation and non-negative matrix factorization based clustering techniques. *Journal of Information Science*, v. 42, n. 2, p. 200–212, 2016. Disponível em: <https://doi.org/10.1177/0165551515587854>. Citado na página 32.

SRIDHAR, V. K. R. Unsupervised topic modeling for short texts using distributed representations of words. In: *Proceedings of the 1st workshop on vector space modeling for natural language processing*. [S.l.: s.n.], 2015. p. 192–200. Citado na página 27.

SUBALALITHA, C. N. Information extraction framework for kurunthogai. *Sadhana*, Springer Science and Business Media LLC, v. 44, n. 7, jul. 2019. Citado na página 23.

SUMIKAWA, Y.; JATOWT, A.; DüRING, M. Digital history meets microblogging: Analyzing collective memories in twitter. In: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2018. (JCDL '18), p. 213–222. ISBN 9781450351782. Disponível em: <https://doi.org/10.1145/3197026.3197057>. Citado na página 17.

TEH, Y. et al. Sharing clusters among related groups: Hierarchical dirichlet processes. *Advances in neural information processing systems*, v. 17, 2004. Citado na página 27.

THORP, H. H. Chatgpt is fun, but not an author. *Science*, v. 379, n. 6630, p. 313–313, 2023. Disponível em: <https://www.science.org/doi/abs/10.1126/science.adg7879>. Citado na página 15.

TUTUBALINA, E.; NIKOLENKO, S. Automated prediction of demographic information from medical user reviews. In: PRASATH, R.; GELBUKH, A. (Ed.). *Mining Intelligence and Knowledge Exploration*. Cham: Springer International Publishing, 2017. p. 174–184. ISBN 978-3-319-58130-9. Citado na página 20.

VALDEZ, D.; PICKETT, A. C.; GOODSON, P. Topic modeling: Latent semantic analysis for the social sciences. *Social Science Quarterly*, v. 99, n. 5, p. 1665–1679, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ssqu.12528>. Citado na página 27.

WALLACH, H.; MIMNO, D.; MCCALLUM, A. Rethinking lda: Why priors matter. In: BENGIO, Y. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2009. v. 22. Disponível em: <https://proceedings.neurips.cc/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf>. Citado 2 vezes nas páginas 55 e 56.

WALTER, T. et al. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases. *CoRR*, abs/2108.06295, 2021. Disponível em: <https://arxiv.org/abs/2108.06295>. Citado na página 17.

WANG, Y.-X.; ZHANG, Y.-J. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, v. 25, n. 6, p. 1336–1353, 2013. Citado na página 32.

WENG, J. et al. Twitterrank: Finding topic-sensitive influential twitterers. In: *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2010. (WSDM '10), p. 261–270. ISBN 9781605588896. Disponível em: <https://doi.org/10.1145/1718487.1718520>. Citado na página 28.

XU, X. et al. Understanding narratives from demographic survey data: a comparative study with multiple neural topic models. In: *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*. Abu Dhabi, UAE: Association for Computational Linguistics, 2022. p. 33–38. Disponível em: <https://aclanthology.org/2022.nlpcss-1.4>. Citado na página 19.

XUE, J. et al. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *J Med Internet Res*, v. 22, n. 11, p. e20550, Nov 2020. ISSN 1438-8871. Disponível em: <http://www.jmir.org/2020/11/e20550/>. Citado na página 29.

YAN, X. et al. A biterm topic model for short texts. In: . New York, NY, USA: Association for Computing Machinery, 2013. (WWW '13), p. 1445–1456. ISBN 9781450320351. Disponível em: <https://doi.org/10.1145/2488388.2488514>. Citado 6 vezes nas páginas 27, 28, 30, 31, 32 e 44.

ZHAO, R.; TAN, V. Y. F. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing*, Institute of Electrical and Electronics Engineers (IEEE), v. 65, n. 3, p. 555–570, feb 2017. Disponível em: <https://doi.org/10.1109%2Ftsp.2016.2620967>. Citado na página 45.

ZONG, C.; XIA, R.; ZHANG, J. *Text Data Mining*. 1. ed. Singapore, Singapore: Springer, 2021. Disponível em: <https://doi.org/10.1007/978-981-16-0100-2>. Citado 2 vezes nas páginas 15 e 22.

ŘEHůřEK, R.; SOJKA, P. Software framework for topic modelling with large corpora. In: . [S.l.: s.n.], 2010. p. 45–50. Citado 4 vezes nas páginas 36, 37, 41 e 45.