MÍDIAS SOCIAIS ON-LINE A SERVIÇO DAS HUMANAS: SUGESTÕES DE FERRAMENTAS COMPUTACIONAIS ÚTEIS[137]

MIRELLA M. MORO
ANA PAULA COUTO DA SILVA

Introdução

A internet atual perpassa vários ambientes da vida humana. Há tempos ela deixou de ter seu uso limitado a cientistas e profissionais de tecnologia, estando hoje disponível praticamente no mundo inteiro (com poucas exceções). De fato, o acesso à internet no Brasil alcançou 82% da população em 2021 (CGI.br, 2023 apud CRUZ E BECARI, 2019). Em especial, pessoas se comunicam diariamente através de vários recursos tecnológicos disponíveis em seus telefones celulares (CGI.br, 2023). Além da comunicação individual, os movimentos sociais têm explorado de diferentes maneiras a tecnologia a seu favor. Tal mudança não passa despercebida da comunidade científica, que acompanha essa mudança de perto.

Por exemplo, as pesquisadoras Mundt, Ross, Burnett (2018) estudaram o movimento norte-americano *Black Lives Matter* e destacaram a importância das mídias sociais como ferramenta de expansão que facilita o fortalecimento do movimento (construção coletiva de significado e criação de redes de apoio) e amplia o mesmo, permitindo que grupos locais formem coalizões, estendendo e disseminando seus discursos. As autoras ainda apontaram desafios criados pelo uso das mídias sociais e seus riscos, que vão além das limitações descritas em estudos empíricos existentes (complacência, indefinição ideológica e riscos por vezes físicos para ativistas com presença on-line).

De fato, existem várias publicações que exploram o universo digital (especialmente on-line) do ponto de vista de movimentos sociais, incluindo feminismos. E tais publicações não estão limitadas às ciências sociais e humanas. Por exemplo, no Brasil, merece destaque um esforço contínuo do Comitê Gestor da Internet (CGI.br) para promover discussões e publicações considerando o contexto da internet brasileira, como, por exemplo, a sua série de livros de coletâneas de artigos sobre gênero, raça e diversidade, cuja edição mais recente é organizada pelas conselheiras da organização (BARBOSA, TRESCA, LAUSCHNER, 2023).

Esse casamento entre ciências é definitivamente frutífero. Como apontado por Deriu e Iezzi (2020) na introdução à edição especial sobre análise de texto de uma revista acadêmica da área de Sociologia, que foca no volume de dados à disposição para pesquisa, que é ímpar:

137 O presente estudo faz parte do Projeto "Internet como campo de disputa pela Igualdade de Gênero", realizado no Laboratório de Estudos de Gênero e História da Universidade Federal de Santa Catarina com apoio da Fundação de Amparo à Pesquisa e Inovação de Santa Catarina (Fapesc) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

> Nos últimos anos, em todas as áreas do conhecimento, uma abordagem dirigida por dados se difundiu de acordo com o novo cenário definido pela era do *Big Data*. O chamado *dilúvio de dados* deu início a uma época em que uma quantidade impressionante de dados constitui um valioso material de investigação para pessoas do mundo acadêmico. Nesse novo contexto, a abordagem baseada em dados permite que academia e ciência examinem e organizem dados com o objetivo de aumentar o conhecimento em muitas áreas de investigação. O dilúvio de dados hoje nos permite planejar novas análises sobre uma variedade de dados não estruturados que são produzidos em grande parte pela navegação na Web. Estimativas recentes sustentam que 80% de todos os dados são dados textuais. [...] Os dados são em maioria textuais e produzidos por atividades férteis de comunicação humana e intercâmbio que ocorrem num número crescente de plataformas sociais com um poderoso potencial viral. As relações humanas são moldadas em diferentes níveis de abstração, flutuando no ambiente virtual da Web. (DERIU e IEZZI, 2020, p.1, tradução nossa)

Com essa quantidade de dados, é natural que se recorra a ferramentas computacionais para facilidade e velocidade de processamento. Porém, mesmo para pessoas que estudam aspectos técnicos de dados, o conjunto de ferramentas à disposição é vasto. Dentro desse contexto complexo, nosso objetivo é servir como material didático *auxiliar* e *introdutório* que caracteriza algumas dessas ferramentas e aponta para vários outros materiais, com linguagens e dificuldades diversas. Finalmente, é importante lembrar que existem centenas de bons tutoriais em vários formatos (página *web*, texto, PDF, vídeo) disponíveis na *web* na língua portuguesa (muitos mais ainda na língua inglesa). A principal sugestão aqui é buscar sempre por material desenvolvido por pessoas ou instituições conhecidas, como, por exemplo, o portal *Programming Historian* (PH, 2023), ou ainda publicado por revistas e *websites* especializados (EDITORA GLOBO, 2023; GOOGLE FOR EDUCATION, 2023; TECNOBLOG, 2023).

Redes Sociais On-line

As redes de relacionamento têm se ampliado no universo on-line, alimentadas pelo desenvolvimento da internet, que permitiu o funcionamento da *web* e a criação de aplicativos dos mais variados tipos. Em especial, as redes sociais on-line podem ser consideradas um marco na história recente, por levarem os relacionamentos entre pessoas do mundo físico para o mundo etéreo on-line, conforme ilustrado na Figura 1.

Figura 1. A interação social entre pessoas e grupos (verde acima) mudou-se para o modo on-line viabilizado pela web (cinza ao meio), a qual oferece artefatos de coleta e computação de dados e informação para apoiar pesquisadoras e historiadoras (laranja abaixo)

Fonte: As autoras

Em 2010, cientistas da computação já apontavam o grande valor das redes sociais on-line sob diferentes perspectivas, incluindo a comercial, devido a bilhões de dólares investidos em propagandas nesses meios, e a sociológica, com a criação de oportunidades para estudos dos dados on-line (BENEVENUTO, 2010). Em especial, uma grande vantagem da passagem para o mundo on-line é a possibilidade de literalmente explorar todo o tipo de dado e informação gerado a partir dessas interações, porque, como diz o ditado, "caiu na rede, está armazenado em algum lugar para sempre". É importante notar que redes sociais (como um tipo central de plataforma para mídias sociais) estão sempre sendo atualizadas computacionalmente.

> Algumas plataformas sobre as quais operam as mídias sociais, como o Orkut e o MySpace, foram superadas, enquanto outras, como o Facebook, são constantemente alteradas para se manterem ativas. Sendo assim, tanto nossas definições quanto nossas abordagens apresentam certo dinamismo, pois tratamos de um objeto que se mantém em constante movimento. (MILLER et al., 2019, p. *xi*)

Dentre as várias redes sociais on-line, o Twitter (criado em março de 2003, renomeado para *X* em julho de 2023) está entre as mais usadas até o momento para analisar o comportamento e o desenvolvimento de grupos e indivíduos. Por exemplo, Paiva et al. (2023) estudam a movimentação no Twitter brasileiro durante as eleições presidenciais de 2022. As análises focam no debate on-line de temas sensíveis (como aborto e LGBTfobia) e mostram como as pessoas compartilham suas opiniões e preocupações no mundo on-line. Silva e Faria (2023) também analisam as mesmas eleições, mas considerando a presença de presidenciáveis apenas. Igualmente, Kappaun e Oliveira (2023) utilizam também ferramentas computacionais para mostrar a discussão política on-line para as eleições de 2018 e 2022.

Além de terem sido publicados no principal evento nacional sobre redes sociais, esses três trabalhos analisam as discussões on-line durante as eleições presidenciais brasileiras na plataforma

Twitter. Mais importante, todas as análises são realizadas a partir da aplicação de ferramentas computacionais sobre dados provenientes do Twitter e fornecem visualizações diferentes e complementares. Porém, tais análises são limitadas ao conhecimento técnico das autoras dos artigos (coincidentemente, todas mulheres). Nenhum dos artigos provê análise mais abrangente da perspectiva social ou histórica. Ou seja, existe um vão relevante a ser preenchido pela interdisciplinaridade das Ciências Humanas com a Computação.

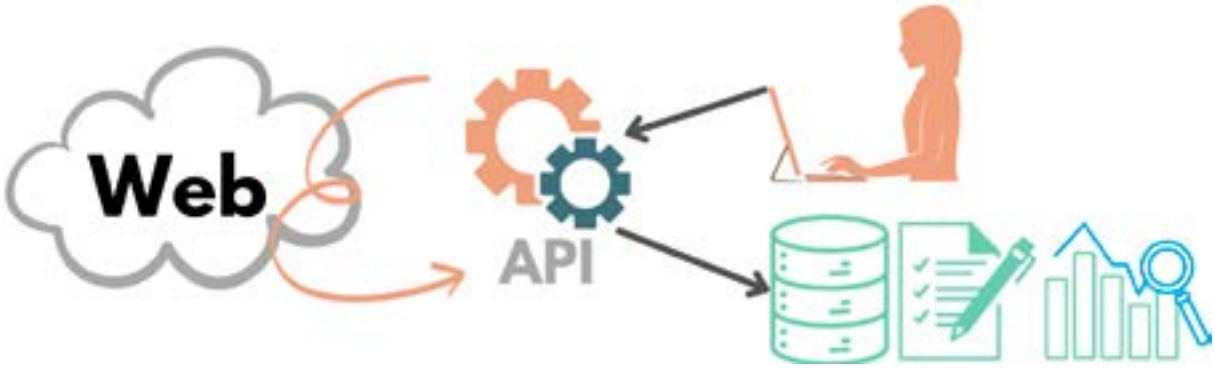

Figura 2. Pesquisadoras interagem com ferramentas *web* através de APIs, as quais retornam dados que podem ser armazenados, conferidos e usados em análises e estatísticas

Fonte: As autoras

De fato, qualquer pessoa pode interagir com ferramentas disponíveis na *web* através de suas interfaces de programação (API – *Application Programming Interface*), as quais disponibilizam dados que podem ser armazenados e usados para diversas análises e estatísticas, conforme ilustrado na Figura 2. De modo geral, a metodologia nesses casos é composta por coleta das mensagens (*tweets*, postagens), imagens ou vídeos, bem como descritores específicos sobre a autoria de cada item. Tal conjunto de dados (organizado ou não) é então utilizado em análises iniciais para obter estatísticas básicas (total de pessoas na plataforma, localização, picos de utilização, etc.) e em análises mais complexas (discutidas a seguir).

Em especial para o Twitter, existem provavelmente centenas de tutoriais e materiais on-line explicando como sua API funciona e como seus dados podem ser explorados. Em especial, Xavier e Souza (2018) apresentam um minicurso sobre ferramentas e métodos de Processamento de Linguagem Natural e Aprendizado de Máquina para coleta e análise de informações semânticas de mensagens no Twitter (lembrando que em 2023, a política de disponibilização de dados da plataforma mudou). É importante notar que, mesmo esse minicurso sendo voltado para pessoas com conhecimento técnico de computação, ele apresenta uma boa ideia sobre o que é possível processar a partir de dados do Twitter. Além disso, várias das técnicas apresentadas podem ser aplicadas a dados provenientes de outras plataformas sociais on-line, com provavelmente pequenas adaptações.

Considerando diferentes redes sociais on-line, Benevenuto (2010) apresenta: conceitos e características das redes sociais no geral e das mais populares (naquela época); principais métricas (clássicas e relevantes atualmente) e análises usadas no estudo dos grafos ou topologias das redes sociais (ou redes complexas); as principais abordagens utilizadas para se obter dados de redes sociais on-line (a maioria ainda válidas hoje); e trabalhos que utilizaram essas técnicas. Apesar de tal minicurso ter sido direcionado para o público de computação, muito do seu conteúdo é apresentado em

linguagem abrangente, com boas referências para quem quiser aprofundar em cada tópico. Mesmo tendo sido publicado em 2010, seu conteúdo continua relevante hoje, especialmente as métricas de estudo das redes sociais complexas.

Finalmente, é importante notar que às vezes é necessário coletar dados de diferentes aplicações on-line. Nesse caso, o minicurso de Batista et al. (2018) é uma excelente fonte técnica para quem quiser saber como funciona o processo complexo de integrar tais dados.

Jupyter Notebook

Provavelmente, uma das ferramentas mais úteis para processamento de dados e sua visualização seja atualmente a plataforma Jupyter Notebook. Em vez de ser limitada a apenas uma linguagem de programação e poucos recursos para construção de gráficos interativos (como a maioria dos ambientes de programação), Jupyter Notebook permite *programação literária interativa*, e ainda conta com vasta literatura disponível gratuitamente on-line, com muitos exemplos em diferentes contextos.

> O paradigma de programação literária busca ajudar na comunicação de programas através da alternância de texto em linguagem natural formatada, pedaços de código executáveis, e resultados de computações. O texto em linguagem natural é usado tanto para explicar o código quanto para comentar o resultado obtido [Perkel 2018]. A interatividade do Jupyter permite que este paradigma seja usado em tempo real para análises de dados, com o processo sendo documentado durante o desenvolvimento, resultados exibidos de forma instantânea e discutidos imediatamente em linguagem natural. (PIMENTEL et al., 2021, p. 14)[138]

Do ponto de vista de profissionais de História, Dombrowski, Gniady, Kloster (2019) apresentam um argumento robusto para a utilização de Jupyter Notebooks.

> E se você pudesse publicar sua pesquisa em um formato que desse um peso equilibrado entre a prosa e o código? A realidade das atuais diretrizes de publicação acadêmica significa que a separação forçosa do seu código e da argumentação pode ser uma necessidade, e sua reunificação pode ser impossível sem que se navegue por numerosos obstáculos. Atualmente o código é tipicamente publicado em separado no GitHub ou em outro repositório, caso no qual os leitores têm que procurar uma nota de rodapé no texto para descobrir quais scripts estão sendo referenciados, encontrar a URL do repositório, acessar a URL, procurar os scripts, baixá-los e também os ficheiro(s) de dados associados, e então executar os códigos. No entanto, se você tiver os direitos e permissões necessários para republicar o texto de sua pesquisa em outro formato, o Jupyter Notebook fornece um ambiente onde código e prosa podem ser justapostos e apresentados com igual peso e valor. (DOMBROWSKI; GNIADY; KLOSTER, 2019)

O Jupyter é uma iniciativa de código aberto e sem fins lucrativos, desenvolvido por uma comunidade de pessoas programadoras. É uma plataforma gratuita e está disponível para qualquer pessoa utilizar (Jupyter, 2023). Para download, recomendamos instalar via Anaconda (2023), que além de instalar o Jupyter, instala as principais bibliotecas Python utilizadas em várias tarefas de processamento e ciência de dados. Essa opção de instalação também é sugerida e bem detalhada por Dombrowski, Gniady, Kloster (2019).

138 Referência dentro do trecho: [Perkel 2018] Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, 563:145–146.

De modo geral, para as pessoas que não têm conhecimento em programação, o Jupyter Notebook pode ser compreendido como uma versão avançada e programável de outras ferramentas mais simples, como planilhas eletrônicas. Ele permite o acesso a dados em diversos formatos (incluindo planilhas) e oferece vários recursos para manipulação de dados, incluindo funções e procedimentos personalizados. A apresentação de resultados é direta e acessível, tornando a interpretação dos dados mais clara para todas as pessoas envolvidas.

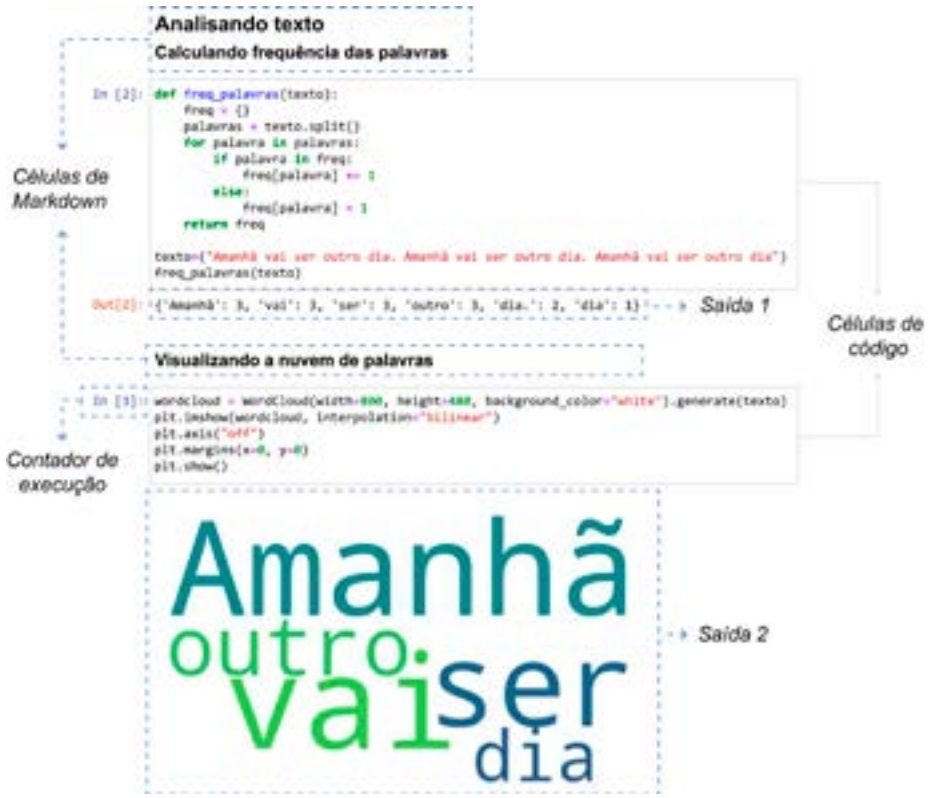

Figura 3. Componentes e exemplo de funcionamento do Jupyter Notebook

Fonte:As autoras

A Figura 3 ilustra um exemplo de execução de notebook Jupyter e seus componentes: texto (*markdown*), código e contador, saída (ou resultado). Texto pode ser adicionado a células do tipo *markdown* (marcação) e formatado de maneira simples e intuitiva. Essas células permitem criar títulos, listas, links, bem como adicionar imagens (entre outras possibilidades mais avançadas). Trechos de código são adicionados em células específicas, as quais possuem um contador (aumenta um a cada vez que a célula é executada). Cada trecho de código pode ser executado individualmente, produzindo uma saída específica. No exemplo ilustrado, o resultado (*Saída 2*) é uma figura contendo uma nuvem de palavras.

De fato, o Jupyter Notebook oferece flexibilidade para realizar análises complexas. Ele suporta a integração de várias bibliotecas e *frameworks*[139] de análise de dados, permitindo explorar dados de diferentes maneiras, úteis para questões interdisciplinares que requerem variedade de métodos analíticos. Além disso, o Jupyter Notebook possibilita a criação de gráficos e visualizações de dados de

139 Em desenvolvimento de software, um *framework* é uma abstração que une códigos comuns entre vários projetos de software provendo uma funcionalidade genérica.

maneira interativa e personalizada; crucial quando a apresentação visual de resultados desempenha papel central na comunicação de descobertas.

Porém, a real "magia" do Jupyter Notebook está em sua interatividade. Sempre que ocorrerem mudanças nos dados ou nas operações, basta executar novamente o notebook (i.e., inteiro ou apenas as células necessárias), e os resultados serão automaticamente atualizados. Isso economiza tempo e esforço, permitindo que analistas ajustem suas abordagens de acordo com as necessidades em constante evolução. Esse recurso, como evidenciado por Pimentel *et al.* (2021), promove uma colaboração mais eficaz e uma compreensão mais profunda dos dados, uma vez que os *insights* podem ser compartilhados e debatidos instantaneamente.

Outro benefício é a promoção da reprodutibilidade e colaboração. Os notebooks podem ser compartilhados facilmente com colegas e grupos de interesse, permitindo que qualquer pessoa reproduza as análises e valide os resultados. Isso é fundamental para garantir a confiabilidade da pesquisa e a integridade dos dados, valores centrais na pesquisa acadêmica.

Um minicurso recente, acessível e completo (desde introdução aos conceitos até análises complexas) é apresentado por Pimentel et al. (2021). O funcionamento e a programação do Jupyter são ensinados considerando um conjunto de dados relativamente simples, relacionado às paradas de sucesso da indústria da música. O minicurso explora várias funções úteis para Ciência de Dados, mas que podem ser facilmente utilizadas em outros contextos que precisam de dados. Ele possui uma página on-line com texto, *slides*, dados e link para os vídeos com apresentação na Escola Inverno da Universidade Federal Fluminense (LINKTREE, 2023). Outro vídeo (de seis horas, em duas partes) desse mesmo curso está disponível em SBBD, 2023.

Análises de Dados Textuais

Existem várias formas de analisar dados textuais. O livro organizado pelas pesquisadoras Caseli e Nunes (2023) é um ótimo material para Processamento de Linguagem Natural (PLN) em português brasileiro que cobre vários aspectos técnicos introdutórios e complexos do ponto de vista computacional. O livro estende bastante um minicurso da primeira autora (CASELI; FREITAS; VIOLA, 2022). Agora, considerando especialmente o contexto de dados disponíveis na *web*, esta seção resume ferramentas para auxílio em: visualização rápida de conteúdo textual; identificação de tópicos; e análise de sentimentos.

Visualização rápida de conteúdo

Com tantos textos e dados disponíveis on-line, às vezes se deseja apenas ter uma rápida ideia do que aquele conteúdo apresenta, como um resumo automaticamente gerado. Essa também é uma tarefa de PLN chamada de *sumarização textual*, ou seja, "gerar um texto mais curto que o original e que ainda seja fluente e fiel ao texto-fonte" (CASELI e NUNES, 2023, p. 407). De fato, resumir texto ainda é uma tarefa complexa computacionalmente, mas existe uma alternativa relativamente bem mais simples e interessante: *resumo visual de palavras-chave*. A partir de um texto, um algoritmo identifica palavras, as agrupa e conta, para então apresentar visualmente de acordo com a quantidade de cada uma. A técnica mais conhecida é, provavelmente, a criação de nuvem de

palavras (*wordcloud*), uma representação visual da frequência e da importância de termos e palavras em um texto. A nuvem de palavras auxilia no reconhecimento fácil das palavras mais usadas e ainda apresenta tudo de maneira mais intuitiva (e.g., quanto maior a fonte da palavra, maior sua frequência no texto).

Existem diferentes implementações e ferramentas para criar nuvem de palavras, inclusive on-line e gratuitas, como os *websites wordclouds.com* e *wordart.com*. A criação de nuvens nesses *websites* é relativamente fácil: primeiro, deve-se entrar com o texto (sendo que *WordClouds* também permite importar os dados direto de uma planilha, extrair de arquivo texto, documento PDF e de páginas *web* direto); depois, pode-se editar a lista de palavras extraída (inclusive agrupando termos manualmente); para então configurar a apresentação visual (selecionar uma forma, e.g., nuvem; selecionar fonte, direção e cor das palavras; e muito mais). A Figura 4 ilustra dois exemplos de nuvens de palavras geradas a partir da letra da música *Apesar de Você*, de Chico Buarque (por simplicidade, ambas as nuvens apresentam apenas palavras que aparecem pelo menos duas vezes na letra).

Figura 4. Exemplos de nuvens de palavras geradas a partir da letra da música *Apesar de Você*, de Chico Buarque, utilizando *WordClouds* (esquerda) e *WordArt* (direita).

Fonte: As autoras

Identificação de tópicos

Megan R. Brett (2012) faz uma excelente introdução à Modelagem de Tópicos destinada a profissionais de História, explicando que:

> A modelagem de tópicos é uma forma de mineração de texto, uma forma de identificar padrões em um corpus. Você pega seu corpus e o executa através de uma ferramenta que agrupa palavras em "tópicos". [...] O que é então um tópico? Uma definição [...] descreveu um tópico como "um padrão recorrente de palavras co-ocorrentes". Uma ferramenta de modelagem de tópicos procura esses grupos de palavras em um corpus e os agrupa por um processo de similaridade [...] Em um bom modelo de tópicos, as palavras do tópico fazem sentido, por exemplo, "marinha, navio, capitão" e "tabaco, fazenda, colheita". (BRETT, 2012, tradução nossa)

Um método relativamente fácil de usar para modelagem de tópicos é o *Latent Dirichlet Allocation* (LDA), que compara a ocorrência de tópicos em um documento a como as mesmas palavras têm sido usadas em outros documentos a fim de encontrar a melhor correspondência (BLEY; NG; JORDAN, 2003). Existem diferentes ferramentas para modelagem de tópicos, a

maioria para a língua inglesa (BRETT, 2012). Porém, o LDA tem mostrado sucesso ao identificar tópicos na língua portuguesa. Por exemplo, em (BATISTA, 2020), a autora escolhe utilizar o LDA para encontrar a temática das ementas das proposições legislativas iniciadas na Câmara dos Deputados de 1995 a 2014; e em (CORRÊA e DE FARIA, 2021), as autoras o utilizam para analisar automaticamente relatos, disponíveis publicamente na internet (sites de notícias, redes sociais, blogs), de vítimas de violência contra mulher no Brasil.

Uma alternativa mais moderna é o BERTopic (GROOTENDORST, 2020a, 2020b), cujos resultados podem ser mais difíceis de serem interpretados, vide o tutorial (em inglês) de Mansurova (2023). Ainda, exemplos de trabalhos da Computação usando o BERTopic com textos em português incluem (CAPELLARO, 2021; VIANNA e DE MOURA, 2022).

Análise de sentimento

Análise de sentimento é um problema que pode ser trabalhado computacionalmente dentro de PLN (LIU, 2010; CASELI e NUNES, 2023), embora muitas vezes envolva pesquisa multidisciplinar com Psicologia e Ciências Sociais. A análise de sentimento visa desenvolver métodos e ferramentas computacionais para extrair e classificar opiniões e emoções expressadas por pessoas em redes sociais, blogs, fóruns e similares (PEREIRA, 2021). A análise de sentimentos pode ser dividida em duas tarefas: (*i*) mineração de opinião, que identifica polaridade como um grau de positividade ou negatividade; e (*ii*) mineração de emoções, que se refere a sentimento de emoção, como felicidade, tristeza e raiva (YADOLLAHI et al., 2017 apud PEREIRA, 2021, p. 1089).

Especialmente, como benefícios para a sociedade, Benevenuto, Ribeiro, Araújo (2015) apontam que "Opiniões nas redes sociais, se devidamente recolhidas e analisadas, permitem não só compreender e explicar diversos fenômenos sociais complexos, mas também prevê-los." Dentre vários conceitos, os autores ressaltam os seguintes:

Polaridade representa o grau de positividade e negatividade de um texto ou trecho de texto. Em PLN, vários métodos de análise de sentimentos retornam polaridade como um resultado discreto binário (positivo ou negativo) ou ternário (positivo, negativo ou neutro). Por exemplo, a frase "Adorei o resultado das eleições" é positiva; "O pior candidato foi eleito" é negativa; e "Amanhã é 7 de setembro" é neutra.

Força do sentimento representa a intensidade de um sentimento ou da polaridade. Pode ser definida como ponto flutuante entre –1 e 1 (ou, $-\infty$ e $+\infty$), e usada como um limiar que define se uma frase é positiva ou negativa.

Sentimento/Emoção indica um sentimento específico presente em uma frase (ex., felicidade, tristeza e raiva).

Além desses termos, é importante introduzir os principais elementos da análise de sentimentos quando realizada através de uma abordagem computacional.

Pré-processamento: identifica palavras-chave para destacar a mensagem central do texto. Inclui pelo menos três passos: tokenização – divide uma frase em vários elementos ou *tokens*; lematização – converte palavras em sua forma raiz (e.g., de "estou" para "estar"); e remoção de palavras irrelevantes – filtra palavras que não agregam valor à frase (e.g., artigos e preposições).

Análise de palavras-chave: define pontuação de sentimento para cada palavra identificada no pré-processamento (e.g., próximo a zero para insatisfação, e próximo a 10 para satisfação completa).

Dicionário léxico: classifica um grande conjunto de palavras (que são frequentes em documentos) em categorias como "positiva" e "negativa". Porém, um dicionário léxico sozinho não é capaz de classificar sentenças de maneira eficaz, e apenas somar a pontuação de cada uma das palavras pode retornar resultados fracos.

Alguns métodos computacionais são capazes de identificar um sentimento específico representado por uma frase. Por exemplo, para a língua inglesa, o EmoLex (Mohammad e Turney, 2013) identifica até nove sentimentos diferentes (alegria, tristeza, raiva, medo, confiança, nojo, surpresa, antecipação, positivo, negativo). Ainda, os autores apresentam uma lista de 14 possíveis aplicações (com devidas referências científicas) para análise de sentimento, incluindo: gerência de relacionamento com clientes, rastreamento de sentimentos em relação a filmes e pessoas da política, e prevenção de suicídios.

Já Pereira (2021) apresenta um levantamento dos esforços feitos especificamente para abordar a análise de sentimentos na língua portuguesa. Ele organiza e descreve trabalhos recentes com abordagens distintas para cada uma das tarefas de análise de sentimento, bem como ferramentas de PLN, léxicos, *corpora*, ontologias e conjuntos de dados.

Existem várias ferramentas para análise de sentimento, a maioria para texto em inglês. Uma das mais utilizadas é, provavelmente, a *vaderSentiment – Valence Aware Dictionary and sEntiment Reasoner* (HUTTO e GILBERT, 2014; HUTTO, 2016), uma biblioteca do Python de código aberto construída para ser usada em tarefas de análise de sentimentos, principalmente aquelas que envolvem dados de mídias sociais. É importante notar que existem adaptações dessa biblioteca para o português, como o LeIA – Léxico para Inferência Adaptada (Almeida, 2018).

Sobre material didático, o minicurso em português de Benevenuto, Ribeiro, Araújo (2015) apresenta visão geral sobre análise de sentimentos e suas aplicações mais populares; discute os principais métodos e técnicas, suas características e formas de execução; e compara tais métodos (vantagens, desvantagens e possíveis limitações). Outro tutorial on-line disponível é o do portal *Programming Historian* em (Saldaña, 2023). Também merece destaque o capítulo de Santana e de Freitas (2023), o qual discute aplicações específicas de análise de texto no contexto de redes sociais e resume um conjunto atual de técnicas e ferramentas disponíveis para aplicações de Detecção de Discurso de Ódio e Linguagem Ofensiva, Análise de Sentimento, Detecção de Notícias Falsas, e Detecção de Ironia/Sarcasmo/Humor.

Considerações Finais

A humanidade está produzindo dados on-line em velocidade surpreendente, de modo que é impossível processar tudo manualmente, ao mesmo tempo que a quantidade que pode ser processada manualmente corresponde a apenas um grão no vasto universo on-line. Este capítulo pode ser lido como uma breve coletânea de definições e referências para diferentes técnicas e ambientes computacionais que estão disponíveis para o processamento de tais dados, especialmente dados textuais. Entende-se que a maioria desses materiais tem o público técnico ou de ciências exatas em mente; ainda assim, foi possível identificar e referenciar materiais em linguagem mais abrangente para pessoas de outras áreas, como as Humanas.

Nota-se ainda que existe muito a ser explorado e criado para tornar tais ferramentas mais acessíveis para todos os públicos. Mesmo com avanços tecnológicos de grande impacto (e.g., Inteligência Artificial), as máquinas ainda estão longe de serem autônomas o suficiente para coletar os dados, processá-los, organizá-los e ainda interpretá-los de maneira adequada. Então, além de pessoas para operá-las adequadamente, as máquinas ainda exigem cientistas para verificar resultados, interpretá-los corretamente e tomar decisões frente a novas descobertas. De modo geral, a sugestão aqui é buscar equipes interdisciplinares que possam se comunicar claramente nos diversos pontos-de-vista exigidos pelos tipos de pesquisa mencionados, como, por exemplo, Computação, História, Ciências Sociais, Antropologia e Psicologia.

Referências

ALMEIDA, Rafael José de Alencar. Análise de Sentimentos em Português. *GitHub*, 2018. Disponível em: https://github.com/rafjaa/LeIA. Acesso em 25/10/2023.

ANACONDA. *Anaconda Distribution*: Free Download. Disponível em: https://www.anaconda.com/download. Acesso em: 28/08/2023.

BATISTA, Mariana. QUAIS POLÍTICAS IMPORTAM? Usando ênfases na agenda legislativa para mensurar saliência. *Revista Brasileira de Ciências Sociais*, v. 35, n. 104, 2020. DOI: 10.1590/3510411/2020.

BATISTA, Natércia A.; BRANDÃO, Michele A.; PINHEIRO, Michele B.; DALIP, Daniel H.; MORO, Mirella M. "Dados de Múltiplas Fontes da Web: Coleta, Integração e Pré-processamento". *In*: Roesler, V.; Kronbauer, A.; Neto, M. C. M.; Novais, R.; Willrich, R. (ed.). *Minicursos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre: Sociedade Brasileira de Computação, 2018. p. 153-192. DOI: 10.5753/sbc.455.7.05

BARBOSA, Bia; TRESCA, Laura; LAUSCHNER, Tanara (orgs.). *3a Coletânea de Artigos TIC, Governança da Internet, Gênero, Raça e Diversidade*. São Paulo: Comitê Gestor da Internet do Brasil, 2023.

BENEVENUTO, F. "Redes Sociais On-line: Técnicas de Coleta e Abordagens de Medição". *In*: PEREIRA, A. C. M.; WINCKLER, M.; GOMES, R. L. (ed.). *Tópicos em Sistemas Colaborativos, Interativos, Multimídia, Web e Banco de Dados*: Minicursos do VII SBSC, XVI WebMedia, IX IHC e XXV SBBD. Porto Alegre: Sociedade Brasileira de Computação, 2010. p. 41-70.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. "Métodos para Análise de Sentimentos em mídias sociais". *In*: FILETO, R.; DA SILVA, A. S.; CRISTO, M.; DE OLIVEIRA, D. F. (ed.). *Minicursos do XXI Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre: Sociedade Brasileira de Computação, 2015. p. 31-59.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, v. 3, p. 993-1022, 2003.

BRETT, Megan R. Topic Modeling: A Basic Introduction. *Journal of Digital Humanities*, v. 2, n. 1, Winter, 2012.

CAPELLARO, Leonardo. *Análise de polaridade e de tópicos em tweets no domínio da política no Brasil*. 2021. 49 f. TCC (Graduação em Curso de Engenharia de Computação) - Universidade Federal de São Carlos, São Carlos, 2021. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/15138/TCC_Leonardo_Capellaro.pdf. Acesso em: 14/11/2023.

CASELI, Helena M.; FREITAS, Cláudia; VIOLA, Roberta. "Processamento de Linguagem Natural". *In*: DA SILVA, T. L. C.; OGASAWARA, E.; SOUZA, D.; LIFSCHITZ, S. (ed.). *Tópicos em Gerenciamento de*

Dados e Informações: Minicursos do SBBD 2022. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 1-25. DOI: 10.5753/sbc.10309.7.1

CASELI, Helena M.; NUNES, Maria da Graça V. (org.) *Processamento de Linguagem Natural*: Conceitos, Técnicas e Aplicações em Português. [*S.l.*]: BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln. Acesso em: 27/10/2023.

CGI.br. *Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros*: TIC Domicílios 2022. Núcleo de Informação e Coordenação do Ponto BR. 1. ed. São Paulo: Comitê Gestor da Internet no Brasil, 2023.

CORRÊA, Isabella Tannús; DE FARIA, Elaine Ribeiro. An analysis of violence against women based on victims' reports. *In*: BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS (SBSI), 12., Uberlândia. *Proceedings…* New York: ACM, 2021. DOI: 10.1145/3466933.3466968

CRUZ, Francisco Brito; BECARI, Jade. Um guia da dieta de mídia digital brasileira. *InternetLab*, 2019. Disponível em: https://internetlab.org.br/pt/pesquisa/um-guia-da-dieta-de-midia-digital-brasileira. Acesso em: 08/11/2023.

DERIU, Fiorenza; IEZZI, Domenica Fioredistella. Text Analytics in Gender Studies. Introduction., *International Review of Sociology*, v. 30, n. 1, p. 1-5, 2020.

DOMBROWSKI, Quinn; GNIADY, Tassie; KLOSTER, David. Introdução ao Jupyter Notebook. *Programming Historian*, 2019. Disponível em: https://programminghistorian.org/pt/licoes/introducao-jupyter-notebooks. Acesso em: 30/10/2023.

EDITORA GLOBO. *TechTudo*. Disponível em: https://www.techtudo.com.br. Acesso em: 27/08/2023.

GOOGLE FOR EDUCATION. *Receba treinamento e suporte para o que você precisar*. Disponível em: https://edu.google.com/intl/ALL_br/get-started/get-product-help. Acesso em 25/08/2023.

GROOTENDORST, Maarten R. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv* abs/2203.05794, 2022a. DOI: 10.48550/arXiv.2203.05794

GROOTENDORST, Maarten R. BERTopic. *GitHub*, 2022b. Disponível em https://github.com/MaartenGr/BERTopic. Acesso em 25/10/2023.

HUTTO, C. J. VADER-Sentiment-Analysis. *GitHub*, 2016. Disponível em: https://github.com/cjhutto/vaderSentiment. Acesso em 25/10/2023.

HUTTO, C. J., GILBERT, Eric. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *In*: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA (ICWSM), 8., 2014, Ann Arbor, Michigan, USA. *Proceedings…* AAAI Press, 2014.

JUPYTER. *About Us*: Project Jupyter's origins and governance. Disponível em: https://jupyter.org/about. Acesso em: 28/05/2023.

KAPPAUN, Andressa; OLIVEIRA, Jonice. Análise sobre Viés de Gênero no Youtube: Um Estudo sobre as Eleições Presidenciais de 2018 e 2022. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 12. , 2023, João Pessoa/PB. *Anais* ... Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 127-138.

LINKTREE. *Ciência de Dados + Jupyter*. Disponível em: https://linktr.ee/uffjupyter. Acesso em: 28/08/2023.

LIU, Bing. "Sentiment Analysis and Subjectivity". *In*: Indurkhya, N. and Damerauthe, F.J. *Handbook of Natural Language Processing*. London: Chapman and Hall/CRC, 2010. p. 1-38.

MANSUROVA, Mariya. Topics per Class Using BERTopic: How to understand the differences in texts by categories. *Towards Data Science*, 2023. Disponível em: https://towardsdatascience.com/topics-per-

class-using-bertopic-252314f2640. Acesso em: 30/10/2023.

MILLER, D.; COSTA, E.; HAYNES, N.; McDONALD, T.; NICOLESCU, R.; SINANAN, J.; SPYER, J.; VENKATRAMAN, S.; WANG, X. *Como o Mundo Mudou as Mídias Sociais*. London: UCL Press, 2019. DOI: 10.14324/ 111.9781787356542.

MOHAMMAD, S. M.; TURNEY, P. D. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, v. 29, p. 436-465, 2013.

MUNDT, Marcia; ROSS, Karla; BURNETT, Charla M. Scaling Social Movements Through Social Media: The Case of Black Lives Matter. *Social Media + Society*, v. 4, n. 4, nov. 2018. DOI: 10.1177/2056305118807911

PAIVA, Beatriz F.; BARBOSA, Beatriz R. G.; SILVA, Ana Paula Couto da; MORO, Mirella M. O debate do feminismo no Twitter: Um estudo de caso das eleições brasileiras de 2022. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 12. , 2023, João Pessoa/PB. *Anais*... Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 103-114.

PEREIRA, D.A. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, v. 54, p. 1087-1115, 2021.

PH. *Programming Historian*. Disponível em: http://programminghistorian.org. Acesso em: 28/08/2023.

PIMENTEL, João Felipe; OLIVEIRA, Gabriel P.; SILVA, Mariana O.; SEUFITELLI, Danilo B.; MORO, Mirella M. "Ciência de Dados com Reprodutibilidade usando Jupyter". *In*: ANDRADE, A. M. S. Andrade; WAZLAWICK, R. S. (ed.). *Jornada de Atualização em Informática 2021*. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 13-62

SALDAÑA, Zoe W. Análise de sentimento para exploração de dados. *Programming Historian*, 2018. Disponível em: https://programminghistorian.org/pt/licoes/analise-sentimento-exploracao-dados. Acesso em: 30/10/2023.

SANTANA, Brenda S.; DE FREITAS, Larissa A. "PLN em Redes Sociais". *In*: CASELI, Helena M.; NUNES, Maria da Graça V. (ed.). *Processamento de Linguagem Natural*: Conceitos, Técnicas e Aplicações em Português. [*S.l.*]: BPLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln/1a-edicao/parte9/cap23/cap23.html. Acesso em: 27/10/2023.

SBBD. *Minicurso 1: Ciência de Dados com Reprodutibilidade usando Jupyter*. Disponível em: https://sbbd.org.br/2021/short-course-1; https://www.youtube.com/@SBBD-/playlists. Acesso em: 28/08/2023.

SILVA, Sarah Maria Braga; FARIA, Elaine Ribeiro de. Análise de sentimentos expressos no Twitter em relação aos candidatos da eleição presidencial de 2022. *In*: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING (BRASNAM), 12., 2023, João Pessoa/PB. *Anais* ... Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 79-90.

TECNOBLOG. [*Site institucional*]. Disponível em: https://tecnoblog.net. Acesso em: 27/08/2023.

VIANNA, Daniela; DE MOURA, Edleno Silva. Organizing Portuguese Legal Documents through Topic Discovery. *In*: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR '22), 45., 2022. *Proceedings*... New York: ACM, 2022. p. 3388-3392. DOI: 10.1145/3477495.3536329

XAVIER, Clarissa Castellã; SOUZA, Marlo. "Extração e Classificação de Dados Semânticos do Twitter". *In*: ROESLER, V.; KRONBAUER, A.; NETO, M. C. M.; NOVAIS, R.; WILLRICH, R. (ed.). *Minicursos do XXIV Simpósio Brasileiro de Sistemas Multimídia e Web*. Porto Alegre: Sociedade Brasileira de Computação, 2018. p. 39-65.