Gustavo Cardoso e Rita Sepúlveda (Orgs.)

Manual de Métodos para Pesquisa Digital

editora
MUNDOS SOCIAIS

Manual de Métodos para Pesquisa Digital

Com o apoio da

IBERIFIER
Iberian Digital Media
Observatory

Gustavo Cardoso e Rita Sepúlveda (Organizadores)

MANUAL DE MÉTODOS PARA PESQUISA DIGITAL

editora MUNDOS SOCIAIS

LISBOA, 2025

Gustavo Cardoso e Rita Sepúlveda (Organizadores)
Manual de Métodos para Pesquisa Digital

Primeira edição: março de 2025

ISBN: 978-989-8536-94-5

Composição em carateres Palatino, corpo 10
Conceção gráfica e composição: Lina Cardoso
Capa: Lina Cardoso
Revisão de texto: Ana Valentim

Este livro foi objeto de avaliação científica

Editora Mundos Sociais, CIES-Iscte, Iscte – Instituto Universitário de Lisboa,
Av. das Forças Armadas, edifício CVTT, sala 007, 1649-026 Lisboa
Tel.: (+351) 217 903 2380
E-mail: editora.cies@iscte-iul.pt
Site: http://mundossociais.com

Índice

Índice de figuras e quadros

Figuras

Quadros

Sobre os autores

Este manual une, na autoria de vários capítulos, colegas de investigação. Têm em comum, entre outros aspetos, participarem na investigação desenvolvida no quadro do MediaLab CIES-Iscte através do qual, entre outras possibilidades e atividades, se realiza investigação e se colocam à prova métodos, desenho de pesquisa e ferramentas digitais. Tal tem permitido aplicar os métodos digitais em diferentes projetos de investigação (Monitorização de propaganda e desinformação nas redes sociais, Covidcheck, Eumeplat, Desinformação política em pré-campanha e campanha eleitoral – 2024, financiados por entidades nacionais (Fundação Calouste Gulbenkian – Gulbenkian Soluções Digitais – Portugal; CNE – Comissão Nacional de Eleições; LUSA) e internacionais (Democracy Reporting International; European Union's Horizon, 2020)).

Os interesses dos autores são diferenciados, nos quais se incluem áreas como a desinformação, a comunicação política, o *online dating*, a plataformização da comunicação ou os estudos de género, mas encontram-se ligados pelo que há de comum na sua investigação: o objeto que é a comunicação e o seu estudo com recurso aos métodos digitais.

Desse conjunto de interesses, simultaneamente convergentes e divergentes, em torno das ciências da comunicação surgiu o projeto MetDigi – Metodologias de Pesquisa Digital, no âmbito do qual surge este manual.

Gustavo Cardoso é professor catedrático de Ciências da Comunicação no Departamento de Sociologia do Iscte-IUL e investigador do CIES-Iscte. Dirige o doutoramento em Ciências da Comunicação e as pós-graduações em Jornalismo e em Comunicação e Assessoria Política. Coordena o MediaLab CIES-Iscte, o OberCom (Observatório da Comunicação) e a participação portuguesa no IBERIFIER – Iberian Digital Media Observatory. A nível internacional é membro do Steering Commitee do Observatory on Information and Democracy e investigador associado no CADIS (Centre d'Analyse et d'Interventions Sociologiques). É membro da Academia de Ciências de Lisboa.

Rita Sepúlveda é investigadora integrada no ICNOVA – Instituto da Comunicação da NOVA, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa. É doutorada em Ciências da Comunicação pelo Iscte-IUL. O seu trabalho tem-se centrado na transformação da intimidade no contexto da apropriação de plataformas digitais com foco nas plataformas de *online dating*. É docente convidada no Iscte-IUL e na Escola Superior de Educação de Coimbra, lecionando unidades curriculares no âmbito da comunicação e das metodologias de investigação. É autora de vários artigos científicos, publicados em diversas revistas indexadas, capítulos de livros e do livro *Swipe, Match, Date* (Arena | Penguin Random House).[1]

Ana Pinto-Martinho é editora do Observatório Europeu do Jornalismo (OEJ), é professora convidada do Iscte-IUL, investigadora do OberCom – Observatório da Comunicação e assistente de investigação do CIES-Iscte. Tem trabalhado em projetos de investigação nacionais e internacionais como o Newsreel, CoMMPASS, Barómetro de Notícias (Newsmetter), Digital News Report, do Reuters Institute for the Study of Journalism, e vários estudos da rede EJO. É doutoranda no Iscte-IUL, com mestrado em Comunicação, Cultura e Tecnologias da Informação (Iscte-IUL) e licenciatura em Comunicação Social (Universidade da Beira Interior).

Cláudia Álvares é professora associada no Departamento de Sociologia do Iscte-IUL e foi eleita, em 2022, para a Comissão Científica do Centro de Investigação e Estudos de Sociologia (CIES). Preside, desde 2024, à Comissão de Ética da Escola de Sociologia e Políticas Públicas do Iscte e integra, desde 2024, a Comissão de Ética da European Communication Research and Education Association, associação à qual presidiu entre 2012 e 2016. O seu trabalho, que abarca áreas como Teorias da Comunicação, Estudos dos Media, Práticas Discursivas, Estudos de Género e Populismo Político, foi publicado nas revistas *Journalism Practice, Women's Studies, Feminist Media Studies, Empedocles: European Journal for the Philosophy of Communication, European Journal of Communication, International Commmunication Gazette, Javnost – The Public e The International Journal of Iberian Studies*. Foi editora da coletânea Routledge Studies in European Communication, integra o Conselho Editorial da *Revista Q1 Feminist Media Studies* e exerce funções regulares enquanto avaliadora.

Inês Narciso é mestre em Criminologia pela Universidade de Leicester e doutoranda em Comunicação no Iscte-IUL. É especialista em OSINT (Open Source Intelligence) e tem desenvolvido trabalho de investigação no MediaLab Iscte-IUL na área da desinformação e das operações de influência. Participa em diversos grupos de trabalho europeus e internacionais nestas áreas, nomeadamente o

---

1 A autora recebeu financiamento por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto no âmbito do projeto "Is that dating related? Mapping dating apps affordances and identifying dating idioms of practices in the Portuguese context" (Referência 2023.09023.ceecind/cp2836/ct0022) com o identificador
DOI: https://doi.org/10.54499/2023.09023.ceecind/cp2836/ct0022.

Code of Practice for Disinformation (CoP), liderado pela Comissão Europeia, e a Partnership for Countering Influence Operations do Carnegie Endowment for International Peace. É também coautora das *Osint Guideline*s que tentam aproximar as práticas de investigação digital em OSINT dos compromissos éticos e de rigor metodológico das metodologias digitais. É voluntária ativa da VOST Europe, uma organização de voluntários digitais, através do qual tem assento no CoP. Desempenhou funções de técnica especialista e de adjunta no XXII e XXIII governos constitucionais de Portugal.

José Moreno é assistente de investigação no CIES – Centro de Investigação e Estudos de Sociologia, e doutorando em Ciências da Comunicação no Iscte-IUL. Estuda as transformações das tecnologias de informação e comunicação na passagem dos *mass media* para a sociedade em rede e as manifestações sociais dessa transição, nomeadamente nas redes sociais. Explora os métodos e ferramentas digitais na investigação sobre comunicação e tem trabalhado predominantemente os fenómenos de desinformação no contexto político e económico. Completou o mestrado em Comunicação, Cultura e Tecnologias de Informação também no Iscte-IUL, com uma dissertação sobre os modelos de negócio dos meios de comunicação social na era digital. Licenciou-se em Comunicação Social no ISCSP – Instituto Superior de Ciências Sociais e Políticas em 1993 e trabalhou como jornalista entre 1992 e 2017, incluindo vários cargos de direção.

Sofia Ferro-Santos é professora auxiliar no IADE e professora convidada no Iscte-IUL e IPPS-Iscte. Licenciada em Ciências da Comunicação, na Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa, e mestre em Gestão pela Nova School of Business and Economics, é doutorada em Ciências da Comunicação pelo Iscte-IUL e investigadora no CIES-Iscte. Participou nos projetos europeus IBERIFIER e EUMEPLAT e na elaboração do Relatório de Comentário Político nos Media 2023 do MediaLab CIES-Iscte. Tem experiência em consultoria de estratégia e comunicação política.

Tiago Lapa é professor auxiliar e investigador integrado no CIES-Iscte, na área da comunicação, no Iscte-IUL. Exerceu ainda funções como professor convidado noutras instituições como o Instituto Politécnico de Leiria ou o ISCSP – Universidade de Lisboa. Leciona em áreas de estudo relativas à sociologia digital e aos métodos de inquirição *online*. Tem participado em redes científicas de âmbito internacional como o World Internet Project e o European Media Coach Initiative, e colaborado em investigações, como as patrocinadas pela fundação "laCaixa" sobre ecrãs móveis e bem-estar subjetivo, entre outras relacionadas com os estudos de *internet*, a divisão digital e a literacia dos novos *media*. Também pertence ao conselho consultivo do Centro Internet Segura da Fundação para a Ciência e Tecnologia (FCT). O produto do seu trabalho científico tem sido publicado em formato de livros, capítulos e artigos em revistas indexadas de circulação nacional e internacional.

Introdução ao manual
Objetivos, desafios e inovação na pesquisa digital

Gustavo Cardoso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

Os métodos digitais são, muitas vezes, ainda apresentados como um subsetor dos métodos e técnicas de investigação. Tal surge, numa analogia à forma como, ainda há alguns anos, falávamos do "digital" para o diferenciar do "analógico" na comunicação. No entanto, da mesma forma que o digital se tornou a norma no contexto comunicacional, também os métodos digitais, muito provavelmente, virão a tornar-se na norma da investigação social e não apenas no quadro das ciências da comunicação.

Não deixa de ser entusiasmante, mas também desafiante, escrever um manual com propostas sobre como realizar investigação no âmbito das plataformas de redes e *media* sociais digitais e na pesquisa digital. Entusiasmante por sentirmos que estamos a contribuir para os próximos passos da ciência, pensando e reimaginando como fazer pesquisa, e desafiante por todas as mudanças que vão surgindo e que, por sua vez, afetam essas novas propostas. Há muito que estudar as interações sociais, que acontecem em contexto *online* ou que tomam lugar através do recurso a plataformas digitais, deixou de ser visto como algo separado da vida "real" e que a presença e relevância daquelas na vida quotidiana deixaram de ser questionadas.

Evidências como a variedade de plataformas disponíveis que atendem a diversas necessidades e propósitos sociais (exemplo: redes sociais *online*, aplicações de bem-estar, plataformas de mensagens, etc.), a transversalidade da presença dessas plataformas em múltiplas atividades do quotidiano e os dados gerados pela sua utilização ajudam a fundamentar o porquê da importância do seu estudo, bem como da necessidade de um manual a si dedicado.

Os dados relativos ao uso de plataformas digitais, nomeadamente das redes sociais *e media* sociais como o Facebook, WhatsApp, Instagram ou o Twitter/X, não surgem num vácuo. Os dados estão ligados a acontecimentos que podem ter origem nas plataformas, mas também influenciam o que ocorre fora destas. As plataformas devem assim ser encaradas como meios tão relevantes como a televisão, rádio e jornais, ou em alguns contextos específicos ainda mais relevantes, para o estudo de variados temas e onde a comunicação acontece. Os estudos dos fenómenos que, de alguma forma, se expressam através de plataformas digitais exigem, por um lado, que os mesmos sejam devidamente enquadrados. Tal poderá passar por

identificar assuntos, sujeitos ou formas de expressão relacionados com esses fenómenos. Por outro lado, exige compreender o meio, isto é, o cenário onde tais fenómenos ocorrem.

Estudar o que acontece em contexto *online* — que pode ter diferentes pontos de partida, variados formatos e múltiplas formas de expressão ou até perspetivas — exige que, intrinsecamente, se estude a plataforma onde tal comunicação ocorre. Esta necessidade deve-se a um motivo específico: as plataformas moldam as formas de expressão e por conseguinte as interações e as práticas comunicativas dos seus utilizadores.

Assumindo que pessoas são a mensagem (Cardoso, 2023) e entendendo que, da mesma forma que as tecnologias fazem coisas, elas também fazem coisas à tecnologia (Bucher, 2018), contornando por vezes imposições e conferindo outros significados a funcionalidades apresentadas e disponibilizadas pelas plataformas, investigar recorrendo aos métodos digitais implica conhecer a plataforma, isto é, conhecer o ambiente de estudo é fundamental para uma adequada compreensão do objeto de estudo.

Nas plataformas digitais, as práticas acontecem através das *affordances*, conceito entendido como o tipo de ações possibilitadas aos participantes na plataforma numa lógica estabelecida entre o *design* da funcionalidade e o contexto do seu uso (Gibson, 1977). Obviamente, sem descurar também as gramáticas ou vernáculos próprios das plataformas (McVeigh-Schultz e Baym, 2015). Referimo-nos às ações de gostar, comentar, partilhar, visualizar, mas também ao recurso a objetos digitais como *hashtags* ou à escolha de formatos para comunicar com imagens (como, por exemplo, *stories*, vídeos) e texto. Tais gramáticas podem ser comuns e transversais a diferentes plataformas (veja-se a presença de *hashtag*s no Facebook, Twitter/X, Instagram ou TikTok), mas também específicas à plataforma, como por exemplo a ação de "dislike" no YouTube.

As especificidades e o âmbito das plataformas digitais sociais, assim como a quantidade de dados gerados pela sua utilização e interações que nelas tomam lugar num contexto de dataficação da comunicação (Cardoso, 2023), conduziram à necessidade de pensar, desenvolver, testar e ensinar novas metodologias de investigação.

As novas metodologias têm em conta aspetos como: 1) os dados serem gerados em contextos muito específicos, idealizados pelos criadores e operacionalizados pelos desenvolvedores das plataformas, com propósitos variados e que não estão necessariamente associados à investigação académica; 2) os dados não serem gerados, propositadamente, como resposta a qualquer estudo que se esteja a desenvolver, como acontece na aplicação de entrevistas ou questionários; 3) poderem envolver uma grande variedade de dados nos quais se incluem imagens, texto, datas, geolocalização, áudio, bem como metadados; 4) serem gerados grandes volumes de dados a uma grande velocidade, o que leva à necessidade de questionar sobre a capacidade de lidar com os mesmos, mas também sobre o(s) momento(s) exato(s) da(s) recolha(s) e espectro temporal da própria análise e; 5) ser necessário ponderar sobre a validade dos dados para responder às perguntas colocadas, seja em termos de amostra, e a sua (in)capacidade de ser representativa, seja na

ausência de determinados dados, dos quais os demográficos são os mais normalmente ausentes.

É no decorrer do conjunto de desafios metodológicos relacionados com o estudo de plataformas digitais sociais que os métodos digitais (Rogers, 2013, 2019) se impuseram como metodologia ao fornecer orientação para a obtenção de respostas. Os métodos digitais têm-se afirmado como uma cultura metodológica para investigar dinâmicas no contexto das plataformas digitais e compreender a sociedade através das mesmas (Sepúlveda *et al.*, 2024).

Os métodos digitais apresentam-se não como meras ferramentas para recolher e processar dados ou como uma adaptação de métodos tradicionais ao contexto digital, mas sim como uma metodologia que situa a pesquisa na plataforma em si. O meio onde a comunicação acontece, o *design*, a lógica e as dinâmicas de funcionamento das plataformas, isto é, as especificidades de cada uma delas, têm de ser tomadas em conta como ambiente explicativo do tópico ou do sujeito que se pretende estudar (Omena, 2019; Rogers, 2013).

Investigar determinados acontecimentos que tomam lugar ou ecoam em plataformas digitais exige, do ponto de vista dos métodos digitais, o conhecimento das plataformas onde se pretende estudar determinado tópico. Seja de uma perspetiva mais técnica, em que se inclui, por exemplo, compreender a arquitetura da rede, as suas funcionalidades, conhecer a interface ou reconhecer as *affordances*, mas também de uma perspetiva sociológica da comunicação, refletindo sobre a apropriação e domesticação das plataformas e o significado que lhes é dado aquando do uso. Esse conhecimento das plataformas é fundamental na hora de definir as perguntas de investigação, desenhar a pesquisa, determinar que ferramentas usar para recolher dados, compreender e analisar esses mesmos dados. É natural que a abordagem metodológica tenha de ser adaptada em função da plataforma, mesmo quando o tópico de investigação é o mesmo (Flores e Sepúlveda, 2020).

Adicionalmente, os métodos digitais como metodologia e parte do desenho de pesquisa não se extinguem numa ótica meramente qualitativa ou quantitativa, mas sim como uma prática qualiquantitativa (Omena, 2019). Os métodos digitais não entram em confronto com os métodos de utilização mais tradicional, nos quais se incluem entrevistas, questionários, observação, diário de campo, entre outros (Silva e Pinto, 1986).

Entendemos também que a possível aplicação dos métodos digitais exige, necessariamente, um moldar no que à sua definição e alcance diz respeito. Tal surge como resposta às mudanças e exigências das plataformas digitais relativamente às abordagens possíveis ou permitidas sobre como investigar um determinado tópico. Não obstante, esse moldar jamais desvirtua a essência da sua definição e aplicação como metodologia.

Os métodos digitais abrangem, assim, uma ampla gama de técnicas, incluindo nestas a mineração de dados, isto é, a técnica assistida por computador usada em análises para processar e explorar grandes conjuntos de dados, a análise de redes e *media* sociais *online*, a etnografia digital (ou netnografia), o *web scraping* ou a análise de texto, para mencionar apenas algumas das mais comuns.

Destacamos como principais vantagens do uso de métodos digitais para realizar investigação: a possibilidade de aceder a grandes conjuntos de dados (*big data*), mas não invalidando o acesso a pequenos volumes; a oportunidade de automatizar processos como a recolha, a análise e a visualização de dados, permitindo economizar tempo em relação aos métodos mais tradicionais; a flexibilidade e possibilidade de adaptação do desenho da pesquisa de acordo com mudanças nas perguntas de investigação; e a hipótese de analisar vários tipos de dados e diversos formatos (textos, imagens, vídeos, *links*, reações, entre outras), podendo assim obter-se uma compreensão mais abrangente do objeto de estudo.

Não obstante, os métodos digitais apresentam também algumas limitações, entre as quais se destacam: a necessidade de competências e recursos técnicos, incluindo acesso a *software*, *hardware* ou, em alguns casos, conhecimento de programação; a amostra ser normalmente obtida por conveniência, não podendo, consequentemente, os resultados ser apresentados como representativos, limitando a generalização dos mesmos; e as fontes de dados poderem, em si mesmas, comprometer a fiabilidade dos resultados, devido, por exemplo, à presença de contas falsas ou *bots* ou ainda aos efeitos da ação algorítmica das próprias plataformas e na forma como apresentam, classificam ou recomendam conteúdo. Tanto as vantagens como as suas limitações devem estar presentes na hora de escolher ou não os métodos digitais e fazer investigação em ambientes digitais.

Fruto da vontade, mas também da necessidade, de investigar em contexto *online*, de ultrapassar os desafios associados, de responder a questões colocadas por alunos, por colegas investigadores e por revisores de artigos científicos relativas a sobre como recolher, analisar e estudar dados provenientes de plataformas digitais, surgiu este manual. Trata-se de um manual pensado para todos os que têm curiosidade sobre, ou desejam, fazer investigação no âmbito de plataformas digitais e/ou sobre as expressões de dimensão social que com elas se relacionam.

Os autores deste manual, em algum momento do seu trabalho como investigadores, também se questionaram sobre o processo de estudar plataformas digitais. A necessidade de obter respostas a essas perguntas levou ao estudo das plataformas digitais, tais como redes e *media* sociais, motores de busca, plataformas de mensagens ou a multitude de plataformas temáticas que tem vindo a manifestar-se na comunicação em rede, com o objetivo de entender em que consistem, saber mais sobre o seu funcionamento, explorar as suas lógicas de uso e conhecer culturas ou práticas inerentes às mesmas.

Esse processo levou também à exploração de métodos e técnicas de investigação, além daqueles e daquelas entendidas como tradicionais e, consequentemente, à experimentação com ferramentas e *softwares* diversos e ao desenho da pesquisa, tendo em conta as especificidades do meio. Tal processo pode ser caracterizado pela criatividade, persistência, questionamento e permanente atualização necessárias, podendo afirmar-se que são essas as características que moldam a cultura da descoberta e inovação nos métodos digitais.

Este manual constitui parte do processo atrás descrito. A ideia, e grande desafio, deste manual consiste em conseguir sintetizar noções básicas sobre como fazer investigação no âmbito das plataformas digitais através de ferramentas, mas

obviamente sem nunca esquecer todo o contexto onde essa investigação se desenvolve e as questões associadas ao mesmo.

Para tal, propomos aos leitores começarem por um conjunto de capítulos com um teor mais teórico e contextual, fundamentais para entender as questões associadas a fazer investigação no âmbito de plataformas digitais. Através dos capítulos "Plataformas digitais, algoritmos e dados", "Enquadramento ético para a investigação digital" e "Desenhar, planear e estruturar a pesquisa", a pesquisa digital é contextualizada. São apresentados conceitos teóricos como a dataficação da comunicação (Cardoso, 2023) e a plataformização (van Dijck *et al.*, 2018), os quais encerram muita da discussão relativa ao papel das plataformas, expondo a capacidade destas de quantificar as ações dos seus utilizadores ou a presença das mesmas, e o seu impacto nas diversas dimensões da vida. São abordados conceitos, como *big data*, algoritmos ou inteligência artificial, numa ótica de os desconstruir, na tentativa de melhor entender o seu significado. Igualmente, é colocada a importância da ética no âmbito da pesquisa digital. Numa discussão, ainda sem resposta clara, sobre se os dados disponíveis nas plataformas devem ser encarados como exclusivamente públicos ou privados, é necessário tomar decisões informadas no âmbito da investigação. Importa, assim, debater a importância do desenho da pesquisa, apontando o planeamento e a estruturação como passos fundamentais, visto que serão esses a influenciar as ferramentas a usar e como o fazer.

Os capítulos de maior incidência teórica são complementados com outros centrados numa abordagem prática, sendo apresentados numa lógica "how to". A sua leitura é guiada passo por passo, relativamente ao uso de determinadas ferramentas e à recolha de dados através das mesmas, para que quem o deseje os possa reproduzir. A organização deste segundo conjunto de capítulos está feita segundo cada plataforma. A seleção das plataformas apresentadas está diretamente relacionada com a sua popularidade, relevância e ferramentas disponíveis para recolher dados passíveis de usar numa "ótica de utilizador", isto é, sem conhecimentos específicos de codificação ou programação.

Quanto às técnicas ou estratégias de investigação apresentadas neste manual, podem ser comuns ou aplicadas unicamente a distintas plataformas. No entanto, sempre que possível, tentámos apresentar várias opções para um maior conhecimento das possibilidades à disposição do investigador. Portanto, as abordagens apresentadas devem ser entendidas como complementares entre si e não encaradas como únicas.

Por último, numa lógica complementar em relação ao conjunto de capítulos mais práticos, apresenta-se o capítulo "A análise de dados: propostas, exemplos e sugestões". Como se pode antecipar, as possibilidades são múltiplas. Com base na experiência dos autores, são apresentados alguns exemplos, tendo por base estudos publicados e sugestões que vão sendo testadas. Novamente, estas análises não são únicas nem exclusivas, antes procuram sensibilizar pedagogicamente os leitores para as características do universo dos métodos digitais. Reconhecemos que a escrita de um manual para esta área de estudo e a apresentação de uma proposta metodológica estão associadas a vários desafios e riscos. Estes prendem-se com o acesso às ferramentas, ao funcionamento das mesmas, à sua manutenção ou até ao

permanente surgimento de novas alternativas. Tais riscos estão intimamente ligados à "vontade" das plataformas digitais de permitirem ou não recolhas, definirem acessos à sua API, mas também dos legisladores e organismos reguladores encararem uma determinada plataforma de uma dada forma, desenvolvendo "jurisprudência" sobre o seu comportamento quanto ao acesso de dados. Assim, é importante ter consciência de que a informação que consta neste manual necessita de uma atualização permanente, não compatível com a edição de mais uma versão atualizada de um livro, mas sim tendo mais que ver com a lógica de versões, tal como na prática do desenvolvimento de *software*. A aprendizagem no âmbito da investigação digital com recurso aos métodos digitais é contínua. Essa é a ideia que norteia o capítulo final deste manual: "Prospetiva e pesquisa digital: futuros, desafios, oportunidades e tendências".

Desde que os autores começaram a trabalhar neste manual, ocorreram vários desenvolvimentos em diversos campos com impacto nos *media* sociais. De repente, passámos a discutir qual o papel do ChatGPT, dos grandes modelos de linguagem (em inglês Large Language Models — LLM) e outros serviços de inteligência artificial na análise de dados, escrita e revisão de artigos científicos. Vimos tornar-se objeto de atenção pública a discussão e avaliação do impacto do Digital Services Act (DSA), o Regulamento Europeu sobre os Serviços Digitais. Isto é, o regulamento sobre como as redes e *media* sociais ou as lojas de aplicações móveis devem levar em conta os interesses mais vastos da sociedade e dos cidadãos que neles participam. Sendo o DSA implementado numa ótica de proteção dos participantes nas plataformas, passou a ter implicações diretas no grau de restrição ou abertura por parte dessas quanto ao acesso dos investigadores à recolha de dados. Surgiram igualmente, e continuam a surgir, novas revistas académicas que incidem especificamente sobre plataformas, criando novas possibilidades sobre onde publicar trabalhos sobre os temas dos métodos digitais. Nas universidades e politécnicos, novas unidades curriculares são conceptualizadas, *workshops* e formações são oferecidas a cada novo semestre iniciado.

Estes acontecimentos mostram como todos aqueles que pretendem estudar plataformas sociais digitais, e investigar fenómenos que nelas tomam lugar, deparam com um campo de estudo em perpétuo movimento e com a necessidade de regularmente repensar metodologicamente como estudá-lo.

De facto, os fenómenos no âmbito ou relacionados com os meios sociais digitais desafiam constantemente os investigadores. Este desafio pode estar relacionado com o desenho da pesquisa, com as ferramentas que vão ser utilizadas, com os dados aos quais estas têm acesso, como os recolhem e como esses dados serão analisados. Procurámos, assim, ao longo dos vários capítulos deste manual, tornar todos esses desafios tão claros quanto possível.

Os autores partilham aqui o manual que eles próprios gostariam de ter encontrado quando iniciaram as suas investigações nesta área das ciências da comunicação. Os autores reúnem, assim, neste manual contributos que acreditam poder fazer a diferença no pensar a investigação no contexto particular do digital.

Este manual procura dotar de competências de investigação os investigadores com curiosidade neste tipo de métodos e técnicas, mas é também a assunção de que muito proximamente os "métodos digitais" de hoje passarão a ser apenas os

normais "métodos" de investigação do amanhã, quando terminar o seu período de adoção por todas as ciências sociais e quando forem ensinados a todos os alunos dos primeiros anos de licenciaturas. Nesse momento, este manual deixará de ser necessário. Até lá, esperamos com o nosso contributo ajudar a que se continue a percorrer a estrada da ciência, moldando os ventos do contexto, dados pelas plataformas, e navegando com as ferramentas que existem e, entretanto, surgirem para a investigação.

Referências

Bucher, T. (2018), *If... Then: Algorithmic Power and Politics*, Oxford University Press.

Cardoso, G. (2023), *A Comunicação da Comunicação. As Pessoas São a Mensagem*, Lisboa, Mundos Sociais.

Flores, A.M. e Sepúlveda, R. (2021), "Métodos digitais e educação: uma proposta de investigação", em Ana Nobre, Ana Mouraz, Marina Duarte (eds.), *Portas Que o Digital Abriu na Investigação em Educação*, Universidade Aberta, pp. 226-255, 10.34627/uab.edel.15.11.

Gibson, J. (1977), "The theory of affordances", em R. Shaw e J. Bransford (eds.), *Perceiving, Acting, and Knowing. Toward an Ecological Psychology* Lawrence Erlbaum, pp. 67-82.

McVeigh-Schultz, J., e Baym, N. (2015), "Thinking of you: vernacular affordance in the context of the microsocial relationship app, couple", *Social Media + Society*, pp. 1-13, DOI: 10.1177/2056305115604649.

Omena, J.J. (ed.) (2019), *Métodos Digitais. Teoria-Prática-Crítica*, Livros ICNOVA.

Rogers, R. (2013), *Digital Methods*, MIT Press.

Rogers, R. (2019), *Doing Digital Methods*, Sage.

Silva, S. A., e Pinto, M. J. (1986), *Metodologia das Ciências Sociais*, Edições Afrontamento.

Sepúlveda, R., Narciso, I., Moreno, J. e Palma, N. (2024), "Pensar, desenhar e realizar investigação no contexto da nova comunicação", em G. Cardoso (coord.), *A Nova Comunicação*, Almedina, pp. 71-86.

van Dijck, J., T. Poell, e M. De Waal (2018), *The Platform Society. Public Values in a Connective World*, Oxford University Press.

Parte 1 | Conceptualizar a pesquisa digital

Capítulo 1

Plataformas, algoritmos e dados

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Estudar plataformas digitais

O primeiro capítulo deste manual obriga, necessariamente, a conceptualizar o que são plataformas digitais. Largamente adotadas como meio através do qual grande parte da população mundial realiza um conjunto de atividades quotidianas, as plataformas digitais são infraestruturas *online* que facilitam as interações entre utilizadores e fornecedores, permitindo a troca de informações, bens, serviços e interações sociais através da *internet*. Talvez as plataformas digitais mais evidentes, pela sua dimensão e utilização, sejam as plataformas de *social media*, isto é o Facebook, o Instagram, o TikTok ou o X, para mencionar aquelas que acumulam maior número de utilizadores mundialmente. Porém, existem outras tipologias de plataformas digitais, cada uma servindo diferentes propósitos e funções, que permitem a realização das mais variadas tarefas em diversas áreas. Assim, às plataformas de *social media* juntam-se plataformas de *e-commerce* (exemplo: Amazon, eBay ou Alibaba), de conteúdo (exemplo: Netflix ou Spotify), de mensagens (exemplo: WhatsApp ou Telegram), de serviços (exemplo: Uber ou Airbnb), as colaborativas (exemplo: GitHub ou Wikipédia), as educativas (exemplo: Coursera ou Duolingo) ou as financeiras (exemplo: PayPal ou Stripe).

De forma geral, estas plataformas oferecem uma estrutura, que pode ser no formato *site* ou aplicação móvel (*app*), através da qual são permitidas diversas atividades, tais como a comunicação, o comércio ou a criação de conteúdos, aproveitando as vantagens das tecnologias digitais e conectando partes distintas. Assim, entre as suas vantagens destacamos o facto de melhorarem a conectividade, o acesso à informação, as oportunidades económicas, a inovação ou o envolvimento social. Têm a capacidade de moldar a forma como os indivíduos interagem, aprendem, trabalham e participam na sociedade.

Podemos destacar, como características-chave das plataformas digitais, a sua capacidade de intermediação, uma vez que as plataformas atuam como intermediárias entre diferentes grupos de utilizadores, como os consumidores, produtores, vendedores, ou criadores de conteúdo e públicos; o efeito de rede, traduzindo-se no

aumento do valor da plataforma à medida que mais utilizadores aderem e participam; e a sua escalabilidade, podendo estas ser dimensionadas, acomodando um número crescente de utilizadores e respetivas interações.

Os efeitos de rede gerados pelas plataformas digitais são um elemento a destacar, uma vez que ajudam a explicar o carácter centrípeto que tendem a exercer sobre os vários tipos de utilizadores e a sua tendência para resultarem em monopólios ou quase monopólios. Os efeitos de rede podem ser decompostos em:

— efeitos de rede diretos: quanto mais pessoas usarem uma plataforma, mais ela é atrativa para todos os utilizadores;
— efeitos de rede indiretos: quanto mais pessoas usarem uma plataforma, mais ela é atrativa para todos os serviços complementares (desenvolvedores, criadores de conteúdos, ferramentas, etc.);
— efeitos de rede relacionados com os dados: quanto mais pessoas usarem uma plataforma (e quanto mais intensamente o fizerem), mais dados serão gerados, os quais poderão alimentar outros serviços ou mesmo plataformas, como a publicidade, por exemplo (Nielsen e Ganter, 2022).

A conjugação desses diferentes efeitos é o que produz a atratividade das plataformas digitais para os diversos utilizadores. A Lei de Metcalfe, cunhada por Robert Metcalfe (2013) no final dos anos 1980 para explicar a expansão da *ethernet*, postula que o valor de uma rede é o quadrado do número de utilizadores, o que significa que o valor da rede aumenta exponencialmente em relação ao número de utilizadores. O que, por sua vez, incrementa os chamados "switching costs", estando por isso na base da tendência monopolística das plataformas digitais.

É isso que converte as plataformas digitais nos espaços preferenciais de comunicação para a maioria das pessoas (Newman *et al.*, 2023) e, portanto, também em espaços fundamentais para o estudo dos fluxos de comunicação que aí são gerados.

Não obstante, é também fundamental reconhecer os desafios e potenciais desvantagens associadas às plataformas digitais. Entre eles destacamos: 1) as questões no âmbito da privacidade e segurança de dados; 2) o papel das plataformas como meio para propagar desinformação e notícias falsas e a sua influência na política e na opinião; 3) o impacto na saúde mental e no bem-estar, nomeadamente associados ao uso excessivo ou à comparação social; 4) assim como o impacto nos relacionamentos e interações sociais, remetendo para relações superficiais e comodificação de emoções; e 5) as preocupações éticas e a necessidades de regulamentação sobre o funcionamento, recolha e uso de dados. Perante os desafios apresentados, os investigadores, os decisores políticos e as partes interessadas devem trabalhar em conjunto no sentido de garantir o potencial positivo das plataformas digitais.

A relevância do conceito de plataformização

A conceptualização do estudo no âmbito das plataformas digitais e a inerente presença de infraestruturas digitais económicas conduz-nos ao conceito de plataformização. Este refere-se ao processo através do qual vários aspetos da vida económica, social e cultural são reorganizados em torno de plataformas digitais (van Dijck *et al.*, 2018). Estas plataformas, muitas vezes fornecidas por empresas como a Google, a Meta, a Amazon ou a Uber, servem como intermediárias que facilitam as interações entre utilizadores, empresas e outras entidades.

O conceito de plataformização abrange uma vasta gama de fenómenos, um dos principais prende-se com as mudanças nos modelos de negócio, nas subjacentes dinâmicas de mercado e nas implicações que tais modelos têm nas práticas laborais. Através das plataformas digitais, novos modelos e formas de fazer negócio foram surgindo. Veja-se por exemplo o impacto de plataformas como a Amazon e o Alibaba no setor de retalho. O abandono do modelo convencional assente em lojas físicas em favor de um mercado digital levou a um aumento da acessibilidade, mas também introduziu desafios como a monopolização do mercado e a redução do controlo por parte dos vendedores individuais, que têm de obedecer às regras e algoritmos dessas plataformas.

Um outro exemplo de como a plataformização teve impactos significativos no mercado de trabalho encontra-se em plataformas como a Uber. Esta facilita a ligação entre prestadores de serviços e consumidores, permitindo acordos de trabalho flexíveis e serviços a pedido. Embora esta dinâmica possa proporcionar maior flexibilidade e autonomia aos trabalhadores, muitas vezes ocorre à custa da segurança no emprego, dos benefícios e das proteções laborais tradicionais. Esse exemplo ilustra claramente o conceito de *Gig economy* (Vallas e Schor, 2020) que expõe a dupla natureza da plataformização: oferece oportunidades e desafios, como a flexibilidade e a autonomia dos trabalhadores, mas remete para as potenciais desvantagens, como a insegurança no emprego e a falta de benefícios.

De um ponto de vista estrutural, o funcionamento atual das plataformas digitais é fundamentalmente influenciado, por um lado, pelas políticas e algoritmos de gestão dos dados e, por outro, pelos modelos de negócio associados e para os quais a arquitetura das plataformas é pensada. Van Dijck (2018) distingue dois tipos de plataformas: as plataformas infraestruturais — que compõem a estrutura fundamental do que a autora chama a "Sociedade das Plataformas" e que inclui as principais plataformas de redes sociais — e as plataformas setoriais — focadas em nichos de mercado específicos, como a Amazon ou a Uber. Tarleton Gillespie (2010), por seu lado, afirma que as plataformas podem ser entendidas num sentido arquitetural — como o plano comum no qual vários agentes interagem — e num sentido computacional — como o conjunto de programas e aplicações que regem a informação e os dados. Inicialmente, as plataformas eram entendidas como neutrais, igualitárias e abertas à participação de todos, atuando como distribuidoras de informação em vez de serem produtoras, baseando a sua oferta nos conteúdos gerados pelos utilizadores. Segundo Benkler (2006), o que as plataformas fazem é precisamente captar o valor gerado pelos utilizadores, funcionando como "two"

ou "multisided markets". É precisamente a captura desse valor que constitui o modelo de negócio das plataformas de redes sociais que aqui nos interessam. Deste modo, as plataformas digitais são, hoje em dia, por um lado, construções técnico-culturais, com a participação dos utilizadores, alimentando-se dos conteúdos gerados por esses utilizadores, moderadas por várias tecnologias, incluindo os algoritmos, e, por outro lado, estruturas socioeconómicas, com proprietários específicos, regras de moderação determinadas por esses proprietários (exemplo: termos de uso ou condições de utilização) e modelos de negócio desenhados para maximizar o lucro dos mesmos (van Dijck, 2013).

No entanto, é por aqui que passam a maior parte dos fluxos comunicacionais atuais. As plataformas digitais são atualmente os principais reguladores de mercado, seja em termos de produtos, seja em termos de comunicação e partilha de informação. O que significa que é neste quadro complexo que o investigador interessado nas plataformas digitais se tem de mover, tendo em atenção que nada na arquitetura das mesmas foi concebido a pensar na investigação e que terá de ser esta a adaptar-se às contingências da plataforma e tê-las em consideração, tanto no desenho da pesquisa como na apresentação de dados e resultados.

Dataficação e a importância dos dados

As plataformas digitais dependem fortemente de conteúdos gerados pelos utilizadores e da participação ativa destes para funcionarem de forma eficaz. De facto, as pessoas constituem o maior valor das plataformas, e é na sua utilização que fazem destas que reside a sua relevância. Neste contexto, uma outra característica-chave das plataformas digitais é a sua capacidade de recolher grandes quantidades de dados gerados pelas interações dos utilizadores e analisá-los. O objetivo da recolha e análise de dados não é só aprimorar o serviço do ponto de vista do utilizador, mas também se centra na conversão. A conversão pode, em função da plataforma, ser traduzida na frequência da utilização, no facto de as pessoas passarem mais tempo *online*, na compra ou na venda de algo através da plataforma ou na interação com outras pessoas e conteúdos, para mencionar apenas algumas possibilidades.

Essa capacidade de recolha e análise remete para o termo cunhado de *dataficação* (Mayer-Schönberger e Cukier, 2013). Este refere-se ao processo através do qual vários aspetos da vida e das atividades em sociedade são transformados em dados quantificáveis. Tal processo é obviamente impulsionado pela proliferação de tecnologias digitais, da sua presença na realização das mais diversas tarefas diárias, do seu *design* e funcionamento que permitem capturar, analisar e utilizar os dados de formas que anteriormente não eram possíveis.

Aspetos-chave da dataficação envolvem a capacidade de as plataformas digitais transformarem a utilização que os utilizadores fazem das suas plataformas em dados. Isto é, a capacidade que têm em recolher qualquer tipo de interação que acontece na plataforma e transformá-la num dado quantificável. Essa recolha pode ser relativa a interações menos óbvias como, por exemplo, os passos dados para adquirir um determinado livro na Amazon (exemplo: quanto tempo demorou a

decisão de comprar, os outros livros que consultou, aqueles que colocou no carrinho, mas acabou por eliminar, os tópicos pelos quais realizou a sua pesquisa) como aquelas mais óbvias e que acontecem nas plataformas de redes sociais (exemplo: quantos amigos tem, as páginas/perfis que segue, os gostos que dá ou que recebe, as partilhas que realiza, os comentários que faz ou que acumula). A transformação das interações, comportamentos e preferências dos utilizadores em dados pode ser utilizada para diversos fins, como publicidade personalizada, recomendação de conteúdo e análise social.

Com o desenvolvimento e crescente adoção das plataformas digitais como locais por onde passam a maioria dos fluxos comunicativos em sociedade, o que as empresas que gerem essas plataformas fazem é dirigir a atenção, a comunicação e as interações entre os utilizadores, maximizando a sua permanência na plataforma e a geração de receita. Por isso, os processos comunicativos são tecnicamente envolvidos em métricas de sociabilidade como o "gosto", os "amigos", a partilha ou o comentário. Trata-se no fundo de uma sociabilidade "plataformizada" ou "programada" (Bucher, 2018; van Dijck, 2013). E são essas métricas que geram os dados a que o investigador tem acesso e a partir dos quais pode desenvolver a sua investigação. É nesse processo de dataficação que também reside parte das oportunidades do trabalho de investigação no contexto deste manual.

Algoritmos e *big data*

No coração do funcionamento das plataformas estão os algoritmos. Um algoritmo pode ser entendido como um procedimento lógico e codificado para transformar um determinado *input* num *output* igualmente determinado (Gillespie, 2014). Embora todos os computadores funcionem com algoritmo, um algoritmo não tem necessariamente de ser um programa de computador. Uma receita de culinária, por exemplo, pode ser entendida como um algoritmo.

Trabalhando sobre dados, os algoritmos das plataformas funcionam melhor quanto mais dados tiverem disponíveis, incluindo não só os conteúdos produzidos pelas plataformas, mas também todos os pontos de interação entre os utilizadores e a plataforma ou entre os próprios utilizadores (Domingos, 2015). Ou seja, é do interesse da plataforma que os utilizadores gerem a maior quantidade possível de dados e, portanto, que permaneçam na plataforma e sejam ativos enquanto lá estão. Daí que as métricas de interação sejam particularmente importantes para as plataformas.

Por outro lado, ao operar sobre a multitude de dados gerados pelos utilizadores de uma plataforma, os algoritmos priorizam, classificam, associam e filtram esses dados (Diakopoulos, 2014). Para o investigador, o que é relevante ter em consideração é que cada uma dessas ações tem influência nos dados e na interação dos utilizadores com eles. A priorização é uma das mais importantes. Considerando o grande número de utilizadores das plataformas digitais e o enorme volume de dados que a sua ação gera (*big data*), estas plataformas são obrigadas a selecionar os conteúdos que mostram aos seus utilizadores. Ou seja, irão dar prioridade a uns

conteúdos em detrimentos de outros. O que significa que os utilizadores irão ver — e interagir — com esses conteúdos e não com outros. Para o investigador, isto é um fator importante a considerar.

A classificação refere-se ao processo pelo qual os algoritmos das plataformas associam determinados conteúdos ou dados a categorias específicas e não a outras. Por exemplo, uma publicação numa rede social pode ser associada ao género do autor nuns casos, mas não o ser noutros, o que significa que esse elemento pode ser tido em conta na investigação, mas com esta ressalva. A associação, por sua vez, diz respeito às relações estabelecidas entre entidades diferentes. Por exemplo, a associação entre um utilizador e as contas que segue, também gerida pelo algoritmo, está disponível nalguns casos, mas não noutros. Por fim, a filtragem relaciona-se com a moderação de conteúdos e refere-se à seleção de conteúdos que a política de moderação de uma plataforma decida mostrar (ou não mostrar) aos utilizadores, determinando, assim, com quais poderão (ou não) interagir. É este efeito do algoritmo que, combinado com o primeiro, dá origem àquilo a que se convencionou chamar *filter bubbles* ou filtro-bolha (Pariser, 2011).

Como referido acima, este "viés algorítmico" deve ser tido em conta pelo investigador na hora de desenhar a sua pesquisa e na hora de recolher e analisar os dados.

E, embora saibamos que estas são as formas pelas quais os algoritmos influenciam a disponibilidade e a relevância dos dados, a verdade é que os códigos específicos que regulam esses procedimentos não são conhecidos e são guardados pelas empresas que gerem as plataformas como verdadeiros segredos industriais, convertendo os algoritmos em verdadeiras "black boxes" inacessíveis para os investigadores.

Em suma, a moderação de conteúdos através de algoritmos é a essência das plataformas digitais e, na verdade, o seu "produto" principal (Gillespie, 2018). O que necessariamente coloca em causa o mito da "objetividade algorítmica". Os dados gerados nas e pelas plataformas parecem objetivos, mas na realidade são influenciados e manipulados em função de todos os elementos descritos acima.

Uma das grandes vantagens do estudo das plataformas digitais é obviamente a quantidade massiva de dados disponíveis para estudo, sejam relativos a conteúdos comunicativos, sejam relativos aos fluxos dessa comunicação ou da sociabilidade em rede. Mas, perante o que ficou dito atrás, torna-se também claro que, para o investigador, é fundamental ter em conta o contexto — da plataforma e da interação dos utilizadores com a plataforma — em que ocorre a produção dos dados. Isso é fundamental para que outros investigadores possam fazer a reprodução da metodologia e verificar os resultados.

Como veremos ao longo deste manual, a investigação com metodologias digitais é fortemente influenciada pelos dados disponíveis e pelas ferramentas existentes para os recolher e analisar. Por isso, é fundamental uma descrição exaustiva do tipo de dados recolhidos e das ferramentas utilizadas, mas também do contexto — da plataforma e da interação com a plataforma — em que os dados são gerados. Muitas vezes, isso obriga também a trabalhar com diferentes métodos e chamar à colação diferentes áreas científicas, nomeadamente ligadas à computação e às

linguagens de programação. Por isso, a abertura à interdisciplinaridade deve fazer parte da abordagem de investigação quando adotamos métodos digitais.

Tendo em conta a relação entre plataformas, dados e métodos digitais, chamamos a atenção para o contributo de Salganik (2019) que identifica 10 características dos dados digitais que devem ser tidas em conta pelos investigadores:

1. abundância: a utilização de plataformas digitais gera grandes quantidades de dados passíveis de investigação;
2. permanência: os dados das plataformas digitais estão sempre a ser gerados, permitindo a comparação diacrónica e o estudo de fenómenos em tempo real;
3. não reatividade: os fenómenos estudados nas plataformas digitais não são afetados pela presença do investigador, mantendo, desse ponto de vista, a sua pureza;
4. carácter incompleto: os dados gerados em plataformas digitais são gerados com fins diversos dos objetivos da investigação, pelo que poderão não corresponder exatamente às necessidades do investigador;
5. inacessibilidade: o investigador não tem acesso direto aos dados, mas apenas um acesso indireto ao fluxo de dados que os responsáveis decidem tornar acessível e no modo em que o fazem;
6. sem carácter representativo: os dados gerados em plataformas digitais não são representativos da população (porque nem todas as pessoas participam — ou participam da mesma forma em plataformas digitais), não podendo, por isso, em muitos casos, gerar generalização dos resultados;[1]
7. fontes mutáveis: os dados gerados em plataformas digitais mudam com frequência, seja pela forma como são recolhidos, seja pela forma como são disponibilizados ao investigador;
8. viés algorítmico: a ação dos algoritmos sobre os dados gerados em plataformas digitais influencia a produção desses dados e, portanto, a sua "pureza" face à investigação;
9. fácil "contaminação": a recolha de grandes volumes de dados gerados em plataformas digitais, muitas vezes em modo automático, pode incluir dados que não correspondem ao objeto de pesquisa e que, assim, "contaminam" a amostra;
10. dados sensíveis ou não autorizados: os dados gerados pelos utilizadores em plataformas digitais podem incluir dados sensíveis e não são gerados para fins de investigação, o que pode limitar a sua utilização.

Deste modo, as investigações envolvendo métodos digitais incluem uma vantagem evidente em relação aos métodos tradicionais, que se prende com o facto de ser possível observar o comportamento dos utilizadores através da abundância de

1 Embora, em certos casos, os métodos digitais possam proporcionar o acesso à totalidade do universo de estudo, dependendo da forma como esse universo é definido. Por exemplo, se optarmos por estudar a comunicação de um determinado agente político numa determinada rede social, então podemos analisar todas as publicações realizadas por esse agente político nessa rede social.

dados deixados como rasto desse comportamento. Cada clique, cada comentário e cada partilha — de cada utilizador — fica registado nos dados das plataformas digitais. Isto difere significativamente do que acontece nos métodos de investigação tradicionais, em que os dados comportamentais são escassos e normalmente só acessíveis através de inquirição. Mas, em contraponto, os métodos de investigação usando dados de plataformas digitais envolvem outro tipo de desafios metodológicos que devem ser tidos em conta.

Referências bibliográficas

Benkler, Y. (2006), *The Wealth of Networks. How Social Production Transforms Markets and Freedom*, Yale University Press.

Bucher, T. (2018), *If... Then. Algorithmic Power and Politics*, Oxford University Press.

Diakopoulos, N. (2014), "Algorithmic-accountability: the investigation of black boxes", *Tow Center for Digital Journalism*, https://doi. org/10.7916/D8ZK5TW2.

Domingos, P. (2015), *The Master Algorithm. How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books.

Gillespie, T. (2010), "The politics of 'platforms'", *New Media & Society*, 12 (3), pp. 347-364, https://doi.org/10.1177/14614448093427.

Gillespie, T. (2014), "The relevance of algorithms", em Tarleton Gillespie, Pablo J. Boczkowski, e Kirsten A. Foot (eds.), *Media Technologies: Essays on Communication, Materiality, and Society*, https://doi. org/10.7551/mitpress/9780262525374.003.0009.

Gillespie, T. (2018), *Custodians of the Internet. Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press.

Mayer-Schönberger, V., e Cukier, K. (2013), *Big Data: a Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt.

Metcalfe, B. (2013), "Metcalfe's law after 40 years of ethernet", *Computer*, 46 (12), pp. 26-31, https://doi.org/10.1109/MC.2013.374.

Newman, N., R. Fletcher, K. Eddy, C.T. Robertson, R.K. Nielsen (2023), *Digital News Report 2023*, Reuters Institute for the Study of Journalism, disponível em https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

Nielsen, R. K., e S.A. Ganter (2022), *The Power of Platforms. Shaping Media and Society*, Oxford University Press.

Pariser, E. (2011), *The Filter Bubble. How the New Personalized Web Is Changing What We Read and How We Think*, Penguin.

Salganik, M. J. (2019), *Bit by Bit. Social Research in the Digital Age*, Princeton University Press.

Vallas, S., e J.B. Schor (2020), "What do platforms do? Understanding the gig economy", *Annual Review of Sociology*, 46, pp. 273-294, https://doi.org/10.1146/annurev-soc-121919-054857.

van Dijck, J. (2013), *The Culture of Connectivity: a Critical History of Social Media*, Oxford University Press.

van Dijck, J., T. Poell, e M. De Waal (2018), *The Platform Society: Public Values in a Connective World*, Oxford University Press.

Capítulo 2

Enquadramento ético para a investigação digital
Um exercício de reflexão com base num caso prático

Cláudia Álvares
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Introdução

A rápida expansão dos *media* digitais e a integração de dados na vida quotidiana apresentam desafios éticos particulares às metodologias tradicionais de investigação nas ciências sociais. A complexidade da tomada de decisões éticas na pesquisa digital exige uma abordagem flexível e adaptativa, como destacado por Markham e Buchanan (2012). Estas autoras valorizam o recurso a princípios éticos que sejam indutivos e baseados em casos específicos, permitindo aos investigadores explorar os contextos diversificados dos ambientes digitais. Essa abordagem facilita a reflexão sobre procedimentos metodológicos em cenários virtuais, em que a rigidez de procedimentos originalmente pensados muito antes do surgimento da *Net* pode não abarcar todas as modalidades passíveis de integrarem a investigação na atualidade.

O objetivo deste capítulo é refletir sobre os desafios éticos específicos à investigação digital, relevando-se aspetos relacionados com a privacidade, consentimento informado e gestão de dados pessoais. Pretende-se analisar as diretrizes e recomendações de vários órgãos de ética, escrutinando-se a sua aplicabilidade e concomitantes limitações em contexto digital. Além disso, o capítulo discute as implicações da implementação, na União Europeia, do Regulamento Geral de Proteção de Dados (UE, RGPD, 2023) e da Diretiva ePrivacy (European Commission, 2017), bem como as consequências de se aplicarem esses regulamentos em pesquisa que utiliza as redes sociais e outras plataformas *online*. Para responder a esses objetivos, utilizaremos um caso prático para (re)pensar os principais parâmetros éticos que regem a investigação nas ciências sociais na atualidade à luz dos desafios que decorrem da utilização do digital.

O estudo de caso apresentado neste capítulo incide sobre a sugestão de atualização de um questionário sobre o tratamento de dados pessoais disponibilizado *online* pelo Conselho de Ética do Iscte (https://www.iscte-iul.pt/contents/iscte/organizacao/rgaos-consultivos/1345/conselho-de-etica), instituição de ensino superior pública em Portugal, visando adaptá-lo às necessidades e desafios da investigação digital da atualidade. Este questionário serve como ferramenta para garantir que

os procedimentos de recolha e gestão de dados estejam em consonância com os parâmetros éticos e legais prevalecentes. A atualização do questionário inclui reflexão sobre questões relacionadas com consentimento informado em ambientes digitais, técnicas de anonimização e pseudonimização, e a gestão de dados sensíveis, em sintonia com as orientações do RGPD e outros regulamentos relevantes. As recomendações que decorrem da análise deste estudo de caso poderão contribuir para esclarecer algumas dúvidas dos investigadores que trabalham na área dos métodos digitais, embora suscitem, muito possivelmente, mais dúvidas do que certezas, como aliás é apanágio das questões levantadas no domínio da ética.

O desafio da proteção de dados em espaço digital

No âmago da pesquisa digital estão as preocupações com a privacidade, conforme destacado pelas diretrizes da Universidade de Oxford (2020). Se bem que as redes sociais *online*, como o Twitter ou Facebook, sejam ricas em dados, a extração de material dessas plataformas apresenta problemas éticos devido à sua natureza pessoal e sensível. Essas considerações incluem a necessidade de respeitar a privacidade dos indivíduos, mesmo quando optam por tornar os seus dados públicos. Os investigadores devem ponderar cuidadosamente as implicações do uso desses dados, especialmente em contextos em que informações sensíveis possam ser recontextualizadas de forma não previamente antecipada pelos utilizadores, correndo-se o risco de lhes causar danos potenciais, como exposição ao *trolling* ou outras formas de prejuízo (cf. Maldonado, Castellanos, e Barrios, 2023).

A *Association of Internet Researchers* (AoIR, 2019) oferece um amplo quadro para enfrentar desafios éticos que possam surgir no âmbito da investigação *online*. As suas indicações apontam para a necessidade de se assegurar o uso ético de dados publicamente disponíveis. Recomenda-se, assim, que os investigadores considerem as expectativas de privacidade dos indivíduos, mesmo quando os dados são publicamente acessíveis. Isso implica reconhecer que os indivíduos possam não antecipar que os seus dados sejam usados para fins de pesquisa, e os investigadores devem esforçar-se para minimizar qualquer potencial dano ou desconforto causado pelos seus estudos (*British Psychological Society*, 2021; Markham e Buchanan, 2012).

A expectativa razoável de privacidade desempenha um papel crítico na formulação de práticas éticas de pesquisa, pois as pessoas geralmente têm entendimentos implícitos sobre como as suas informações devem ser tratadas, mesmo em fóruns públicos (Townsend e Wallace, 2016; Markham e Buchanan, 2012), já que não esperam que os seus *posts* sejam analisados e utilizados em pesquisas sem o seu consentimento explícito. Essa discrepância entre a disponibilidade técnica dos dados e as expectativas de privacidade dos utilizadores impõe uma responsabilidade adicional aos investigadores para garantir que as suas práticas de pesquisa respeitem os direitos e as perceções de privacidade dos participantes.

As figuras públicas, devido ao seu estatuto, têm expectativas de privacidade diferentes em comparação aos indivíduos privados. Efetivamente, as publicações de figuras públicas nas redes sociais online destinam-se, habitualmente, a

suscitar interação e consumo alargados. Assim, o seu conteúdo é geralmente considerado como pertencente ao domínio público, particularmente quando se relaciona com desempenho ou o exercício de atividades públicas. No entanto, os investigadores devem procurar equilibrar o direito do público à informação, isto é, o interesse público, com o direito à privacidade do indivíduo, especialmente no tocante a conteúdos pessoais ou sensíveis (AoIR, 2019; Nesh, 2019).

As plataformas de redes sociais possuem termos e condições específicos que definem o que constitui conteúdo público. Esses termos geralmente estipulam que qualquer conteúdo definido como "público" pelo utilizador possa ser acedido, visto e utilizado por qualquer pessoa, incluindo investigadores e terceiros. Por exemplo, os termos de serviço do Facebook especificam que, quando os utilizadores tornam o seu conteúdo público, concedem ao Facebook e àqueles que acedem ao conteúdo permissão para usá-lo e compartilhá-lo (*Facebook Terms of Service*, 2024). Esta configuração pública implica que os utilizadores abram mão, em certa medida, do controlo sobre como as suas informações são usadas, embora tal não elimine a necessidade de considerações éticas na pesquisa (Universidade de Columbia, 2023).

As diretrizes éticas sugerem que, sempre que possível, os dados devam ser anonimizados e que o consentimento informado possa ser obtido para minimizar riscos inerentes à identificação (Markham e Buchanan, 2012). Todavia, a obtenção do consentimento informado na pesquisa digital representa um sério desafio. Conforme apontado pela Comissão de Ética para a Investigação da *London School of Economics* (LSE) (2022), o consentimento nem sempre é necessário para dados publicamente disponíveis, embora considerações éticas devam guiar os investigadores na determinação das situações em que o consentimento seja recomendado para recolher dados em espaços *online* privados ou semiprivados. Assim, embora indicações claras sobre os objetivos do estudo e o uso pretendido dos dados recolhidos possam assegurar a transparência e fomentar a confiança (LSE Research Ethics Committee, 2022), o espaço digital rege-se por alguma falta de clareza relativamente a assuntos que, na pesquisa *offline*, já são considerados como dados adquiridos. Efetivamente, há alguma discordância quanto à necessidade de consentimento informado no domínio *online*, especialmente em pesquisas que envolvam figuras públicas ou dados publicamente acessíveis.

Outra questão que suscita dúvidas diz respeito ao recurso à anonimização, pseudonimização e confidencialidade na investigação em meios digitais, técnicas essas fundamentais na pesquisa *offline*. A anonimização remove informações identificáveis, a pseudonimização usa identificadores falsos e a confidencialidade garante que os dados não sejam divulgados a partes não autorizadas (Universidade de Oxford, 2021). No domínio digital, torna-se difícil aplicar técnicas que se destinam à remoção de informação que permita identificar os titulares dos dados, pois trata-se da colação de dados volumosos, sendo improvável que se consiga eliminar todos os dados pessoais.

Em suma, dada a nebulosidade da fronteira entre público e privado que rege o estatuto dos dados recolhidos *online*, podemos considerar que existem duas perspetivas sobre a prática da ética que subjaz à investigação digital: a primeira consiste

numa prática que se centra na proteção dos dados dos utilizadores, em que se tende a considerar o consentimento informado, a anonimização e o direito ao esquecimento como ferramentas importantes para a manutenção da privacidade; a segunda remete para uma prática orientada para a utilização de dados como integrando o domínio público, pressupondo que os benefícios sociais que daí advêm compensam quaisquer riscos envolvidos (Lukito, 2024). Em suma, o cerne da questão prende-se com o entendimento, da parte dos utilizadores, de que os seus dados possam vir a ser usados para fins de investigação. Markham e Buchanan (2012) tal como Townsend e Wallace (2016) defendem uma abordagem diferenciada para avaliar potenciais prejuízos, levando em consideração o contexto em que os dados são utilizados e as potenciais vulnerabilidades dos participantes. Esta perspetiva é crucial para entender todo o espectro de implicações éticas em ambientes digitais, em que o impacto da pesquisa pode estender-se para lá do estudo inicial (Markham e Buchanan, 2012; Townsend e Wallace, 2016).

Além do mais, os procedimentos éticos para pesquisa *online* não têm de ser forçosamente lineares. Por exemplo, um investigador pode defender que os dados recolhidos pertencem ao domínio público para determinadas finalidades, como o *data mining*. Simultaneamente, o mesmo investigador pode reivindicar a proteção de dados para outras finalidades, recorrendo a medidas como a anonimização ou pseudonimização contra acessos não autorizados, bem como técnicas de armazenamento seguro mediante *software* acessível por palavra-passe, ou então por encriptação (Lukito, 2024; Universidade de Oxford, 2021).

Enquadramento jurídico europeu para a proteção de dados

O Regulamento Geral de Proteção de Dados (UE RGPD, 2023) fornece, desde 2018, o quadro jurídico que garante a privacidade digital dos cidadãos na União Europeia. Ao aceder à *internet*, os utilizadores frequentemente confiam informações pessoais importantes ao seu Provedor de Serviços de Internet (ISP) e aos *site*s que utilizam, levantando preocupações sobre a segurança e o uso indevido dos dados (European Commission, 2023). O RGPD assegura que os dados pessoais só possam ser coligidos sob condições restritas e para fins legítimos. Segundo as diretrizes do RGDP, as organizações devem adotar medidas técnicas e organizacionais adequadas para garantir um nível de segurança apropriado ao risco (EUR-Lex, 2022; UE RGDP, 2023; GDPR.EU, 2024).

A base jurídica do RGPD para tratamento de dados pessoais corresponde aos artigos 6.º e 9.º, podendo o primeiro ser utilizado em várias situações de tratamento de dados que não incluam dados de grande sensibilidade, enquanto o segundo recai precisamente sobre o tratamento de categorias específicas de informações pessoais conhecidas como "dados sensíveis".

Efetivamente, o artigo 6.º estabelece as fundamentações legais para o processamento de informações pessoais, autorizando a sua utilização com base em uma das seis opções previstas: consentimento do indivíduo titular dos dados, necessidade para execução de um contrato, cumprimento de obrigações legais, proteção

dos interesses vitais do titular dos dados (alíneas a)-f), n.º 1 do artigo 6.º do RGPD, disponível em https://www.privacy-regulation.eu/pt/6.htm).

O artigo 9.º, por sua vez, incide sobre "dados pessoais que revelem a origem racial ou étnica, as opiniões políticas, as convicções religiosas ou filosóficas, ou a filiação sindical, bem como o tratamento de dados genéticos, dados biométricos para identificar uma pessoa de forma inequívoca, dados relativos à saúde ou dados relativos à vida sexual ou orientação sexual de uma pessoa" (n.º 1 do artigo 9.º do RGPD, disponível em https://www.privacy-regulation.eu/pt/9.htm). Normalmente, é proibido realizar o processamento desses dados, salvo em situações de exceção. Por exemplo, quando o titular dos dados expressa consentimento explícito para tal ou quando há uma obrigação legal de os tratar em contexto do direito do trabalho (alíneas a) e b) do n.º 2 do artigo 9.º, disponível em https://www.privacy-regulation.eu/pt/9.htm). Geralmente falando, a natureza sensível desses dados requer uma justificativa mais robusta e medidas de proteção adicionais do que no caso do enquadramento proporcionado pelo artigo 6.º do RGPD.

A Diretiva ePrivacy (Directive 2002/58EC) e o regulamento que dela decorre desde 2017 abordam a privacidade e a proteção dos dados nas comunicações eletrónicas. Este quadro legal aplica-se principalmente aos operadores de telecomunicações e aos provedores de serviços de *internet* (ISP), assegurando que todas as comunicações através de redes públicas respeitem os direitos fundamentais de privacidade e proteção de dados. Assim, a diretiva e-Privacy complementa o RGPD, exigindo, por exemplo, que os países da UE garantam que os utilizadores concedem consentimento antes de os *cookies* (pequenos arquivos de texto armazenados no navegador *web* do utilizador) serem armazenados e acedidos em dispositivos conectados à *internet*.

O novo regulamento ePrivacy, proposto em 2017 (*European Commission*, 2017), introduz ainda o conceito de "privacidade por *design*" (EDRi, 2024; artigo 25.º do GDPR), permitindo aos utilizadores escolherem o nível desejado de privacidade. O conceito de "privacidade por *design*" (EDRi, 2024) pressupõe que a privacidade seja integrada no desenvolvimento e funcionamento dos sistemas de informação desde o início, podendo assim ter impacto nas plataformas de redes sociais como a Meta, conduzindo a possíveis alterações operacionais e técnicas. Efetivamente, esse regulamento obrigaria as plataformas a garantirem que os utilizadores controlem as suas configurações de privacidade e que a recolha e o uso de dados sejam minimizados e protegidos (i-Scoop, 2019; SecurePrivacy, 2022). Tal incluiria a garantia de que os padrões de privacidade mais altos seriam aplicados automaticamente, a menos que o utilizador optasse por configurá-los de forma diferente. A Meta seria obrigada a ser transparente sobre como os dados dos utilizadores são recolhidos, utilizados e compartilhados. O consentimento informado deveria ainda ser obtido de maneira clara e não ambígua, especialmente para o uso de *cookies* e tecnologias de rastreamento. Além do mais, o novo regulamento prevê a implementação de medidas que asseguram a eliminação de dados de forma segura, para que não sejam conservados por mais tempo do que o necessário (*European Commission*, 2017).

No entanto, até ao momento, a proposta de regulamento ainda está a ser alvo de discussão entre o Conselho da União Europeia e o Parlamento Europeu,

enfrentando vários desafios e *lobbies* de diversas indústrias, incluindo a publicidade e as telecomunicações, que temem as implicações restritivas da privacidade para os seus modelos de negócios (ITGovernance, 2024). Assim, enquanto o RGPD foi plenamente implementado, à data da redação deste capítulo, o regulamento ePrivacy ainda se encontrava em fase de negociação, (cf. *European Parliamen*t, 2024).

Estudo de caso sobre proteção de dados em contexto digital

Conforme se viu na primeira secção, há determinadas caraterísticas da investigação *online* que por vezes dificultam a aplicação dos parâmetros habitualmente implementados pela investigação *offline*, suscitando o imperativo de uma reflexão sobre os mesmos. Entre estas dificuldades, incluem-se a aplicação de termo de consentimento quando se analisam dados públicos *online* e o recurso a técnicas que impeçam a identificação de dados pessoais, temas esses referidos na introdução deste capítulo.

Junte-se ainda a estes dois potenciais problemas um terceiro, o qual diz respeito à dificuldade em se proceder a uma recolha de amostra representativa para inquéritos ministrados *online*, que esteja em consonância com critérios preestabelecidos. Apesar de esta última observação ser pertinente para a ética da investigação digital, já que pode influenciar a condução da própria investigação e os resultados obtidos, este tema por si só daria o mote para a redação de outro capítulo relacionado com a ética na pesquisa digital, pelo que o deixaremos, por ora, fora do presente trabalho.

Tentaremos, ao invés de nos debruçarmos sobre todos os ângulos deste complexo tema, pôr em prática o conselho do *Markkula Center for Applied Ethics* (2024) da Universidade de Santa Clara, nos EUA, o qual desafia os investigadores a refletirem criticamente sobre as implicações éticas dos seus métodos a partir de estudos de caso concretos. No nosso contexto, tomaremos como ponto de partida o questionário sobre tratamento de dados pessoais disponibilizado publicamente pelo Conselho de Ética do Iscte-IUL (disponível em https://www.iscte-iul.pt/contents/iscte/organizacao/rgaos-consultivos/1345/conselho-de-etica), ao qual os investigadores devem responder no caso de utilizarem dados pessoais na sua investigação.

Este questionário de tratamento de dados ilustra precisamente os desafios com os quais a investigação nas ciências sociais, recorrendo a meios digitais, se depara, já que as questões colocadas no questionário, tendo em vista suscitar reflexão da parte dos investigadores relativamente ao modo como utilizam os dados recolhidos, não contemplam as dificuldades que decorrem das especificidades dos novos meios digitais. A adaptação ao novo contexto digital, abrangendo redes sociais e outras plataformas *online*, do questionário sobre tratamento de dados pessoais deverá ter em conta, a nosso ver, os aspetos que iremos desenvolver de seguida.

Considera-se, antes de mais, que a categoria "dados públicos" deva contemplar dados provindos de redes sociais e de outros recursos *online*, especificando-se que os dados pessoais possam ter origem nessas fontes. Tal implicaria que a base

legal para o tratamento de dados pessoais contemplasse os termos de uso das plataformas (C.3), além dos artigos 6.º e 9.º do RGPD, já que a participação nessas plataformas pressupõe a aceitação desses termos. O reconhecimento dos termos e condições das plataformas como base legal para a recolha de dados alinharia o questionário com as práticas comuns em pesquisas que utilizam dados provenientes de redes sociais e outras plataformas *online*.

Neste âmbito, recomenda-se a especificação do número (aproximado) de titulares de dados envolvidos no contexto das redes sociais e plataformas *online*. A este propósito, a Biblioteca de Conteúdos e a API (Interface de Programação de Aplicações) da Biblioteca de Conteúdos da Meta disponibilizam acesso ao arquivo com todos os conteúdos públicos do Facebook e do Instagram, com informação sobre o proprietário da publicação e o número de interações e de partilhas. Entre os conteúdos definidos como públicos estão, no Facebook, publicações partilhadas em páginas, grupos e eventos e, no Instagram, publicações partilhadas por contas comerciais e de criador de informações sobre essas contas. Contempla-se, sob a categoria de informações públicas, contas pessoais que tenham sido definidas como públicas, com selo de verificação indicando a sua autenticidade, ou que tenham um mínimo de 50 mil seguidores (*Transparency Center*, 2024).

Haverá ainda que detalhar o uso de dados públicos acedidos *online*, incluindo-se, por exemplo, uma questão direcionada ao tratamento de dados publicamente disponíveis em redes sociais (E.5), em que se abordam as considerações éticas relacionadas com essas fontes. Efetivamente, a natureza dos dados pessoais a tratar poderia também contemplar dados públicos, como *posts* de redes sociais *online*, enquanto base de recolha de material potencialmente sensível.

Sugere-se também que se proceda à atualização das medidas de proteção de dados pessoais em ambiente digital, incluindo-se medidas específicas, além da anonimização e da pseudonimização, para responder à dificuldade da manutenção da privacidade da informação pessoal em dados públicos digitais, mediante introdução da categoria da confidencialidade (H.3), limitando-se assim o acesso apenas aos dados mais sensíveis, protegendo-os contra acessos não autorizados e uso indevido.

Os termos de consentimento informado terão, em contexto de investigação digital, de assegurar clareza relativamente ao enquadramento jurídico de recolha de dados pessoais. Tais termos devem ainda apontar os riscos decorrentes da possibilidade de reidentificação de dados quando inseridos em novo contexto. Efetivamente, o Quadro 2 do questionário atual já prevê riscos inerentes à criação de perfis, resultante do agrupamento ou busca de padrões em conjuntos de dados, por poderem contribuir, segundo o Considerando 71 do artigo 22.º do RGPD, para "efeitos discriminatórios contra pessoas singulares em razão da sua origem racial ou étnica, opinião política, religião ou convicções, filiação sindical, estado genético ou de saúde ou orientação sexual" (https://gdpr-text.com/pt/read/recital-71/).

Considerações finais

Diversos orgãos responsavéis por diretrizes éticas enfatizam a importância da transparência na divulgação de métodos e resultados de pesquisa, para promover a confiança nos processos de investigação académicos, ao mesmo tempo que se aumenta a possibilidade de replicação dos estudos. Este capítulo demonstra, todavia, que, mesmo dentro da área das ciências sociais, podem existir nuances de diferença importantes nos procedimentos aplicados na investigação que recorre a métodos digitais, por um lado, e naquela que se baseia em métodos tradicionais, por outro. Essa diferença na aplicabilidade de procedimentos levanta questões éticas interessantes sobre as quais se pretendeu aqui refletir. O exercício de reflexão aqui conduzido foi coadjuvado por um caso prático, nomeadamente a tentativa de se pensar como atualizar o questionário sobre tratamento de dados pessoais disponibilizado pelo Conselho de Ética do Iscte, tornando-o aplicável tanto à investigação digital, como à investigação mais tradicional.

Em suma, a investigação digital, especialmente quando envolve redes sociais *online* e outras plataformas eletrónicas, levanta novos desafios éticos, assentes em questões para as quais as respostas habituais não chegam. Destina-se esta proposta de reformulação do questionário existente a responder a novas especificidades da investigação nas ciências sociais, garantindo-se assim que a pesquisa esteja em conformidade com as regulamentações e proteção dos direitos dos participantes.

Referências bibliográficas

British Psychological Society (2021), *Ethics Guidelines for Internet-Mediated Research*, disponível em https://www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002, disponível em https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32002L0058.

EUR-Lex (2022), *General Data Protection Regulation* (GDPR), disponível em https://eur-lex.europa.eu/EN/legal-content/summary/general-data-protection-regulation-gdpr.html.

European Commission (2023), *Shaping Europe's Digital Future: Digital Privacy*, disponível em https://digital-strategy.ec.europa.eu/en/policies/digital-privacy.

European Commission (2017). *Proposal for an ePrivacy Regulation*, disponível em https://digital-strategy.ec.europa.eu/en/policies/eprivacy-regulation.

European Parliament (2024), *Legislative Train Schedule*, disponível em https://www.europarl.europa.eu/legislative-train/theme-connected-digital-single-market/file-jd-e-privacy-reform.

Facebook Terms of Service (2024), disponível em https://www.facebook.com/terms.php.

GDPR EU (2024), "What is DGPR, the EU's new data protection law?", disponível em gdpr.eu/what-is-gdpr/.

i-Scoop (2019), "The new EU ePrivacy Regulation: what you need to know", disponível em https://www.i-scoop.eu/gdpr/eu-eprivacy-regulation/.

ITGovernance (2024), "The EU ePR (ePrivacy Regulation)", disponível em https://www.itgovernance.co.uk/eprivacy-regulation-epr.

London School of Economics (LSE) (LSE Research Ethics Committee) (2022), "Using data from the internet and social media in research: ethics & consent", disponível em https://info.lse.ac.uk/staff/divisions/research-and-innovation/research/Assets/Documents/PDF/ethics-Using-internet-and-Social-media-data-v8.pdf

Lukito, J. (2024), "Platform research ethics for academic research", Center for Media Engagement, Moody College of Communication, The University of Texas at Austin, disponível em https://mediaengagement.org/research/platform-research-ethics.

Maldonado, C., P. Castellanos, e R. Barrios (2023), "Ethical issues when using digital platforms to perform interviews in qualitative health research", *International Journal of Qualitative Methods,* 22 (4), https://doi.org/10.1177/16094069231165949.

Markham, A., e E. Buchanan (2012), "Ethical decision-making and Internet research: recommendations from the AoIR Ethics Working Committee", Association of Internet Researchers, disponível em http://www.aoir.org/reports/ethics.pdf.

Markkula Center for Applied Ethics. (2024), "Internet ethics cases", Santa Clara University, disponível em https://www.scu.edu/ethics/focus-areas/internet-ethics/cases/.

Questionário sobre o tratamento de dados pessoais (s.d.), Conselho de Ética do Iscte, disponível em https://www.iscte-iul.pt/contents/iscte/organizacao/rgaos-consultivos/1345/conselho-de-etica.

SecurePrivacy (2022), "EU's ePrivacy Regulation: 2022 Updates", disponível em https://secureprivacy.ai/blog/eu-eprivacy-regulation-2022-updates.

Townsend, L., e C. Wallace (2016), "Social media research: a guide to ethics", University of Aberdeen, disponível em https://www.gla.ac.uk/media/Media_487729_smxx.pdf.

Transparency Center (2024), "Biblioteca de conteúdos e API da Meta", disponível em https://transparency.meta.com/en-gb/researchtools/meta-content-library.

UE RGPD (2023), *UE Regulamento Geral sobre a Proteção de Dados*, disponível em https://www.privacy-regulation.eu/pt/.

Universidade de Columbia (Global Freedom of Expression) (2023), *Privacy and Freedom of Expression*, disponível em https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2023/01/Privacy-and-Freedom-of-Expression-4.pdf.

Universidade de Oxford (Central University Research Ethics Committee (CUREC)) (2021), *Best Practice Guidance, Internet-Mediated Research*, disponível https://researchsupport.web.ox.ac.uk/files/bpg-06-internet-based-research-ibr.

Capítulo 3

Desenhar, planear e estruturar a pesquisa

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

Inês Narciso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia(CIES-Iscte), Lisboa, Portugal

Conceptualizar o desenho da pesquisa

Escrever sobre propostas metodológicas para realizar investigação implica escrever sobre o desenho da pesquisa. Este desempenha um papel essencial, já que fornece uma estrutura para planear, organizar e conduzir investigação sistemática de fenómenos sociais que tomam lugar nas plataformas sociais digitais. O desenho da pesquisa serve como um roteiro que auxilia os investigadores a terem os seus objetivos mais claros e orienta na formulação das perguntas de investigação, na identificação de métodos apropriados, na escolha de ferramentas para recolher dados, na definição de como se procederá à recolha e à realização da análise e até na elaboração de conclusões, baseadas na discussão dos dados, permitindo garantir a validade e fiabilidade da investigação.

A formulação das perguntas de investigação é um passo considerado essencial em qualquer desenho de pesquisa, mas particular quando estamos a estudar acontecimentos, sujeitos ou assuntos no contexto das plataformas sociais digitais.

É importante ter em conta que a formulação das perguntas de investigação no âmbito dos métodos digitais deve equacionar a plataforma sobre a qual vai incidir a investigação, isto é, onde os acontecimentos tomam lugar, os sujeitos interagem e os assuntos são abordados. Essa importância deve-se ao facto de as especificidades das plataformas sociais digitais, isto é, o *design*, o funcionamento, as *affordances* e as gramáticas próprias, terem impacto nas possibilidades de como e o que perguntar, logo na formulação da pergunta. Como também terão impacto na forma como as respostas serão obtidas. Adicionalmente, nessa formulação das perguntas é também importante ter em conta as ideologias de *media* (Gershon, 2010) e as culturas de uso da plataforma, isto é, como a plataforma é entendida e apropriada pelos utilizadores e como tal se reflete nas suas práticas. Tal não só terá impacto em como perguntar, como também na discussão dos resultados e na contextualização destes em função do meio.

Ainda que existam tópicos que possam ser transversais a várias plataformas e, portanto, possam ser estudados em diferentes meios — veja-se por exemplo o tópico da desinformação — existem especificidades das plataformas que permitem (ou exigem) questionar sobre esses tópicos e consequentemente estudá-los de formas distintas. Por exemplo, fará menos sentido questionar sobre representações visuais no X (antigo Twitter), e mais sentido fazê-lo relativamente ao Instagram ou ao TikTok. Adicionalmente fará mais sentido estudar assuntos através de *hashtag*s (#) no X ou no Instagram, mas menos no Facebook ou no YouTube. Ainda que a funcionalidade *hashtag* seja comum a essas plataformas, o seu uso como forma de expressão nas plataformas foi entendido de forma distinta e apropriado com propósitos diferenciados. Fará sentido pensar sobre vídeos no contexto do YouTube e TikTok, mas menos no âmbito do X. Estas possibilidades da plataforma devem, como mencionado anteriormente, refletir-se na formulação da pergunta de investigação, mas também em como questionar a plataforma sobre o objeto de estudo.

Essa necessidade de pensar sobre como questionar a plataforma relativamente ao objeto de estudo remete-nos para o conceito "query design" (Rogers, 2017). Este refere-se à técnica de construir listas de palavras, contas, URL de perfis ou de *posts*, ID de vídeos ou canais, como ponto de entrada para questionar a plataforma. Tal implica a capacidade de o investigador identificar e definir que palavras, contas, URL de perfis ou de *posts*, ID de vídeos ou canais se tornaram chave (*keywords*) relativamente aos tópicos que está a estudar e nas plataformas sociais digitais em que pretende fazê-lo.

Nalguns casos, o foco da investigação poderá estar nas publicações e discussões realizadas acerca de um determinado tema nas plataformas de redes sociais. Nesse caso, importa definir quais são as palavras que expressam esse tema nas publicações nessas plataformas. Mas, noutros casos, a investigação pode escolher focar-se naquilo que foi publicado por um ator específico ou por um conjunto de atores. E, nesse caso, o que terá de determinar é quem são esses atores e porquê esses e não outros. Por fim, também é possível combinar as duas abordagens e procurar analisar o que determinado ator ou atores publicaram sobre um tema específico. Seja como for, esse propósito da investigação irá sempre influenciar a forma como iremos inquirir a plataforma (ou plataformas) e, portanto, a definição das "keywords".

Como indica Rogers (2017: 83), "Na formulação de *queries*, é pertinente considerar as palavras-chave como partes de programas, antiprogramas ou esforços de neutralidade, pois esta perspetiva permite ao investigador estudar tendências, compromissos e alinhamentos entre os atores".

A identificação e a definição dessas "keywords" implicam que o investigador procure, monitorize ou siga o objeto ou tópico de estudo no meio em que o pretende fazer, isto é, que se envolva com o mesmo na ou nas plataformas sociais digitais. Deverá, claro, fazê-lo sempre com uma postura crítica, sem esquecer o seu papel como investigador e com a pergunta de investigação e objetivos presentes. Ainda na ótica da construção do "query design", não se pretende que as "keywords" sejam equivalentes ou sinónimas, mas sim específicas tendo em conta o tema que se está a estudar. No âmbito do projeto Eumeplat — European Media Platforms:

Assessing Positive and Negative Externalities for European Culture, concretamente nos *work packages* 2 e 4, dedicados, respetivamente, à plataformização das notícias e às representações das questões de género e das migrações, houve necessidade de serem definidas listas de palavras capazes de funcionarem como "keywords" para que a pesquisa fosse executada.[1] Essa necessidade obrigou a uma pesquisa exploratória exaustiva, para procurar identificar que palavras eram mais frequentemente usadas na expressão desses temas nas plataformas de redes sociais, assim como palavras relacionadas.

No processo de desenvolvimento das *queries* contendo essas "keywords", dois passos foram fundamentais para assegurar a consistência da metodologia: a definição das palavras a usar em cada dimensão de análise e a pesquisa de conceitos semelhantes em diferentes países. No *work package* 2 do projeto Eumeplat escolhemos observar publicações nas redes sociais *online* (Facebook, Twitter/X e YouTube) acerca de quatro dimensões: Europa, economia, saúde e ambiente. O *work package* 4, por seu lado, focou a sua atenção nas questões de género e das migrações. Para cada uma dessas dimensões foram escolhidas palavras iniciais — por exemplo, "Europa" — que foram objeto de pesquisa em motores de busca e em redes sociais. Em cada uma dessas pesquisas foram identificadas as palavras mais frequentemente associadas, que posteriormente eram elas próprias o ponto inicial de outra iteração de pesquisa. Este processo foi repetido várias vezes. O resultado foi, em qualquer dos casos, uma *query* de múltiplos termos (incluindo binómios e trinómios) para cada uma das dimensões de análise recolhidas.

O segundo passo consistiu na adaptação dessa *query* às 11 línguas envolvidas no projeto, contando para isso com a colaboração dos parceiros internacionais falantes nativos de cada língua. Igualmente, por razões metodológicas, optámos por uma adaptação em vez de uma tradução, pedindo aos parceiros para identificarem quais as palavras que melhor expressavam cada um dos conceitos constantes na respetiva língua. Por exemplo, "Serviço Nacional de Saúde" não tem uma tradução direta noutras línguas, mas pode ter uma adaptação diferente para cada país.

Igualmente, foi deixado espaço na *query* para que cada parceiro pudesse incluir alguns termos considerados particularmente relevantes nesse país, em cada uma das dimensões de análise, de acordo com a mesma metodologia descrita acima. Devido à abrangência dos temas escolhidos e por envolver 10 países e 11 línguas, este projeto ilustra bem a complexidade de desenhar uma *query* de pesquisa capaz de responder ao objeto de estudo e às perguntas de partida.

No desenho da pesquisa e no que à formulação de perguntas diz respeito, encontramo-nos perante dois tipos de necessidades distintas, mas ligadas entre si: definir a pergunta de investigação e determinar como se vai questionar a plataforma, sendo relevante ter em atenção que esses passos têm a capacidade de moldar as etapas seguintes do desenho da pesquisa.

Uma dessas etapas é relativa à abordagem metodológica ao auxiliar o desenho na escolha de *softwares*, ferramentas e técnicas mais apropriadas para responder à

1 https://www.eumeplat.eu/

pergunta de investigação. Neste caso concreto, a abordagem metodológica não recairá sobre a escolha de métodos qualitativos, quantitativos ou mistos, isto porque a abordagem metodológica já está determinada, tratando-se de investigação com recurso aos métodos digitais. Assim, neste âmbito, o desenho de pesquisa auxilia os investigadores na escolha, considerando fatores como a natureza do fenómeno, mas com especial ênfase nos recursos, nomeadamente nos recursos técnicos, disponíveis no que à recolha de dados diz respeito.

As ferramentas de recolha de dados

No desenvolvimento de um plano de pesquisa, é essencial fazer um levantamento das ferramentas existentes, que poderão permitir a recolha dos dados necessários. Esse levantamento deve considerar que se trata de um setor em contínua evolução e adaptação às alterações das próprias plataformas. A ferramenta escolhida hoje pode não funcionar da mesma forma amanhã, pelo que muitas vezes pode ser prudente antecipar a recolha dos dados ou ter em especial consideração estas limitações num estudo longitudinal. Igualmente importante na escolha dos recursos técnicos de apoio é considerar algumas características diferenciadoras. As ferramentas são projetadas com objetivos diversos e em poucos casos são construídas exclusivamente para pesquisa académica. As ferramentas desenvolvidas para investigação académica apresentam, muitas vezes, maior transparência relativamente aos seus métodos e limitações de recolha de dados, e, geralmente, estão disponíveis gratuitamente. Um bom exemplo de um repositório deste tipo de ferramentas é o projeto da Digital Methods Initiative, que dispõe não só de várias ferramentas, como de toda a documentação metodológica de apoio e exemplos de investigação realizada com as mesmas.[2]

No entanto, os contínuos ajustes técnicos exigidos pela natureza dinâmica das plataformas tornam desafiante o desenvolvimento e manutenção deste tipo de ferramentas, puramente centradas na investigação. Consequentemente, os investigadores recorrem frequentemente a ferramentas comerciais concebidas para monitorização de marcas ou fins semelhantes (Poell *et al.*, 2021). É crucial que os investigadores compreendam que o objetivo inicial de uma ferramenta pode afetar a priorização e recolha de dados, influenciando potencialmente os resultados da investigação (Mneimneh *et al.*, 2021). Por outro lado, considerando que estas ferramentas não são especificamente desenvolvidas com fins académicos, o tipo de dados disponibilizados e a forma como são apresentados também podem não corresponder exatamente àquelas que seriam as necessidades do investigador. Tal poderá ter como consequência a necessidade de preparar e organizar os dados como passo intermédio.

O custo também varia significativamente, existindo ferramentas a preços inacessíveis ao investigador individualmente, até soluções gratuitas, passando pelos

2 https://www.digitalmethods.net/Dmi/ToolDatabase

modelos *freemium* que oferecem serviços básicos gratuitamente, mas cobram por recursos avançados. As ferramentas pagas têm de proporcionar uma experiência amiga do utilizador, o que nem sempre as ferramentas gratuitas conseguem garantir, podendo ser necessário ter conhecimentos técnicos mais avançados para que possam ser utilizadas de forma eficaz. A disponibilidade de API, que podem ser gratuitas ou pagas, complica ainda mais este cenário. Por exemplo, a diferença de acessibilidade entre a API aberta do Telegram e as opções mais restritas e dispendiosas do Twitter (X) exemplifica como os modelos de preços podem limitar o âmbito da investigação, especialmente em plataformas que impõem taxas elevadas para acesso a dados.

A transparência nos métodos de recolha de dados é outro fator crítico. O grau de transparência afeta a objetividade e a confiabilidade dos dados coletados. Uma ferramenta de *marketing* digital que faz *social listening*, baseada nos EUA, por exemplo, pode ter uma cobertura muito mais extensa das páginas e grupos americanos no Facebook, potencialmente favorecendo esses resultados quando pesquisamos por tópicos gerais. As ferramentas comerciais raramente têm a transparência e a documentação metodológica de apoio das ferramentas académicas, omitindo detalhes sobre como funciona o seu algoritmo e como tem impactos na sua recolha de dados, o que pode introduzir desafios metodológicos. É muito importante considerar estes aspetos, a fim de garantir a integridade das suas conclusões.

Por último, o volume de dados que as ferramentas podem recolher também varia significativamente. Os serviços *freemium* normalmente limitam a quantidade de dados acessíveis, o que pode restringir o âmbito e a extensão da pesquisa. Esta limitação pode ser crítica para investigadores que queiram recolher dados em larga escala.

Ou seja, a escolha da(s) ferramenta(s) a utilizar para a recolha de dados na investigação académica deve ser feita compreendendo a finalidade da conceção da ferramenta, o seu custo, a transparência metodológica e as limitações na recolha de dados.

A amostra

Uma vez determinada a ferramenta, as opções inerentes e as suas limitações, o desenho de pesquisa também orienta quanto aos procedimentos da recolha de dados. O grande propósito é que esses procedimentos sejam sistemáticos, padronizados e consistentes, tendo em conta os fenómenos e as plataformas. Tal poderá passar por, entre outras opções, determinar o momento em que é efetuada ou a frequência com que acontece. Tais aspetos também nos remetem para a orientação e seleção de estratégias na obtenção da amostra.

A amostra ou *corpus* de análise de estudos com recurso aos métodos digitais pode ter dimensões muito variadas. Pode ser composta por um único objeto digital (como um *post* ou um vídeo), assim como por centenas ou até milhares, já numa ótica de *big data*. Não obstante a dimensão, é comum que a amostra seja obtida por conveniência, isto é, aquela que está acessível. Não é então resultado de um critério

estatístico aplicado, mas sim da disponibilidade dos dados. Isso significa que não se trata de uma amostra representativa, uma vez que não é possível determinar um universo a representar. E a limitação daí decorrente é que a generalização dos resultados será sempre mais limitada e sem o mesmo grau de confiança que teria numa amostra representativa. Se uma amostra representativa permite fazer a generalização dos resultados exatamente porque representa o universo, numa amostra de conveniência como aquelas que se podem obter neste tipo de pesquisa, os resultados apenas se podem referir à própria amostra (ex.: "na nossa amostra"; "no conjunto de publicações analisadas"). Ainda assim, podem ser instituídos critérios de inclusão. Estes podem até já estar refletidos no *query design* ou serem adicionados após a recolha e antes da análise, podendo ser, por exemplo, um determinado número de *posts* com maior ou menor *engagement*, conteúdos publicados num determinado período de tempo, conteúdos contendo determinadas palavras, conteúdos publicados por determinados sujeitos, etc.

No âmbito do desenho da pesquisa, a pergunta e os objetivos de investigação definidos informam sobre a seleção de técnicas e métodos de análise e inerentes procedimentos para interpretar os resultados obtidos. Assim, o desenho de pesquisa auxiliará a escolher métodos quantitativos, qualitativos ou uma combinação dos dois para analisar os dados recolhidos, testar hipóteses e obter conclusões significativas. Poderemos estar a referir-nos a análises estatísticas, à análise temática, seja esta visual, textual ou ambas, análise de conteúdo, análise de redes, entre outras possibilidades.

É importante que, através do desenho da pesquisa, as técnicas de análise escolhidas estejam alinhadas com as questões de pesquisa, tipos de dados e abordagens metodológicas.

Um desenho da pesquisa bem concebido terá impacto na investigação ao auxiliá-la na sua eficiência. Um cuidado planeamento pode, por um lado, otimizar recursos e, por outro, minimizar o número de etapas necessárias, maximizando os esforços com vista ao resultado final. No projeto de investigação Eumeplat já referido, por exemplo, o desenho de pesquisa desenvolveu um *framework* metodológico detalhado, especificando e planeando quais as redes sociais a investigar, quais os temas a pesquisar e respetivas palavras-chave, quais os períodos temporais da recolha, quais os *outputs* gerados e como analisar esses dados (Cardoso *et al.*, 2021).

Um desenho da pesquisa bem desenvolvido e estruturado para uma investigação no âmbito digital deverá ser capaz de clarificar:

— Qual a pergunta de partida ou as hipóteses em estudo?
— Quais os objetivos na investigação?
— Quais os métodos que serão utilizados para atingir esses objetivos?
— Quais os recursos que são necessários para implementar esses métodos. Por exemplo, que ferramentas vão ser usadas e como?
— O estudo está devidamente defendido dos pontos de vista ético e legal? Os dados a recolher podem ser, efetivamente, recolhidos?
— A base de dados resultante da recolha de dados é suficiente para permitir conclusões válidas e fundamentadas?

— Através dos dados recolhidos é possível responder à pergunta de partida?
— Que dados serão usados na análise?
— Se forem utilizadas métricas, quais são essas métricas?

Não obstante a necessidade de precisão e detalhe, o desenho da pesquisa deverá permitir flexibilidade e adaptabilidade na resposta a determinados desafios ou mudanças imprevistas que podem surgir, tanto no ambiente de investigação, como no objeto de estudo, sem esquecer aqueles relacionados com as ferramentas (exemplo: a sua disponibilidade, os dados a que deixa de se poder aceder). Deverá, então, haver espaço para ajustar o desenho de pesquisa, mas mantendo a integridade e o rigor da mesma.

Note-se que o desenho da pesquisa é relevante, uma vez que estabelece as bases para a produção de resultados válidos, fiáveis e significativos que contribuem para o avanço do conhecimento. Uma investigação metodologicamente sólida pode fazer contribuições significativas para o campo de estudo, não só pelos resultados obtidos, mas também pelas opções metodológicas em si.

Referências bibliográficas

Cardoso, G., C. Álvares, J. Moreno, R. Sepúlveda, M. Crespo, e C. Foà (2021), "Deliverable D2.1 A framework and methodological protocol for analyzing the platformization of news", Eumeplat, disponível em https://www.eumeplat.eu/results/deliverables/.

Gershon, I. (2010), "Media ideologies: an introduction", *Journal of Linguistic Anthropology*, 20, pp. 283-293, doi:10.1111/j.1548-1395.2010.01070.x.

Mneimneh, Z., J. Pasek, L. Singh, R. Best, L. Bode, E. Bruch, e S. Wojcik (2021), "Data acquisition, sampling, and data preparation considerations for quantitative social science research using social media data", disponível em https://www.jonathanmladd.com/uploads/5/3/6/6/5366295/ dataacquisition.pdf.

Poell, T., D. B. Nieborg, e B. E Duffy (2021), *Platforms and Cultural Production*, John Wiley e Sons.

Rogers, R. (2017), "Foundations of digital methods: query design", em Mirko Tobias Schäfer, Karin van Es (eds.), *The Datafied Society: Studying Culture through Data*, Amesterdão, Amsterdam University Press, pp. 75-94, https://doi.org/10.25969/mediarep/125.

Parte 2 | Fazer investigação nas redes sociais *online*

Capítulo 4

Instagram
Perfis, comentários e *hashtag*s

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

O Instagram, lançado em 2010 e adquirido pela Meta em 2012, é uma rede social *online* na qual as imagens são o formato preferencial e principal para comunicar.[1] As publicações são assim maioritariamente visuais, passando o texto para um segundo plano.

Caracteriza-se por ser uma das plataformas, de redes sociais *online*, mais populares em todo o mundo, acumulando 1,4 mil milhões de utilizadores a nível global (Dixon, 2023). A sua apropriação tornou a plataforma uma parte relevante da sua vida quotidiana dos seus utilizadores (Highfield e Leaver, 2016), sendo múltiplas as razões apontadas para a sua utilização. Estas incluem motivos tão variados como entretenimento, conveniência, autoexpressão, interação social interpessoal, documentação/arquivo ou criatividade (Romero Saletti, Van den Broucke e Van Beggelaer, 2022) para mencionar alguns.

Tornou-se uma plataforma atrativa para marcas, instituições ou profissionais de diversas áreas, sendo um palco onde vários tópicos são abordados e debatidos. Ativismo ambiental, feminismo, saúde mental ou questões relacionadas com discriminação são alguns exemplos. A plataforma tem sofrido uma transformação, passando a deixar de ser encarada apenas como um espaço fútil e polido, tornando-se mais politizada, ativista e educacional (Al-Rawi, 2021; Childs, 2022; Drüeke *et al.*, 2022).

Através da criação de um perfil, os utilizadores podem expressar-se, ao publicar conteúdo e ao interagir com outros utilizadores. Não obstante, essas formas e formatos de expressão e interação estão convencionados pelas *affordances* da plataforma, isto é, o tipo de ações que são possibilitadas aos seus utilizadores. Assim, o desenho, arquitetura e funcionalidades do Instagram desempenham um papel fundamental no processo comunicacional, merecendo reflexão aquando do estudo da plataforma ou de fenómenos que nela tomam lugar.

Estudar o Instagram como plataforma e como meio de comunicação requer uma abordagem metodológica específica. Não só devido ao acesso aos dados e como estes são recolhidos, mas também porque as imagens não são estruturas isoladas. Estas devem ser estudadas tendo em conta a informação contextual envolvente (Leaver *et al.*,

1 https://about.meta.com

2020). Informações essas que estão presentes nas legendas das publicações: o texto. Assim, as publicações devem ser percebidas e estudadas como um todo.

A legenda, que pode ou não acompanhar uma publicação, também entendida como a descrição do *post*, pode conter até 2200 caracteres. Nesta é comum identificar a presença de *hashtag*s, isto é, expressão precedida do símbolo #. Apesar das *hashtag*s não serem exclusivas desta plataforma, foram largamente adotadas. A sua utilização pode ter diversas funções, incluindo a categorização de conteúdo, a indicação de uma posição, ideologia ou comportamento para chamar a atenção para um problema, comunicar um acontecimento ou associar-se a uma causa.

Através de *hashtag*s, o discurso é organizado (Daer *et al.*, 2014; Zappavigna, 2015), tendo estas o potencial de enquadrar determinado tópico (Meraz, 2017; Meraz e Papacharissi, 2013). Assim, o seu estudo permite entender o contexto das publicações, as posições de quem recorre às *hashtag*s, o significado desse recurso, revelando também atores (perfis) associados àquelas.

Perfis, *posts* e *hashtag*s constituem pontos de partida válidos, ao mesmo tempo que se revelam abordagens complementares para estudar o conteúdo que está disponível no Instagram. Em seguida, partilhamos algumas dessas abordagens. Algumas delas fazem sentido serem combinadas entre si. Qual/is a/s mais indicada/s? Dependerá da pergunta de partida e dos objetivos da investigação que está a realizar.

Recolha de dados

Ferramenta. PhantomBuster

Para a recolha de dados, vamos usar o PhantomBuster, um *software* que realiza ações automatizadas tais como a recolha de dados de plataformas sociais, entre as quais o Instagram. A recolha de dados acontece através da API das plataformas. Não é ilegal desde que a mesma ocorra e cumpra as diretrizes das práticas recomendadas. Note-se que o PhantomBuster apenas recolhe dados de contas públicas, definidas como tal, pelos utilizadores, no campo relativo à privacidade. Assim, dados de contas/perfis/conteúdo definidos como privados não farão parte da base de dados. Esta informação é relevante para quando refletir sobre questões éticas no que diz respeito à ferramenta e à metodologia.

É importante ter em conta que o PhantomBuster é uma ferramenta de carácter *freemium*. Isto significa que poderá ser necessário pagar para ter acesso a determinadas funcionalidades, para ter mais tempo de utilização ou aumentar o volume de dados que pretende recolher.

O PhantomBuster não foi desenvolvido para realizar investigação científica, mas sim para dar resposta a necessidades no âmbito do *marketing*. Assim, irá deparar com várias limitações, uma das quais é, por exemplo, não ser possível estabelecer balizas temporais aquando da recolha, o que terá consequências no modo como irá desenhar a sua pesquisa, requerendo adaptações relativamente à recolha e tratamento dos dados.

Para usar o PhantomBuster terá de ter uma conta criada. Adicionalmente, também é necessário ter uma conta criada na plataforma digital da qual pretende recolher dados, neste caso específico no Instagram.

Criar conta no PhantomBuster

1. Ir a https://phantombuster.com/signup;
2. indicar *email* e *password*;
3. indicar dados solicitados (nome, apelido, empresa);
4. confirmar dados de registo através do email que recebeu na caixa de correio que usou para criar a conta;
5. instalar a extensão do PhantomBuster no seu *browser*. Como *browser*, vamos usar, preferencialmente, o Chrome.

Usar o PhantomBuster

O PhantomBuster funciona numa lógica de ações/módulos, intitulados *phantoms* e *flows*. Recolhe, no caso do Instagram, os dados diretamente da versão *website* da rede social. Sim, *website* e não diretamente da versão *app* móvel.

Explorar a ferramenta

No campo "Solutions", encontrará todos os *phantom*s disponíveis, bem como uma explicação sobre cada um deles, incluindo o tipo de dados que recolhe, os *inputs* necessários e os *outputs* gerados.

Neste capítulo, a nossa atenção recairá sobre o Instagram e três *phantom*s específicos:

— Instagram Profile Post Extractor — recolhe *posts* de uma conta ou de uma lista de contas do Instagram;
— Instagram Post Commenter Export — recolhe comentários de *posts* do Instagram;
— Instagram Hashtags Search Export — recolhe os *posts* mais populares do Instagram associados a uma *hashtag* (ou a uma localização).

Inputs *necessários*

Após ter selecionado, em função da pergunta de investigação e objetivos do estudo, o *phantom* que vai utilizar, a ferramenta requer, de forma geral, um conjunto de *inputs*/informações, para realizar a recolha. No quadro 4.1 indicam-se os *inputs* e a sua descrição em função dos três *phantom*s que vão ser explorados.

Uma vez estabelecidos todos os critérios de recolha, deverá clicar no botão "Launch". Em primeiro lugar aparecerá uma informação sobre o sucesso da autenticação e, em segundo lugar, a indicação da percentagem correspondente ao progresso da recolha.

Quadro 4.1 *Inputs* necessários e a sua descrição em função do *phantom*

Input	Descrição	Instagram Profile Post Extractor	Instagram Post Commenter Export	Instagram Hashtags Search Export
Connect to Instagram	Através dos *cookies* da sessão do Instagram.	X	X	X
Profile URL	URL do perfil, do qual pretende recolher *posts*.	X		
Post URL	URL do *post* do qual pretende recolher comentários.		X	
Hashtag	*Hashtag* através do qual quer realizar recolha.			X
Number	Indicar amostra.	X	X	X
Name your results file	Atribuir um nome ao ficheiro de resultados.	X	X	X
Launch settings	Aconselhamos que comece por frequência "once" e "Launch manually".	X	X	X

Fonte: elaboração própria da autora.

Obter resultados

Os resultados aparecerão na mesma página *web* do PhantomBuster. Basta deslizar para baixo. Pode explorá-los diretamente na página *web*, mas, para analisar, será preferível fazer *download* dos mesmos:

— Clique no botão "Download Results".
Automaticamente, um ficheiro CSV, com os resultados da recolha, irá ser transferido para o seu computador. Um ficheiro CSV (*comma separated value*) poderá ser aberto através do Excel ou do Google Sheets. Por questões de possibilidades e funcionalidades, que nos vão ser úteis para a exploração dos dados, iremos optar pelo Google Sheets. Para tal, é preciso fazer o *upload* do ficheiro para o Google Drive e selecionar a opção "Abrir com" Google Sheets;
— Os resultados serão apresentados organizados por linhas e colunas. Cada linha corresponde a um resultado específico (*post* de um perfil, um comentário de um *post* ou um *post* associado a uma *hashtag*). As colunas fornecerão dados específicos sobre esse *post* de um perfil, um *post* associado a uma *hashtag* ou um comentário de um *post*.

Importante: caso não tenha a versão *premium* do PhantomBuster, encontrará limitações para fazer o *download* do ficheiro dos resultados. Poderá, em alternativa, copiar os dados que são apresentados na página *web* e colá-los diretamente num ficheiro Google Sheet. Demorará mais tempo, mas é uma alternativa. Se recolheu um grande volume de dados, não aparecerão todos na página *web*, pelo que a única solução para obter todos os resultados será a versão *premium*.

Outputs *gerados*

Os *outputs* gerados, isto é, os dados que estarão presentes no ficheiro dos resultados, a base de dados, serão específicos do *phantom* utilizado. No quadro 4.2 reúnem-se os *outputs* em função dos três *phantom*s que merecem a nossa atenção.

Quadro 4.2 *Outputs* (campos de dados) e a sua descrição em função dos *phantoms*

Dado	Descrição	Instagram Profile Post Extractor	Instagram Post Commenter Export	Instagram Hashtags Search Export
comment	Comentário propriamente dito		X	
commentCount	Número de comentários	X		X
commentDate	Data do comentário	X	X	
commentId	ID do comentário		X	
description	Legenda do *post*. Texto que acompanha a imagem	X		X
fullName	Nome do autor do *post*	X		X
imgUrl	URL da imagem	X		X
isSidecar	Relativo a se é um *post* formato carrossel	X		X
likeCount	Número de *likes*	X	X	X
likedByViewer	Relativamente a se quem viu, gostou	X		
location	Localização. Surge se quem postou a associou	X		X
locationId	ID da localização	X		
ownerId	ID do utilizador		X	X
playCount	Número de vezes que o vídeo começa a ser reproduzido	X		
postId	ID do *post*	X		X
postUrl	URL do *post*	X		X
profilePictureUrl	URL da fotografia de perfil		X	
profileUrl	URL do perfil	X	X	X
pubDate	Data da publicação do *post*	X		X
query	Dados através dos quais a pesquisa foi realizada (perfil, *hashtag*, *post*)	X	X	X
replyCount	Número de respostas ao comentário		X	
sidecarMedias	Quantas fotos/imagens fazem parte do *post* caso este seja carrossel	X		X
timestamp	Data e hora a que a recolha de dados foi realizada	X	X	X
type	Formato do *post* (foto, vídeo)	X		X
username	Nome do utilizador	X	X	X
videoDuration	Duração do vídeo	X		
videoUrl	URL do vídeo	X		X
viewCount	Número de visualizações			X

Fonte: elaboração própria da autora.

Apagar phantoms *utilizados*

Poderá querer usar vários *phantoms* para o mesmo ou distintos tópicos de pesquisa. Caso não tenha conta *premium*, o número de *phantoms* que vai poder usar estará limitado. Para poder realizar outras recolhas, terá de apagar aquelas realizadas até ao momento:

— no menu principal, vá a "Dashboard". Aparecerão todos os *phantoms* que já utilizou;
— no canto superior direito de cada *phantom* utilizado, terá três pontos (...);
— clique nesses três pontos e escolha a opção "Delete". Esta ação apaga do PhantomBuster os dados e ficheiros associados a esse *phantom*.

Receitas

As receitas abaixo foram desenhadas com o objetivo de familiarizar-se com a recolha de dados através do PhantomBuster, assim como com as possibilidades e limitações da ferramenta e de como esta pode ser utilizada no âmbito da sua investigação.

Ponto de partida: perfil

Esta receita irá guiá-lo através do processo de recolha de dados através do *phantom* Instagram Profile Post Extractor:

— faça *login* no PhantomBuster;
— faça *login* no Instagram (versão *browser*);
— escolha o *phantom* Instagram Profile Post Extractor;
— siga o passo a passo:
  — clique no botão "Connect to Instagram";
  — indique o URL do perfil ou perfis dos quais deseja recolher dados. Para este exercício será utilizado o perfil do Instagram do Serviço Nacional de Saúde português (https://www.instagram.com/sns_pt/);
  — defina a sua amostra na secção "Number of posts to extract per profile", indique se deseja, no caso dos *posts* carrossel, recolher apenas o primeiro *post*, e que nome pretende atribuir ao ficheiro de resultados: "Name your results file";

Duas considerações:

I. Pense no número de *posts* apropriados para responder à sua questão de pesquisa, isto é, o número de *posts* que comporá a sua amostra. Poderão existir vários fatores que influenciam a decisão, tais como: a popularidade do perfil, a análise concentrar-se num período específico, a análise ser longitudinal. Aconselhamos que comece por obter uma amostra de pequena dimensão e que a explore.

Note que extrair um grande número de *posts* poderá levar mais tempo ou até ser limitado pelo Instagram.

II. Atribua uma nomenclatura ao seu ficheiro que o ajude a identificá-lo e da qual não se esqueça. No contexto da pesquisa digital é natural realizar várias recolhas. Uma estratégia que poderá funcionar é: Nome da rede_nome do perfil_n_data da recolha. Neste caso Instagram_SNSPT_100_080224. Adicionalmente, recorde-se onde, no seu computador ou *online*, vai guardar o ficheiro. Será lógico que crie uma pasta específica para a investigação que está a realizar. Ter os ficheiros organizados é de extrema ajuda:

— defina os "Launch settings": "once" e "Launch manually";
— clique no botão "Launch". Poderá acompanhar a recolha através da barra e do valor da percentagem;
— clique em "Download results" e abra o ficheiro de resultados CSV com o Google Sheets.

Caso a recolha não tenha progredido, verifique:

— se tem o *login* feito no Instagram num separador do *browser*;
— se tem a extensão do PhantomBuster instalada no *browser*;
— se está a recolher dados de um perfil público;
— caso esteja a recolher dados de mais do que um perfil, e tendo que indicar o URL de um ficheiro Google Sheets, certifique-se de que a privacidade desse ficheiro está definida como pública;
— experimente com uma amostra mais pequena;
— experimente com outro *browser*;
— verifique os critérios da recolha.

Analisar top posts *do perfil*

— Abra o ficheiro de resultados no Google Sheets;
Faça *upload* no Google Drive, clique com o botão do lado direito do rato, escolha "abrir com" e selecione Google Sheets;
— uma vez o ficheiro aberto, realize uma cópia da folha de resultados. O objetivo é garantir que mantém a versão original da recolha inalterada. Trabalhe nessa nova folha que criou.
Crie um filtro na folha:
  — selecione o cabeçalho da folha;
  — clique no ícone correspondente ao "Adicionar filtro" (assemelha-se a um funil) ou no menu "Dados" selecione a opção "Criar um filtro". Uma vez o filtro aplicado, aparece um triângulo (invertido) na primeira célula das colunas do cabeçalho da folha;
  — os filtros permitirão ordenar e filtrar os dados por qualquer uma das colunas dos cabeçalhos.

— ordene os *posts* pelo número de *likes*:
  — clique no triângulo do cabeçalho da coluna "Like count" e depois "Ordenar A a Z". Agora pode ver os *posts* ordenados de acordo com o número de *likes* que receberam.

O que caracteriza esses *posts*? Que assunto ou assuntos focam? Qual o tom dos mesmos? Em que data foram publicados?

O processo anterior pode ser usado também para explorar comentários. Para tal:

— Na coluna "Comment count", já com o filtro criado, clique no triângulo e depois "Ordenar A a Z". Agora pode ver os *posts* de acordo com o número de *comments* que receberam.

Quais os *posts* que obtiveram o maior número de comentários? E os que obtiveram menos? Quais as diferenças entre eles? Há diferenças entre *posts* que obtiveram maior número de *likes*, mas menor número de *comments*? O que pode concluir sobre eles?

Obter imagens associadas aos posts

Verificará que o ficheiro de resultados (base de dados) não tem as imagens dos *posts* em formato JPG ou PNG. Aconselhamos que, caso queira obter as imagens publicadas pelos perfis que está a analisar, siga os seguintes procedimentos logo após obter o ficheiro de resultados. Isto porque as imagens estão alojadas, temporariamente, num URL que expira rapidamente.

Passo prévio: caso ainda não tenha realizado os dois primeiros passos do procedimento anterior, "Analisar *top posts* do perfil", realize-os antes de iniciar o processo de obter as imagens.

a. Adicione uma coluna ao lado da coluna "ImgURL":
  — selecione a coluna "ImgURL";
  — clique no botão do lado direito do rato e selecione a opção "inserir uma coluna à direita";
  — atribua o nome "Image" a essa coluna.

Esse URL vai ser o ponto de entrada. Por fim, clique em "Enter". Obterá na coluna "Image" uma pré-visualização da imagem associada ao *post*.

Para aplicar a função às outras células da coluna "Image", clique na célula onde a pré-visualização da primeira imagem aparece. Verá que no canto inferior direito surge um círculo (também poderá ser um quadrado verde). Clique duas vezes nesse círculo/quadrado e a função será aplicada às demais células da coluna.

Outra opção é, em vez de clicar duas vezes nesse círculo/quadrado, arrastá-lo para baixo ao longo das células da coluna "Image".

Verá que a pré-visualização das imagens aparece nas demais células da coluna "Image". Poderá redimensionar a largura e a altura da coluna "Image" para ter uma melhor visualização de cada uma das imagens.

Caso não consiga obter uma pré-visualização, verifique:

— se a fórmula está bem aplicada;
— se está a selecionar o URL correto, o da coluna "ImgURL", que começa por https://scontent-;
— se não passou demasiado tempo entre a recolha de dados e o passo a passo para a pré-visualização.

Após pré-visualizar as imagens associadas aos *posts*, pode fazer o *download* das mesmas:

— copie os dados da coluna "ImgURL";
— abra um programa editor de texto (Word não é válido). No Mac use o "Editor de texto". Em ambiente Windows/PC pode usar o "Bloco de Notas";
— crie um novo documento;
— no menu "Formatação", escolha a opção "Converter em texto simples" (versão Mac);
— cole os dados da coluna "ImgURL" que tinha copiado;
— guarde o ficheiro. Use uma nomenclatura apropriada e da qual se irá recordar. Por exemplo:
Nome da rede_nome do perfil_n_data da recolha_nome_coluna.
Neste caso, o nome seria:
Instagram_SNSPT_100_080224_URL_IMAGENS.

Os procedimentos seguintes são realizados através da extensão DownThemAll.[2] Uma vez instalada, siga os passos:

— no *browser*, clique no ícone (seta amarela) que corresponde ao DownThemAll;
— clique em "Gestor" (ou "Manager", caso o idioma esteja definido para inglês);
— clique no botão do lado direito do rato e selecione "Importar do ficheiro". O ficheiro a selecionar é aquele onde colou os dados da coluna "Img URL" — Instagram_SNSPT_100_080224_URL_IMAGENS — e clique em abrir;
— uma nova janela surgirá e nessa verifique que:
  — no campo "Filtros" estão selecionadas as opções: "Todos os ficheiros", "Images JPG" e "Videos". Caso não estejam, selecione;
  — no campo "Subpasta", indique o nome de uma pasta a ser criada e onde as imagens dos *posts* serão guardadas. Nota: não precisa de criar previamente essa pasta no seu computador. A pasta será criada automaticamente na recolha.

2 Pode instalá-la através do *link* https://www.downthemall.net/

Sugerimos:
Nome da rede_nome do perfil_n_data da recolha_nome_IMAGENS_POSTS.
Neste caso: Instagram_SNSPT_100_080224_IMAGENS_POSTS

— no campo "Máscara" escrever o seguinte comando *idx*.*ext*;
— clicar em transferir.

Verá o resultado da transferência através dos dados que vão aparecendo no ecrã, nomeadamente através do campo "progresso". Cada linha corresponderá a um *post*. A ordem pela qual aparecem será reflexo da ordem pela qual os *posts* estavam no ficheiro de resultados e a ordem pela qual foram copiados e colados no editor de texto. Caso alguma das barras de progresso apareça a vermelho, significará que a imagem já não está disponível para *download*.

Uma vez terminado o *download* poderá explorar a pasta onde as imagens foram automaticamente guardadas. Neste caso:
Instagram_SNSPT_100_080224_IMAGENS_POSTS para as explorar e analisar.

Nota: para um novo *download* de imagens de *posts*, através do DownThemAll, deverá eliminar o *download* anterior. Para tal:

— no *browser*, clique na seta correspondente ao DownThemAll;
— clique em "Gestor". Aparecerá o resultado da recolha anterior;
— clique e escolha a opção "Selecionar tudo";
— clique e escolha a opção "Remover transferência".

Pode ser útil, de forma que se explorem as imagens, que estas sejam organizadas cromaticamente. O programa ImageSorter permitirá fazê-lo.[3]

Nota: o procedimento para a obtenção de imagens não é exclusivo para o *phantom* Instagram Profile Post Extractor ou imagens provenientes de *posts*. Pode ser reproduzido para recolhas cujo ponto de partida sejam *hashtag*s e que iremos explorar mais à frente neste capítulo.

Outras possibilidades de investigação

— Como é que os *posts* evoluem ao longo do tempo? Explore a coluna "Pub Date";
— existe um formato preferencial de publicação? Explore a coluna "Type";
— qual o assunto abordado nos *posts*? Explore também a coluna "Description". Além de uma possível análise de conteúdo ou de discurso, também poderá explorar o *software* WORDij.[4] Este cria uma rede semântica.

3 https://appnee.com/imagesorter/
4 https://wordij.net/

Em Sepúlveda e Espanha (2022), o perfil de Instagram do Serviço Nacional de Saúde português constituiu o objeto de estudo. Através da análise foi possível concluir quanto à tipologia de mensagens de saúde visualmente apelativas, quer quanto aos temas com os quais os utilizadores mais interagiram. As conclusões reforçavam a importância do Instagram como meio de comunicação para obter informação, revelando-se particularmente importante em tempos de crise sanitária.

Já em Sepúlveda e Crespo (2021), os perfis oficiais no Instagram dos candidatos às Presidenciais 2021 foram alvo de análise quantitativa (publicações, seguidores, fotos e vídeos partilhados, comentários), como longitudinal (tentando caracterizar a evolução temporal), e qualitativa (os mais comentados, padrões visuais, caracterização de publicações mais recorrentes, *emojis* e *hashtag*s mais utilizados).

Ponto de partida: comentários

Esta receita irá guiá-lo através do processo de recolha de dados com recurso ao *phantom* "Instagram Post Commenter Export":

— faça *login* no PhantomBuster;
— faça *login* no Instagram (versão *browser*);
— escolha o *phantom* Instagram Post Commenter Export;
— siga o passo a passo:
  a. clique no botão "Connect to Instagram";
  b. indicar o URL do *post* ou *posts* dos quais deseja recolher comentários. Para este exercício, vamos usar o *post* do jornal *Público* no qual é dada a notícia sobre o, naquele momento, presidente da Federação Espanhola de Futebol, Luis Rubiales, ter beijado a jogadora Jenni Hermoso durante as comemorações da vitória no mundial feminino. Pode ver o *post* em: https://www.instagram.com/p/CwNg8QpN7x9/;
  c. defina a sua amostra na secção "Number of comments to extract per post" e indique qual o nome do seu ficheiro de resultados "Name your results file". Usámos: Nome da rede_ID post_data da recolha, ou seja, Instagram _CwNg8QpN7x9_09022024;
  d. defina os "Launch settings": "once" e "Launch manually";
  e. clique no botão "Launch". Poderá acompanhar a recolha através da barra que está em baixo e do valor da percentagem.
— clique em "download results" e abra o ficheiro de resultados CSV com o Google Sheets.

Caso a recolha não tenha progredido, verifique as possibilidades mencionadas anteriormente.

Recolher emojis dos comentários

Os *emojis* podem ser elementos de extrema utilidade para, por exemplo, análise de sentimentos. Para obter os *emojis* presentes nos comentários, siga os seguintes passos:[5]

— abra o seguinte *link* https://labs.polsys.net/tools/textanalysis/;
— copie o conteúdo da coluna "Comment";
— cole o conteúdo da coluna "Comment" no campo "Paste text" do *software* Textanalysis;
— clique em enviar;
— no campo "Emoji stats", terá o resultado dos emojis presentes nos comentários, o seu nome, a sua representação visual e a sua frequência.
— copie os dados no campo "Emoji stats" e cole num ficheiro Google Sheet;
— organize os dados separando-os por: ",". Para tal, utilize a função "Dividir texto em colunas" presente no menu "Dados" do Google Sheet.[6]

Outras possibilidades de investigação

— quem são aqueles que comentam o *post*? Pessoas comuns? *Influencers*? Políticos? Líderes de opinião? Explore a coluna "Profile URL" e categorize os comentadores;
— que tipo de comentários são realizados? Ofensivos? Opinativos? A defenderem algum dos intervenientes? De natureza misógina? Explore a coluna "Comment";
— algum comentário se destaca pelo número de *likes* ou *reply*? Explore as colunas "Like count" ou "Reply count";
— qual a data dos comentários? Perduraram no tempo? Explore a coluna "Comment Date".

Ponto de partida: hashtag

Esta receita irá guiá-lo através do processo de recolha de dados com recurso ao *phantom* "Instagram Hashtags Search Export".

— Faça *login* no PhantomBuster;
— faça *login* no Instagram (versão *browser*);
— escolha o *phantom* Instagram *Hashtags* Search Export;

---

5 Este processo também pode ser aplicado a outros dados textuais como, por exemplo, os da coluna "Description" — relativa à legenda da publicação.

6 Se deseja analisar dados por tempo, as colunas "Pub Date" (Instagram Profile Post Extractor e Instagram Hashtags Search Export) ou "Comment Date" (Instagram Post Commenter Export) são particularmente úteis. Porém, a data aparece no formato "2023-08-21T15:19:12.000Z". Utilize a função "Dividir texto em colunas". Escolha a opção "Personalizado" e escreva "-".

— siga o passo a passo:
  a. clique no botão "Connect to Instagram";
  b. indique a *hashtag* (ou *hashtag*s) através das quais deseja recolher dados. Para este exercício, vamos usar a *hashtag* #legislativas2024;
  c. defina a amostra no campo "Number of posts to extract per hashtag" e indique qual o nome do seu ficheiro de resultados "Name your results file". Usámos: Nome da rede_hashtag_data da recolha, ou seja, Instagram_#legislativas2024_09022024;
  d. defina os "Launch settings": "once" e "Launch manually";
  e. clique no botão "Launch". Poderá acompanhar a recolha através da barra e do valor da percentagem;
— abra o ficheiro de resultados CSV com o Google Sheets.

Caso a recolha não tenha progredido, lembre-se de verificar as hipóteses mencionadas anteriormente.

Explorar contas associadas à/s hashtag/s

As colunas "Full name" e "Profile URL" permitem, entre outras possibilidades, não só explorar que utilizadores publicaram conteúdo associado ao #legislativas2024, como aferir quanto à sua frequência:

— abra o ficheiro de resultados no Google Sheets;
— faça uma cópia da folha de resultados. Trabalhe nessa nova folha que criou;
— crie um filtro na folha;
— ordene a coluna "Full name". Clique no triângulo invertido para escolher "Ordenar A a Z";
— use a função "Countif" para contar quantas vezes os agentes se repetem:
  a. selecione a coluna "Full name";
  b. clique no botão do lado direito e escolha "Adicionar coluna à direita";
  c. escreva a função "COUNTIF": =COUNTIF (intervalo de valores; "condição") ou seja =COUNTIF (intervalo de células onde estão os dados; "nome do agente". Exemplo: COUNTIF(D2:D29; "ECO").
  Como resultado, na coluna que adicionou, surge o número de vezes que aquele agente se repete. Uma vez terminado, ordene a coluna COUNTIF de Z a A.

Explorar outras hashtags

Para explorar outras *hashtag*s presentes no resultado, trabalhe com os dados da coluna "Description". Essa exploração implicará que os dados dessa coluna sejam tratados. Pode usar a ferramenta Hashtag Extractor.[7] Através desta irá obter uma lista das *hashtag*s que são usadas.

Para tal:

— copie e cole os dados da coluna "Description" da base de dados no campo "input" da ferramenta Hashtag Extractor;
— clique em "submit";
— copie os dados do campo "output" e cole numa folha de um ficheiro Google Sheets;
— use a função "COUNTIF" para contabilizar qual e quantas vezes determinada *hashtag* aparece.

Outras possibilidades de investigação

— Quais os *top posts* associados às *hashtag*s?
Reproduza o processo "Analisar *top posts* do perfil" indicado na receita "Análise de um perfil";
— quais os assuntos abordados nos *posts* associados com #legislativas2024?
Explore a coluna "Description". Caso seja interessante, poderá recolher e analisar os *emojis* presentes no texto;
— quais são as imagens associadas às diferentes *hashtag*s?
Explore as colunas "Description" e "Img URL". O passo a passo "Obter imagens dos *posts*" será útil;
— que imagens são publicadas pelos perfis?
Use a função "criar filtro" do Google Sheet para criar listas separadas;
— como é que os *posts* associados à *hashtag* evoluem ao longo do tempo?
Explore a coluna "Pub date", converta a data e crie um gráfico. O recurso ao RawGraphs poderá ser útil.

Referências bibliográficas

Al-Rawi, A. (2021), "Political memes and fake news discourses on Instagram", *Media and Communication*, 9 (1), pp. 276-290, https://doi.org/10.17645/MAC.V9I1.3533.

Childs, K. M. (2022), "'The shade of it all': how black women use Instagram and YouTube to contest colorism in the beauty industry", *Social Media + Society*, 8 (2), pp. 1-15, https://doi.org/10.1777/20563051221107634.

Daer, A.R., R.F. Hoffman, e S. Goodman (2014), "Rhetorical functions of *hashtag* forms across social media applications", *Proceedings of the 32nd* ACM International Conference on the Design of Communication, CD-ROM (SIGDOC '14), ACM, Nova Iorque.

Dixon, S.J. (2023), "Number of Instagram users worldwide from 2020 to 2025", *Statista*, disponível em https://www.statista.com/statistics/183585/instagram-number-of-global-users/.

7 https://www.text-utils.com/hashtag-extractor/

Drüeke, R., C. Peil, e M. Schreiber (2022), "How do Black lives matter? Zur visuellen Konstruktion von Protest in deutschsprachigen Tageszeitungen und auf Instagram How do Black Lives Matter? On the visual construction of protest in German-language daily newspapers and on Instagram", *Studies in Communication Sciences*, 22 (1), pp. 1-19, https://doi.org/10.24434/j.scoms.2022.01.2982.

Flores, A.M. e R. Sepúlveda (2021), "Métodos digitais e educação: uma proposta de investigação", em A. Nobre, A. Mouraz e M. Duarte (eds), *Portas Que o Digital Abriu na Investigação em Educação*, pp. 226-255, DOI: 10.34627/uab.edel.15.11.

Highfield, T., e T. Leaver (2016), "Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji", *Commun. Res. Pract*, 2 (1), pp. 47-62.

Leaver, T., T. Highfield, e C. Abidin (2020), *Instagram: Visual Social Media Cultures*, Digital Media and Society Series, Cambridge, Reino Unido, Polity.

Meraz, S. (2017), "Hashtag wars and networked framing", em A.S. Tellería (ed.), *Between the Public and Private in Mobile Communication*, Routledge, pp. 303-323.

Meraz, S., e Z. Papacharissi (2013), "networked gatekeeping and networked framing on #Egypt", *Int. J. Press-Polit*. 18 (2), pp. 138-166.

Romero Saletti, S. M., S. Van den Broucke, e W. Van Beggelaer (2022), "Understanding motives, usage patterns and effects of instagram use in youths: a qualitative study", *Emerging Adulthood,* 10 (6), pp. 1376-1394, https://doi.org/10.1177/21676968221114251.

Sepúlveda, R. e M. Crespo (2021), "Presidenciais 2021 no Instagram — de 17 de dezembro a 17 de janeiro", *MediaLab*, disponível em https://medialab.iscte-iul.pt/presidenciais-2021-no-instagram-de-17-de-dezembro-a-17-de-janeiro/.

Sepúlveda, R. e R. Espanha (2022), "Online social media and public health institutions: an analysis of the Portuguese National Health Service on Instagram", em H. Martins (coord.), *Cadernos de Saúde Societal*, pp. 63-73, disponível em www.iscte-iul.pt/assets/files/2022/11/24/1669293875177_cadernos_Saude_Societal_03_2022_PT.pdf

Zappavigna, M. (2015), "Searchable talk: the linguistic functions of hashtags", *Soc. Semiot*, 25 (3), pp. 274.291.

Capítulo 5

Twitter/X

Publicações, conteúdos e análise de redes

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Sofia Ferro-Santos
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

O Twitter é tradicionalmente uma das plataformas de redes sociais mais usadas para o estudo de temas de relevância social e política. Esta plataforma foi criada em 2006, fundada por Jack Dorsey, Evan Williams e Biz Stone, com o fim de replicar, na *internet*, os populares serviços de mensagens curtas existentes nos telefones. O primeiro *tweet* – o nome pelo qual ficaram conhecidas as publicações do Twitter – foi publicado em 21 de março de 2006 pelo próprio Jack Dorsey.[1]

Inicialmente limitado a 140 caracteres de texto, o Twitter duplicou essa quantidade de texto em 2017 – a pedido dos utilizadores – mas manteve, em parte devido a essa limitação, uma forma de comunicação rápida e incisiva, na qual os utilizadores procuram condensar o essencial da sua mensagem em duas ou três frases. Embora esta plataforma tenha evoluído muito ao longo dos anos (e os utilizadores pagantes tenham já acesso a uma maior quantidade de caracteres nas suas publicações) essa característica manteve-se como um dos traços marcantes do Twitter.

Como curiosidade, refira-se também que foi no Twitter que surgiu a popular *hashtag*, em 2007, aliás correspondendo à sugestão de um dos seus primeiros utilizadores – Chris Messina – que aventou a hipótese de ligar as mesmas palavras usadas em diferentes *tweet*s através do símbolo cardinal.[2]

Uma das principais características do Twitter em comparação com outras plataformas de redes sociais é permitir a "autocomunicação de massas" (Castells, 2009), no sentido em que permite que um utilizador comunique, sem intermediários, para muitos outros, ou "multicast" porque o Twitter não exige que a relação entre utilizadores seja recíproca (Murphy, 2013). Ou seja, um utilizador pode seguir poucas contas e ser seguido por muitas. Além disso, ao contrário de outras plataformas de redes sociais, o conteúdo que cada utilizador vê no Twitter não é apenas o conteúdo que "subscreveu" ou que foi partilhado por utilizadores que segue. Quer através da utilização e visualização de conteúdos com *hashtag*s, quer pela recomendação do algoritmo, é muito comum o utilizador ver conteúdo de contas que não segue e que não foi partilhado por estas.

1 https://twitter.com/jack/status/20
2 https://www.cnbc.com/2018/04/30/chris-messina-hashtag-inventor.html

O Twitter foi uma das primeiras plataformas de redes sociais a ser amplamente adotada pelos meios de comunicação social e pelos políticos para aumentar o alcance das suas notícias e das suas mensagens. E isso também acabou por caracterizar a rede. A conta de Twitter de Barack Obama, ex-presidente dos Estados Unidos da América, foi durante muitos anos aquela que contava com mais seguidores, num *ranking* que incluía – e ainda inclui – muitos outros políticos e empresários, como Narendra Modi, Donald Trump ou Bill Gates e meios de comunicação como a CNN, o *New York Times* ou a BBC (Kemp, 2024).

Em outubro de 2022, na sequência de um processo negocial longo, o empresário Elon Musk comprou a totalidade das ações do Twitter e retirou a empresa da bolsa, passando desde então a ser uma empresa privada, sem obrigações de prestação de contas. Em 2023, a plataforma passou a chamar-se oficialmente X.

Atualmente, o Twitter/X chega a cerca de 610 milhões de pessoas em todo o mundo e tem uma predominância de utilizadores do sexo masculino e com idades entre os 18 e os 34 anos (Kemp, 2024). Em Portugal, o Twitter/X chega a mais de dois milhões de utilizadores, também predominantemente do sexo masculino (Kemp, 2024b).

Embora o Twitter/X não seja uma das plataformas de redes sociais com mais utilizadores em Portugal, é aquela em que o consumo de notícias e conteúdos de relevância social e política tende a ser mais referido nos estudos existentes (Cardoso *et al.*, 2023). A nível da sua adoção, apesar de ser uma plataforma usada por apenas 14,9% da população portuguesa, a sua penetração tem aumentado na população mais jovem (Cardoso *et al.*, 2023). A sua utilização também tem crescido entre os atores políticos, por exemplo pelos deputados: em 2020 apenas 41% dos deputados tinham uma conta na plataforma (Haman e kolnik, 2021) e em 2022 já eram 56% (Ferro-Santos *et al.*, 2024a). Daí que esta seja igualmente uma plataforma muito relevante para a pesquisa e investigação recorrendo a métodos digitais.

Atualmente, a plataforma permite acompanhar os textos das publicações com fotos, vídeos ou GIF (além dos *links*) e cada publicação – agora chamada *post* – pode ser objeto de comentários, "gostos" ou partilhas, quer de forma pública – os *retweets* ou *quote-retweets* – ou privada, através das mensagens privadas. Os conteúdos disponíveis na plataforma são maioritariamente públicos (e podem ser vistos mesmo por quem não tem conta), embora possam ser marcados como privados, quer através de tornar o perfil privado, quer bloqueando alguns utilizadores (apesar de esta opção poder vir a ser alterada em breve). Cada utilizador pode igualmente limitar o tipo de utilizador que pode responder às publicações (qualquer utilizador, contas que segue, contas verificadas ou apenas contas que são mencionadas na publicação).

Há estudos sobre comunicação no Twitter/X nas mais variadas áreas e com os mais variados atores, desde política (Ferro-Santos *et al.*, 2024a, Ferro-Santos *et al.*, 2024b), desporto (Gouveia *et al.*, 2018), celebridades e *fandom* (Highfield *et al.*, 2013), além de outros temas. Há também estudos que se focam em analisar efeitos sociais e algorítmicos que podem acontecer na plataforma como a criação de *filter bubbles* (Bruns, 2019; Pariser, 2011) e *echo chambers* (Habermas, 2022; Sunstein, 2006). Há também estudos que se focam não só no Twitter/X, mas no seu uso comparativo com outras plataformas na abordagem de um tema ou de um grupo de atores (Moreno *et al.*, 2024). Por fim, há também diferentes objetos de estudo no Twitter/X,

desde a sua adoção, ao seu conteúdo, passando pelas redes de interação que se criam ou mesmo a decisão de deixar de o usar.

Recolha de dados

Ferramenta: SentiOne Listen

Historicamente, a primeira forma de recolher dados das redes sociais de uma forma automatizada ocorreu através do *scraping*, uma técnica que usa o código HTML de uma determinada página *web* – neste caso uma visualização de uma rede social num determinado *browser* – para extrair parte dos dados identificados nesse HTML. Com o desenvolvimento das plataformas de redes sociais e para responder às crescentes solicitações de recolha de dados, as empresas de plataformas de redes sociais começaram a desenvolver API – *Advanced Programmers Interfaces* (interfaces avançadas para programadores) – para disponibilizar esses dados. O Twitter/X teve durante algum tempo uma API especialmente dedicada a investigadores académicos, mas atualmente apenas dispõe da API paga, à qual acedem as empresas que querem monitorizar a sua presença nas redes sociais ou as empresas que desenvolvem sistemas de *software* para fornecer a essas empresas mediante uma *fee* de utilização.

O SentiOne é uma dessas empresas. Tal como outros pacotes de *software online* semelhantes – Brandwatch, Talkwalker, Brand24, NewsWhip, Mention, etc. – o SentiOne faculta o acesso pago a um produto – o SentiOne Listen – que permite monitorizar várias redes sociais diferentes, segundo parâmetros de configuração altamente customizáveis, o que permite responder a vários tipos de utilização.[3] E embora a sua vocação seja, como nos casos acima, servir as estratégias de *marketing* das empresas, o facto de se constituir como uma ferramenta de recolha de dados das redes sociais possibilita a sua utilização para fins de pesquisa académica. O SentiOne funciona por "projects". Ou seja, cada pesquisa que quisermos realizar vai corresponder a um "project". E cada projeto pode dirigir-se a uma ou a várias redes sociais e incidir sobre determinados atores dessa rede ou sobre uma *query* – um conjunto de palavras-chave – cuja utilização nessa rede social queremos explorar. Neste capítulo iremos dirigir o nosso SentiOne para o Twitter/X e iremos explorar várias formas de recolher dados dessa plataforma usando esta ferramenta. Um ponto importante: embora o SentiOne e as ferramentas semelhantes enunciadas acima funcionem todas segundo as mesmas regras básicas – todas recolhem dados através da API e todas permitem uma grande amplitude de configurações – existem diferenças entre elas, que podem ir do tipo de dados recolhidos até ao modo como são apresentados. Por isso, embora possam existir ligeiras diferenças na forma de implementar as receitas seguintes noutras ferramentas semelhantes ao SentiOne, com a configuração todas elas permitem implementar o mesmo tipo de análise. O que significa que o SentiOne é aqui usado como mero exemplo. Também é importante sublinhar que embora o SentiOne, como os outros *softwares* semelhantes, apresente grandes potencialidades em termos de investigação

3 https://sentione.com/features/listen

académica, revela também importantes limitações, uma vez que a recolha e apresentação de dados é pensada para uma utilização em campanhas de *marketing* digital e não em projetos de investigação científica.

Criar conta no SentiOne Listen

O SentiOne é um produto pago e dispõe de vários patamares de preço, cada um com o seu custo e as suas funcionalidades. Atualmente, a gama de preços começa nos 300 dólares por mês. A criação de conta envolve o registo com *email* empresarial e *password*, além do pagamento do produto, obviamente.

Criar um projeto no SentiOne Listen

No SentiOne Listen, um projeto pode ser definido por uma *query* e/ou por um conjunto de perfis ou contas em redes sociais e ambos podem ser filtrados por vários parâmetros, com destaque para a língua e para o país. A configuração de um projeto é uma fase importante, uma vez que vai determinar que tipo de resultados vão ser recolhidos e como eles vão ser apresentados. Na definição de projetos no SentiOne podemos optar por focar a análise numa marca, num conjunto de perfis sociais associados a essa marca ou numa área "advanced", que é onde se pode melhor explorar as potencialidades da ferramenta para a investigação académica. Nessa área, a diversidade de configurações possíveis é bastante alargada. Começa com a atribuição de um nome ao projeto e termina com o guardar desse projeto (o número de projetos disponíveis na conta SentiOne é um dos fatores constitutivos do preço). A configuração de um projeto avançado no SentiOne passa pela escolha de uma série de regras ("rules", no original) de monitorização, que especificam que tipo de conteúdos vamos tirar.

Nesta fase de configuração do projeto no SentiOne podemos escolher recolher publicações a partir de uma *query*, ou seja, a partir de um conjunto articulado de *keywords*, que irá trazer até ao acervo da investigação todas as publicações que correspondam a essa *query*. Como neste exemplo:

("Alterações climáticas" OR "Crise climática" OR "Ativismo climático" OR "Ativistas do Clima" OR "Ativistas Climáticos" OR "Greve Climática Estudantil" OR "Climáximo")

Mas também podemos recolher todas as publicações que sejam originárias de uma conta ou de um conjunto de contas existentes na rede. Como neste exemplo:

https://twitter.com/antoniocostapm
https://twitter.com/PNSpedronuno
https://twitter.com/LMontenegroPSD
https://twitter.com/AndreCVentura
https://twitter.com/ruirochaliberal
https://twitter.com/MRMortagua
https://twitter.com/InesSousaReal
https://twitter.com/ruitavares

Quadro 5.1 Regras de monitorização do SentiOne para criar um novo projeto

Regra	Descrição(*)
Query	
Keywords avançadas	Palavras-chave incluídas nos conteúdos a recolher
Autores	Identificação dos autores a monitorizar
Query avançada	*Query* usando operadores booleanos
Perfis sociais	
Páginas de Facebook	Apenas *posts* públicos de páginas de Facebook, incluindo comentários
Instagram	*Posts* e *reels* de Instagram
Hashtags de Instagram	*Hashtags* de Instagram
X (Twitter)	*Tweets*, respostas e menções
YouTube	Vídeos e *shorts*: título, descrição e comentários
TikTok	Vídeos públicos e comentários
Reddit	Posts públicos e comentários
Fonte	
Domínio	*Websites*
URL	Endereços web
País	Quando disponível na API
Google reviews	*Reviews* da Google
LinkedIn (conta própria)	*Posts* e comentários da própria conta
Língua	Quando disponível na API

(*) https://listen.help.sentione.com/docs/getting-started-1
Fonte: elaboração própria dos autores.

Por fim, também é possível fazer uma combinação das duas abordagens e criar uma *query* que seja aplicada apenas numa conta ou num conjunto de contas.

A vantagem das ferramentas abrangentes como o SentiOne é precisamente permitirem uma ampla gama de utilizações. Neste capítulo, iremos limitar-nos a usar o SentiOne para recolher dados do Twitter/X, mas usando simultaneamente uma *query* e um conjunto de contas.

Explorar os dados no SentiOne Listen

Depois de criado o projeto, os dados podem ser explorados em duas áreas do SentiOne: o separador "Mentions" e o separador "Analysis". O separador "Mentions" é onde o SentiOne nos mostra todas as publicações correspondentes à configuração (que *query* ou que canais) que configurámos na preparação do projeto – e, portanto, é onde, na maior parte dos casos, iremos recolher os nossos dados. Nesse separador, podemos ordenar todas as publicações correspondentes à *query,* à conta ou ao conjunto de contas monitorizadas pelos critérios "newest", "oldest", "influence score", "engagement rate", "likes", "shares" e "followers". Note-se que todos os

critérios de ordenação resultam de métricas existentes na própria rede e, portanto, fornecidos através da API, com exceção do "influence score", que é uma métrica desenvolvida pelo próprio SentiOne com base nos seguidores de uma conta e nas visualizações e partilhas de um *post*. As ferramentas como o SentiOne possuem frequentemente este tipo de métricas exclusivas e que podem variar de ferramenta para ferramenta.

No separador "Mentions" podemos igualmente definir o período temporal em relação ao qual pretendemos realizar a recolha de publicações. No SentiOne, os chamados "dados históricos" – ou seja, a retroatividade permitida à recolha de dados – podem variar entre 12 e 36 meses, consoante o plano subscrito, o que também poderá ser uma limitação importante à investigação. Além do separador "Mentions", cada projeto dispõe ainda do separador "Analysis", que, como o próprio nome indica, realiza um conjunto de análises pré-definidas a partir dos dados resultantes da *query* ou da lista de contas que estamos a seguir: evolução temporal, nuvem de palavras, autores mais ativos, análise de sentimento, análise demográfica, etc. Nalguns casos, os dados trabalhados para essas análises parcelares podem ser descarregados em formato CSV, XLS ou PNG, noutros caso apenas podem ser visualizados. De qualquer forma, convém recordar que os dados são os mesmos que estão agrupados no separador "Mentions", e que, por isso, é deste separador que devemos prioritariamente recolher os dados que pretendemos.

Outputs *gerados no SentiOne Listen*

No SentiOne Listen, os dados totais constantes do separador "Mentions" podem ser descarregados em formatos CSV e XLS para serem analisados e manipulados noutros programas, como o Google Sheets ou o Microsoft Excel, por exemplo.

De recordar que o SentiOne permite recolher dados de várias plataformas de redes sociais no mesmo projeto. Por isso, os *outputs* gerados, em CSV ou XLS, listados abaixo, são comuns a todas as redes (ver quadro 5.2).

Uma vez que a estrutura deste *output* é a mesma, independentemente da rede social de onde provêm os dados, isso significa que, em qualquer extração, haverá colunas que estarão vazias, correspondentes a dados que apenas existem em outras redes sociais que não aquela que estamos a explorar. Ou seja, podem existir diferentes designações para dados semelhantes provenientes de diferentes plataformas. O que significa que, como precaução metodológica, é importante conhecer o significado de cada métrica na rede social que estamos a estudar e de onde provêm os dados, independentemente da forma como são apresentados neste *output*.

Quadro 5.2 *Outputs* gerados e metadados sobre as publicações no SentiOne

Dimensão	Significado	Dimensão	Significado
ID	Identificação do *post*	Shares	Número de partilhas do *post*
Specific type	Tipo de publicação (*post*, *tweet*, vídeo, etc.)	Wow	Número de *wows* (Facebook)
Title	Título (quando existe)	Love	N.º de *loves* (Facebook)
Author	Nome do autor	Like	N.º de *likes* (Facebook)
Author ID	ID do autor	Haha	N.º de *hahas* (Facebook)
Content of posts	Conteúdo do *post*	Sad	N.º de *sads* (Facebook)
Created	Data e hora de criação do *post*	Angry	N.º de *angry* (Facebook)
Added to system	Data e hora de adição aos servidores do SentiOne	Thankful	N.º de *thankful* (Facebook)
Context	Texto adicional (quando existe)	Uniquevievs	Visualizações únicas (quando determinado)
Link to the source	*Link* para o *post* original	Fans	Número de fãs
Domain	Domínio ou rede social de origem do *post*	Facebook page category	Categoria da página de Facebook (quando aplicável)
Sentiment	Sentimento dominante (positivo, negativo ou neutro)	Retweet	Número de *retweets* (quando aplicável)
Sentiment points	Pontos atribuídos ao sentimento	Favs	Número de *favs* (quando aplicável)
Domain group	Rede social de origem	Hearts	Número de corações (quando aplicável)
Quality points	Métrica para *posts* próprios (*marketing*)	Likes	Número de *likes* (quando aplicável)
Tag	Categorização de *posts* próprios (*marketing*)	Dislikes	Número de *dislikes* (quando aplicável)
Keywords	Palavras-chave	Followers	Número de seguidores
Gender	Género do autor (quando determinado)	Geolocation	Geolocalização (quando determinada)
Project name	Projeto de origem dos dados	Language	Língua
Domain category	Categoria do domínio (*website*, portal, etc.)	Country	País (quando determinado)
Influence score	Indicador de influência (seguidores e métricas do *post*)	Rating	Classificação interna do *post* (*marketing*)
Comments	Número de comentários no *post*	Thread ID	Identificação única do *post*
Views	Número de visualizações do *post*		

Fonte: elaboração própria dos autores.

Receitas

Ponto de partida. Conteúdo de vários tweets *com as mesmas* keywords

Um dos primeiros passos numa investigação é decidir que conteúdo queremos recolher e quais as limitações que temos. Por exemplo, usando o SentiOne há limitação temporal (só podemos recolher *tweet*s até 3 anos atrás). Analisando os *outputs* passíveis de retirar do SentiOne podemos responder a diferentes questões de investigação, mas temos de decidir que *query* é a mais correta para obter o *output* pretendido: queremos todas as publicações feitas em Portugal durante um período de tempo? Queremos todas as publicações feitas durante um período de tempo que usaram uma determinada *hashtag* ou *keyword* ou *keywords*? Queremos todas as publicações feitas durante um período de tempo por uma determinada lista de utilizadores ou por um determinado utilizador único?

Depois de decidirmos o que queremos investigar podemos começar a recolher os dados. Nesta receita, vamos exemplificar como recolher todos os *tweet*s publicados nos últimos 30 dias antes da extração, mencionando uma palavra-chave, que neste caso ilustramos com o nome "Passos Coelho".

Recolher dados

Partindo do princípio que já tem acesso a uma conta na plataforma SentiOne, o procedimento é o seguinte:

1. Vá ao separador "Projects" clique no botão "Create project";
2. das várias formas de criar "Projects", escolha o modo "Advanced";
3. dê um nome ao seu projeto. Por exemplo: "Twitter_PassosCoelho";
4. dentro do separador "Keywords" escolha a opção "Advanced keywords" e escreva:
   "Passos Coelho" (é importante manter as aspas, que vão funcionar como um operador booleano que restringe a procura à expressão integral "Passos Coelho");
5. dentro do separador "Source" escolha a opção "Domains" e inscreva "twitter.com" no formulário correspondente para pesquisar a *keyword* acima, mas apenas em publicações no Twitter/X;
6. faça "Enter" e verifique, na pré-visualização de resultados, se estes correspondem ao objetivo da sua pesquisa;
7. grave o projeto e mude para o separador "Mentions", onde poderá ver todos os *tweet*s correspondentes ao projeto, ordenados por "Influence score" (mas poderão ser ordenados por outros critérios);
8. as datas pré-definidas estarão configuradas para os últimos 30 dias, mas podem ser selecionados outros períodos de tempo;
9. faça *scroll* até ao fundo da coluna da direita, onde encontra os botões para fazer o *download* dos resultados em formato CSV ou XML.

O SentiOne irá descarregar para o seu disco um ficheiro no formato solicitado, contendo um *output* semelhante ao descrito no quadro 5.2, com um *tweet* por cada linha e todas as informações sobre o *tweet*, incluindo o texto do *tweet*, assim como vários metadados sobre o mesmo. Por definição, o SentiOne permite descarregar um máximo de 10 mil registos em cada extração.

Algumas possibilidades de investigação

O *output* desta pesquisa com todas as publicações do Twitter/X com uma determinada expressão (neste exemplo, "Passos Coelho") permite várias possibilidades de investigação como, por exemplo:

- Que temas são mais comuns nas publicações?
- Que palavras são mais comuns nas publicações? – Há formas rápidas de fazer algumas análises ao conteúdo das publicações e de as visualizar, como fazer uma nuvem de palavras (*word cloud)* que apresente de forma visual as palavras mais comuns nas diferentes publicações (ficando as mais comuns maiores).
- Que temas são abordados por mais utilizadores?
- Que temas são mais abordados em conjunto?
- Que utilizadores abordam mais um determinado tema?
- Que tipo de utilizadores fizeram publicações usando esta expressão?
- Qual a evolução temporal do número de publicações com esta expressão?
- Visualização de redes:
  Usando o *output* do SentiOne com o conteúdo das publicações é possível construir uma rede de interações, mas é preciso fazer um passo intermédio. Com base no texto das publicações conseguimos perceber que contas são mencionadas por outras contas (as contas são identificadas com um "@"), quer no início da publicação – quando se trata de uma resposta, quer a meio da publicação – quando é uma menção. Com base nesta informação desenvolvemos uma matriz para criar a rede, que pode ser feita em Excel ou no table2net.[4] Esta matriz deve indicar os *nodes* (as contas) e os *edges* (a interação entre as contas).
  Através do Gephi é possível fazer diferentes análises e visualizações de redes de interação e seguidores. O Gephi (Bastian *et al.*, 2009) é uma ferramenta de *open source* utilizada para fazer análise e visualização de redes sociais (Social Network Analysis – SNA). Utilizando esta ferramenta conseguimos visualizar a rede de interação entre diferentes contas do Twitter/X e adquirir alguma informação sobre essa rede, por exemplo a centralidade de uma conta na rede ou que conta é alvo de mais interações. Para usar o Gephi será preciso fazer *download* do mesmo, de forma gratuita, em https://gephi.org/.

4 https://medialab.github.io/table2net/

- Exemplos de redes de interação (por exemplo, repostas e menções).
- Que contas interagiram em resposta a uma publicação? Um exemplo desta linha de investigação é o estudo de Bruns (2011).
- Como é que várias contas interagem com uma conta em específico? (por exemplo, com uma celebridade).
- Como é que várias contas interagem umas com as outras em relação a um tema? (neste caso "Passos Coelho").

Apesar do SentiOne não permitir retirar uma lista de seguidores, se for usada outra ferramenta para a extração de dados, depois o Gephi também permite analisar e visualizar redes de seguidores. Por exemplo, mapear uma comunidade de pessoas com um interesse em comum que podem ser identificadas de diferentes formas, por exemplo uma palavra específica na sua *bio* ou por terem usado uma determinada *hashtag* (Grandjean, 2016). Outro exemplo de linha de investigação com redes de interação e seguidores seria identificar uma lista de utilizadores com base na sua interação/relação *offline* (por exemplo, trabalharem no mesmo sítio) e analisar a rede de seguidores e interação dessa comunidade *offline* no mundo *online*, como foi feito por Ferro-Santos *et al.* (2024b) com os deputados da Assembleia da República.

Referências bibliográficas

Bastian, M., S. Heymann, e S. Jacomy (2009), "Gephi: an open source software for exploring and manipulating networks", *Proceedings of the International AAAI Conference on Web and Social Media*, 3 (1), pp. 361-362, https://doi.org/10.1609/icwsm.v3i1.13937.

Bruns, A. (2011). "How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi", *Information, Communication & Society*, 15 (9), pp. 1323-1351, https://doi.org/10.1080/1369118X.2011.635214.

Bruns, A. (2019), *Are Filter Bubbles Real*, Polity Press.

Cardoso, G., M. Paisana, e A. Pinto-Martinho (2023), *Digital News Report Portugal 2023*, OberCom – Reuters Institute for the Study of Journalism, disponível em https://obercom.pt/digital-news-report-portugal-2023/.

Castells, M. (2009), *Communication Power*, Oxford University Press.

Ferro-Santos, S., G. Cardoso, e S. Santos (2024a), "What do Portuguese MPs use Twitter for? A case study on political communication in a country with a low Twitter adoption rate", *Cuadernos.info*, 58, pp. 184-207, https://doi.org/10.7764/cdi.58.72109.

Ferro-Santos, S., G. Cardoso, e S. Santos (2024b), "Bursting the (filter) bubble: interactions of members of Parliament on Twitter", *Media e Jornalismo*, 24 (44), article e4403, https://doi.org/10.14195/2183-5462_44_3.

Gouveia, C., T. Lapa, e B.D. Fátima (2018), "Benfica vs Sporting: o *derby* visto a partir do Twitter", *Observatorio (OBS*)*, 12 (2), https://doi.org/10.15847/obsOBS12220181228.

Grandjean, M. (2016), "A social network analysis of Twitter: mapping the digital humanities community", *Cogent Arts & Humanities*, 3 (1), https://doi.org/10.1080/23311983.2016.1171458.

Habermas, J. (2022), “Reflections and hypotheses on a further structural transformation of the political public sphere", *Theory, Culture & Society*, 39 (4), DOI: 10.1177/02632764221112341.

Haman, M., e M. Skolník (2021), “Politicians on social media. The online database of members of national parliaments on Twitter", *Profesional de la información*, 30 (2), https://doi.org/10.3145/epi.2021.mar.17.

Highfield, T., S. Harrington, e A. Bruns (2013), “Twitter as a technology for audiencing and fandom: the #Eurovision phenomenon", *Information, Communication & Society*, 16 (3), pp. 315-339, https://doi.org/10.1080/1369118X.2012.756053.

Kemp, S. (2024a), “Digital 2024: Global Overview Report”, disponível em https://datareportal.com/reports/digital-2024-global-overview-report.

Kemp, S. (2024b), “Digital 2024: Portugal”, disponível em https://datareportal.com/reports/digital-2024-portugal.

Moreno, J. C., S. Ferro-Santos, e R. Sepúlveda (2024), “Taking Europe home: how political agents stand out in their approach to Europe on social media", *Observatorio (OBS*)*, 17 (5).

Murphy, D. (2013), *Twitter: Social Communication in the Twitter Age*, Polity Press.

Pariser, E. (2011), *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.

Sunstein, C. R. (2006), *Infotopia – How Many Minds Produce Knowledge*, Oxford University Press.

Capítulo 6

Facebook
Posts, interações e comentários

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Introdução

O Facebook foi criado em fevereiro de 2004, por Mark Zuckerberg, então estudante na universidade de Harvard, nos EUA, como uma versão digital dos conhecidos "livros de curso", em que os alunos incluíam, para a posteridade, uma foto e algumas informações sobre os seus gostos e preferências. O objetivo era estabelecer conexões entre os indivíduos com conta no *website*, para que pudessem comunicar uns com os outros (Coutinho, 2018). Inicialmente restrito aos alunos de Harvard e depois estendido às universidades de Columbia, Stanford e Yale, o Facebook passou a ser aberto ao público em geral em 2006, permitindo que qualquer pessoa com mais de 13 anos e um *email* válido criasse uma conta na plataforma.

De então para cá, a plataforma Facebook tornou-se cada vez mais sofisticada e massificada. O crescimento do número de utilizadores foi constante e vertiginoso entre 2006 e 2020 (menos vertiginoso depois) e chegou a mais de três mil milhões de utilizadores ativos mensais em 2023, cerca de 37,7% da população mundial (Kemp, 2024a). Em Portugal, o Facebook tem atualmente 5,95 milhões de utilizadores ativos mensalmente, o que corresponde a 58,1% da população (Kemp, 2024b).

Além de ser a rede social com maior número de utilizadores ativos, no mundo e em Portugal, o Facebook tornou-se também a principal fonte de tráfego para os *website*s dos meios de comunicação social e a principal fonte de notícias para os utilizadores de *internet* (Newman *et al.*, 2023), tanto globalmente como em Portugal. No nosso país, 40% por cento dos inquiridos afirmaram receber as suas notícias através do Facebook (Cardoso *et al.*, 2023).

Paralelamente, ao longo dos anos, o Facebook desenvolveu todo um ecossistema de relacionamentos de *marketing* entre a plataforma, os utilizadores e as marcas, com contribuição dos desenvolvedores e criadores de conteúdos (Helmond *et al.*, 2019). Essa sofisticação da plataforma converte-a num exemplo paradigmático daquilo que foi designado por van Dijck (2018) como a "Sociedade das plataformas", uma evolução moderna da "Sociedade em rede" de Castells (2009).

Por todas as razões acima, o Facebook converteu-se, ao longo dos anos, numa das plataformas de eleição para o estudo dos fluxos de comunicação através dos

métodos digitais. No entanto, a relação dos investigadores com a plataforma nem sempre foi pacífica ou fácil (Perriam *et al*., 2020). Depois dos primórdios muito dependentes do *scraping*, o Facebook passou por duas fases principais no que concerne ao acesso dos investigadores aos dados gerados pelos utilizadores da plataforma, ambas baseadas em API. Entre 2010 e 2018, a principal ferramenta de recolha de dados do Facebook foi o Netvizz, um *software* desenvolvido por Bernhard Rieder, investigador da Universidade de Amesterdão, que funcionava através da API autorizada pela plataforma (Rieder, 2013). Uma busca no Google Scholar devolve mais de 2390 artigos científicos que mencionam a ferramenta ou a usam na investigação de conteúdos no Facebook. Em 2018, na sequência do escândalo de exploração abusiva de dados por parte da Cambridge Analytica, o Facebook restringiu as possibilidades de acesso à API e o Netvizz deixou de funcionar.[1]

Mais tarde, em 2019, a empresa começou de novo a facultar o acesso dos investigadores aos seus dados através de uma ferramenta chamada CrowdTangle, um *software* de monitorização de redes sociais que o Facebook tinha comprado em 2016 e cujo acesso, gratuito, tinha até então sido limitado às empresas de *media* presentes na plataforma. O CrowdTangle recolhia dados do Facebook, do Instagram e do Reddit.[2] Do mesmo modo, se procurarmos no Google Scholar referências ao CrowdTangle, encontramos 2600 artigos científicos que se referem a ele ou o usam nos métodos de investigação. Mas também o CrowdTangle acabou por ser afetado pela pressão da regulação sobre proteção de dados e a Meta, proprietária do Facebook e do Instagram, anunciou, em março de 2024, que o CrowdTangle deixaria de funcionar em agosto do mesmo ano.[3] Ao mesmo tempo, a empresa anunciou o desenvolvimento de um acesso especificamente destinado aos investigadores académicos, a que chamou Content Library API.[4] Esse acesso está, na altura da escrita deste capítulo, em fase de implementação, mas ainda sem reflexos claros na melhoria da facilidade de acesso dos investigadores aos dados do Facebook e Instagram.

Como forma de contornar as limitações de acesso ao Facebook e a outras plataformas de redes sociais, muitos investigadores recorrem a ferramentas pagas de acesso aos dados, desenvolvidas com o objetivo de monitorizar a presença de marcas nas redes sociais e associadas às estratégias de *marketing*. Ferramentas como o Brandwatch, o Brand24, o Talkwalker, o Newswhip ou o SentiOne.[5] Por causa do objetivo comercial que está na sua base, os dados que este tipo de ferramentas disponibiliza e a forma como os apresentam não corresponde, por vezes, às necessidades académicas. Mas em muitos casos permitem um acesso aos dados que não estaria disponível de outra forma, sendo por isso uma alternativa válida para a

1 https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought
2 https://help.crowdtangle.com/en/articles/4201940-about-us
3 https://help.crowdtangle.com/en/articles/9014544-important-update-to-crowdtangle- march-2024
4 https://transparency.meta.com/researchtools/meta-content-library/
5 https://www.brandwatch.com/, https://brand24.com/, https://www.talkwalker.com/, https://www.newswhip.com/, https://sentione.com/features/listen.

investigação académica. Neste capítulo iremos explorar como recolher dados do Facebook usando precisamente o SentiOne.

Sendo hoje em dia uma plataforma com um grande número de utilizadores e uma grande diversidade de funcionalidades, o Facebook oferece a quem tem conta três tipos de presença na plataforma. O perfil é entendido como uma presença pessoal e não pode ser objeto de recolha de dados (a API não fornece dados sobre as publicações em perfis pessoais). A página é uma presença sempre pública, destinada a marcas, entidades, instituições, autores ou figuras públicas (embora qualquer pessoa possa criar uma página) e pode ser gerida por várias pessoas. Sendo sempre públicos, os dados referentes a uma página (métricas e conteúdos, por exemplo) são o objeto de pesquisa mais frequente no Facebook. Os grupos, pensados para agregarem comunidades à volta de interesses em comum, podem ser públicos, privados ou secretos. Para a pesquisa académica, apenas os grupos públicos podem ser objeto de investigação, uma vez que, mais uma vez, são os únicos coletados pela API.

Recolha de dados

Ferramenta: SentiOne Listen

Tal como referido no capítulo dedicado ao Twitter/X, o SentiOne Listen é um *software* pago, fornecido como plataforma *online*, destinado a monitorizar a presença de uma marca nas redes sociais. E, como muitos outros *softwares* semelhantes, permite também a recolha de dados para fins de investigação académica. Iremos usá-lo aqui para recolher dados a partir do Facebook.

O funcionamento do SentiOne está descrito em detalhe no capítulo anterior, mas, no contexto da utilização que vamos aqui simular, é útil recordar os seguintes aspetos. Em primeiro lugar, o SentiOne funciona por projetos. É através da configuração de cada projeto que definimos que redes sociais vão ser monitorizadas, que palavras-chave ou que contas vão ser analisadas e que filtros vão ser aplicados a essa pesquisa. Em segundo lugar, depois de configurado o projeto, os resultados podem ser analisados em dois separadores: o "Mentions" e o "Analysis". O separador "Mentions" exibe todas as publicações correspondentes à *query* e permite, entre outras coisas, filtrar esses resultados por data e por vários outros critérios, assim como fazer o seu *download* em CSV ou XLS. O separador "Analysis" usa vários *widgets* para explorar diferentes visualizações dos mesmos dados, a maioria dos quais também podem ser descarregados em CSV, XLS ou em formatos de imagem. Em terceiro lugar, convém recordar que o *output* CSV ou XLS que nos dá a totalidade dos dados recolhidos a partir do separador "Mentions" é igual, qualquer que seja a rede social analisada, o que significa que, dependendo da plataforma, algumas colunas serão preenchidas enquanto outras não.

Receitas

Neste capítulo iremos explorar duas receitas para recolher dados do Facebook, uma centrada num conjunto de páginas e outra centrada na exploração dos comentários a um *post* em particular. De sublinhar, no entanto, que, dada a versatilidade de uma ferramenta como o SentiOne, seria possível igualmente pesquisar por *keywords*, em grupos em vez de páginas e com uma grande variedade de filtros.

Comparar o engagement *de diferentes páginas de Facebook*

Nesta primeira receita, o objetivo da investigação é comparar o desempenho de várias páginas de Facebook, neste caso páginas de meios de comunicação social portugueses, em termos do *engagement* (ou interação) que as suas publicações obtêm num determinado período de tempo.

Obviamente, a primeira decisão metodológica importante é definir quais as páginas a serem analisadas. O primeiro passo, prévio ao SentiOne, é justamente recolher o URL completo das páginas de Facebook a incluir (exemplo: https://www.facebook.com/Publico/). É aconselhável fazer essa pesquisa previamente para verificar a existência e autenticidade das páginas a analisar e, por exemplo, registar o URL (e outros dados que considere relevantes) dessas páginas num Excel ou CSV. Depois de definida essa lista, o processo começa com a criação de um projeto.

Criação de um projeto no SentiOne

— Faça *login* no SentiOne;
— clique no separador "Projects" e depois clique em "Create project";
— escolha a opção "Social profiles" e, das escolhas disponíveis, selecione "Facebook";
— copie o URL da página (por exemplo: https://www.facebook.com/Publico/), cole esse URL no formulário correspondente e espere uns segundos para que o SentiOne consiga localizar a página que procura. Se isso não acontecer, pode significar que o SentiOne não está a ser capaz de localizar a página. Nesse caso deve verificar o URL e, se isso não resolver, deve reportar a falha ao suporte do SentiOne, uma vez que a página poderá não estar rastreada no sistema;
— se a página estiver rastreada pelo SentiOne, o respetivo ícone, o nome e domínio irão aparecer. Quando clica nesse nome, o SentiOne regista a página e mostra, no ecrã à direita, uma previsão dos conteúdos correspondentes;
— nessa previsão poderá reparar que existem *posts* do *Público*, mas também comentários aos *posts* (e eventualmente outros tipos de conteúdos da página);
— como aquilo que pretendemos nesta receita são apenas os *posts* da própria página, deve clicar no "Show mention type" que aparece por baixo da identificação da página. Mostra as várias opções ("Private messages", "Posts", "Comments", "Shares", "Reviews" e "Mentions"). Como, neste caso, aquilo

que nos interessa são apenas os *posts* da página, desmarque todas as outras opções exceto essa e repare que a previsão de resultados vai começar a mostrar apenas *posts* do *Público*, que é aquilo que pretendemos;
— se estiver confortável com o resultado, repita o processo para todas as outras páginas, selecionado a opção "Add another social account +" e seguindo todos os passos anteriores (não esquecendo de restringir a pesquisa aos *posts*). Repare que, por cada nova página que adiciona, os respetivos *posts* vão sendo adicionados à pré-visualização;
— quando tiver adicionado todas as páginas que pretende incluir na análise e seguido todos os procedimentos indicados, clique no botão "Continue";
— dê um nome ao projeto e clique em "Save project". Isso irá guardar o projeto com esse nome e irá transferir o utilizador imediatamente para o separador "Mentions", onde poderá analisar os resultados.

Análise de resultados e download *de dados no separador "Mentions"*

No SentiOne, a análise de resultados pode ser feita de várias formas. Tal como foi dito no ponto 1, a maior parte dos dados é extraída a partir do separador "Mentions", mas também é possível recolher dados (e visualizações) do separador Analysis. Nesta receita iremos descrever os passos para fazer ambas as coisas.

Uma vez que aquilo que nos interessa é comparar o *engagement* (ou interação) das publicações das páginas escolhidas, é isso que iremos fazer de seguida.

— Uma vez guardado o projeto, o separador "Mentions" irá mostrar-lhe todos os *posts* publicados pelas páginas monitorizadas nos últimos 30 dias, ordenados dos mais recentes aos mais antigos. No topo da página, o SentiOne informa o total de *posts* recolhidos para a filtragem atual. Se alterar a filtragem, esse total também vai mudar;
— primeiro definimos as datas. No canto superior direito podemos escolher um dos períodos pré-definidos ou customizar um período temporal. Neste exemplo, vamos escolher os últimos 7 dias. Repare que a ativação desse filtro temporal (como de outros filtros que aplique) aparece mencionada no topo do monitor;
— os *posts* aparecem ordenados por "Newest", ou seja, dos que têm data de publicação mais recente para aqueles que têm data de publicação mais antiga, dentro do limite temporal definido. Uma vez que queremos analisar as publicações com mais *engagemen*t (mais interações) vamos escolher a ordenação "Engagement rate", que pondera o número de reações, comentários e partilhas pelo número de seguidores da página.[6] Isto permite-nos visualizar os *posts* com melhor taxa de interação antes de descarregar os dados;
— na coluna da direita encontramos vários filtros que podem ser aplicados à recolha de *posts* feita, nomeadamente por autor (neste caso cada página),

6 https://listen.help.sentione.com/docs/engagement-rate

palavra-chave, língua ou país. Neste exemplo não vamos ativar nenhum desses filtros. Assim, no final desta coluna encontramos dois botões de *output* dos dados, em CSV ou em XLS. Basta clicar num desses botões para o registo de todos os *posts* recolhidos ser descarregado para o seu disco, no formato pretendido;

— nesse ficheiro CSV ou XLS irá encontrar um *post* em cada linha e os atributos de cada *post* (data de publicação, conteúdo, métricas, etc.) em cada coluna. Tal como foi exportado, o ficheiro terá a lista ordenada de publicações com melhor taxa de interação.

Obviamente, uma vez exportado este ficheiro — com todas as publicações passíveis de recolha tendo em conta a filtragem — ele pode ser manipulado de várias formas para analisar os dados. Por exemplo, poderá ser interessante isolar as publicações de cada um dos meios de comunicação social e somar as interações para ver qual deles conseguiu obter mais interações no período considerado. Ou cruzar esse dado com o número de publicações realizadas para saber qual deles registou mais interações, em média, por cada *post* publicado.

Análise de resultados e download *de dados no separador "Analysis"*

Existe outra área do SentiOne — o separador "Analysis" — onde é possível usar vários *widgets*, cada um com algumas destas configurações já preparadas. Vejamos um exemplo:

— dentro do mesmo projeto, clique no separador "Analysis" no topo da página ou no menu da esquerda;

— o SentiOne irá então mostrar-nos um conjunto de *widgets* de análise dos dados, de análises temporais de *posts* e interações a análises de sentimento, passando por repartição de género ou palavras ou *hashtag*s mais frequentes. Tanto o conjunto dos *widgets* como cada um deles individualmente podem ser configurados em função do tempo e de vários outros critérios;

— desça na página até encontrar um *widget* chamado "Top authors". Esse *widget*, que é aquele que nos interessa, faz uma hierarquização das páginas (a que chama "authors") somando os *posts* publicados e as interações obtidas;

— todos os *widgets* do separador "Analysis" estão por definição configurados para devolver resultados dos últimos 30 dias, mas neste exemplo interessam-nos os últimos 7 dias. Para isso, clique no botão "Edit settings", no canto superior direito da página, e escolha os últimos 7 dias e carregue no botão "Apply";

— de seguida, no botão ao lado, terá as opções de *download* dos dados assim configurados, em CSV ou XLS. Mais uma vez, o SentiOne irá descarregar para o seu disco um ficheiro, no formato solicitado, contendo os resultados agregados por cada página (ou meio de comunicação social), permitindo-lhe analisá-los comparativamente quanto ao número de publicações realizadas e número de interações obtidas.

Ou seja, usando estes dois *outputs* do SentiOne — todos os *posts* do separador "Mentions" e os resultados agregados por página do separador "Analysis" — poderá ter uma noção de quais as publicações das páginas analisadas que geram mais interações e quais dessas páginas obtêm um melhor resultado agregado em termos de interações (*engagement*) no período considerado, respondendo assim ao que era desejado.

Outras possibilidades de investigação

— Os filtros existentes no separador "Mentions" permitem uma grande variedade de utilizações. Por exemplo, a aplicação de filtros "Author" (por exemplo, "Público" para ver apenas os conteúdos da página https://www.facebook.com/Publico/) permite analisar e descarregar apenas os *posts* de cada uma das páginas. Se quisermos fazer uma análise individual do conteúdo de cada página, esta é uma possibilidade;
— também pode ser interessante explorar a utilização de certas palavras-chave nas publicações de cada página ou nas publicações do conjunto das páginas monitorizadas. Isso pode ser feito usando o filtro "Included words";
— no separador "Analysis" também temos um *widget* para as palavras-chave mais usadas no projeto, assim como outro para as *hashtag*s mais usadas. Qualquer deles pode dar outras possibilidades interessantes de investigação para o mesmo projeto, por exemplo.

Analisar os comentários a um post *de Facebook*

Nesta segunda receita para o Facebook, exemplificamos uma possibilidade de investigação, possibilitada pelo SentiOne, que pode proporcionar resultados muito interessantes: a análise dos comentários a um *post*. É de recordar que a API do Facebook faculta acesso aos *posts* das páginas (que são sempre públicas) e, dependendo da ferramenta, também aos comentários. No entanto, isso não acontece em todas as ferramentas (o CrowdTangle, por exemplo, não proporcionava acesso aos comentários). A razão dessa limitação prende-se com o facto de os comentários, feitos por perfis do Facebook (é importante distinguir entre perfis e páginas de Facebook) poderem ser considerados privados mesmo que sejam realizados em páginas públicas. Outras ferramentas, como o SentiOne, proporcionam acesso ao conteúdo dos comentários, mas omitem a identificação da autoria dos mesmos. Como veremos, quando trabalhamos com comentários no SentiOne, esses podem ser apresentados, mas o autor do comentário não é identificado, por esta razão.

No SentiOne, os comentários são parte integrante do acervo que é possível pesquisar no Facebook. Como vimos na configuração do SentiOne para a receita anterior, quer a nossa *query* seja uma página, um conjunto de páginas, uma palavra-chave ou um conjunto de palavras-chave, os comentários fazem parte do conteúdo que é possível recolher. O que significa que, se recolhermos o conteúdo de uma página, os comentários poderão fazer parte dessa recolha, se assim o desejarmos. E, do mesmo modo, se pesquisarmos por uma palavra-chave, essa pesquisa

pode recolher igualmente os comentários que usam essa palavra, mesmo que o *post* que comentam não o faça.

Para esta receita, vamos adotar uma abordagem diferente dos comentários, partindo, não de uma página ou de uma *keyword*, mas de um *post* específico. Ou seja, o objetivo da investigação, neste exemplo, é partir de um *post* específico e fazer a análise dos comentários ao mesmo. Obviamente, a primeira preocupação metodológica, anterior à implementação da pesquisa, é definir o critério para a escolha do *post* (ou *posts*) a analisar.

Criação do projeto e extração dos dados — Opção A

Para recolher os comentários de um *post* específico de Facebook, podemos usar mais do que um método. O primeiro recorre ao registo da página para recolher *posts* e comentários, enquanto o segundo usa o modo "Advanced". O primeiro desses procedimentos é o seguinte:

— faça *login* no SentiOne;
— dentro do separador "Projects", clique em "Create project" e depois escolha a opção "Social profiles";
— no separador para o Facebook, insira o URL da página de Facebook da qual pretende retirar o *post*. Neste exemplo, será: https://www.facebook.com/jornalnoticias. Espere que o SentiOne localize essa página e depois clique para a registar no projeto;
— no "Mention type", por baixo do nome da página, ative o *dropdown* e marque apenas as opções "Posts" e "Comments";
— de seguida clique em "Continue", dê um nome ao projeto e guarde-o, o que o levará para o separador "Mentions"
— no separador "Mentions", use o módulo de data no canto superior direito para escolher para início do período de pesquisa o dia de publicação do *post* que pretende analisar. Deve deixar a data de final do período suficientemente distante dessa data (por exemplo, uma semana ou um mês);
— de seguida, do *dropdown* à direita, ordene os *posts* por "Oldest";
— a seguir, use os botões CSV ou XLS para descarregar os resultados para o seu disco. O que estará a descarregar é o conjunto de *posts* publicados entre as datas selecionadas, ordenados por data de publicação (válida para *posts* e comentários);
— abra o CSV ou XLS. Repare que a coluna B — "Specific type", indica se o registo em cada linha é um *post* ou um comentário. Mas vamos usar o URL como caminho para identificar o *post* que queremos e os respetivos comentários.
— ordene o ficheiro CSV ou XLS pela coluna J — "Link to the source". Esta coluna contém o *link* único de cada *post* e de cada comentário;
— a seguir identifique o ID específico do *post* que pretende. Neste *post* — https://www.facebook.com/100064621156693/posts/844452984385399 — o ID da página é 100064621156693 e o ID do post é 844452984385399;
— procure este *post* no CSV ou XLS. Vai reparar que, por causa da ordenação pelo *link*, o *post* estará seguido de todos os seus comentários. Os comentários

terão um URL como este: https://www.facebook.com/610267314470635/posts/844452984385399?comment_id=954926279436410, em que 954926279436410 é o ID individual do comentário;
— no CSV ou XLS, elimine todos os registos anteriores ao *post* que pretende e todos aqueles a seguir ao último comentário a esse *post*;
— o CSV ou XLS que fica terá então apenas o *post* pretendido e todos os seus comentários.

Criação do projeto e extração dos dados — Opção B

O segundo método de recolher comentários a um *post* de Facebook usa o modo "Advanced" e envolve o seguinte procedimento:

— faça *login* no SentiOne;
— dentro do separador "Projects", clique em "Create project" e depois escolha a opção "Advanced";
— dê um nome ao projeto e, dentro da opção "Keywords", escolha a terceira alternativa: "Advanced query". Isto irá abrir uma janela para inserção da *query*. Neste caso, esta será uma *query* específica apontando para o *post* em relação ao qual queremos recolher os comentários;
— tomemos como exemplo este *post*: https://www.facebook.com/588374336720321/posts/872710591620026, um *post* da página do *Jornal de Notícias* com o título "Migrações: Hungria vai deixar de aplicar regras de asilo da UE" e com 60 comentários;
— na janela de inserção de *query* insira a seguinte *query*: (*_all*: "Migrações: Hungria vai deixar de aplicar regras de asilo da UE") *AND sourceURLdomain*: www.facebook.com. A componente *AND* é um operador booleano básico e as componentes *_all* e *sourceURLdomain* são operadores específicos do SentiOne. É importante notar que, muitas vezes, as ferramentas como o SentiOne têm formas de pesquisar os dados que são específicas e que requerem aprendizagem antes de poderem ser utilizadas. É o caso destes operadores;
— verifique os resultados obtidos na pré-visualização e, se estiverem corretos, guarde o projeto;
— no separador "Mentions", selecione a opção de ordenação "Oldest" e verifique se aparece primeiro o *post* e em seguida todos os seus comentários;
— se sim, então use os botões CSV ou XLS para descarregar os resultados para o seu disco.

Tal como foi dito anteriormente, quer use um método ou outro para recolher os comentários a um *post* de Facebook, poderá confirmar que nenhum dos dois *outputs* terá a identificação de quem comentou (coluna D), a não ser nos casos (raros) em que o comentário é feito por uma página. Mas terá todas as restantes informações sobre o conteúdo — neste caso, comentários — incluindo as métricas de popularidade.

Outras possibilidades de investigação

— Considerando que a extração dos comentários a um *post* inclui as métricas desses mesmos comentários, torna-se possível, por exemplo, ordenar os comentários por uma das métricas de popularidade (ou por uma combinação de várias delas, como as interações) e analisar quais os comentários mais populares entre os utilizadores que interagiram com o *post* ou com os comentários.
— Uma vez que a extração realizada incluiu igualmente o conteúdo integral de cada comentário, uma possibilidade adicional é fazer uma análise desse conteúdo, usando outras ferramentas específicas para esse fim.

Referências bibliográficas

Cardoso, G., M. Paisana, e A. Pinto-Martinho (2023), *Digital News Report Portugal 2023*, OberCom — Reuters Institute for the Study of Journalism, disponível em https://obercom.pt/digital-news-report-portugal-2023/.

Castells, M. (2009), *The Rise of the Network Society*, Wiley-Blackwell.

Coutinho, V. (2018), *The Social Book — Tudo o Que Precisa de Saber sobre o Facebook*, Leya.

Helmond, A., D. B. Nieborg, e F.N. van der Vlist (2019), "Facebook's evolution: development of a platform-as-infrastructure", *Internet Histories*, 3 (2), pp. 123-146, https://doi.org/10.1080/24701475.2019.1593667.

Kemp, S. (2024a), "Digital 2024: Global Overview Report", disponível em https://datareportal.com/reports/digital-2024-global-overview-report.

Kemp, S. (2024b), "Digital 2024: Portugal", disponível em https://datareportal.com/reports/digital-2024-portugal.

Newman, N., R. Fletcher, K. Eddy, C.T. Robertson, R.K. Nielsen (2023), *Digital News Report 2023*, Reuters Institute for the Study of Journalism, disponível em https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

Perriam, J., A. Birkbak, e A. Freeman (2020), "Digital methods in a post-API environment", *International Journal of Social Research Methodology*, 23 (3), pp. 277-290, https://doi.org/10.1080/13645579.2019.1682840.

Rieder, B. (2013), "Studying Facebook via data extraction: the Netvizz application", *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 346-355.

van Dijck, J., T. Poell, e M. De Waal (2018), *The Platform Society: Public Values in a Connective World*, Oxford University Press.

Capítulo 7

TikTok

Algoritmo, conteúdo e interação

Inês Narciso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

O TikTok é uma aplicação de redes sociais, desenvolvida pela ByteDance, que foi lançada internacionalmente em 2018. A plataforma cresceu rapidamente, diferenciando-se da concorrência pelo seu foco em vídeos de curta duração. A variedade de elementos multimédia, como música, filtros e texto, que a aplicação integra, permite aos utilizadores criar conteúdos mais distintivos, envolventes e criativos. Embora utilize o formato de vídeo, à semelhança de plataformas como o YouTube, o TikTok distingue-se pela sua ênfase em conteúdos breves e frequentemente virais. Este formato foi replicado pelo Instagram com os *reels*, mas a eficácia algorítmica do TikTok promoveu uma cultura única de tendências e desafios ainda sem paralelo noutras plataformas.

O TikTok posiciona-se como uma plataforma orientada para a comunidade, incentivando a interação entre os utilizadores através de funcionalidades como gostos, comentários, partilhas e a possibilidade de seguir criadores de conteúdos (Anderson e Rainie, 2020). O registo na plataforma é necessário para a criação de conteúdos e a participação ativa, mas não para a visualização, tornando os seus conteúdos mais acessíveis relativamente a outras plataformas. Contrariamente ao YouTube ou Instagram, o TikTok também permite fazer o *download* dos vídeos, com som, deixando uma marca de água que contribui para a divulgação da plataforma. A base de utilizadores desta plataforma tem vindo a expandir-se, atingindo uma estimativa de 1,5 mil milhões de utilizadores mensais ativos em 2023, com expectativas de crescimento para 1,8 mil milhões até ao final de 2024 (Iqbal, 2024). O TikTok também conta com um envolvimento dos utilizadores acima da média: 55 minutos diários. Em 2023, o público-alvo do TikTok tornou-se mais heterogéneo: visto inicialmente como plataforma do público jovem, o TikTok tem vindo a crescer sobretudo entre adultos, e, entre estes, sobretudo mulheres (Ceci, 2024).

O conteúdo do TikTok é bastante variado, abrangendo géneros como o entretenimento, a educação e o estilo de vida, reflexo da diversificação da sua base de utilizadores. A plataforma é particularmente conhecida pela sua janela "For You" que, orientada por algoritmos, faz a curadoria de conteúdo com base nas interações do utilizador, nas suas preferências de visualização e padrões de interação, oferecendo *feeds* de conteúdo personalizados (Rogers, 2024; Zhao *et al.*, 2021). Este sistema de recomendação, embora seja fundamental para a popularidade da aplicação,

também suscitou discussões sobre a utilização de dados dos utilizadores e preocupações com a sua privacidade (Trifiro, 2023).

Comercialmente, o TikTok capitalizou a sua capacidade de gerar interações e reter utilizadores na plataforma, oferecendo vários formatos de publicidade, e fazendo parcerias com criadores para colaborações com marcas. Surgiu, então, o conceito de *tiktoker*: indivíduos que capitalizam o seu conteúdo, fidelizando um grande número de seguidores, ecoando a dinâmica de influenciadores observada noutras plataformas (Bishop, 2020).

Do ponto de vista académico, o TikTok apresenta um terreno fértil para a investigação, oferecendo perspetivas sobre a cultura digital contemporânea, a viralidade dos conteúdos e o impacto da curadoria algorítmica. Várias investigações têm-se centrado em aspetos como as implicações sociológicas das tendências do TikTok, o papel da plataforma nos ecossistemas digitais globais, ou os efeitos psicológicos do consumo de conteúdos crescentemente curados à nossa medida (McCashin e Murphy, 2023). Assim, o TikTok, como fenómeno cultural e social cada vez mais relevante, constitui também um tema de crescente interesse académico, no qual o recurso aos métodos digitais é relevante.

Recolha de dados

Ferramenta: API do TikTok

Para aceder à API do TikTok enquanto investigador é necessário cumprir vários requisitos definidos pela plataforma: apresentar comprovativos de formação académica de experiência na área de investigação relativa ao tema que se pretende explorar, bem como de filiação a uma instituição académica sem fins lucrativos nos EUA ou na Europa. O tema da investigação deve destinar-se a fins não comerciais, e não devem existir conflitos de interesses, como, por exemplo, um trabalhador-estudante que deseja fazer um estudo sobre uma tendência ou marca ligada à sua entidade empregadora.

Uma vez autorizado o acesso à API, é possível extrair vários dados, mas sempre recorrendo a linguagem Python, porque não existe nenhuma versão amiga do utilizador, pelo que pode ser necessário pedir apoio técnico para a seleção e extração de dados. A API permite a recolha de dados sobre os vídeos e utilizadores, desde que ligados a contas públicas detidas por maiores de 18 anos. Sobre os vídeos, a API permite recolha de métricas de interação e visualização, comentários, data e hora e região associados a vídeos; relativamente aos utilizadores, a API permite aceder a listas de seguidores e contas seguidas, métricas do seu conteúdo, *bio*, descrição, e *links*, bem como saber se a conta é verificada.

É importante destacar que existem várias ferramentas pagas com acesso à API do TikTok que permitem a pesquisa e extração de todos estes dados. Centros de investigação como o CIES do Iscte-IUL, através do MediaLab, têm acesso a este tipo de ferramentas, por exemplo, o SentiOne. O SentiOne é uma ferramenta de escuta social e acompanhamento de marcas que permite a pesquisa de todas as métricas acima

referidas, por palavras-chave, *hashtag*s e por nome de utilizador, para algumas plataformas, nomeadamente o TikTok. O SentiOne dispõe ainda de filtros temporais, de interação, entre outros, sendo a extração feita em formato CSV e XLS.

Ferramenta: Zeeschuimer

O Zeeschuimer é uma extensão gratuita para navegadores *web* que permite aos investigadores recolher dados das redes sociais. Funciona através da análise de tráfego de um utilizador enquanto este navega, contornando assim as limitações de recolha impostas pelas plataformas. Foi desenvolvida por Stijn Peeters, professor da Universidade de Amsterdão e colaborador do Digital Methods Initiative, e está licenciada ao abrigo da Mozilla Public License 2.0. Para o utilizar, é necessário instalar a respetiva extensão do *browser*, ativar a opção de recolha de dados, navegar na plataforma e, no final, exportar os elementos recolhidos.

O Zeeschuimer permite reunir informações exclusivas dos vídeos do TikTok que podem não ser obtidas diretamente através da API, ou de ferramentas que usam a API. O Zeeschuimer consegue, por exemplo, identificar se um vídeo é um anúncio, parte de um dueto, ou um desafio. Por outro lado, o Zeeschuimer restringe as possibilidades de pesquisa no TikTok, sendo apenas possível pesquisar uma coisa de cada vez, e com uma restrição temporal, só sendo possível aceder a conteúdo atualmente disponível na plataforma.

O Zeeschuimer é uma ferramenta gratuita e de fácil utilização, particularmente útil para a análise de redes sociais sob a perspetiva do utilizador. No caso do TikTok, a extensão permite navegar na plataforma e exportar posteriormente uma lista de todo o conteúdo que vimos durante o período que estivemos a navegar. Partindo de um perfil recém-criado, e sem muitos dados, o Digital Methods Initiative procurou compreender de que forma o algoritmo conduziu o novo utilizador a teorias conspiratórias, após pesquisar por algumas palavras-chave (Aguillar *et al.*, 2023). A ferramenta é compatível com várias plataformas e permite a exportação de dados como JSON, ou para a plataforma 4CAT, também do Digital Methods Initiative, para análise posterior.

A ferramenta permite ainda recolher dados do LinkedIn, Instagram, 9GAG, LinkedIn, Instagram, 9GAG, Imgur, Twitter/X, Gab e Douyinn, tendo, por isso, uma utilidade significativa para investigadores que pretendam utilizar métodos digitais.

Aceder e explorar a ferramenta:

1. Ir a https://github.com/digitalmethodsinitiative/zeeschuimer através de um navegador Firefox.[1]
2. Ir ao separador *releases*, e selecionar o ficheiro XPI, que será automaticamente adicionado à sua biblioteca de extensões do Firefox.[2] Na biblioteca

1 À data da publicação deste manual, a extensão não funciona em outros navegadores.

Quadro 7.1 Menu de opções oferecido pela versão 1.9 do Zeeschuimer

Recolha	Plataforma	Itens recolhidos	Exportação		
ON / OFF	TikTok	0	Apagar	JSON	4CAT
ON / OFF	Instagram	0	Apagar	JSON	4CAT
ON / OFF	LinkedIn	0	Apagar	JSON	4CAT
ON / OFF	9GAG	0	Apagar	JSON	4CAT
ON / OFF	Imgur	0	Apagar	JSON	4CAT
ON / OFF	Twitter	0	Apagar	JSON	4CAT
ON / OFF	Douyin	0	Apagar	JSON	4CAT

Fonte: elaboração própria da autora.

de extensões do Firefox, representado na peça de puzzle ou no menu sanduíche no canto superior direito do seu navegador, clique na extensão Zeeschuimer. Automaticamente, abre-se um novo separador, com um menu para as diversas opções que a extensão oferece, replicadas no quadro 7.1.[3]

Outputs gerados

Para a recolha de dados do TikTok, deve-se ativar essa opção colocando o botão no On e iniciar a navegação na versão *desktop* do domínio tiktok.com. Uma vez ativa, a ferramenta Zeeschuimer irá recolher elementos sobre os resultados que forem sendo vistos pelo utilizador na sua página "For You", ou em sequência de uma pesquisa. Ao terminar o processo, a ferramenta indica quantos elementos foram recolhidos e permite a exportação em formato JSON ou a sua integração na plataforma analítica 4CAT, tal como surge no quadro 7.1.

O Zeeschuimer recolhe, para cada vídeo visualizado (mesmo que não na totalidade), os elementos indicados no quadro 7.2. Destacam-se alguns elementos de maior interesse para os investigadores, como as métricas de interação com o conteúdo, textos associados, e dados sobre a data, hora, música e autor do vídeo.

Destaca-se que os valores apresentados pelo Zeeschuimer correspondem às métricas referentes à data da visualização do conteúdo, devendo, por isso, o momento da extração também ser considerado nas avaliações metodológicas.

2 https://github.com/digitalmethodsinitiative/zeeschuimer/releases
3 Versão 1.9, a mais recente disponível à data da publicação deste manual.

Quadro 7.2 Elementos sobre vídeos do TikTok recolhidos pelo Zeeschuimer

Tipo de elemento	Descrição
ID	Número identificativo do conteúdo
Thread ID	Número identificativo do fio do conteúdo
Author	Nome de utilizador do autor
Author_full	Nome do autor
Author_followers	Número de seguidores do autor
Author_likes	Número total de gostos do autor
Author_videos	Número total de vídeos do autor
Author_avatar	Imagem de perfil do autor (não acessível via *link*)
Body	Texto da publicação
Timestamp	Data da publicação apresentada como dia da semana (3 letras) — Mês (3 letras) dia e ano.(*)
Unix_timestamp	Data e hora da publicação em formato unix.(**)
Is_duet	É um dueto? Sim / Não
Is_ad	É um anúncio? Sim / Não
Music_name	Nome da música
Music_id	Número de identificação da música
Music_URL	URL da música (não acessível)
Music_thumbnail	Imagem de perfil da música (não acessível)
Music_author	Autor da música
Vídeo_URL	URL do vídeo (não acessível)
Tiktok_URL	URL do vídeo no TikTok (não acessível)
Thumbnail_URL	URL da imagem miniatura do vídeo (não acessível)
Likes	Número de gostos do vídeo
Comments	Número de comentários do vídeo
Shares	Número de partilhas do vídeo
Plays	Número de visualizações do vídeo
Hashtags	As hashtags usadas
Challenges	Desafios usados
Diversification_labels	Etiquetas de diversificação (normalmente vem sem dados)
Location_created	Localização associada (normalmente vem sem dados)
Stikers	Texto introduzido como autocolante sobre o vídeo
Effects	Efeitos (normalmente em branco)
Warning	Aviso (normalmente em branco)

(*) Pode ser desafiante reformular a data para um formato reconhecido pelos filtros do Google Sheets ou do Excel. Poderá sempre recorrer a ferramentas *online* como o Epoch Converter (https://www.epochconverter.com/) ou usar ChatGPT, fazendo a respetiva validação dos resultados.
(**) Há uma série de conversores *online* para este formato de dados. Também pode recorrer a uma ferramenta de inteligência artificial para fazer a conversão.

Fonte: elaboração própria da autora.

Receitas

Condições prévias

Antes de iniciar a sua pesquisa, existem alguns passos importantes a seguir a fim de não comprometer a objetividade dos dados recolhidos, considerando que o Zeeschuimer só funciona com um *login* do TikTok ativo. Para perceber a importância desses passos, relembramos que o TikTok é uma plataforma que recolhe uma variedade de dados sobre o seu utilizador, que lhe permitem afinar o seu algoritmo. Quando usamos um perfil pessoal, ou um navegador não sanitizado, esses dados podem enviesar os resultados, comprometendo a nossa investigação. Vamos imaginar que, a título pessoal, tem uma paixão enorme por carpintaria, consumindo conteúdos sobre o tema com regularidade, enquanto na qualidade de investigador, quer estudar a forma como as instituições bancárias comunicam nas novas redes sociais, e perceber que tipo de resultados surgem para o utilizador português comum quando este procura pela palavra-chave "banco" no TikTok. Se para essa pesquisa utilizar o seu perfil pessoal ou um navegador não sanitizado, é natural que lhe surjam muito mais resultados de bancos de madeira do que ao utilizador regular, comprometendo, assim, a sua análise.

Considerando esta realidade, é importante seguir os seguintes passos:

1. limpe o histórico de *cookies* e de cache no seu navegador Firefox antes de criar uma nova conta. Para isso, vá primeiro até ao menu do navegador, selecione "Histórico" e faça a limpeza de histórico. Depois, vá às definições, segurança e privacidade, e faça a limpeza da cache e das *cookies*. Agora estará pronto para criar a sua conta TikTok;
2. vá a www.tiktok.com e selecione a opção de criar uma conta de TikTok nova a partir do *email*. Use um *email* recém-criado a que tenha acesso ou que raramente usa, uma vez que o TikTok faz cruzamento de contactos, e também isso pode interferir com o algoritmo e, portanto, com os resultados das suas pesquisas nesta plataforma. O TikTok não pergunta o género, mas pede data de nascimento por motivos legais. Se saltar a opção de criar um nome de utilizador, o TikTok automaticamente criará um para si;
3. o TikTok abrirá, então, automaticamente a página "For You" no seu navegador. Está pronto para iniciar a sua recolha de dados.

Além do *email* e da data de nascimento, existem alguns dados que o TikTok vai sempre recolher. Fatores como a sua localização, inferida pela língua, e fuso horário do sistema operativo e do navegador e o IP são difíceis de alterar sem apoio técnico, sendo necessário o uso de máquinas virtuais e VPN. Considere estas questões na sua análise de resultados. Pode também fazer um desenho de pesquisa que analise variações nos resultados na sequência de alterações a estas variáveis. Mas, para o presente manual, o enfoque é numa pesquisa mais generalista.

Tipos de recolha

Destacamos no quadro 7.3 três tipos diferentes de recolha dos dados, que partem de pontos de partida e tipos de pergunta distintos, com exemplos ilustrativos para melhor compreensão.

Estes diferentes tipos de recolha não são mutuamente exclusivos, podendo até ser complementares na sua investigação.

Recolher dados

Explicamos de seguida os passos a tomar para a recolha de dados, partindo de uma pesquisa por palavra-chave ou *hashtag*.

1. Ative o Zeeschuimer e coloque o termo de pesquisa que lhe interessa na barra de pesquisas do TikTok.
2. Após surgirem os resultados, vá fazendo *scroll down* até sentir que recolheu dados suficientes.
3. Quando sentir que já tem dados suficientes, vá até ao Zeeschuimer, colocando a recolha em *off*. Na coluna itens recolhidos (ver quadro 7.1) terá o número de vídeos recolhidos enquanto fez *scroll down*.

Quadro 7.3 Diferentes tipos de recolha com o Zeeschuimer

Ponto de partida	Palavra-chave ou *hashtag*	Pesquisas sobre um tema	Comparação dados sociodemográficos
Tipo de perfil	Perfil sanitizado	Perfil sanitizado	Perfil com dados sociodemográficos
Local de recolha	Resultados da pesquisa, separador "vídeos"	Separador "For You"	Separador "For You" ou resultados da pesquisa
Tipo de pergunta	Que conteúdo apresenta a plataforma na sequência da pesquisa por uma palavra-chave ou *hashtag*?	Que conteúdo apresenta a plataforma no separador "For You" a um utilizador que pesquise por determinados temas?	Que diferenças existem no tipo de conteúdo apresentado pela plataforma no separador "For You", ou nos resultados de pesquisa, para diferentes tipos de utilizador?
Exemplo	Análise em março de 2024 da *hashtag* #legislativas2024 para análise de conteúdo sobre a campanha para as Legislativas de 10/03/2024.	Análise do conteúdo da página "For You" de um perfil sanitizado após pesquisas por dietas ou outros programas de redução de peso.	Análise dos resultados de uma pesquisa por tatuagens entre dois perfis com idades muito diferentes.

Fonte: elaboração própria da autora.

4. Faça a exportação de resultados, selecionando a opção em formato JSON.[4] O ficheiro aparecerá pouco depois na sua pasta de transferências, com o nome "zeeschuimer", com a data e hora da extração e o sufixo .ndjson.
5. Grave o ficheiro numa pasta com uma indicação que lhe permita identificar posteriormente a que tipo de recolha corresponde.
6. Caso esteja interessado em recolher dados do separador "For You", ou recolher dados em outra data ou com outro tipo de pesquisa, repita o mesmo processo.

Conversão para CSV

O Zeeschuimer exporta os dados em JSON, o que, para a maioria dos investigadores, pode ser um formato de ficheiro difícil de trabalhar. Neste contexto, os mesmos criadores do Zeeschuimer criaram o Zeehaven, que converte ficheiros JSON para formato CSV. Explica-se de seguida os passos para esta conversão.

1. Navegue até https://publicdatalab.github.io/zeehaven/ e abra a ferramenta Zeehaven.
2. Arraste até à zona de *drop* o ficheiro que exportou do Zeeschuimer (terá o sufixo .ndjson).
3. Automaticamente o Zeehaven faz a conversão do ficheiro e faz o descarregamento automático, com o mesmo nome, do ficheiro CSV.
4. Devido a problemas de compatibilidade do Microsoft Excel, aconselhamos que copie, sem abrir, o ficheiro para uma Drive da Google (www.google.com/drive/) e o abra, pela primeira vez, dentro da Drive, convertendo para Google Sheets, ou, caso utilize um Macintosh, que utilize a aplicação Numbers.
5. No ficheiro CSV serão visíveis, nas linhas, os itens visualizados e recolhidos e, nas colunas, as características do *output* já listado no quadro 7.2.

Algumas possibilidades de investigação

Antes de começar a explorar o seu ficheiro Google Sheets ou Numbers (Mac) lembre-se de realizar uma cópia e trabalhar nessa mesma cópia, mantendo a recolha original inalterada, funcionando esta como um repositório de dados.

Vamos regressar aos exemplos dados num ponto anterior sobre os diferentes tipos de recolha. São três exemplos com diferentes pontos de partida, que o podem inspirar acerca do tipo de perguntas a que pode responder com os dados que o Zeeschuimer consegue recolher.

---

4 Neste manual, optou-se por não desenvolver a opção de exportação para a ferramenta 4CAT, uma ferramenta gratuita de análise, também da Digital Methods Initiative, com aplicabilidade transversal, que vale a pena explorar caso esteja a recolher elementos de mais do que uma plataforma. Pode saber mais sobre o 4CAT em: https://4cat.nl/

Quadro 7.4 Exemplos com diferentes pontos de partida

Exemplo 1	Exemplo 2	Exemplo 3
Que conteúdo apresenta a plataforma na sequência da pesquisa por uma palavra-chave ou *hashtag*?	Que conteúdo apresenta a plataforma no separador "For You" a um utilizador que pesquise por determinados temas?	Que diferenças existem no tipo de conteúdo apresentado pela plataforma no separador "For You", ou nos resultados de pesquisa, para diferentes tipos de utilizador?
Análise em março de 2024 da *hashtag* #legislativas2024 para análise de conteúdo sobre a campanha para as Legislativas de 10/03/2024.	Análise do conteúdo da página "For You" de um perfil sanitizado após pesquisas por dietas ou outros programas de redução de peso.	Análise dos resultados de uma pesquisa por tatuagens entre dois perfis com idades muito diferentes.

Fonte: elaboração própria da autora.

Exemplo 1

Durante os primeiros dias de março de 2024, em três dias diferentes, foram recolhidos resultados do separador de vídeos para a pesquisa pela *hashtag* #legislativas2024. No total foram extraídos dados sobre 227 vídeos diferentes visualizados. A informação recolhida permite-nos fazer várias análises, entre as quais:

— que tipo de temas surgem nestes vídeos? Uma análise das *hashtag*s extraídas permite construir uma nuvem de palavras (figura 7.1);

Figura 7.1 Nuvem de palavras das *hashtag*s dos vídeos extraídos

Fonte: elaboração própria da autora.

— quais os autores com mais alcance a produzir conteúdo com esta *hashtag*? Uma soma do número de visualizações dos diferentes vídeos extraídos, por autor, permite compreender que alguns autores têm um destaque maior para esta *hashtag*;
— existiram anúncios relacionados com esta *hashtag* durante o período de campanha eleitoral? Uma simples verificação na coluna "is_ ad" revelou a existência de conteúdo patrocinado com esta *hashtag*, o que pode ser considerado uma infração da regulação eleitoral.

Exemplo 2

Foi criado um perfil sanitizado e foram feitas pesquisas iniciais por palavras-chave relacionadas com dietas e programas de perda de peso, nomeadamente "emagrecer", "dieta" e "perder peso". Foram visualizados os primeiros resultados de cada pesquisa sem dar destaque a nenhum vídeo em particular, procurando ver os primeiros 5-10 segundo de cada vídeo. Nos três dias seguintes, abriu-se o mesmo perfil e consultou-se a página "For You", procurando novamente não dar especial destaque a nenhum tipo de vídeo, e visualizando diariamente os primeiros resultados, tendo-se extraído, no total, 108 vídeos.

— Qual a diversidade de conteúdo apresentada pelo TikTok a um utilizador que já pesquisou por um tópico? Dos 108 vídeos extraídos, apenas 28 não incluíam conteúdos sobre dietas, imagem corporal, hábitos de alimentação, hábitos de exercício físico ou outros temas relacionados. No terceiro dia, este tipo de conteúdo correspondeu a mais de 95% do conteúdo apresentado.
— Qual a orientação que as nossas pesquisas podem ter no conteúdo publicitário apresentado? 100% dos anúncios apresentados estavam relacionados com o tópico pesquisado.

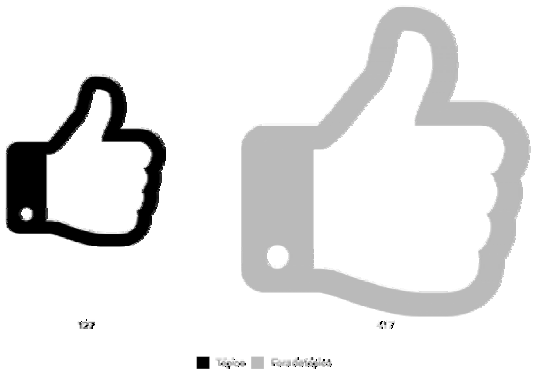

Figura 7.2 Média dos *gostos* entre o conteúdo dentro e fora do tópico dietas
Fonte: elaboração própria da autora

— A plataforma procura manter o utilizador na sua bolha? As métricas extraídas dizem-nos que as interações do conteúdo sobre o tópico pesquisado são, em média, significativamente mais baixas do que a do conteúdo não relacionado. Isto indica que o TikTok apresenta alguns dos seus vídeos mais virais sobre outros temas, para tentar alargar e identificar o espectro de gostos do utilizador.

Exemplo 3

No terceiro caso foram criados dois perfis sanitizados com idades muito diferentes, um deles com 16 anos, e outro com 40. De seguida, fizeram-se pesquisas pelas mesmas palavras-chave: "tatuagem", "tatuagens", "tatuador", e "tatuadora". Foram visualizados os primeiros resultados de cada pesquisa sem dar destaque a nenhum vídeo em particular, procurando ver os primeiros 5-10 segundos de cada vídeo. De seguida, foram extraídos os resultados apresentados no separador "For You". A pesquisa e visualização foi feita durante os mesmos dias para ambos os perfis, usando o Firefox em modo anónimo, a fim de reduzir o cruzamento de dados. Foram extraídos 164 vídeos, com alguns elementos interessantes.

— Que diferenças foram identificadas nos vídeos extraídos que podem resultar da diferença de idades?
  — Maior diferença entre conteúdo sobre o tópico *vs.* fora de tópico com o utilizador mais velho a ter menos variedade e mais conteúdo sobre o tema pesquisado.
  — Mais conteúdo com músicas, ou áudios de terceiros, para o utilizador mais novo, com o utilizador mais velho a ver mais conteúdo com áudio original.
  — Mais desafios para o utilizador mais novo.

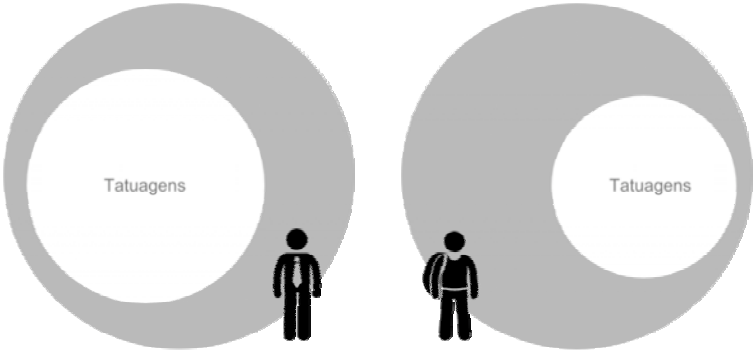

Figura 7.2 As diferenças entre o volume de conteúdo sobre o tópico entre o perfil adulto e o perfil jovem
Fonte: elaboração própria da autora.

- Que diferenças foram identificadas nos vídeos extraídos sobre tatuagens que podem resultar da diferença de idades?
  - Análise do tipo de linguagem no texto que acompanha os vídeos. Utilizando o ChatGPT, foram identificadas diferenças nos tempos verbais, na formalidade do tom e na forma de abordar o utilizador.
- Que conteúdos existem sobre tatuagens no TikTok, para diferentes idades?
  - Uma codificação por 4 categorias principais — ideias, técnicas, apresentação de serviços, histórias pessoais — permite perceber que não existe grande diferença entre os tipos de categorias apresentadas, com mais histórias pessoais a surgir ao utilizador adulto e mais técnicas ao utilizador novo.

Os presentes exemplos de como os diferentes exemplos e a respetiva recolha de dados permitem responder a diversos tipos de perguntas pretendem apenas ser uma apresentação das possibilidades de investigação que uma ferramenta como o Zeeschuimer oferece.

Relembramos que o Zeeschuimer também pode ser instalado no navegador de terceiros, com o devido consentimento informado dos mesmos, considerando que, no processo, não extrai dados sobre o utilizador em si, apenas sobre os vídeos visualizados. Tendo em conta os vieses que advêm do uso de um perfil não sanitizado na apresentação de resultados, esta possibilidade abre ainda diversas outras linhas de investigação, sobretudo no estudo desses mesmos vieses, ou, com uma amostra de diversos utilizadores, como um determinado tema surge perante uma determinada comunidade.

Referências bibliográficas

Aguilar, G. K., Schueler, M., Teggin, A., Brennan, M., Brossman, B., Caroleo, L., .. e Yuzawa, M. (2023), "The divine online? Mapping algorithmic conspirituality on TikTok", *Digital Methods Initiative*, https://wiki.digitalmethods.net/Dmi/TheDivineOnline

Anderson, M., e Rainie, L. (2020), "The future of digital spaces and their role in democracy", *Pew Research Cente*r. https://www.pewresearch.org/internet/2020/07/13/the-future-of-digital-spaces-and-their-role-in-democracy

Bishop, S. (2020), "TikTok and the virtual scene of the screen: on virtual camera, affective proximity, and other ways of seeing", *Convergence*, 26 (5-6), pp. 1124-1137, doi:10.1177/1354856520923966.

Ceci, L. (2024), "TikTok distribution of global audience 2024", *Statista*, disponível em https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution.

Iqbal, M. (2024), "TikTok revenue and usage statistics", *Business of Apps*, disponível em https://www.businessofapps.com/data/tik-tok-statistics

McCashin, D. e C.M. Murphy (2023), "Using TikTok for public and youth mental health — a systematic review and content analysis", *Clinical Child Psychology and Psychiatry*, 28 (1), pp. 279-306, doi:10.1177/13591045221106608.

Perrin, A., e M. Anderson (2019), "Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018", *Pew Research Center*, disponível em https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Rogers, R. (2024), *Doing Digital Methods*. Sage.

Trifiro, B. M. (2023), "Breaking Your Boundaries: how TikTok use impacts privacy concerns among influencers", *Mass Communication and Society*, 26 (6), pp. 1014-1037, https://doi.org/10.1080/15205436.2022.2149414.

Zhao, Y., X. Zhou, K. Zhang, e F. Wang (2021), "TikTok as a health information source: assessment of the quality of information in diabetes-related videos", *JMIR Diabetes*, 6 (3), e30409, doi:10.2196/30409.

Capítulo 8

YouTube

Canais, vídeos e comentários

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

O YouTube, plataforma social e *website* criado em 2005 e adquirido em 2006 pela Google, caracteriza-se por apresentar o vídeo como formato comunicacional, complementado pelas descrições textuais dos vídeos e dos canais, dos grafismos usados para identificar ambos e, ainda, dos comentários aos vídeos. Embora, atualmente, o vídeo não seja um formato exclusivo desta plataforma, uma vez que outras o têm adotado, como o Facebook, Instagram ou TikTok, no YouTube esse é o formato nativo, tendo esta sido, historicamente, a plataforma que primeiro deu grande projeção ao vídeo como formato de comunicação digital.

Apresentando-se como uma plataforma dedicada à partilha de vídeos pelos utilizadores (Burgess e Green, 2018), permite, numa ótica do utilizador-produtor, criar canais em que se podem armazenar e difundir vídeos próprios, e, numa ótica de utilizador-consumidor, procurar vídeos, vê-los, interagir com o seu conteúdo (através de ações como gostar, não gostar, comentar, partilhar, transferir, entre outras) e seguir os produtores desses vídeos através da subscrição do canal. Trata-se de um serviço gratuito. Porém, para fazer o *upload* de vídeos, o utilizador terá de se registar, enquanto para ver conteúdo isso não é necessário.

Presente em 100 países (Dean, 2024) e acumulando 2,51 mil milhões de utilizadores ativos mensalmente, o YouTube é a segunda plataforma social mais popular. A maioria dos seus utilizadores tem entre 25 e 34 anos (20,7%), 45,6% pertencem ao género feminino e 54,4% ao género masculino. Em média, os seus utilizadores estão 23 horas por mês *online*, sendo a segunda plataforma social em que os utilizadores de *internet* passam mais tempo (Kemp, 2023).

As pesquisas mais comuns no YouTube estão relacionadas com termos como "song", "movie", "film", "Dj" (Kemp, 2023). Inclusivamente, esta plataforma é muitas vezes apresentada como o segundo motor de busca mais utilizado no mundo, a seguir ao Google Search, precisamente na procura de conteúdos informativos ou de entretenimento em vídeo. [1] O canal mais popular na plataforma intitula-se "T-series", acumulando 233 milhões de subscritores.[2] O vídeo mais visualizado é "Pink Fong — Baby Shark Dance", reproduzido mais de 12 mil milhões de vezes

1 Fonte: https://www.searchenginejournal.com/seo/meet-search-engines/
2 T-series é a maior gravadora musical e estúdio de filmes da Índia.

(Kemp, 2023). Porém, o YouTube não se resume a essas tipologias. De facto, as 15 categorias de canais da plataforma são bastante diversas, incluindo educação, notícias e política, jogos, comédia, pessoas e blogues, entre outras. Estas estão organizadas de forma que os utilizadores descubram outros vídeos (Burgess e Green, 2018). Nesta lógica, por cada vídeo visto, o YouTube recomenda outros vídeos. Essas recomendações podem ser em termos do conteúdo ou do canal.

Essa dinâmica e lógica de recomendação são produto do algoritmo do YouTube que explica, em parte, a popularidade da plataforma. Na sua composição, caso o utilizador tenha feito *login*, estão dados relativos aos seus comportamentos. Caso o utilizador não tenha feito *login*, dados como o número de vezes que o vídeo foi reproduzido, a interação desses vídeos, a atualidade dos mesmos ou do tema e o número de subscrições do canal desempenham um papel importante (Rieder, Matamoros-Fernández e Coromina, 2018). Essa lógica algorítmica de recomendação também explica, em parte, polémicas às quais o YouTube tem sido associado (Chen *et al.* 2021), nomeadamente introduzindo o conceito de "rabbit hole" associado à sequência polarizadora dessas recomendações (Ledwich e Zaitsev, 2019).

O YouTube baseia o seu sucesso comercial na atividade dos utilizadores e na publicidade. A monetização da plataforma aproveita as vantagens da "datificação", levando à disseminação de conteúdos com objetivo de rentabilidade (Holland, 2016). A monetização também se baseia na criação e distribuição de conteúdos produzidos por diversos atores. Nesse contexto, o termo *youtuber* — um criador de conteúdo que grava vídeos sobre si ou o seu contexto — tornou-se popular (Berzosa, 2017). Este não só pode ser compreendido como uma marca em si mesmo, como ser atrativo para marcas comerciais enquanto influenciador e prescritor (De-Aguilera-Moyano *et al.*, 2018). O YouTube foi a primeira plataforma a implementar um sistema de incentivo aos criadores para criação de conteúdos, através do YouTube Partner Program, criado em 2007 e desenvolvido a partir daí, mediante o qual a plataforma passou a partilhar uma parte das receitas de publicidade com esses criadores de conteúdo.[3] Posteriormente muitas outras plataformas de redes sociais implementaram programas semelhantes.

O YouTube e os seus vídeos, como objeto de estudo, contemplam os mais diversos temas (questões de saúde, moda, etc.) e as mais diversas questões, como o funcionamento e arquitetura da plataforma, a ação do algoritmo ou o sistema de recomendação. Assim o YouTube torna-se interessante como palco de investigação, não só pelos conteúdos que disponibiliza, mas também pela forma como os distribui e pelos agentes envolvidos.

3 https://support.google.com/youtube/answer/72851?hl=pt

Recolha de dados

Ferramenta: YouTube Data Tools

O YouTube Data Tools é uma ferramenta criada em 2015 por Bernhard Rieder.[4] Recolhe dados através da API do YouTube, não tendo um intuito comercial e sendo de utilização livre. Tem sido empregue em variados estudos que se centram em diferentes temas e questões, mostrando não só a sua importância como ferramenta, mas também o seu potencial.

No projeto Eumeplat, especificamente no âmbito do WP2, dedicado ao tema "Platformization of News", a ferramenta YouTube Data Tools foi de extrema utilidade para a recolha de dados, relativos a várias dimensões, tópicos de análise e nos dez países envolvidos.[5] O desenho da pesquisa (Cardoso *et al.*, 2021), mais complexo do que o exemplo aqui apresentado, pode servir de orientação para o desenho de outras pesquisas.

Tal como outras ferramentas, já apresentadas em capítulos deste livro, o YouTube Data Tools também apresenta uma lógica de recolha de dados através de módulos. Atualmente estão seis módulos disponíveis. A sua denominação e função resumem-se no quadro 8.1.

Quadro 8.1 Módulos disponíveis no YouTube Data Tools

Módulo	Descrição
Channel info	Permite obter estatísticas relativas a um canal do YouTube.
Channel list	Permite obter uma lista de informações e estatísticas sobre um canal.
Channel network	Permite obter uma rede de canais ligados através da função "featured channels" e subscrições.
Video list	Permite obter uma lista de informações e estatísticas de vídeos.
Video co-commenting network	Permite obter uma rede de vídeos baseados numa lógica de comentário.
Video comments	Permite obter dados sobre um determinado vídeo (exemplo: informação geral, estatísticas, comentários, informação sobre quem comentou, interações entre utilizadores na secção dos comentários).

Fonte: elaboração própria dos autores.

Aceder e explorar a ferramenta

— Ir a https://ytdt.digitalmethods.net/
Ao contrário de outros serviços, não precisa de criar conta.
Uma vez que tenha acedido à ferramenta, pode escolher o método que pretende explorar, clicando no nome do mesmo. A recolha de dados requer o fornecimento de um conjunto de *inputs*. No quadro 8.2 resumem-se os *inputs* requeridos, e a sua respetiva explicação, de acordo com os diferentes métodos.

4 https://twitter.com/RiederB
5 https://www.eumeplat.eu/

Quadro 8.2 *Inputs* requeridos pelo YouTube Data Tools, e a sua descrição, em função do método

Input	Descrição	Channel info	Channel list	Channel network	Video list	Video co-commenting network	Video info & comments
Channel ID ou URL	O ID ou URL do canal	X			X		
Search query	Expressão ou expressões através das quais se realiza a pesquisa		X	X	X	X	
Iterations	Refere-se à repetição das ações		X	X	X	X	
Language (optional)	O código, de duas letras, do idioma no qual se apresentam os resultados.		X		X	X	
Region code (optional)	O código, de duas letras, correspondente ao país do qual se pretende fazer recolha		X		X	X	
Published	Possibilidade de limitar a pesquisa numa baliza temporal específica		X		X	X	
Rank by(*)	Indicar qual o critério através do qual os dados são ordenados		X	X	X	X	
Location	Pesquisar por vídeos que especifiquem a localização nos metadados				X		
Manual selection(**)	Possibilidade de pesquisa através dos ID dos canais		X	X	X	X	
File format	Escolher entre CSV ou TAB para o *output* de dados		X		X		
Subscriptions	Possibilidade de usar na pesquisa a opção "subscrição" e "featured channels"			X			
Crawl depth	Número de níveis a aplicar na pesquisa de canais relacionados			X			
Comments	Número de comentários a retirar					X	
Playlist ID	O ID da *playlist*				X		
Video ID	Identificação do vídeo						X
Limit to	Número de comentários						X
reCAPTCHA	Selecionar para provar que não é um robô	X	X	X	X	X	X

(*) Existem seis critérios distintos: *relevance,date, rating, title, video count, view coun*t**.**

(**) Também chamado “seeds”.

Fonte: elaboração própria dos autores.

Outputs gerados

Após ter selecionado o módulo que pretende explorar, ter introduzido os *inputs* requeridos, selecionado a função "reCAPTCHA" e ter clicado na opção "enviar", a ferramenta começará a recolher os dados. O ficheiro com os resultados aparecerá na página do módulo que está a usar. Não obstante, em função do módulo, esse ficheiro terá diferentes formatos. No quadro 8.3 resumimos a tipologia do formato do ficheiro de resultados em função do módulo.

Quadro 8.3 Formato do ficheiro de resultados em função do módulo

Módulo	Formato do ficheiro de resultados
Channel info	TAB Pode ser aberto com Excel. Caso deseje trabalhar no Google Drive, faça *upload* do ficheiro, formato Excel, e abra com Google Sheets.
Channel list	CSV ou TAB
Channel network	GDF Pode ser aberto através do *software* Gephi.
Video list	CSV ou TAB
Video co-commenting network	GDF
Video info and comments	Opção HTML Opção Pseudonymize Opção CSV ou TAB

Fonte: elaboração própria dos autores.

Receitas

Utilizar o módulo "Video list"

Através do módulo "Video list", irá conseguir obter um conjunto de resultados — a sua base de dados — pesquisando através de um canal, de uma *playlist* ou de uma expressão/termo ou conjunto articulado de expressões ou termos (*query*). Iremos exemplificar o uso deste módulo, através da terceira opção: pesquisa através de *query*.

Recolher dados

— Ir a https://ytdt.digitalmethods.net/;
— selecionar o módulo "Video list";
— preencher os diferentes campos.

Para este exemplo, vamos realizar uma pesquisa pelo termo "corrupção". Pode utilizar outros termos, de acordo com os seus interesses de pesquisa ou investigação. Tenha em conta que, quanto mais precisa for a sua *query*, mais distintos serão os resultados. Não só quanto ao número de vídeos, mas também aos vídeos propriamente ditos. Assim, pesquisar por "corrupção" ou pesquisar por "corrupção portugal" originará resultados diferentes. No quadro 8.4 indicamos os campos, *inputs* e parâmetros de pesquisa estabelecidos para a recolha através do módulo "Video list".

Quadro 8.4 Resumo dos campos e *inputs* para pesquisa através do módulo “Video list”

Campo	*Input*
Search query	Indicar a expressão de pesquisa no campo "query". Exemplo usado: "corrupção".
Relevant language	Indicar o código de duas letras relativo ao idioma. Indicámos "pt".
Region code	Indicar o código de duas letras relativo ao país. Indicámos "PT".
Iterations	Indicar o número de *iterations*. Uma vez que cada *iteration* fornece 50 resultados, outros valores devem ser usados com prudência. Seja devido ao tempo que a recolha demorará como à capacidade que teremos para analisar os mesmos. Neste caso, indicámos 1 *iteration*.
Make a search for each day of the timeframe (can yield many more videos, use wisely)	Selecionar caso pretenda que a pesquisa seja feita para cada um dos dias da baliza temporal definida. Não selecionámos Pode eleger entre seis opções diferentes: "Relevance"; "Date"; "Rating"; "Title"; "viewCount".
Rank by	Selecionar no campo "Rank by" o critério pelo qual quer ordenar os dados. Selecionámos "Relevance".
Limit search to videos published in a specific timeframe (format: yyyy-mm-ddThh:mm:ssZ — timezone: UTC)	Indicar o limite temporal no qual os vídeos foram publicados. Colocámos After: 2023-01-01T00:00:00Z Before: 2024-01-01T00:00:00Z
Location	Permite pesquisar por vídeos que especifiquem uma determinada localização nos metadados. Não selecionámos/indicámos localização.
File format	Indicar formato do ficheiro de output. Selecionámos CSV
Run	Selecionar reCAPTCHA para provar que não é um robot.

Fonte: elaboração própria dos autores.

— Clique no botão “Submit” ou “enviar”. A ferramenta começará a recolher os dados. A mensagem “Processing” aparecerá, seguida de “executing searches” e “Getting video details”.
— Uma vez que a recolha tenha terminado, o ficheiro de resultados surgirá no final dessa mesma página. Clique no nome do ficheiro. Este estará destacado a amarelo e o *download* será realizado de forma automática. Note que não é possível atribuir um nome ao ficheiro de resultados. Este é definido automaticamente pela ferramenta.

Explorar os dados

— Os resultados serão apresentados num ficheiro CSV uma vez que esse foi o formato selecionado. Poderá abri-lo fazendo *upload* do mesmo para o Google Drive

e selecionando a opção de abrir com Google Sheets.[6]

— Uma vez aberto o ficheiro, mude-lhe o nome. Como mencionado noutros capítulos, é importante que se lembre do nome do ficheiro e que este corresponda à pesquisa realizada.
Sugerimos a seguinte nomenclatura: Rede social_termo da pesquisa_módulo_data da recolha. Para este exemplo ficaria: YouTube_corrupção_videolist_26022024. A data de recolha dos dados é importante, uma vez que a disponibilidade de dados — neste caso, vídeos ou comentários, por exemplo — pode mudar com o tempo. Na descrição da metodologia, será importante referir a data de recolha dos dados.

— Uma vez aberto o ficheiro de resultados, verá várias linhas e colunas com dados. Cada linha corresponderá a um vídeo e cada coluna, a um dado específico sobre cada um desses vídeos. Entre esses dados encontra, por exemplo: "channelId", "channelTitle", "videoId", "publishedAt", "videoTitle", "videoDescription", "tags", "videoCategoryId", "videoCategoryLabel", "duration", "viewCount", "likeCount", "dislikeCount", "favoriteCount" e "commentCount". A nomenclatura é bastante autoexplicativa e permite identificar o conteúdo de cada coluna. De notar que os dados aqui recolhidos são aqueles que estão disponíveis publicamente na plataforma através da API e que a ferramenta utilizada permite recolher. Outras API ou outras ferramentas poderão gerar diferentes "datapoints".

Algumas possibilidades de investigação

Antes de começar a explorar a base de dados, lembre-se de realizar uma cópia da folha de resultados e trabalhar nessa mesma cópia, mantendo a recolha original inalterada, funcionando como um repositório de dados.

— A que canais pertencem os vídeos?
Através da coluna "channelTitle" terá informação sobre o nome do canal. Pode explorá-los um a um ou ordená-los, por exemplo, por ordem alfabética para perceber se estes se repetem. Para tal basta criar um filtro e ordenar os dados.[7]
Há algum canal que se destaca pelo número de vídeos com os quais contribui para o tema e que o YouTube considera como relevante?

— A que categorias pertencem os vídeos sobre corrupção?
Poderá explorar a tipologia de categorias de que os vídeos fazem parte através dos dados da coluna "videoCategoryLabel" (tendo em atenção, no entanto, que esta é uma categorização atribuída automaticamente pelo próprio YouTube. Deverá verificar essa categorização). A função "COUNTIF" será

6 Caso tenha escolhido a opção TAB, pode abrir o mesmo através do Excel. Clique no ficheiro de resultados com o botão do lado direito, eleja a opção "Abrir com" e aí escolha a opção "Excel".

7 Selecione a linha 1 (clique na primeira célula da linha 1), em seguida clique no símbolo do filtro — aquele que se parece com um funil.

útil para contar quantas vezes as categorias se repetem.[8] Como é que os vídeos se distribuem ao longo do tempo?
Para explorar o número de vídeos ao longo do tempo, a coluna *"publishedAt"* será o ponto de partida e a função filtros será bastante útil.

— Quais os assuntos abordados nos vídeos?
Pode começar por explorar a coluna "videoDescription". O texto poderá conferir informações sobre os tópicos abordados no vídeo.
Para uma análise mais detalhada, deverá visualizar os vídeos. Para tal, a coluna "videoId" será importante uma vez que ela contém o ID de cada vídeo. Copie o ID dos vídeos que pretende analisar, abra o YouTube e no campo "pesquisa" cole esse ID. Em geral, o primeiro vídeo dos resultados corresponderá ao vídeo que pretende analisar. Ainda assim, confirme através do título e da descrição.

— Quais os vídeos que receberam mais ou menos atenção dos utilizadores do YouTube?
Por atenção compreendemos o número de visualizações e o número de *likes* e para tal poderá explorar as colunas "viewCount" e "likecount". Pode fazê-lo de forma individual ou até somar os dados das mesmas.[9]

— Quais os vídeos que geraram mais comentários?
Os comentários também podem ser agrupados numa lógica de atenção conferida ao vídeo. Porém, o comentário, quando comparado com o *like*, revela um maior envolvimento por parte da audiência.
Através da coluna "commentCount", poderá ordenar os vídeos em função do número de comentários e, por exemplo, realizar uma análise dos vídeos em função dessa métrica.
O número de comentários poderá também ser um ponto de partida para outra análise: a dos comentários em si. Em seguida, exploraremos essa possibilidade.

Utilizar o módulo "Video Info and Comments"

Partindo dos resultados da recolha anterior, realizada através do módulo "video list", centraremos esta receita no vídeo que recebeu o maior número de comentários. Para tal, ordenamos os resultados pelos dados da coluna "commentCount".[10] Note que poderá, obviamente, usar o módulo "Video info and comments" sem ter usado o módulo "Video list".

Recolher dados

— Ir a https://ytdt.digitalmethods.net/;
— selecionar o módulo "Video info and comments";
— preencher os diferentes campos. No quadro 8.5 indicamos os campos e *inputs* facultados.

---

8 Consulte o procedimento "Explorar contas associadas à/s *hashtag*/s" no capítulo dedicado ao Instagram para relembrar como usar a função "COUNTIF".
9 Aconselhamos a que crie uma coluna, na base de dados, que seja específica para essa soma.
10 Basta criar um filtro e selecionar a opção de ordenar de maior a menor.

Quadro 8.5 Resumo dos campos, *inputs* e parâmetros para pesquisa através do módulo “Video info and comments”

Campo	*Input*
Video ID	Indicar o ID do vídeo do qual se pretende recolher comentários. Exemplo usado: "OmVca2R32wk". Era aquele que acumulava o maior número de comentários.
Limit to	Número de comentários. Indicámos 100. No momento da análise, o vídeo somava 17 486 comentários.
HTML output	Não selecionámos a opção.
Pseudonymize	Não selecionámos a opção uma vez que poderia ser útil explorar quem comentou o vídeo. Ainda assim, considerações éticas seriam tidas em conta no momento de análise e partilha de resultados.
File format	Indicar formato do ficheiro. Selecionámos CSV.
Run	Selecionar reCAPTCHA para provar que não é um robô.

Fonte: elaboração própria dos autores.

Quadro 8.6 Ficheiros de resultados provenientes da recolha através do módulo “Video info and comments”

Nome "tipo" do ficheiro	Conteúdo	Formato
videoinfo_ID video_Date_basicinfo.csv	Ficheiro com informação sobre o vídeo (Exemplo: ID, published, title, description, channelId, channelTitle, duration, dimension, definition, viewCount, likeCount, dislikeCount)	CSV ou TAB
videoinfo_ID video_Date_comments.csv	Ficheiro com informação sobre os comentários (replyCount, likeCount, publishedAt, authorName, authorChannelId, isReply, isReplyTo) e o texto dos comentários propriamente ditos (text).	CSV ou TAB
videoinfo_ID video_Date_authors.csv	Ficheiro com informação sobre os autores tal como o handle e número de comentários (author; count).	CSV ou TAB
videoinfo_ID video_Date_commentnetwork.gdf	Ficheiro que permite criar uma rede de comentários.	GDF

Fonte: elaboração própria dos autores.

— Clique no botão “Submit” ou “enviar”. A ferramenta irá começar a recolher os dados. Uma vez que a recolha esteja terminada, quatro ficheiros serão gerados e aparecem no ecrã. No quadro 8.6 são indicados os ficheiros de resultados e o seu conteúdo. De novo, o nome do ficheiro será atribuído pela ferramenta. Note que o que muda, na nomenclatura tipo dos mesmos, é a última palavra. Clique no nome dos ficheiros e o *download* será realizado de forma automática.

Explorar os dados

Antes de explorar os dados, altere o nome dos ficheiros de forma que se lembre deles e da respetiva pesquisa. Sugerimos a seguinte nomenclatura: Rede social_id do vídeo_módulo utilizado_output gerado_data da recolha. Para este exemplo, e referente ao ficheiro com os comentários, ficaria: YouTube_OmVca2R32wk_videocomment_comments_28022024.

Pode, obviamente, usar outra nomenclatura que considere mais adequada em termos de organização da sua pesquisa, mas, mais uma vez, reforçamos como desejável que adote uma nomenclatura que seja clara para si e que inclua a data de recolha dos dados.

Algumas possibilidades de investigação

— Que assuntos são abordados nos comentários?
Através do ficheiro terminado em "comments", no caso aqui exemplificado YouTube_OmVca2R32wk_videocomment_comments_28022024, a coluna "text" será aquela que merecerá atenção.
A tipologia de análise dependerá da sua pergunta de pesquisa e objetivos. Uma análise de conteúdo ou temática poderá ser útil. Poderá, por exemplo, classificar ou categorizar os comentários em termos de assunto, relativamente à tipologia (informativos, ofensivos...), realizar uma análise de sentimentos ou verificar o uso de *emojis*.[11]

— Quem comenta?
O autor dos comentários poderá também ser de interesse para a pesquisa. Os dados das colunas "authorName" e "authorChannelId" conferem dados para responder a essa pergunta. Permitem saber quem são os comentadores e adicionalmente procurá-los no YouTube para, por exemplo, caracterizá-los. Poderá também combinar esses dados com os do ficheiro terminado em "author". Esse apresenta métricas relativas ao número de comentários realizados por cada um dos intervenientes.

— Qual a dinâmica de comentários?
A dinâmica inerente aos comentários poderá auxiliar a caracterizar a discussão dos sujeitos envolvidos e da relevância do vídeo em si. Desta forma, colunas do ficheiro dos resultados "comments" como "isReply", "isReplyTo"/"isReplyToName" permitirão saber se o comentário é original, ou seja, um comentário ao vídeo, ou se se trata de um comentário a um outro comentário. Adicionalmente, a coluna "publishedAt" também permitirá refletir sobre se os comentários perduram no tempo ou se são muito próximos da data de publicação. Ao perdurarem no tempo, poderá indicar a contínua relevância do vídeo como referência sobre o tópico.
Ainda no âmbito da dinâmica dos comentários, poderá recorrer ao ficheiro

11 A ferramenta https://labs.polsys.net/tools/textanalysis/ será útil.

GDF. Este permite criar uma rede de comentários. Para tal, o *software* Gephi será uma ajuda.

Utilizar o módulo "Channel network"

Através do módulo "Channel network" será possível obter uma rede de canais ligados através da função "Featured channels" e subscrições. A função "Featured channels" é uma opção que permite aos proprietários de um canal destacar outros canais e a função "Subscrições" identifica canais que são subscritos pelo canal a ser analisado. Deste modo é possível estabelecer uma "rede de recomendações" dos próprios canais.

Recolher dados

— Ir a https://tools.digitalmethods.net/netvizz/youtube/;
— Selecionar o módulo "Channel network";
— Preencher os diferentes campos. No quadro 8.7 resumimos esses campos.

Quadro 8.7 Resumo dos campos, *inputs* e parâmetros para pesquisa através do módulo "Channel network"

Campo	*Input*
Search query	Indicar a expressão através da qual realizar pesquisa. Colocámos: "corrupção"
Iterations	Número de *iterations*. Indicámos 1.
Rank by	Selecionar no campo "Rank" o critério pelo qual quer recolher os dados. Selecionámos "Relevance".
Seeds	Pode optar por pesquisar por ID dos canais em vez de por uma expressão. Mantivemos a opção de pesquisa por expressão.
Subscriptions	Canais ligados por funcionalidade "Subscrição". Selecionámos essa opção.
Crawl depth	Número de níveis a aplicar na pesquisa de canais relacionados. Estabelecemos 1.
Run	Selecionar reCAPTCHA para provar que não é um robot.

Fonte: elaboração própria dos autores.

— Clique no botão "Submit" ou "enviar". A ferramenta irá começar a recolher os dados. Uma vez terminada a recolha, um ficheiro é gerado e aparecerá no ecrã. Clique no nome do ficheiro, que é atribuído pela ferramenta, e o *download* do mesmo será realizado de forma automática. Neste caso o ficheiro gerado é no formato GDF.

Explorar os dados

Como referido anteriormente, antes da exploração dos dados, altere o nome do ficheiro de forma que se lembre que dados ele inclui e da respetiva pesquisa. Mantenha o ficheiro original e trabalhe numa cópia. O ficheiro GDF pode ser aberto através do *software* Gephi. Este *software*, de uso livre, requer instalação no computador. A partir do mesmo será possível, neste caso concreto, analisar e visualizar a rede de canais formada a partir da palavra "corrupção".

Para conseguir realizar a análise e tirar o máximo partido da mesma, caso não esteja familiarizado com o Gephi, será necessário que tome conhecimento do *software* e das suas funcionalidades. No *site* https://gephi.org/users, encontrará vários tutoriais que poderão ser úteis caso deseje explorar redes de canais ou, por exemplo, comentários de vídeos (módulo "Video comments").

Algumas possibilidades de investigação

— Como é que a rede é composta?
Pode não só tirar conclusões quanto ao número de canais (os *nodes* — nós da rede), como também analisar a rede desde o centro para a periferia, concluindo sobre a distribuição de canais (vídeos ou comentários, noutros exemplos).
— Como é que os diferentes agentes se distribuem na rede ou onde, na rede, se situam os vídeos que acumulam o maior número de visualizações ou canais que acumulam maior número de subscrições?
— Quais são os vídeos mais relevantes?
Na lógica de funcionamento do YouTube, a relevância não é uma decorrência direta do *engagement*. A medida "Betweenness centrality" será importante para explorar a relevância.

Nos artigos de Moreno e Sepúlveda (2021) e Sepúlveda (2021), pode explorar como, em temas diferentes, a análise de redes ajudou a compreender a dinâmica dos agentes, dos vídeos, mas também a própria lógica do sistema de recomendação algorítmica do YouTube. Em Moreno e Sepúlveda (2021), foi possível identificar os canais relevantes, segundo a lógica de funcionamento do YouTube, quando o tema referente ao Artigo 13.º era discutido naquela plataforma. Um tópico bastante relevante para os *youtubers* e em que a atenção que lhe foi prestada pelos mesmos influenciou a agenda dos *media* tradicionais. Em Sepúlveda (2021), a pesquisa pela palavra-chave "Tinder" revelou que os vídeos com *engagement* mais elevado não eram os mais relevantes. Foi também possível identificar quais eram os vídeos relevantes, como a rede estava organizada por *clusters* temáticos específicos e que temas eram esses.

Referências bibliográficas

Berzosa, M. (2017), *Youtubers y Otras Especies: El Fenómeno que Ha Cambiado la Manera de Entender los Contenidos Audiovisuales,* Barcelona, Ariel.

Burgess, J., e J. Green (2018), *YouTube: Online Video and Participatory Culture*, Polity Press.

Cardoso, C., C. Álvares, J. Moreno, R. Sepúlveda, M. Crespo, e C. Foà (2021), "Amethodological framework for analyzing platform journalism", *Eumeplat*, disponível em https://www.eumeplat.eu/results/deliverables/.Chen, A., B. Nyhan, J. Reifler, R. Robertson, e C. Wilson, (2021), "Exposure to alternative e extremist content on YouTube", *Center for Technology & Society, Anti-Defamation League*, disponível em https://www.adl.org/*sites*/default/files/pdfs/2022-05/FINAL_FINAL_ADL-Report-Single-Final-Design.pdf.

De-Aguilera-Moyano, M., A. Castro-Higueras, e J.P. Pérez-Rufí (2018), "Between broadcast yourself and broadcast whatever: YouTube's homepage as a synthesis of its business strategy", *El Profesional de la Información*, 28 (2), e280206, https://doi.org/10.3145/epi.2019.mar.06.

Dean, B. (2024), "How many people use YouTube", *BlackLinko,* disponível em https://backlinko.com/youtube-users.

Holland, Margaret (2016), "How YouTube developed into a successful platform for user-generated contente", *Elon Journal of Undergraduate Research in Communications*, 7 (1), pp. 52-59.

Kemp, S. (2023), "Digital 2023: global overview report", disponível em https://datareportal.com/reports/digital-2023-global-overview-report.

Ledwich, M., e A. Zaitsev (2019), "Algorithmic extremism: examining YouTube's rabbit hole of radicalization", *arXiv*, 1912.11211.

Moreno, J. e R. Sepu´lveda (2021), "Article 13 on social media and news media: disintermediation and reintermediation on the modern media landscape", *Communication & Society*, 34 (2), pp. 141-157.

Rieder, B., A. Matamoros-Fernández, e O. Coromina (2018), "From ranking algorithms to 'ranking cultures': investigating the modulation of visibility in YouTube search results", *Convergence: The International Journal of Research into New Media Technologies*, 24 (1), pp. 50-68, https://doi.org/10.1177/1354856517736982.

Sepúlveda, R. (2021), "Engagement y contenido en videos sobre el aplicativo de citas Tinder en YouTube", *Palabra Clave*, 24 (4), e2446, https://doi.org/10.5294/pacla.2021.24.4.6.

Parte 3| Outras abordagens para fazer investigação digital

Capítulo 9

Pesquisa *online*
Google e Google Trends

Ana Pinto-Martinho
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

O número estimado de utilizadores da *internet*, em 2023, era de 5,4 mil milhões, o que representa 67% da população mundial (Petrosyan, 2024). A conectividade *online* está amplamente disseminada por todo o mundo. O uso da *internet* veio mudar comportamentos, mas também possibilitar o acesso à informação sobre os seus utilizadores de formas nunca antes conseguidas. As grandes quantidades de dados que as empresas das grandes plataformas, como a Alphabet, Meta, Amazon, obtêm através dos padrões de navegação e comportamentos *online* são incomparáveis a quaisquer outros tempos na história da humanidade.

Os motores de pesquisa são uma parte muito importante deste "ecossistema" *online* ligado em rede e povoado de enormes quantidades de conteúdos e grandes volumes de dados, em que a escolha da informação se torna desafiante para qualquer utilizador. Estas ferramentas tornaram-se essenciais para a navegação na *internet*, uma vez que apresentam os resultados que mais se aproximam dos interesses de pesquisa que os utilizadores expressam, quando as utilizam, tendo sempre em conta que o seu funcionamento é baseado em algoritmos. Saliente-se também que, na generalidade, os utilizadores desconhecem a forma como estes algoritmos funcionam.

Quando falamos de motores de pesquisa, o Google é incontornável pois assumiu-se como o mais utilizado a nível mundial, com exceção de alguns territórios como a China, que tem os seus próprios motores de pesquisa e plataformas *online*. Dados do Statista (Bianchi, 2024) mostram que o Google tinha 91,47% da quota de mercado a nível mundial. E, segundo Prater (2023), embora a Google não revele o número de pesquisas realizadas no seu motor de busca, as estimativas apontam para que ele processe cerca de 99 mil pesquisas por segundo. Isto traduz-se em 8,5 mil milhões de pesquisas por dia, com uma média de entre três e quatro pesquisas por dia, por pessoa.

No atual panorama da investigação académica, os motores de pesquisa *online* desempenham um importante papel, oferecendo aos investigadores duas vertentes fundamentais: a) a recolha de dados e informações através de pesquisas que ajudem a orientar o seu trabalho e b) a análise dos dados sobre as próprias pesquisas realizadas pelos utilizadores. Esta dualidade proporciona uma perspetiva abrangente e multifacetada para os académicos, enriquecendo significativamente o processo de investigação.

Salientamos que neste capítulo não são abordados especificamente motores de pesquisa especializados, embora grande parte das técnicas apresentadas na primeira parte do artigo possam ser utilizadas também nestas ferramentas de busca.

Recolha de informação estruturada e focada. Motores de pesquisa generalistas

Embora não sejam ferramentas exclusivamente académicas, motores de pesquisa como o Google podem desempenhar um papel importante nas fases iniciais da investigação académica. Eles oferecem um ponto de partida valioso para investigadores, docentes e estudantes, facilitando a exploração preliminar de tópicos e a descoberta de recursos diversificados. Por exemplo, um estudante a iniciar um trabalho sobre energias renováveis pode utilizar o Google para encontrar artigos de divulgação científica, relatórios governamentais e *site*s de organizações relevantes, proporcionando assim um contexto inicial para o seu trabalho.

Estas ferramentas podem também ajudar na identificação de termos-chave e conceitos relacionados com o tema em estudo. Por exemplo, um investigador a explorar um novo campo na área da inteligência artificial pode utilizar o Google para descobrir terminologia específica e áreas de aplicação emergentes, orientando assim as suas pesquisas subsequentes em bases de dados académicas mais especializadas.

Os motores de busca também são úteis para localizar recursos não tradicionais que podem enriquecer a investigação. Um historiador, por exemplo, pode encontrar blogues de especialistas, fóruns de discussão ou arquivos digitais que oferecem perspetivas únicas ou fontes primárias relevantes para o seu trabalho. Adicionalmente, podem facilitar a identificação de instituições, investigadores e grupos de investigação que trabalham um determinado campo. Um estudante de mestrado em biologia marinha pode, por exemplo, utilizar o Google para descobrir laboratórios importantes para o seu trabalho ou projetos de investigação em curso, podendo levar a oportunidades de colaboração ou a recursos académicos mais especializados.

Os motores de pesquisa generalistas podem também ser uma ferramenta valiosa para a contextualização sociocultural de tópicos de investigação. Um sociólogo a estudar movimentos sociais contemporâneos pode utilizar o motor de busca para aceder a notícias recentes, comentários e reações públicas, proporcionando uma compreensão mais ampla do impacto social do seu objeto de estudo.

Ou seja, embora não substituam as bases de dados académicas especializadas, os motores de busca generalistas, como o Google, Bing, ou DuckDuckGo, podem ser ferramentas importantes nas fases iniciais da investigação académica. Poderão ajudar na exploração preliminar de tópicos, na descoberta de recursos diversos e na contextualização de temas de investigação, estabelecendo assim uma base sólida para pesquisas mais aprofundadas.

A eficácia destas ferramentas reside na sua capacidade de indexar e categorizar informações de forma rápida e precisa, facilitando a localização de estudos relevantes, dados estatísticos e outras informações pertinentes para a investigação em curso.

Para otimizar o uso destes motores, os investigadores devem dominar técnicas e estratégias avançadas de pesquisa, como por exemplo: a utilização de operadores booleanos (AND, OR, NOT); a pesquisa por frases exatas usando aspas; ou a filtragem por tipo de ficheiro, domínio ou data de publicação — como veremos mais adiante.

Estas técnicas permitem refinar os resultados e aceder a informações mais específicas e relevantes para o tema em estudo.

Estratégias de pesquisa eficazes

No contexto da investigação académica, a capacidade de realizar pesquisas eficazes e precisas é fundamental. Com o vasto volume de informação disponível *online*, é crucial que os investigadores dominem técnicas avançadas de pesquisa para filtrar e localizar as informações mais relevantes para os seus estudos. Uma das estratégias mais poderosas e versáteis neste âmbito é o uso de operadores booleanos.

A pesquisa booleana

Realizamos uma pesquisa booleana quando usamos os chamados operadores booleanos que não são mais do que palavras e símbolos cujo objetivo é aumentar ou restringir o resultado de pesquisas. Esse conjunto de "comandos" pode ser utilizado em quase todos os mecanismos de pesquisa, dos motores de busca às bases de dados, catálogos *online* ou *website*s que tenham a funcionalidade de pesquisa.

Neste sentido, os operadores booleanos são ferramentas essenciais que permitem aos investigadores refinar e focalizar as suas pesquisas. Os três operadores booleanos básicos são: AND, NOT e OR — que podem ser combinados com muitos outros para tornar mais fechados ou mais abertos os resultados (ver quadro 9.1).

Olhemos alguns exemplos de utilização destes operadores para o trabalho académico. Se quisermos refinar tópicos de pesquisa e, por exemplo, queremos encontrar resultados sobre "educação ambiental" AND "ensino superior". Esta pesquisa vai devolver-nos resultados que contêm ambos os termos, focando especificamente a educação ambiental no contexto do ensino superior.

Já se o nosso objetivo for a inclusão de sinónimos ou termos relacionados, podemos olhar para o exemplo "alterações climáticas" OR "aquecimento global". Esta *query* vai expandir os resultados para incluir conteúdos que usem qualquer um destes termos, capturando uma quantidade de conteúdos mais ampla.

Por outro lado, se o que queremos é excluir termos irrelevantes, por exemplo, "inteligência artificial" NOT "ficção científica". Ao usar o NOT vamos excluir os resultados relacionados com ficção científica, focando-se apenas em conteúdos sobre inteligência artificial.

Podemos ainda usar os operadores para combinações complexas. Por exemplo esta pesquisa: ("política energética" OR "transição energética") AND "Portugal" NOT "Espanha", vai devolver-nos resultados sobre política ou transição energética em Portugal, excluindo referências a Espanha.

Quadro 9.1 Operadores booleanos básicos

Operadores booleanos básicos		
AND	NOT	OR
Usado para combinar termos, garantindo que todos eles estejam presentes nos resultados.	Empregue para excluir termos específicos dos resultados da pesquisa. O NOT pode ser substituído pelo sinal menos (-), desde que fique junto (sem espaço) antes da palavra a excluir.	Utilizado para procurar um ou outro termo, ampliando o resultado da pesquisa.
Diminui o número de resultados	Diminui o número de resultados	Aumenta o número de resultados

Fonte: elaboração própria da autora.

Quadro 9.2 Operadores de pesquisa

Operadores de pesquisa		
Operador	Para que serve	Exemplo
* (asterisco)	Substitui qualquer palavra, ou parte de palavra, que esteja antes, depois ou após outra palavra, dependendo da localização do asterisco	Migra* Aqui poderá ter resultados como migrações, migrantes, etc.
() (parênteses)	Junta dois ou mais termos de pesquisa num grupo lógico e separado	(casa OR moradia) AND imobiliário A pesquisa devolve resultados que contêm a palavra "casa" ou a palavra "moradia" com a palavra "imobiliário".
“” (aspas)	Devolve resultados em que a expressão entre aspas aparece integral.	“cogito ergo sum” Os resultados da pesquisa serão apenas dos sítios onde se encontra a expressão exata pesquisada.
(euro) ou $ (cifrão)	Associa preços à pesquisa.	VW Golf 10.000 euros O resultado irá associar o VW Golf ao valor
Define	Devolve uma definição	define:saudade Vai procurar definições de "saudade"
Filetype	Procura um determinado tipo de ficheiro	filetype:pdf Pesquisa Google Vai dar-nos ficheiros PDF sobre Pesquisa Google. PDF pode ser substituído por qualquer tipo de ficheiro
Site	Procura dentro de um site	site:publico.pt iefp O resultado aqui seriam páginas do site do Público em que fosse mencionado IEFP
Related	Procura sites relacionados com o mencionado	related:publico.pt Dá-nos sites relacionados com o site do Público
Intitle/Allintitle	Procura palavras nos títulos	intitle:desinformação O resultado seria artigos ou páginas cujos títulos tivessem a palavra desinformação
Intext/Allintext	Procura resultados no texto	intext:iefp O resultado seria artigos ou páginas que tivessem a palavra desinformação nos seus textos.

Fonte: elaboração própria da autora.

Vamos agora conhecer alguns operadores de pesquisa para o Google. Aqui, além dos básicos, podemos usar ainda outros operadores que podem ser combinados (no quadro 9.2 encontra alguns).

Além dos operadores que podem ser utilizados, o Google tem disponíveis quatro locais onde pode levar a cabo pesquisas avançadas bastante focadas:

— páginas *web* e ficheiros (https://www.google.com/advanced_search);
— imagens (https://www.google.com/advanced_image_search);
— vídeos (https://www.google.com/advanced_video_search);
— livros (https://books.google.com/advanced_book_search).

Aqui, se seguir as ligações, apenas terá de ir preenchendo as caixas, afunilando a sua pesquisa.

Exercícios

Podemos experimentar operadores em pesquisas relacionadas com comunicação e jornalismo, por exemplo. Aqui ficam alguns exercícios para praticar.

Exercício de refinamento
Tema: jornalismo digital
Tarefa: refine a pesquisa para focar especificamente no uso de dados no jornalismo *online*. Solução possível: "jornalismo digital" AND ("jornalismo de dados" OR "data journalism") AND "online".

Exercício de expansão
Tema: redes sociais na comunicação
Tarefa: expanda a pesquisa para incluir diferentes plataformas de redes sociais e seu impacto na comunicação. Solução possível: ("redes sociais" OR "social media") AND (comunicação OR jornalismo) AND (Facebook OR Twitter OR Instagram OR LinkedIn OR TikTok).

Exercício de exclusão
Tema: ética no jornalismo
Tarefa: pesquise sobre ética no jornalismo, excluindo resultados relacionados com jornalismo de guerra. Solução possível: "ética" AND "jornalismo" NOT ("jornalismo de guerra" OR "correspondentes de guerra").

Exercício de pesquisa complexa
Tema: desinformação e *fake news* em períodos eleitorais
Tarefa: crie uma pesquisa que abranja vários aspetos da desinformação em eleições, focando estudos de caso europeus, mas excluindo um país específico. Solução possível: ("desinformação" OR "fake news" OR "notícias falsas") AND ("eleições" OR "campanha eleitoral") AND ("Europa" OR "União Europeia") AND (verificação OR fact-checking) NOT Reino Unido.

Exercício de pesquisa temporal
Tema: evolução da publicidade nos meios de comunicação
Tarefa: pesquise sobre a evolução da publicidade na última década, focando em meios digitais. Solução possível: ("publicidade digital" OR "marketing digital") AND ("meios de comunicação" OR "media") AND (evolução OR tendências) AND (2014.2024).

Exercício de pesquisa por formato
Tema: *podcasts* no jornalismo
Tarefa: encontre estudos académicos sobre o uso de *podcasts* no jornalismo, incluindo apenas documentos em formato PDF. Solução possível: "podcasts" AND "jornalismo" AND (investigação OR estudo OR análise) filetype:pdf.

Ao praticar regularmente com este tipo de exercícios, os investigadores podem melhorar significativamente as suas competências de pesquisa, tornando-se mais eficientes e eficazes na localização de informações relevantes para os seus estudos.

A facilidade de uso destas técnicas de pesquisa avançada não só economiza tempo valioso, mas também melhora a qualidade e a abrangência da investigação académica, permitindo descobrir fontes e perspetivas que poderiam passar despercebidas com métodos de pesquisa menos sofisticados.

Análise de dados sobre tendências de pesquisas dos utilizadores: Google Trends

A quantidade de informação que a Google recolhe tendo por base estas pesquisas é enorme, tanto em quantidade como em diversidade, estamos no domínio da *big data*. Põe-se, então, a questão da disponibilização e utilização desses dados em várias áreas, nomeadamente na investigação científica.

Em 2006, a Google lançou a Google Trends (Google Press, 2006), uma ferramenta que permitia explorar milhares de milhões de pesquisas efetuadas no Google, obtendo informações sobre padrões de pesquisa gerais ao longo do tempo. Com o passar do tempo, foram sendo acrescentadas outras funcionalidades (Rogers, 2016). Atualmente, a Google Trends permite, além da evolução temporal, o acesso a uma amostra não filtrada de pesquisas reais efetuadas através do motor de pesquisa Google. A amostra é anónima, categorizada (determinando o tópico de uma consulta de pesquisa) e agregada (agrupada), o que permite mostrar o interesse num determinado tópico em todo o mundo, por país, região do país ou até, caso haja dados suficientes, ao nível geográfico de uma cidade. Mais adiante falaremos de forma mais pormenorizada sobre as características dos dados que a ferramenta nos fornece.

Neste capítulo abordamos duas formas de aceder aos dados das tendências de pesquisa: o *website* Google Trends e a API Google Trends.

Ambas permitem aos investigadores executar as mesmas funções básicas, mas têm algumas diferenças que é importante apontar, nomeadamente em relação ao acesso. No quadro 9.3 comparam-se as duas formas de aceder.

Quadro 9.3 Comparação entre acessos à ferramenta Google Trends

	Website	API
Pagamento	Gratuito	Gratuito
Acesso	Geral	Jornalistas e investigadores académicos
Como aceder	Basta aceder através do website: https://trends.google.com/trends/	É necessário preencher um formulário de acesso *online* disponível no seguinte endereço: https://support.google.com/trends/contact/trends_api
Número de termos que é possível comparar	5	5
Especificidades	Não tem	É necessário ter conhecimento de alguma linguagem de programação como Phyton, por exemplo, e criar um projeto na consola Google Cloud
Estrutura de dados	Escala de 0 a 100 tendo em conta o resultado mais elevado	Escala de 0 a 100 tendo em conta o resultado mais elevado
Diferenças no uso	Não permite fazer *scrapping* dos dados	Permite fazer *scrapping* dos dados, facilitando o seu tratamento e compreensão

Fonte: elaboração própria da autora.

Cabe a quem investiga ponderar qual das duas opções utilizar, tendo em conta que para poder trabalhar com a API terá de se candidatar e ter já um projeto delineado, que poderá ou não ser aceite, terá de ter conhecimento de alguma linguagem de programação e deverá criar um projeto na Google Cloud.[1] Pode ser aconselhável fazer parte do Google Research Programme para ter acesso a API de várias ferramentas Google, que podem vir a ser úteis para as investigações que pretende levar a cabo.[2] Salienta-se ainda que há várias API não oficiais, que têm vindo a ser utilizadas para fazer *scrapping* dos dados da Google Trends, mas aqui apenas são abordadas ferramentas oficiais da Google. Uma das principais razões para não o fazer é que, a qualquer momento, estas API podem deixar de ter acesso aos dados, com todas as consequências que isso pode trazer a uma investigação em curso ou prestes a começar (estas consequências não serão novas para aqueles que trabalham com dados provenientes de plataformas *online*, por isso fica também ao critério de quem investiga se as quer ou não utilizar). Além disso, há ainda que ter em conta as questões relacionadas com o uso de API de terceiros, para tratamento e utilização de dados.

A Google Trends tem aberto as portas a vários tipos de investigações científicas, em áreas como a saúde (Díaz *et al.*, 2023; Lippi, Nocini e Henry, 2022; Pelat *et al.*, 2009), a economia (Cho e Varian, 2012; Orastean *et al*., 2024; Garcia e Schweitzer, 2015) as ciências sociais (Hyndman e Athanasopoulos, 2013; Seung-Pyo *et al.*, 2024), entre outras (Fornaro e Wolf, 2017). Alguns dos artigos publicados têm por base dados retirados da Google Trends, mas também há um número considerável de análises relacionadas com questões metodológicas (Cebrián e Domenech, 2023,

1 https://cloud.google.com/appengine/docs/standard/python3/building-app/creating-gcp-project?hl=pt-br

2 https://transparency.google/researcher-engagement/#learn-researcher-program

2024), tentando perceber qual o melhor caminho para a utilização de dados. Alguns dos primeiros usos foram realizados na área da epidemiologia (ainda antes da pandemia de covid-19 se ter abatido sobre o planeta).

Como veremos, mais adiante, a ferramenta também possibilita retirar o tipo de dados, já mencionados para o YouTube, uma vez que o YouTube é uma plataforma do universo Alphabet/Google.

Ferramenta: Google Trends website

Como mencionado anteriormente, a Google Trends é uma ferramenta disponibilizada de forma gratuita a qualquer utilizador com *internet*. Vamos agora saber um pouco mais sobre os dados aos quais podemos aceder.

Caracterização dos dados

Os dados da Google Trends mostram as pesquisas que as pessoas fazem diariamente no Google, mas não com números absolutos, e esta é uma característica que tem de ser tida em conta aquando do seu uso. A Google salienta que os dados disponibilizados fazem parte de uma amostra considerada representativa, anonimizada e categorizada. A necessidade de trabalhar com amostragem deve-se ao volume de pesquisas ser tão elevado que seria necessário muito tempo de processamento para dar uma resposta rápida. "Através da amostragem de dados, podemos analisar um conjunto de dados representativo de todas as pesquisas do Google e, ao mesmo tempo, encontrar informações que podem ser processadas poucos minutos após a ocorrência de um evento no mundo real" (Google Trends FAQ).[3] Através da Google Trends podemos aceder a dois tipos de amostras, tendo por base o espectro temporal:

— dados em tempo real — amostra que abrange os últimos sete dias;
— dados que não são em tempo real — uma amostra cujo espectro temporal de análise se situa entre 2004 e até 72 horas antes da pesquisa que está a ser realizada (esta amostra é separada da pesquisa em tempo real porque a amostra em tempo real apenas tem em conta os sete dias anteriores à pesquisa que está a ser efetuada).

A Google Trends apenas inclui dados para termos populares, o que quer dizer que os termos de pesquisa com baixo volume acabam por ser filtrados e não aparecem. Outro processo muito importante que devemos ter em conta é que os dados de pesquisa são normalizados. Estes são classificados de forma que a redundância seja diminuída e se eliminem anomalias, o que permite evitar erros e garantir a consistência e integridade de dados, para facilitar as comparações entre termos.

3 https://support.google.com/trends/answer/4365533?hl=en

Segundo a Google, os resultados da pesquisa são normalizados usando a hora e o local de uma consulta, dividindo cada ponto de dados pelo total de pesquisas da área geográfica e do intervalo de tempo em questão, para poder comparar aquilo que é apelidado de "popularidade relativa". Este processo é necessário para evitar que os locais onde há maior volume de pesquisas sejam sempre melhor classificados. Os números obtidos são organizados num intervalo de 0 a 100, tendo por base a proporção de um tópico em relação às pesquisas sobre todos os tópicos, num dado local e intervalo temporal, obtendo assim o Índice de Volume de Pesquisas, uma medida relativa da popularidade de um termo (Cebrián e Domenek, 2024).

Limitações dos dados

Apesar de, como já referido, os dados apresentados pela ferramenta refletirem as pesquisas feitas diariamente no Google, elas podem dizer respeito a atividades de pesquisa irregular, como pesquisas automatizadas que possam estar associadas a tentativas de *spam*. Segundo a Google, embora a ferramenta tenha mecanismos para detetar e filtrar atividades irregulares, e estas pesquisas podem ser retidas no Google Trends como medida de segurança, "filtrá-las do Google Trends ajudaria os autores dessas pesquisas a perceber que as identificámos". Isto tornaria mais difícil manter essa atividade filtrada noutros produtos da pesquisa Google, onde os dados de pesquisa de alta-fidelidade são fundamentais" (Google Trends FAQ).[4] Tendo esta questão em conta, a Google aconselha a que se tenha em conta que os dados da Trends não são um espelho perfeito da atividade de pesquisa, o que poderá influenciar o uso e conclusões retiradas de dados provenientes desta ferramenta.

É, pois, de suma importância destacar que os dados resultantes da Google Trends não devem ser confundidos com dados de sondagens, devem ser sempre considerados como um ponto de dados entre outros antes para que se possam tirar conclusões válidas do ponto científico.

Formato dos dados

Os dados resultantes das *queries* de pesquisa podem ser descarregados em CSV e abertos com qualquer *software* de folha de cálculo, ou ser carregados diretamente para ferramentas de análise e visualização de dados, que o permitam, como Tableau, Flourish, Infogram, Power BI, Raw Graphs, Datawrapper entre outros.

Permissões

Todas informações e dados que constam da Google Trends podem ser utilizados e reutilizados, desde que sejam atribuídas as informações ao Google através de citação, mencionando a fonte: Google Trends. É ainda salientado que estes usos estão sujeitos aos termos de utilização da Google.

---

4 https://support.google.com/trends/answer/4365533?hl=en

É também possível partilhar uma pesquisa na Google Trends, bem como partilhar um gráfico nas redes sociais, através de *email*, etc., ou fazer o *embed* de um gráfico num *site* (esta funcionalidade não está disponível para todos os gráficos).[5]

Aceder e navegar na ferramenta

Quanto entra no *site* da Google Trends (https://trends.google.com/trends/) encontra o logótipo acima do lado esquerdo, e o menu, composto pelos seguintes campos:

— Home (Início).
— Explore (Explorar).
— Trending now (Tendências atuais).
— Portugal (worldwide ou outro país dependendo da localização atual e se está ou não ligado a uma conta Google, no seu navegador).

No entanto, do lado esquerdo do logótipo encontra um pequeno menu colapsado. Se abrir vai encontrar seis itens, os três primeiros itens que encontra são os mesmos que no menu principal, mas há mais três, a saber:

— Year in search (o ano em pesquisa).
— Help (Ajuda).
— Send feedback (enviar *feedback*).

Vamos descobrir o que está em cada um deles.

"Home"

Aqui temos uma caixa de pesquisa onde podemos escrever uma *search query*, carregar em *enter* ou em "Explore" e passamos automaticamente para a mesma página que aparece quando carregamos no item "Explore". Nesta primeira página há também alguma informação sobre o que está a ser mais pesquisado, com informação cruzada com o Google News, mostrando notícias associadas a termos de pesquisa, para dar algum contexto (atenção que, por vezes, o contexto dado por estas notícias pode não ser o mais adequado). Na *homepage* pode ainda subscrever a *newsletter* diária que dá informação sobre as tendências de pesquisa daquele dia.

"Explore"

Ao clicar em "Explore" (explorar), tanto no menu principal, como no menu colapsado, vai deparar com uma caixa onde deverá pôr os seus termos de pesquisa.

5 *Embed* vem do inglês e significa "incorporar", neste caso esta funcionalidade permite, através do código HTML fornecido, incorporar conteúdo de terceiros num *site*, neste caso o gráfico ou lista da Google Trends.

Abaixo desta caixa está um outro menu com os seguintes itens, colapsados:

— País
Aqui vai encontrar opções sobre a geografia da sua pesquisa, pode escolher um país ou "worldwide".
— Espectro temporal
Aqui estão as opções de espectro temporal, a saber: última hora; últimas 4 horas; último dia; últimos 7 dias; últimos 30 dias; últimos 90 dias, últimos 5 anos; entre 2004 e hoje; e no final "Custom time range" (onde temos a divisão entre "Arquivo" e "Semana passada". No primeiro, podemos ter um espectro temporal de pesquisa, desde 2004 até à atualidade, no segundo podemos definir um espectro temporal da semana anterior, incluindo dias e horas). Atenção que por defeito está sempre nos "Past 12 months" e nunca confundir com ano passado, é para os últimos 12 meses.
— Categorias
Dentro das categorias podemos afunilar as pesquisas, tendo em conta uma categorização prévia feita pela Google que inclui variados temas, a lista é muito extensa, mas vai do entretenimento, passando pelas finanças, até aos jogos e saúde, cada uma delas com subcategorias também. Mas devemos perceber que aqui já temos mais uma camada de filtros que podem, de certa forma, condicionar mais a nossa pesquisa.
— Web Search
Neste separador, podemos escolher se queremos saber os padrões de pesquisa no Google "geral", no Google Image Search, no Google News, no Google Shopping ou no YouTube.

"Trending Now"

Em setembro de 2024 esta opção sofreu mudanças significativas, permitindo uma forma mais rápida e fácil de compreender as tendências de pesquisa atuais.

As tendências de pesquisa deixam de ser apresentadas como um *top* 20 das mais pesquisadas, para serem visualizadas numa consola com bastantes possibilidades de escolha quanto à visualização dos dados, bem como quanto ao seu *download* ou cópia para *clipboard*.

A página passou a ter um menu superior com quatro itens em que pode filtrar a informação por país, espectro temporal, tendências e relevância.

No que respeita ao espectro temporal pode filtrar pelas últimas 24 horas (este é o tempo que está predefinido), últimas 4 horas, últimas 48 horas e últimos 7 dias. Em relação às tendências, estão predefinidas "All trends", mas pode escolher ver apenas as que estão ativas no momento da sua consulta ("Show active trends only"). Além disso é agora possível ver a expressão ou palavra pesquisada, bem como o seu volume de pesquisa aproximado, mostrando se a sua tendência de pesquisa está a aumentar ou a diminuir, há quanto tempo se mantém nas tendências de pesquisa e se ainda está a ser pesquisada, e mostra ainda as várias consultas que são variantes da mesma pesquisa ou consideradas relacionadas.

Estes tópicos de pesquisa são apresentados em conjunto com notícias para ajudar a contextualizar o possível aumento nas pesquisas, dando algum contexto aos dados, no entanto, como já foi salientado, é preciso que o investigador olhe para estas pistas com sentido crítico, cruzando com outras fontes de informação para compreender melhor o possível contexto.

O utilizador pode clicar no tópico e visualizá-lo numa coluna que vai abrir do lado direito onde encontra um gráfico que mostra a evolução temporal da pesquisa, os itens relacionados, as notícias de contextualização e a possibilidade de ir para a consola "Explore", onde poderá ficar com uma ideia da distribuição geográfica das tendências de pesquisa, da sua evolução nos últimos sete dias, bem como dos tópicos relacionados e das *queries* relacionadas.

"Country"

Aqui pode mudar a localização ou país, clicando na seta (triângulo invertido) que mostrará uma lista de países.

"Year in Search"

Neste separador tem o resultado das tendências de pesquisa por ano, e por tópicos, como personalidades, pesquisas que começam por "Como.." ou "O que..", entre outros. Aqui também pode mudar o país e mudar o ano que deseja consultar, desde 2011 até ao ano anterior àquele em que se encontra.

Funcionalidade "Explore"

Através da funcionalidade "Explore" podemos perceber as tendências de pesquisa de um ou mais termos (em comparação) ao longo do tempo, bem como de que forma se comportam estas pesquisas do ponto de vista geográfico e também quais os termos de pesquisa relacionados.

Para recolher os dados sobre como se comporta um interesse de pesquisa num determinado espectro temporal, é preciso escolher, nos menus, o separador "Explore". Aqui pode, além de explorar as pesquisas da *internet* em geral ("Web search"), também saber quais as tendências de pesquisa em diferentes áreas da Google, como o News Search, o Image Search, o Google Shopping ou o YouTube. As formas de recolha e a estruturação dos dados são iguais para todos, apenas temos de escolher no separador acima mencionado qual a área sobre a qual a nossa pesquisa incide.

Recolher dados

Quanto estiver no separador "Explore", assim que começar a escrever a sua *query* (com a palavra ou expressão que quer pesquisar) vai aparecer, em baixo, uma lista de tópicos (como acontece com as sugestões Google no motor de pesquisa) que o algoritmo da Google Trends acredita que o utilizador poderá estar a tentar pesquisar.

Essas sugestões têm em baixo uma classificação, algumas terão "Search term" (termo de pesquisa) e outras terão "Topic" (tópico). Caso a *query* sobre o qual está a pesquisar apareça na lista, clique nela para obter os resultados, caso não esteja, opte por utilizar o que diz "Search term" que deverá corresponder à sua *query*. No entanto, se, além da sua *query*, também aparecer a categorização "Topic" é aconselhável que use os tópicos sempre que possível, uma vez que a própria Google os considera mais fiáveis, para pesquisas na Google Trends, pois incluem o termo exato, assim como possíveis erros ortográficos de quem pesquisa e acrónimos. Mas atenção! Pode comparar até cinco *queries* diferentes, mas nunca deve comparar *queries* dos tópicos com *queries* dos termos de pesquisa, uma vez que são agrupados de forma diferente e isso pode trazer diferenças nos resultados e enviesar a comparabilidade. Assim, efetue sempre as suas comparações com categorias do mesmo tipo.

Experimente: se pesquisar um termo de pesquisa e tópico com a mesma *query* vai ver que os resultados serão diferentes.

Além da escolha da sua *query,* não se esqueça de enquadrar a sua pesquisa tendo em conta o menu principal, onde pode escolher o país sobre o qual quer as informações, o espectro temporal da sua pesquisa, e se o que quer são dados sobre o Google "geral", ou outra categorização, como por exemplo o YouTube.

Após definir os parâmetros da sua pesquisa aparecerão os resultados que vão estar estruturados da seguinte forma:

— primeiro gráfico — "Interest over time" (interesse ao longo do tempo) mostra a evolução temporal das tendências de pesquisa escolhidas;
— segundo gráfico — "Interest by sub-region" (interesse por sub-região). Abaixo da evolução temporal, vai encontrar um mapa de Portugal (caso tenha sido essa a sua escolha), com o interesse dividido por sub-região (neste caso corresponde aos distritos portugueses), mas pode também escolher, clicando no pequeno menu do canto superior esquerdo dessa moldura, a cidade (tendo em conta que aqui só encontra algum resultado ou representação caso os números de pesquisa sejam consideráveis, não sabendo a partir de quanto isso acontece);
— terceiro gráfico — "Search topics" (tópicos de pesquisa). Abaixo do mapa do lado direito encontra uma lista que mostra outros tópicos de pesquisa que os utilizadores que pesquisaram o tópico em análise também pesquisaram. Aqui podemos ver alguma afinidade nas pesquisas realizadas, entre temas ou formulações dos tópicos de pesquisa.

Em cada uma destas áreas tem, em cima à direita, várias opções, a primeira permite descarregar os dados em formato CSV, a segunda permite fazer o *embed* do gráfico ou lista num *website*, podendo escolher se quer que continue a ser feita automaticamente a atualização dos dados, e o terceiro permite escolher a partilha direta no Facebook, X (ex-Twitter), LinkedIn ou Tumblr.

Para efeitos de investigação, o que nos interessa é o descarregamento dos dados em CSV que podem ser facilmente transformados em Excel ou outro formato com o qual precisemos trabalhar, uma vez que pode ser lido por máquinas. Ou seja,

é um formato que pode ser importado ou lido por um sistema informático para processamento posterior sem intervenção humana, assegurando simultaneamente que não se perde qualquer significado semântico.

Explorar os dados

Vamos ver como podemos explorar estes dados, funcionalidade a funcionalidade, e o que nos oferecem os resultados.

Interest over time

O que nos dizem estes números?

Os números que aparecem no gráfico da evolução temporal representam o interesse da pesquisa em relação ao ponto mais alto do gráfico para a região (neste caso país) e o espectro temporal escolhido. Um valor de 100 é o pico de popularidade do termo. Um valor de 50 significa que o termo tem metade da popularidade. Uma pontuação de 0 significa que não existem dados suficientes para este termo, ou seja, pode haver pesquisas, mas não em número suficiente para serem consideradas pela ferramenta.

Interest by sub-region (interesse por sub-região)

O que nos dizem esses números?

Aqui conseguimos saber em que região (no caso de Portugal, o distrito) onde o termo foi mais popular durante o período especificado. Aqui, os valores "são calculados numa escala de 0 a 100, em que 100 é a localização com maior popularidade como uma fração do total de pesquisas nessa localização, um valor de 50 indica uma localização que é metade da popularidade. Um valor de 0 indica uma localização onde não existem dados suficientes para este termo".[6]

É importante ter em conta que termos um valor mais elevado significa que este termo tem uma proporção de pesquisa mais elevada entre todas as outras pesquisas, mas não uma contagem absoluta mais elevada. Como é explicado na própria ferramenta, num "pequeno país onde 80% das consultas são para "bananas" obterá o dobro da pontuação de um país gigante onde apenas 40% das consultas são para "bananas".[7]

No entanto, apesar das limitações estes são números interessantes para aferir os interesses e tendências de pesquisa num determinado país, por região. E porque estamos a falar de proporções, ajuda a perceber melhor a relevância destas pesquisas, em relação a todas as outras realizadas.

---

6 Descrição da Google Trends, no *website*: https://support.google.com/trends/answer/4355212.
7 Descrição da Google Trends, *website*: https://support.google.com/trends/answer/4355212

Search Topics (tópicos de pesquisa) e Search queries (consultas de pesquisa)

O que nos dizem estes números?

Tanto no campo "Search topics", como no espaço "Search queries" poderá escolher, em cada um deles num pequeno separador em cima, à direita entre:

— "Top"
— "Rising" (em ascensão)

Ao escolher o separador "Top" vão ser mostrados os tópicos ou *queries* (dependendo de onde se encontre) mais populares. Aqui, a pontuação está numa escala relativa em que um valor de 100 é o tópico mais procurado e um valor de 50 é um tópico procurado com metade da frequência do termo mais popular, e assim por diante.

Caso escolha o separador "Rising", verá os tópicos relacionados com o maior aumento na frequência de pesquisa desde o último período de tempo. Os resultados marcados como "Breakout" tiveram um grande aumento, provavelmente porque estes tópicos são novos e tiveram poucas (ou nenhuma) pesquisas anteriores.

Na prática

Em 2020, o MediaLab-Iscte levou a cabo uma análise da presença do tema "Coronavírus"/"Covid-19" nas pesquisas Google, nas notícias e nas redes sociais, com maior ênfase nos 30 dias anteriores à confirmação oficial dos primeiros casos de covid-19, em Portugal, que ocorreu a 2 de março (Cardoso *et al.*, 2020).

A Google Trends foi a ferramenta usada para proceder à análise das pesquisas. Numa primeira abordagem, foi realizada uma análise ao relatório diário (menu "Trendind now" > "Daily search trends") entre 1 de fevereiro e 2 de março, que mostra o *top* 20 dos termos mais pesquisados em Portugal. A partir daí passou a analisar-se diariamente este *top* 20, procurando por temas de pesquisa ligados à covid-19. Com base nestes temas/palavras, construiu-se uma lista e passou a fazer-se o acompanhamento destas palavras através da funcionalidade "Explore", o que permite perceber a sua evolução ao longo tempo. Por exemplo, percebeu-se que dos 30 dias analisados apenas houve pesquisas relacionadas com o coronavírus em 13 desses dias.

Resumindo, através da análise foi possível determinar a evolução do interesse no tema, ao longo do tempo, quais as palavras e expressões mais usadas para fazer as pesquisas e também perceber que a informação que passava nas redes sociais e nos *media* tradicionais também acabava por influenciar as pesquisas. Por exemplo, quando eram notícias de um novo caso, o número de pesquisas pelos casos também aumentava. Foram também observadas mudanças nos padrões de pesquisa a partir de uma determinada data, passando do interesse inicial em saber de casos da doença para saber quais os "sintomas coronavírus" ou "máscaras de proteção", dando uma clara noção das preocupações e dos interesses informativos em relação à pandemia.

Este é um caso muito simples de uso da ferramenta para aferir interesse num determinado tema, que pode também ser usado para, por exemplo, perceber de que forma os utilizadores pesquisam sobre determinados assuntos e a partir de onde o fazem.

Referências bibliográficas

Bianchi, T. (2024), "Market share of leading search engines worldwide from January 2015 to January 2024 [Chart]", *Statista*, disponível em https://www.statista.com/statistics/1381664/worldwide-all-devices-market-share-of-search-engines/.

Cardoso, G., A. Pinto-Martinho, I. Narciso, J. Moreno, M. Crespo, N. Palma, e R. Sepúlveda (2020), "Informação e desinformação sobre o coronavírus nas notícias e nas redes sociais em Portugal", MediaLab Iscte, disponível em https://medialab.iscte-iul.pt/o-tema-coronavirus-nos-media-e-nas-redes-sociais.

Cebrián, E., e J. Domenech (2023, "Is Google Trends a quality data source?", *Applied Economics Letters*, 30 (6), pp. 811-815, https://doi.org/10.1080/13504851.2021.2023088.

Cebrián, E., e J. Domenech (2024), "Addressing Google Trends inconsistencies", *Technological Forecasting and Social Change*, 202, 123318, https://doi.org/10.1016/j.techfore.2024.123318.

Choi, H., e H. Varian (2012), "Predicting the present with Google Trends", *Economic Record*, 88 (S1), pp. 2-9, https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Díaz F., P. Henríquez, N. Hardy, D. Ponce (2023), "Population well-being and the COVID-19 vaccination program in Chile: evidence from Google Trends", *Public Health*, 219, pp. 22-30, Doi:10.1016/j.puhe.2023.03.007.

Fornaro, P., e M. Wolf (2017), "Nowcasting tourism demand with Google Trends", *Annals of Tourism Research*, 62, pp. 1-11, doi:10.1002/jtr.2137.

Google Press. (2006), "New Google Search technologies make information easier to discover, organize and share", disponível em https://googlepress.blogspot.com/2006/05/new-google-search-technologies-make_10.html.

Hyndman, R. J., e G. Athanasopoulos (2013), "Forecasting with Google Trends: the role of search queries as predictors of future service usage", *International Journal of Forecasting*, 29 (4), pp. 604-614, doi:10.2139/ssrn.1659302.

Lippi G., R. Nocini, B.M. Henry (2022), "Analysis of online search trends suggests that SARS-CoV-2 omicron (b.1.1.529) variant causes different symptoms", *J. Infrect*., 84, pp. 76-77, doi:, 10.1016/j.jinf.2022.02.011.

Lolic, I., M. Matošec, e P. Soric (2024), "DIY Google Trends indicators in social sciences: a methodological note", *Technology in Society*, 77, 102477, doi:0.1016/j.techsoc.2024.102477.

Orastean, R., S.C. Marginean, e R. Sava (2024), "Exploring the relationship between Google Trends and cryptocurrency metrics", *Studies in Business and Economics*, 19 (1), pp. 368-379, https://doi.org/10.2478/sbe-2024-0020.

Pelat, C., C. Turbelin, A. Bar-Hen, A. Flahault, e A.-J. Valleron (2009), "More diseases tracked by using Google Trends", *Emerg. Infect. Diseases*, 15 (8), pp. 1327-1328, doi:10.3201/eid1508.090299.

Petrosyan, A. (2024), "Number of internet users worldwide from 2005 to 2023 [Chart]", *Statista,* disponível em https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/.

Prater, M. (2024), "31 Google search statistics to bookmark ASAP", *Hubspot*, disponível em https://blog.hubspot.com/marketing/google-search-statistics#how-many-people-use-google.

Rogers, S. (2016), "What is Google Trends data — and what does it mean?", *Medium, Google News Lab*, disponível em https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8.

Seung-Pyo, J., S. Y. Hyoung, e C. San (2018), "Ten years of research change using Google Trends: from the perspective of big data utilizations and applications", *Technological Forecasting and Social Change*, 130, pp. 69-87, doi: 10.1016/j.techfore.2017.11.009.

Capítulo 10

Plataformas de mensagens
Comunicação, comunidade e partilhas

Inês Narciso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

As plataformas de mensagens, como o Telegram e o WhatsApp, vieram alterar a forma como usamos os nossos *smartphones* para comunicar, com outra pessoa ou em grupo, ao permitirem a combinação de mensagens instantâneas, reações com *emojis* e a partilha de conteúdos multimédia (*clips* de voz, chamadas de vídeo e envio de fotos, vídeos e *links*).

Neste capítulo focamo-nos no Telegram e no WhatsApp, pela sua dimensão no mercado europeu e crescente relevância como objeto de investigação académico (Baulch, 2024). Tanto o Telegram como o WhatsApp têm as funcionalidades necessárias para a troca de mensagens, tanto nas comunicações individuais como nas de grupo, mas cada uma tem características particulares que atendem a necessidades diversas dos utilizadores.

O WhatsApp, lançado em 2009 e adquirido pelo Facebook (hoje, Meta Platforms) em 2014, é conhecido pela sua simplicidade e ampla utilização (Johns *et al.*, 2024). Em 2023, o WhatsApp contava com mais de 2 mil milhões de utilizadores mensais ativos em todo o mundo, sendo a principal ferramenta de comunicação em mais de 180 países, incluindo Portugal (Dixon, 2023). A sua popularidade é particularmente notável em regiões como a América Latina, Europa e Índia, onde é essencial para comunicações pessoais e profissionais.

O Telegram, fundado em 2013 por Nikolai e Pavel Durov, apresenta-se como uma plataforma alternativa às oferecidas pelos gigantes tecnológicos de Silicon Valley. Tem-se desenvolvido, sobretudo, como espaço de comunicação comunitário, possuindo amplas funcionalidades para conversas em grupo, e tendo sido pioneiro na criação de canais e supergrupos que podem hospedar milhões de utilizadores (Herrero-Solana e Castro-Castro, 2022). O Telegram tem vindo a crescer significativamente, atingindo os 700 milhões de utilizadores mensais ativos em 2023, impulsionado pelas funcionalidades de proteção da privacidade que oferece, bem como pela sua capacidade de lidar com comunicações em grande escala (Ceci, 2023). Esta plataforma é mais utilizada em regiões e comunidades com maiores preocupações com a privacidade, e entre utilizadores com conhecimentos de tecnologia, que dão prioridade à segurança nas suas comunicações. Estas características também tornaram a plataforma bastante ativa para comunidades

tradicionalmente reportadas e banidas das redes sociais (Rogers, 2020; Gerster *et al.*, 2022)

Tanto o WhatsApp como o Telegram garantem encriptação ponto a ponto nas suas comunicações.[1] Ambas as plataformas operam sob um modelo *freemium*, onde os serviços principais são gratuitos, mas recursos adicionais são disponibilizados mediante o pagamento de uma taxa. As duas plataformas têm três principais formatos de comunicação:

1. *chat* direto— comunicação entre dois utilizadores;
2. grupos — comunicação entre um grupo de utilizadores;
3. canais — comunicação unidirecional de um utilizador para um conjunto ilimitado de seguidores.

Adicionalmente, o WhatsApp apresenta o formato Comunidades, um conjunto de grupos interligados sob o chapéu de uma comunidade. O administrador da comunidade pode fazer comunicação unidirecional para todos os membros.

Apesar das semelhanças na sua estrutura base, são várias as características que diferenciam o WhatsApp do Telegram, tal como é destacado no quadro 10.1.

Do ponto de vista académico, estas plataformas de mensagens oferecem uma fonte de dados para o estudo da comunicação e do impacto das tecnologias digitais nas interações sociais. Certas investigações académicas debruçam-se sobre os protocolos de encriptação utilizados por estas plataformas e o seu papel na formação da opinião pública sobre privacidade (Pierson, 2021). O uso extensivo do WhatsApp e do Telegram na organização de movimentos sociais e em ambientes empresariais tem também sido objeto de interesse (Chagas *et al.*, 2022), assim como a desinformação, a criminalidade e os extremismos, presente nestes espaços não moderados e mais fechados que as redes sociais (Cardoso *et al.* 2022; Rogers, 2020).

As diferenças funcionais entre o Telegram e o WhatsApp moldam também a investigação académica. O Telegram, que tem uma API aberta e extensas funcionalidades de pesquisa, oferece aos investigadores mais liberdade na recolha de dados (Golovnin *et al.*, 2023). Nos últimos anos, foram desenvolvidas diversas ferramentas de análise de dados do Telegram, em resultado desta abertura (Baulch, 2024). As diversas funcionalidades permitem também, por exemplo, utilizar as funcionalidades do *bot* do Telegram para divulgar pesquisas ou recolher dados em tempo real, aumentando a eficiência da pesquisa e o envolvimento dos participantes. No geral, a fácil acessibilidade aos canais e grupos do Telegram oferece uma base rica para o estudo das práticas de comunicação de massa e de disseminação de informação, particularmente em movimentos políticos ou sociais (Rossini, 2023).

Em contraste, a API fechada do WhatsApp e as rigorosas políticas de privacidade de dados limitam significativamente a recolha direta de dados, levando os

1 Método de segurança que garante que as mensagens enviadas entre dois utilizadores sejam codificadas de tal maneira que apenas o emissor e o recetor possam decifrá-las. Esta técnica assegura que, mesmo que a mensagem seja intercetada durante a transmissão, não possa ser lida por ninguém, incluindo a própria plataforma que facilita a comunicação.

Quadro 10.1 Algumas das principais diferenças entre o WhatsApp e o Telegram [*]

Característica	WhatsApp	Telegram
Recursos dentro das conversas	Integração muito limitada de *chat-bots*; possibilidade de fazer sondagens; partilha de vídeos até 100 MB e ficheiros até 2 GB; partilha de localização; partilha de *status*; não permite edição de mensagens.	Integração extensa de *chat-bots*; introdução recente de sondagens; partilha de ficheiros até 2 GB; partilha de localização; não oferece partilha de *status*; permite edição de mensagens.
Grupos — acessibilidade	Podem ser acedidos via *link* (caso essa funcionalidade esteja ativa) ou por convite dos membros (caso essa funcionalidade esteja ativa) ou dos administradores do grupo.	Podem ser acedidos via *link* (caso essa funcionalidade esteja ativa), por convite dos membros e administradores ou juntando-se diretamente após uma pesquisa.
Grupos — limite de utilizadores	Até 1024 membros. O WhatsApp permite também o agrupamento de até 100 grupos dentro de uma comunidade, com um limite de 2000 membros por comunidade.	Até 200 mil membros para os supergrupos.
Grupos — descoberta	Os grupos ou comunidades no WhatsApp não podem ser pesquisados dentro da própria aplicação. A única forma de entrar num grupo é receber um convite ou ter *link* do grupo. Dentro de uma comunidade, podem ser visíveis os outros grupos que a compõem.	Os grupos públicos são pesquisáveis dentro da aplicação, o que permite encontrar grupos com base em interesses ou tópicos.
Grupos — privacidade	Todas as mensagens em grupos do WhatsApp são encriptadas, garantindo que as comunicações sejam legíveis apenas pelos membros do grupo. É possível exportar uma conversa de grupo.	O Telegram oferece dois tipos de grupos — os grupos regulares e os *chats* secretos. As mensagens dos grupos regulares não são encriptadas ponto a ponto, ao passo que os *chats* secretos são. É possível exportar uma conversa de grupo.
Canais — características base	Os canais diferem dos grupos, por serem espaços de comunicação unidirecionais, em que o/s administrador/es conseguem enviar mensagens a um largo número de seguidores / subscritores.	
Canais — acessibilidade	Inicialmente apenas por convite. Os participantes precisam de um *link* para aceder ao canal.	Qualquer pessoa consegue entrar num canal público do Telegram. O acesso pode ser feito via *link* ou diretamente na aplicação. O acesso a canais privados só pode ser feito via link e poderá estar dependente de autorização do administrador.
Canais — limite de utilizadores	Sem limite	
Canais — descoberta	É possível encontrar alguns canais através da aplicação, no separador "atualizações". Se o canal for privado, depende da partilha do *link* pelo administrador.	O/s administrador/es do canal pode/m definir um canal como público, tornando-o pesquisável dentro da aplicação.
Canais — privacidade	Todos as mensagens são encriptadas, ponto a ponto. Não é possível exportar, através da plataforma, a comunicação de um canal.	Só os canais privados são encriptados. É possível exportar, através da plataforma, a comunicação de um canal.

[*] À data da publicação deste manual.

Fonte: elaboração própria da autora.

investigadores a depender mais frequentemente de autorrelatos dos participantes e de metodologias qualitativas, tais como entrevistas e grupos focais, para estudar a plataforma (Kligler-Vilenchik, 2021). Nos últimos anos, a necessidade criada pela adoção generalizada do WhatsApp a nível global tem fomentado o desenvolvimento de métodos de recolha não dependentes da API, e uma maior discussão sobre os dilemas éticos associados ao uso de meios alternativos (Baulch *et al.*, 2024; Piaia *et al.*, 2022).

Recolha de dados

Quando construímos um desenho de pesquisa que prevê a recolha de dados de plataformas de mensagens, há que ter em conta os desafios no acesso aos dados, bem como eventuais questões éticas. Isto aplica-se particularmente ao WhatsApp, por se tratar de uma plataforma fechada e totalmente encriptada. Há também que considerar as especificidades do uso da plataforma no contexto em estudo: em Portugal, por exemplo, o uso do WhatsApp é bem mais generalizado do que o uso do Telegram, pelo que a segunda plataforma, apesar de oferecer menos barreiras ao acesso e recolha, poderá não responder satisfatoriamente à questão central em investigação por falta de dados ou de representatividade. Nesse sentido, o primeiro passo no estudo destas plataformas de mensagem é a identificação de espaços de recolha.

Identificar espaços de recolha

Para o Telegram, tal como explicado supra, é possível pesquisar por palavras-chave, por grupos e canais, que são depois passíveis de serem exportados da plataforma. O investigador pode, assim, pesquisar pelo tema que lhe interessa e identificar espaços de recolha diretamente na aplicação.

Já o WhatsApp não permite pesquisas por grupos ou comunidades diretamente na sua plataforma. Permite a pesquisa por canais, mas não permite a sua exportação, pelo que terá de se recorrer a uma ferramenta de terceiros para recolha de dados. Por questões de privacidade, esse tipo de ferramentas não é explorada no presente manual, como é descrito infra. Assim, para acesso e identificação de grupos e/ou comunidades para recolha de dados no WhatsApp, sugerem-se alguns dos seguintes passos.

- Aceder ao espaço de recolha e exportação via um membro do grupo / seguidor do canal, sem acesso direto por parte do investigador.
- Ser convidado, por meio de *link,* por um membro ou administrador do grupo. Esses utilizadores podem ser encontrados via meios de pesquisa mais tradicionais, como, por exemplo, procurar espaços públicos da comunidade que pretende estudar:
  - eventos abertos ao público;
  - perfis ou grupos abertos / públicos nas redes sociais ou em fóruns;
  - organizações, associações ou outro tipo de estruturas ligadas à comunidade;
  - na sua rede de contactos pessoal, possivelmente com recurso a técnicas como o *snowballing* (amostragem em bola de neve).

- Entrar por meio de *link* divulgado em outros espaços *online*:
  - pesquisas booleanas em Google, considerando que os *links* para o WhatsApp têm sempre uma estrutura semelhante a https://chat.whatsapp.com;
  - espaços nas redes sociais e fóruns ligados à comunidade, onde estes *links* poderão ser partilhados sem serem rastreados pelos motores de busca;
  - *site*s que conglomeram *links* para grupos de WhatsApp.

Destaca-se que os grupos com *links* disponíveis para acesso direto tendem, muitas vezes, a estar ligados a fraudes financeiras, vendas ou outro tipo de atividades pouco orgânicas. Poderá ser mais difícil de explicar o objeto de pesquisa e o respetivo consentimento sem um contacto dentro da comunidade / grupo, pelo que o método via membro existente ou convite direto será sempre uma forma mais sustentada de conseguir o acesso ao espaço de recolha.

Ferramenta: funcionalidade de exportação das próprias plataformas

Tanto o WhatsApp como o Telegram, como plataformas de mensagens amplamente utilizadas, oferecem um recurso valioso para investigadores: a capacidade de exportar dados de conversas reais, incluindo de grupos dos quais fazemos parte. Essa exportação é sempre feita em dois grupos. A conversa é exportada em texto (em formato TXT), enquanto o conteúdo multimédia, que podemos ou não descarregar, é exportado separadamente para uma pasta à parte. Esta exportação é gratuita e o único desafio, caso seja uma conversa com muito conteúdo multimédia, é o tempo e o espaço que o descarregamento ocupa.

Embora o presente capítulo não desenvolva, em profundidade, questões éticas, a recolha de dados de plataformas de mensagens exige reflexão e avaliação de argumentos metodológicos e éticos sobre a importância de obter o consentimento dos utilizadores dos quais se vai recolher dados. Victor Piaia e os seus colegas (2022) promovem uma discussão positiva sobre esta matéria num artigo publicado em 2022, que permanece atual, e que poderá ser um bom ponto de partida para esta reflexão. Em grupos mais pequenos, compostos de utilizadores que se conhecem pessoalmente, pode haver uma expectativa de privacidade forte, o que torna esse consentimento absolutamente necessário. Estas considerações poderão ser menos relevantes se se extrair dados de grupos públicos do Telegram. No caso dos canais, como se trata de um meio de comunicação unidireccional público sem contacto com o administrador, estas considerações não se aplicam.[2]

Em termos práticos, é também importante destacar que só consegue recolher dados para exportação a partir da data em que integra o grupo.[3] Poderá, pois, ser

2 Há canais não pesquisáveis no Telegram e no WhatsApp, acessíveis apenas por *link*, mas como não existe forma de contactar o administrador do canal ou de obter os seus dados, nem o mesmo tem acesso aos dados do seguidor, há um obstáculo evidente ao pedido de consentimento. Os canais são meios de comunicação e divulgação de uma entidade, pelo que é importante notar que não haverá, na maioria dos casos, uma expectativa de privacidade.

necessário deixar passar um período relativamente alargado de tempo antes de proceder à exportação de dados. O acesso via um membro existente pode facilitar acesso ao histórico, destacando-se, caso se considere necessário obter consentimento, que se terá acesso a dados e conversas anteriores. Infra sugere-se uma forma de anonimizar o nome dos diversos intervenientes na conversa.

Aceder e explorar a ferramenta: WhatsApp

1. Abra a aplicação do WhatsApp no seu *desktop* ou no seu telefone e abra a conversa que deseja exportar.
2. Toque no nome do grupo / contacto na parte superior da janela da conversa para abrir o separador de informações do grupo / conversa.
3. Vá até ao fim deste separador, onde encontra a opção "Exportar conversa".
4. Surge então a opção de exportar a conversa com ou sem conteúdo multimédia (como imagens e vídeos).
5. Escolha a opção de acordo com o tipo de análise que tenciona fazer. Incluir conteúdo multimédia resultará num tamanho de arquivo significativamente maior.
6. Depois de selecionar uma das duas opções, escolha como quer receber e guardar o arquivo exportado, existindo a hipótese de enviar como *email*, armazenar numa *cloud* ou guardar diretamente no seu dispositivo. É criado / enviado um ficheiro ZIP.

Os dados exportados são fornecidos em formato ZIP, numa pasta com o nome da conversa. Dentro dessa pasta, está um ficheiro TXT, que inclui toda a conversa disponível ao utilizador por ordem cronológica, com a data, hora, autor e conteúdo de cada mensagem de texto. Se o conteúdo multimédia estiver incluído na exportação, os arquivos serão anexados separadamente, nos seus formatos originais, com indicação da data e hora da sua publicação.

Outputs gerados

Existem várias extensões e ferramentas, algumas delas gratuitas, que fazem este tipo de exportação diretamente do WhatsApp, organizando os dados já em CSV ou em ficheiro Excel, e filtrando os ficheiros multimédia, removendo, por exemplo, os duplicados. No entanto, é muito importante destacar que, ao conceder acesso destas ferramentas à sua aplicação de mensagens, não consegue garantir a privacidade dos dados nele contidos, o que é especialmente relevante para o WhatsApp, onde não existem grupos e canais públicos de partilha não encriptada de dados. Neste

3 A não ser que, excecionalmente, o administrador tenha a opção da história do grupo ser acessível a novos membros.

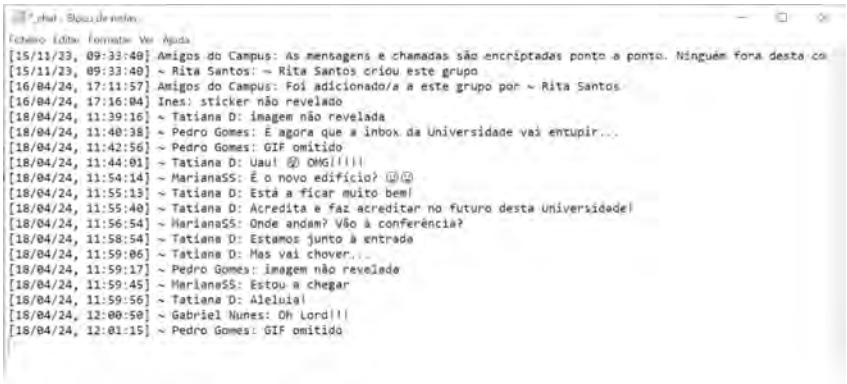

Figura 10.1 Exemplo fictício de ficheiro de exportação em TXT de um *chat* de WhatsApp
Fonte: elaboração própria da autora.

contexto, o presente manual apresenta uma solução que impede a partilha de quaisquer dados privados com terceiros.[4]

O arquivo de TXT contém os seguintes dados, cronologicamente ordenados, da conversa:

— a data de criação do grupo, o nome e o utilizador que a criou;
— uma indicação de data/hora para cada mensagem no formato [DD/MM/AA, HH:MM:SS];
— o nome do remetente, conforme está gravado na sua lista de contactos, ou precedido de ~ caso não tenha esse contacto na sua lista;
— o conteúdo da mensagem em texto, incluindo os *emojis* utilizados;
— referências textuais aos arquivos multimédia, que são integrados na sequência do texto, indicando o tipo de ficheiro enviado.

Os ficheiros multimédia exportados têm no nome o tipo de ficheiro —GIF, *sticker*, *photo*, *video*, PDF — e a data e hora da sua publicação, o que permite a sua integração na cronologia da conversa e melhor interpretação dos dados, quando necessário, como acontece no exemplo da figura 10.1. Exemplo:

— 000000004-STICKER-2024-04-16-17-16-04
— 000000005-PHOTO-2024-04-18-11-39-16

É possível que os dados exportados em TXT sejam mais difíceis de trabalhar, conforme o tipo de análise que se queira fazer. É relativamente fácil converter o ficheiro

4 Este tipo de aplicações poderão ser uma opção caso a recolha esteja a ser feita num perfil de WhatsApp / Telegram sanitizado, sem acesso a outras conversas, para recolha de dados de um canal público, onde não existe expectativa de privacidade.

para o Excel e para o Google Sheets, fazendo importação de dados, texto para colunas e escolhendo os ":" como elemento separador. De seguida, pode usar a função CONCAT ou CONCATENAR para unir as colunas das horas que ficarão divididas, passando a haver 3 colunas: data / hora, remetente e mensagem.

Pode haver uma ou outra mensagem que também fique dividida em mais do que uma coluna por ter um ":" algures, mas basta, novamente, usar a função supra para ficar com 3 colunas. Posteriormente, pode usar também funções para converter o nome de utilizador para um valor anónimo, garantindo a privacidade.

Aceder e explorar a ferramenta: Telegram

1. Descarregue e abra a aplicação do Telegram no seu *desktop* (não funciona para versão *mobile*) e abra a conversa que deseja exportar.
2. Clique nos três pontos no canto superior direito junto ao nome do grupo / contacto para abrir o menu de ações para este grupo / conversa.
3. Selecione a opção "Exportar histórico da conversa".
4. Surge então a opção de exportar a conversa com ou sem conteúdo multimédia (como imagens e vídeos) e em dois formatos de exportação: JSON ou HTML.
5. Escolha a opção de acordo com o tipo de análise que tenciona fazer. Incluir conteúdo multimédia resultará num tamanho de arquivo significativamente maior.
6. Escolha a opção de acordo com o tipo de análise que tenciona fazer. A versão HTML resulta numa visualização direta da conversa, em modo *browser*. A versão em JSON permite mais facilmente a exportação para folhas de cálculo.
7. Depois de selecionar se deseja o conteúdo multimédia e o formato de exportação, inicia-se o *download*, sendo este guardado diretamente no seu dispositivo. É criada uma pasta chamada "Telegram desktop", onde ficam guardadas as conversas descarregadas.

Outputs gerados

Existem, tal como para o WhatsApp, diversas ferramentas que permitem a exportação e análise direta de uma conversa no Telegram. Por serem soluções que nem sempre garantem a privacidade dos dados, optou-se por, neste manual, não as apresentar de forma detalhada. Mas caso esteja interessado em analisar um canal ou um grupo totalmente público, sugere-se a utilização do 4CAT: Capture and Analysis Toolkit[5] ou do TELEPATHY DB.[6]

---

5 O 4CAT é uma ferramenta da Digital Methods Initiatives da Universidade de Amsterdão, que faz recolha e análise de conteúdo do Telegram. Existe a instalação de um servidor próprio, mas é gratuito e tem uma utilidade transversal nos Métodos Digitais. https://github.com/digitalmethodsinitiative/4cat

Tal como descrito, o Telegram apenas permite exportação de conversas em dois formatos: HTML e JSON. O formato HTML cria um *output* que é, visualmente, muito semelhante ao formato das conversas no *desktop*, incluindo as imagens, vídeos e *stickers* partilhados. Já o formato em JSON — uma linguagem de programação — é de leitura menos fácil, mas é tipo de ficheiro mais prático de converter para outros formatos de mais fácil leitura, e assim permitir uma análise de um maior volume de dados. Os ficheiros multimédia, tal como acontece com o WhatsApp, são gravados numa pasta à parte, com a indicação do tipo de ficheiro e a data e a hora de publicação, o que permite a sua integração no encadeamento nas conversas.

Segue-se uma explicação, passo a passo, da conversão do ficheiro da conversa em JSON para CSV. Os passos podem parecer tecnicamente desafiantes, mas são simples de executar:

1. Guardar o ficheiro JSON numa pasta específica:
   a. selecione uma pasta específica onde quer guardar o ficheiro JSON que exportou do Telegram;
   b. o nome do ficheiro deve ser "results". Se o nome for "result" ou outro, faça a alteração;
   c. vá até às propriedades do ficheiro e copie a sua localização.
2. Instalar o *software* Python
   a. vá até à loja do seu sistema operativo, ou até https://www.python.org/ , e instale a última versão do *software* Python disponível. À data da escrita deste manual, a última versão era a 3.12.
3. Guardar o ficheiro Python que faz a conversão
   a. vá até https://github.com/keizerzilla/telegram-chat-parser;
   b. clique no botão do lado direito do ecrã, de cor verde, que diz CODE. Faça *download* do ficheiro ZIP;
   c. faça a extração dos ficheiros, e selecione o ficheiro Python identificado pelo nome "telegram-chat-parser";
   d. grave o ficheiro "telegram-chat-parser.py" na mesma pasta onde gravou o ficheiro "results.json". Confirme que a pasta não tem outros ficheiros além destes dois.
4. Abrir a *prompt* de comando (Windows) ou terminal (MacOS ou Linux)
   a. o *prompt* de comando / terminal é uma espécie de bastidores do sistema operativo, onde podemos dar ordens (comandos). Normalmente tem a aparência de um ecrã negro com letras brancas;

---

6 O Telepathy DB foi criado por Jordan Wildon e tem uma versão paga e uma versão *freemium*, sendo que a *freemium* permite pesquisa e recolha de publicações no Telegram por palavra-chave, até um total de 20 mil resultados. Permite filtrar esses resultados por língua, canal e data e permite a sua exportação direta em CSV.

b. abra o seu *prompt* de comando
   i. no Windows, a forma mais fácil é selecionar a tecla do Windows e o R simultaneamente, e escrever "CMD" na caixa de pesquisa que surge.
   ii. nos Macintosh, pesquise no Finder por "Terminal".

5. Dê os comandos para a conversão do ficheiro que exportou do Telegram para CSV
   a. no ecrã preto deve surgir a identificação do seu diretório de base (por exemplo, c:\Users\ines.narciso);
   b. primeiro, verifique que a sua instalação do Python foi bem-sucedida. Digite "python —version" ou "python3 —version" e pressione "enter" para garantir que o Python está instalado corretamente. Se obtiver uma resposta com Python e a versão, significa que o Pyhton está instalado corretamente, e pode avançar para o ponto seguinte;
   c. identifique a pasta onde guardou os dois ficheiros "results.JSON" e "telegram-chat-parser.py". Para isso vai dar indicação desse novo local depois do >;
   d. escreva cd e a localização da pasta onde guardou os ficheiros (que copiamos no ponto 1.) no formato CD/para/sua/pasta e clique em "enter". No meu caso ficou: C:\Users\ines.narcisocd Downloads\Telegram Desktop\chat_teste;
   e. o comando passa a assumir o novo local, respondendo com … /para/sua/pasta. No meu caso surgiu como C:\Users\ines.narciso\Downloads\Telegram Desktop\chat_teste;
   f. em frente à coloque a linha de comando "python3 telegram-chat-parser.py results.json" (também disponível na página de GitHub do *script* de conversão) e faça "enter":
      … /para/sua/pastapython3 telegram-chat-parser.py results.json. No meu caso ficou:
      C:\Users\ines.narciso*Downloads*\Telegram Desktop\chat_testepython3-telegram-chat-parser.py results.json;
   g. surge a questão "Enter the output file path and name";
   h. responda com results.csv;
   I. a exportação deve ficar terminada e o ficheiro CSV estará gravado na pasta que indicou junto do ficheiro JSON e do ficheiro PY.

Cada ficheiro CSV terá os campos cujas colunas e a sua respetiva descrição estão indicadas s no quadro 10.2.

A coluna "sender" pode ser removida por questões de privacidade, mantendo-se a coluna "sender_ID".

Quadro 10.2 Campos do ficheiro CSV

Coluna	Descrição
msg_id	Identificador único da mensagem
sender	Nome do remetente tal como ele se inscreveu na plataforma
sender_id	Identificador único do remetente
reply_to_msg_id	ID da mensagem original se esta mensagem for uma resposta, ou -1 caso contrário
date	Data e hora da mensagem no formato AAAA-MM-DD HH:MM:SS
msg_type	Tipo de mensagem, texto, *sticker*, algo do arquivo, *link*
msg_content	O conteúdo da mensagem
has_mention	Se existe alguma menção. Se não, devolve o valor 0
has_email	Se existe algum *email*. Se não, devolve o valor 0
has_phone	Se existe algum telefone. Se não, devolve o valor 0
has_hashtag	Se existe alguma *hashtag*. Se não, devolve o valor 0
is_bot_command	Se a mensagem for um comando de um *bot*. Se não, devolve o valor 0

Fonte: elaboração da própria autora.

Possibilidades de investigação

Até aqui, descreveu-se como funcionam as ferramentas de exportação das duas plataformas abordadas — WhatsApp e Telegram —, e como gerar *outputs* de dados que permitam responder às perguntas de partida do investigador. Abaixo destacam-se possibilidades de investigação tendo em conta os dados extraídos das plataformas de mensagens referidas.

O presente capítulo não vai abordar a recolha e exploração de dados, visto que tal é transversal e foi explicada, passo a passo, na apresentação da ferramenta. Considerando que o *output* final dos dados é uma folha de cálculo, sugerem-se abaixo, dentro das possibilidades de investigação apresentadas, as diferentes formas como os dados extraídos podem ser trabalhados.

Foco no conteúdo

Em investigações focadas em conteúdo, a exploração dos dados centra-se sobretudo na coluna correspondente ao conteúdo das mensagens, ou nas pastas com o conteúdo multimédia. Destacam-se algumas possibilidades de investigação.

— Tópico ou tema: Podem ser utilizadas ferramentas, como a mineração de texto, para identificar temas recorrentes, ou a forma como um determinado tópico é apresentado.

Quadro 10.3 Exemplos de investigação que pode ser feita a partir de dados das plataformas de mensagens

Foco	Conteúdo	Padrões e tendências	Redes
Objeto	Texto e conteúdo multimédia	Dinâmicas macro dos grupos, principais tópicos e padrões comportamentais	Padrões de interação, análise de organização social, caracterização da estrutura de uma comunidade. Foco nos utilizadores e na forma como interagem
Possíveis características	Foco no conteúdo e no seu significado, foco, em termos de exportação, nas mensagens e no conteúdo multimédia partilhado. Pode incluir análises mais comportamentais, análise de sentimento, e/ou integração do conteúdo com outras variáveis, como estudos comparativos de conteúdo com outras plataformas.	Identificação de tópicos e tendências emergentes nas conversas ao longo do tempo, que podem ser indicativos de mudanças sociais mais amplas. Os grupos podem ser espaços de interesse etnográfico, oferecendo dados sobre normas, valores e padrões coletivos.	Mapear as linhas de comunicação dentro de um grupo. Esta análise pode revelar quem são os principais influenciadores, nós centrais e *clusters* isolados dentro do grupo, entre outros padrões sociais.
Exemplo	Estudo comparativo da forma como os movimentos sociais adaptam a sua comunicação ao tipo de plataforma, comparando o conteúdo dos canais de três organizações dos direitos humanos com a sua comunicação em outra rede social.	Estudo de como as comunidades reagem perante a difusão de conteúdo desinformativo no WhatsApp.	Estudo sobre dinâmicas de poder dentro de uma claque de futebol, através de análise de rede do grupo onde organizam as mobilizações.

Fonte: elaboração da própria autora.

  - Exemplos: a forma como os canais de Telegram de duas agências noticiosas apresentavam o conflito na Ucrânia (Ptaszek *et al.*, 2024); a forma como os membros de um grupo antivacinas holandês apresentavam os seus argumentos num grupo de Telegram (Schelette *et al.*, 2023).
- Sentimentos: as ferramentas de análise de sentimento podem ajudar a avaliar o tom emocional das mensagens relativas a eventos ou momentos de comunicação específicos.
  - Exemplo: um estudo sobre um método de classificação de mensagens feito com grupos de jovens no WhatsApp, em que prevalece um sentimento neutro ou positivo perante os tópicos discutidos (Roy e Das, 2022).
- Conteúdo multimédia: podemos ter interesse em saber como um determinado tópico é abordado visualmente, ou de que forma as fotos, vídeos, *stickers*, GIF e *emojis* são integrados numa conversa. Modelos avançados de *computer vision* podem ser usados para classificar as imagens ou reconhecer padrões.
  - Exemplo: um estudo sobre desinformação no WhatsApp, recorrendo a imagens partilhadas no Brasil e na Índia, usando ferramentas para verificar se já tinham sido identificadas pelas plataformas de *fact-checking* (Reis *et al.*, 2022).

— Estudos comparativos: comparar dados recolhidos das plataformas de mensagens com elementos de outras plataformas, como fóruns ou redes sociais. Esta comparação permite identificar semelhanças e diferenças no estilo ou conteúdo da comunicação. Por exemplo, um estudo pode comparar a forma como um evento político é discutido no WhatsApp e no Facebook, utilizando tanto a análise de texto, como de conteúdo multimédia.

Foco nos padrões e tendências

Nas investigações focadas em tendências, a exploração dos dados é feita a um nível mais macro, procurando identificar alterações e padrões gerais das diferentes variáveis recolhidas, quer estas se refiram aos principais tópicos, às dinâmicas sociais, ou a comportamentos de comunicação, entre outros. Destacam-se algumas possibilidades de investigação.

— Análise longitudinal: observar e identificar na conversa mudanças nos tópicos, nos comportamentos de comunicação, ou nas interações ou outros, ao longo de um período alargado de tempo.
    — Exemplo: um estudo longitudinal poderia acompanhar as discussões num grupo de WhatsApp de ativistas ambientais para observar como as principais questões debatidas pelo grupo evoluem, levando à identificação, por exemplo, de mudanças de foco das preocupações ambientais locais para as globais ao longo de vários anos.
— Estudo etnográfico: uma análise etnográfica de grupos numa plataforma de mensagens permite aos investigadores estudar comunidades digitais, observar e documentar estilos de comunicação, valores partilhados e normas, entre outros.
    — Exemplo: um estudo que examina o papel dos *emojis* e dos *stickers* personalizados em conversas entre adolescentes no Telegram, oferecendo uma perspetiva sobre novas formas de comunicação não-verbal / textual e a sua relevância na definição de padrões identitários.
— Padrões de comunicação: uma análise da forma e dos *timings* com que o conteúdo é partilhado, procurando compreender, por exemplo, de que forma eventos externos alteram a periodicidade ou volume de conteúdo partilhado, ou como esse mesmo conteúdo se propaga dentro da plataforma, ou no espaço digital e mediático.
    — Exemplo: a investigação sobre desinformação no Paquistão durante a pandemia da covid-19 procurava compreender as dinâmicas de partilha e de propagação dos diferentes conteúdos sobre a doença no WhatsApp (Javed *et al.* 2022).

Foco nas dinâmicas de rede

Em investigações focadas em dinâmicas de rede, a exploração dos dados é feita sobretudo com base nos padrões de comportamento entre os utilizadores: respostas, reações e o papel que cada membro desempenha dentro de uma conversa de grupo. Eis alguns exemplos de possíveis investigações.

- Centralidade— identifica os indivíduos mais influentes dentro de uma rede com base nas suas posições e no número de conexões que possuem.
  - Exemplo: analisar uma comunidade de profissionais da cultura no WhatsApp e os diversos grupos que a compõem para identificar os principais influenciadores. Esta análise poderá ajudar a compreender como a informação se propaga dentro da rede, e de que forma a importância relativa destes indivíduos tem impacto na forma como o conteúdo é percecionado, bem como na forma como os restantes utilizadores interagem com o mesmo.
- Dinâmicas de rede— e compreender como uma rede evolui ao longo do tempo, incluindo a forma como os relacionamentos e as interações se modificam, eventualmente em resposta a determinados estímulos externos.
  - Exemplo: acompanhar um grupo de Telegram de imigrantes, analisando como a rede se transforma à medida que os utilizadores vão permanecendo mais tempo no país, ou as mudanças na dinâmica de uma mesma rede perante uma onda de hostilidade externa contra imigrantes.

Considerando a atual utilização global de plataformas de mensagens, nomeadamente do WhatsApp e do Telegram, espera-se que estas plataformas assumam um papel cada vez mais central na comunicação diária e, consequentemente, se tornem espaços cada vez mais relevantes para a investigação académica. O processo de recolha e análise de dados destas plataformas é desafiante, devido a questões como a privacidade dos dados, e a necessidade de ferramentas técnicas para extração e conversão dos mesmos. Mas o investimento é compensador, uma vez que se ganha acesso a espaços não regulados e semipúblicos ou privados, com dinâmicas diferentes das redes sociais. O investigador pode recolher dados das plataformas de mensagens sobre um leque vasto de temas, da propagação de desinformação, ao fortalecimento dos laços comunitários, à influência da comunicação digital na opinião pública, ou à apropriação destes espaços pelo mundo empresarial. Diferentes tipos de análise — seja de conteúdo, identificação de tendência, ou mapeamento de redes — podem contribuir para uma compreensão mais profunda de um tema.

Referências bibliográficas

Baulch, E., A. Johns, e A. Matamoros-Fernández (2024), "A critical review of media and communications scholarship on messaging apps", *Research Handbook on Social Media and Society*, pp. 270-286, https://doi.org/10.4337/9781800377059.00031.

Ceci, L. (2023), "Telegram messenger monthly users", *Statista*, disponível em https://www.statista.com/statistics/234038/telegram-messenger-mau-users/.

Chagas, V., I. Mitozo, S. Barros, J.G. Santos, e D. Azevedo (2022), "The new age of political participation? WhatsApp and call to action on the Brazilian senate's consultations on the e-cidadania portal", *Journal of Information Technology & Politics*, 19 (3), pp. 253-268, https://doi.org/10.1080/19331681.2021.1962779.

Dixon, S. (2023), "The rise of messaging apps: a global overview", *Statista*, disponível em https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps.

Gerster, L., R. Kuchta, D. Hammer, e C. Schwieter (2022), "Telegram as a buttress: how far-right extremists and conspiracy theorists are expanding their infrastructures via Telegram", *Institute for Strategic Dialogue*, disponível em https://www.isdglobal.org/wp-content/uploads/2022/11/Telegram-as-a-Buttress_How-far-right-extremists-and-conspiracy-theorists-are-expanding-their-infrastructures-via-Telegram.pdf.

Golovnin, O. K., D.E. Pleshanov, e A.A. Stolbova (2023), "Data mining for public channels and groups in telegram Messenger", *2nd International Conference on Computer Applications for Management and Sustainable Development of Production and Industry*, 12564, pp. 20-25.

Herrero-Solana, V., e C. Castro-Castro (2022), "Telegram channels and bots: a ranking of media outlets based in Spain", *Societies*, 12 (6), p. 164, https://doi.org/10.3390/soc12060164

Javed, R. T., M. Usama, W. Iqbal, J. Qadir, G. Tyson, I. Castro, e K. Garimella (2022), "A deep dive into COVID-19-related messages on WhatsApp in Pakistan", *Social Network Analysis and Mining*, 12 (1), p. 5, doi: 10.1007/s13278-021-00833-0.

Johns, A., A. Matamoro-Fernández, e E. Baulch (2024),"WhatsApp: from a one-to-one messaging app to a global communication platform", *Polity Press.*

Kligler-Vilenchik, N. (2022), "Collective social correction: addressing misinformation through group practices of information verification on WhatsApp", *Digital Journalism*, 10 (2), pp. 300-318, https://doi.org/10.1080/21670811.2021.1972020.

Piaia, V., E. Matos, T. Dourado, P. Barboza, e S. Almeida (2022), "Ethical issues in WhatsApp research: notes on political communication studies in Brazil", *Revue Française des Sciences de l'Information et de la Communication*, 25, https://doi.org/10.4000/rfsic.13328.

Pierson, J. (2021), "Digital platforms as entangled infrastructures: addressing public values and trust in messaging apps", *European Journal of Communication*, 36 (4), pp. 349-361, https://doi.org/10.1177/02673231211028374.

Ptaszek, G., B. Yuskiv, e S. Khomych (2024), "War on frames: text mining of conflict in Russian and Ukrainian news agency coverage on Telegram during the Russian invasion of Ukraine in 2022", *Media, War & Conflict*, 17 (1), pp. 41-61, https://doi.org/10.1177/1750635223116632.

Reis, J. C., P. Melo, K. Garimella, J. M. Almeida, D. Eckles, e F. Benevenuto (2020), "A dataset of fact-checked images shared on WhatsApp during the Brazilian and Indian elections", *Proceedings of the International AAAI Conference on Web and Social Media*, 14, pp. 903-908.

Rogers, R. (2020), "Deplatforming: following extreme Internet celebrities to Telegram and alternative social media", *European Journal of Communication*, 35 (3), pp. 213-229, https://doi.org/10.1177/0267323120922066.

Rossini, P. (2023), "Farewell to big data? Studying misinformation in mobile messaging applications", *Political Communication*, 40 (3), pp. 361-366, https://doi.org/10.1080/10584609.2023.2193563.

Roy, B., e S. Das (2022), "Perceptible sentiment analysis of students' WhatsApp group chats in valence, arousal, and dominance space", *Social Network Analysis and Mining*, 13 (1), p. 9, DOI:10.21203/rs.3.rs-2206392/v1.

Schlette, A., J. W. Van Prooijen, A. Blokland, e F. Thijs (2023), "Information, identity, and action: the messages of the Dutch anti-vaccination community on Telegram", *New Media & Society*, pp. 1-20 DOI:10.1177/14614448231215735.

Capítulo 11

Apps e *apps store*s como objeto de estudo

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

Entre as variadas tecnologias de informação e comunicação disponíveis atualmente, as aplicações móveis (*apps*) têm ganhado relevância. Em 2022, 142,6 mil milhões de *apps* foram descarregadas, em todo o mundo, das lojas de *apps* (Iqbal, 2023) com os dados a apontar para uma média de utilização de 5 horas diárias (Ceci, 2023a).

Os milhões de *apps* disponíveis no mercado estão maioritariamente concentradas entre dois *players*: Apple Store e Google Play. Desde 2008, estas distribuem *apps* para iOS e Android, respetivamente, apresentando-se como *gatekeepers*. Não só fazem a ponte entre desenvolvedores das *apps* e utilizadores (Dieter *et al*., 2019), como também ditam regras sobre a estrutura, funcionamento ou comercialização das *apps*.

Estudar *apps* implica não só compreender as regras e lógicas de funcionamento das *apps*, mas também das próprias lojas. Conhecer as suas categorias, os seus mecanismos de recomendação ou de sugestão, ou até as possibilidades de pesquisa é fundamental.

Os milhões de *apps* disponíveis no mercado encontram-se distribuídas por diferentes categorias que podem diferir em tipologia, número e consequente agrupamento dependendo da loja. De forma geral, as categorias "Jogos", "Educação" ou "Business", são as mais populares, às quais se somam "Lifestyle" e "Saúde e Fitness" (Ceci, 2023b, 2023c). De notar que existem *apps* nativas do próprio *smartphone* e aquelas que o utilizador decide descarregar da loja.

O princípio geral de uma *app* é responder ou satisfazer determinada necessidade através do *smartphone*. Estas podem estar relacionadas com atividades mais simples ou até rotineiras como, por exemplo, fazer uma nota ou realizar uma operação matemática, a atividades mais específicas como, por exemplo, acompanhamento da prática de exercício físico ou registar informações de saúde, a atividades lúdicas como jogar ou assistir a um filme. A utilização de *apps* também pode ser numa ótica profissional, pessoal ou até íntima. Já as tarefas ou ações a desempenhar através das mesmas podem ser de carácter individual ou coletivo. São múltiplas e variadas as possibilidades.

É natural, então, que as *apps* estejam presentes em diferentes dimensões da vida quotidiana. Fazem parte das rotinas e hábitos dos utilizadores de *smartphones* (Morris e Elkins, 2015). Tendo em conta tanto a quantidade como a variedade no

que diz respeito à sua oferta, como a apropriação por parte dos utilizadores, as *apps* têm vindo a tornar-se objetos digitais culturalmente relevantes (Gerlitz *et al.*, 2019). Têm impacto nas práticas individuais dos utilizadores e nas dinâmicas sociais e culturais da sociedade.

Veja-se, por exemplo, a relevância que o WhatsApp ganhou.[1] Não só se trata de um meio apropriado para comunicar com família e amigos, como também, no âmbito laboral, entendido como um canal para gerir várias tarefas diárias ou até para realizar vendas.

Outro exemplo são as *apps* de encontros.[2] Consideradas pelos utilizadores como um meio válido para conhecer alguém com quem desenvolver um relacionamento (Sepúlveda, 2023), são apontadas como formatadoras dos relacionamentos, quantificando emoções e tratando-as como *commodities* (Bandinelli e Gandini, 2022; Illouz, 2018).

Já as redes sociais, que se têm diversificado em oferta e continuamente crescido em número de utilizadores, deixaram de ser um meio apenas para socializar, tornando-se relevantes para, por exemplo, a produção e consumo de notícias.

Todas estas possibilidades tomam lugar através de *apps*, instaladas em *smartphones*, refletindo no seu uso um conjunto de *affordances* comunicacionais específicas de tais aparelhos: portabilidade, disponibilidade, localização e multimédia (Schrock, 2015).[3] Aquando da apropriação das *apps*, os utilizadores dividem-se entre *affordances* técnicas (mecanismos), mas também *affordances* sociais (expectativas comportamentais) nas quais se refletem questões culturais (Bucher e Helmond, 2018) e ideologias relativas à *app* em questão. Todos esses fatores podem formatar o comportamento e, por conseguinte, ter impacto nas dinâmicas sociais. Estudar e compreender o uso de *apps* não pode acontecer de forma isolada do aparelho onde estas operam, nem do contexto cultural e significado atribuído à *app*. Veja-se, por exemplo, a imposição no âmbito da instalação e uso das *apps* governamentais relacionadas com a covid-19 durante a pandemia (Dieter *et al.*, 2021).

À medida que as *apps* se têm tornado *softwares* culturais e socialmente significativos, um recente campo de estudo específico emergiu: os estudos de *apps*. Estes apontam para a necessidade de entender as *apps* como intermediários sociais e as práticas dos utilizadores nas mesmas (Gerlitz *et al.*, 2019). Tal campo de estudo, no âmbito da pesquisa digital, traz consigo desafios do ponto de vista metodológico. Em seguida, partilhamos algumas abordagens que têm em comum as lojas de aplicações como ponto de entrada para a pesquisa.

---

1 O WhatsApp é uma aplicação de mensagens instantâneas e chamadas de voz pertencente à Meta. Acumula, globalmente, 2,78 mil milhões de utilizadores (Ceci, 2023d).

2 O Tinder é a *app* de namoro que acumula o maior número de utilizadores mundialmente.

3 Podemos definir *affordances* como tipo de ações possibilitadas.

Recolha de dados

Ferramenta: DMI Google Play App Store and iTunes App Store Scraper

Para a recolha de dados, vamos usar a ferramenta DMI Google Play App Store and iTunes App Store Scraper criada pelo *Digital Methods Initiative*.[4] Esta ferramenta permite explorar a oferta das lojas de *apps*, obtendo informação específica sobre cada uma das *apps*, através de um conjunto de métodos disponíveis. No quadro 11.1 resumem-se esses métodos em função da loja de *apps*.

A ferramenta DMI Google Play App Store and iTunes App Store Scraper é gratuita e não impõe limites de pesquisas, em termos da quantidade de dados ou do tempo de utilização. Numa ótica de colaboração e de *open science*, as recolhas realizadas através da ferramenta ficarão disponíveis, no menu resultados, para todos aqueles que acedam à mesma.

Quadro 11.1 Métodos disponíveis para recolha de dados em função da loja de aplicações

Método	Descrição	Google Play App Store	iTunes App Store
App	Recolhe os detalhes de uma *app*	X	X
List	Recolhe uma lista de *apps* de uma das coleções/categorias da loja		X
Search	Recolhe uma lista de *apps* e os seus detalhes através da pesquisa por uma determinada expressão	X	X
Developer	Recolhe uma lista de *apps* de acordo com o ID de um desenvolvedor	X	X
Similar	Recolhe uma lista de *apps*, semelhantes à qual se está a pesquisar, numa lógica de sugestão ou recomendação da *app*	X	X
Permissions	Recolhe uma lista de permissões à qual a *app* tem acesso	X	

Fonte: elaboração própria da autora.

Aceder e explorar a ferramenta

a. Ir a https://tools.digitalmethods.net/app-scrapers/
b. Solicitar ou introduzir a *password*.

Uma vez que tenha acedido à ferramenta, terá de escolher qual a loja que pretende explorar e da qual quer recolher dados. Para tal, basta clicar no ícone correspondente. Em seguida terá de escolher o método e, em função do método, terá de fornecer um

4 Pode saber mais sobre o Digital Methods Initiative em https://www.digitalmethods.net/Dmi/DmiAbout

Quadro 11.2 *Inputs* requeridos na pesquisa pela Google Play Store em função do método

Input	Descrição	App	Search	Developer	Similar	Permissions
appId	O Google Play ID(s) da app	x			x	x
num (optional)	O número de *apps* a recolher (amostra)		x	x		
fullDetail (optional)	Opção de incluir nos resultados os detalhes da *app* (exemplo: título, descrição, *rating*)		x	x	x	
term	Termo através do qual se pretende realizar a pesquisa. Pode ser o nome de uma *app* ou uma expressão genérica (exemplo: fazer amigos, meteorologia, *dating*)		x			
devId	O ID do desenvolvedor da *app*			x		
short (optional)	No caso de ser selecionado, os nomes das permissões serão indicados.					x

Fonte: elaboração própria da autora

Quadro 11.3 *Inputs* requeridos na pesquisa pela iTunes App Store em função do método

Input	Descrição	App	List	Search	Developer	Similar
appId	O iTunes ID da *app*	x				x
country (optional)	O código, de duas letras, correspondente ao país (loja de *apps*) do qual se pretende fazer recolha	x	x	x		x
collection (optional)	A coleção que pretende explorar		x			
category (optional)	A categoria que pretende explorar		x			
num (optional)	O número de *apps* a recolher (amostra)		x	x		
fullDetail (optional)	Opção de incluir nos resultados os detalhes da *app* (exemplo: título, descrição, *rating*)		x	x		x
term	Termo através do qual se pretende realizar a pesquisa			x		
page (optional)	Número de páginas na qual se apresenta o resultado			x		
lang (optional)	O código, de duas letras, do idioma no qual se apresentam os resultados			x		x
devId	O ID do iTunes do desenvolvedor				x	

Fonte: elaboração própria da autora.

conjunto de *inputs*/informações. Alguns desses *inputs* poderão ser opcionais e/ou determinados por defeito. Deverá indicá-los ou alterá-los caso pretenda outros diferentes. Nos quadros 11.2 e 11.3 indicam-se e definem-se os *inputs* requeridos em função dos diferentes módulos e da loja de *apps*.

É importante compreender que, do ponto de vista do utilizador, as lojas de *apps* podem ser exploradas, na opção pesquisa, através de expressões, tipos de *apps*, nomes de *apps* ou de desenvolvedores (Dieter *et al.*, 2019). Esta ferramenta assenta nessas possibilidades para a recolha de dados.

Outputs *gerados*

Após ter selecionado a loja e o método que pretende explorar e ter introduzido os *inputs* requeridos, os resultados irão aparecer na mesma página. Para os obter basta clicar no formato desejado: CSV (comma separated value) ou JSON. Iremos optar por CSV. O ficheiro irá ser transferido para o seu computador. Para o abrir poderá fazer o *upload* do mesmo ao Google Drive e escolher a opção "abrir com Google Sheets".

Em função do método utilizado, o ficheiro apresentará um conjunto de dados que lhe permitirão aferir e discutir resultados. Poderá ser informação como o título da *app*, a categoria em que está classificada, a versão dos *softwares*, imagens de ecrã, o ícone, estatísticas como *rating* ou *reviews*. Dependendo da sua pergunta de partida e objetivos da pesquisa, esses dados ajudá-lo-ão a obter resultados para a sua investigação.

Receitas

Em seguida, partilhamos algumas receitas que ajudarão a perceber o potencial da ferramenta à medida que conseguirá experimentar alguns dos métodos para recolher dados. Adicionalmente perceberá como diferentes métodos podem ser combinados entre si, de acordo, e se, fizer sentido no âmbito da sua investigação assim como as limitações da ferramenta.

Utilizar o método "Search"

O método "Search" está disponível para ambas as lojas de *apps*. Permite obter, através da pesquisa por uma determinada expressão, uma lista de *apps* e os seus detalhes. Simula assim a ação de pesquisa nas respetivas lojas de *apps*. É um método interessante de ser usado para explorar a oferta de *apps* em função da pesquisa através de um determinado tópico ou expressão (Sepúlveda, 2024), de como estas são categorizadas nas diferentes lojas ou até por quem (empresas, instituições ou desenvolvedores) são oferecidas.

Recolher dados

— Ir a https://tools.digitalmethods.net/app-scrapers/;
— selecionar a loja de *apps*: Google Play Store ou iTunes App Store. Para este exercício escolhemos a segunda;
— clicar na opção "Search" de entre os métodos disponíveis;
— clicar em "Create a new query";
— preencher os diferentes campos.

Vamos, para este exemplo, realizar uma pesquisa através do termo "Health". Pode repetir ou usar outro termo que considere mais apropriado em função dos seus interesses. Lembre-se de que, em função do termo que utilizar, pode obter um maior ou menor número de resultados. Lembre-se de contextualizar a pesquisa. Esta está a ser realizada numa loja de *apps* e nem todos os termos podem devolver resultados como poderia, à partida, ser expectável. No quadro 11.4 resumem-se os campos, os *inputs* e os parâmetros utilizados para a pesquisa por "Health".

Quadro 11.4 Resumo dos campos e *inputs* para pesquisa através do método "Search"

Campo	*Input*
name	Dar um nome à pesquisa. Este será o nome do ficheiro de resultados. Convém ser um nome do qual se lembre e que situe a pesquisa. Sugestão: Nome da loja_Expressão de pesquisa_Data Neste caso atribuímos: iTunes App Store_health_20022024
term	Escrever o termo/expressão através da qual a pesquisa será realizada. Pesquisámos pelo termo "Health"
num	Definir número de *apps* a recolher. Por defeito a ferramenta recolherá 50 *apps*. Estabelecemos 100. Este número constituirá a amostra.
page	Colocar o número de páginas nos quais serão apresentados os resultados. Por defeito está estabelecida uma página. Mantivemos o estabelecido por defeito.
lang	Indicar o(s) código(s) de idioma(s) para o texto do(s) resultado(s). Por defeito está estabelecido para inglês. Mantivemos o apresentado por defeito.
country	Indicar o(s) código(s) de duas letras do país de onde quer obter *apps*. Por defeito está estabelecido para os Estados Unidos da América. Mantivemos o apresentado por defeito.
fullDetail	Estabelecer se deseja que o ficheiro de resultados apresente os detalhes das *apps* recolhidas. É aconselhável que o faça para obter os respetivos dados de cada uma das *apps*. Indicámos "full detail"

Fonte: elaboração própria da autora.

— Clicar no botão "Create". A ferramenta irá começar a recolher os dados. A mensagem "Processing:1 of 1" aparecerá. Quanto mais ambiciosa a amostra estabelecida ("num"), mais tempo a recolha demorará;
— fazer o *download* do ficheiro. Optámos pela opção CSV.

Explorar os dados

— Uma vez feito o *download* do ficheiro de resultados em formato CSV, abra o mesmo com o Excel. Clique no ficheiro de resultados com o botão do lado direito, eleja a opção "Abrir com" e aí escolha a opção "Excel";
— uma vez aberto o ficheiro, vá ao menu "Ficheiro", opção "Guardar como" e no campo "Formato" escolha a opção ".xls" e em seguida "Guardar";
— faça *upload* desse ficheiro para o Google Drive. Abra-o com o Google Sheets. Para tal, clique no ficheiro de resultados com o botão do lado direito e escolha "Abrir com" e, em seguida, "Google Sheets";
— uma vez aberto o ficheiro de resultados, verá várias linhas e colunas com dados. Cada linha corresponderá a uma *app* e, cada coluna, a um dado específico sobre cada uma dessas *apps*. Poderá ser necessário redimensionar as linhas. Já a criação de filtros poderá ser de grande ajuda para exploração e ordenação dos dados.

Algumas possibilidades de investigação

Antes de iniciar qualquer análise, lembre-se de realizar uma cópia da folha de resultados. Trabalhe na mesma, garantindo assim que mantém a versão original da recolha inalterada.

— Quais as *top apps* associadas ao termo de pesquisa?
Podemos interpretar *top apps* como aquelas que têm classificação, atribuída pelos utilizadores, mais elevada. Através da coluna "averageUserRating" poderá organizar os resultados. Para tal, será possível seguir um processo semelhante ao "Analisar *top posts* do perfil" indicado no capítulo dedicado ao Instagram.
— Como se apresentam as *apps* relacionadas ao termo de pesquisa? Quais as suas áreas de atuação? Quais as suas promessas?
Para responder a essas ou outras questões, explorar a coluna "Description" é uma solução.
— Quais os ícones que representam as *apps*? Aqueles que são apresentados no ecrã do telemóvel. As cores ou símbolos usados nas *apps* podem ser, por exemplo, o foco do seu estudo. De facto, logos e símbolos de *apps* podem ser tendenciosos quanto a quem se dirigem e às funcionalidades (Sepúlveda, 2024; Zhang, 2017).
Para tal explore a coluna "artworkUrl512".[5] Siga os passos do procedimento "Obter imagens dos *posts*" do capítulo dedicado ao Instagram, mas neste caso

5 O URL da coluna "artworkUrl512", começa por https://is1-ssl.mzstatic.com

recorrerá à coluna "artworkUrl512" (em vez da "imgUrl") como meio para obter os ícones das *apps*.
Adicionalmente poderá recorrer ao programa ImageSorter[6] para organizar os ícones cromaticamente, facilitando a sua exploração.

— Quais as categorias às quais pertencem as *apps* associadas ao termo de pesquisa?
Explorando a coluna "primaryGenreName" irá obter dados referentes à categoria, estabelecida pela loja digital e escolhida pelo desenvolvedor, na qual a *app* está classificada.
A função "COUNTIF" será útil para contar quantas vezes as categorias se repetem.[7]

— As mesmas *apps* respondem a termos de pesquisa diferentes? Poderá realizar várias recolhas, através de diferentes termos, e combiná-las entre si para analisar resultados.
O artigo "Conhecer pessoas, namoro e sexo: as respostas da App Store", exemplifica como tal pode ser feito.[8] Tendo as *apps* de *dating* como objeto de estudo, compara a oferta da App Store, através da pesquisa por diferentes palavras-chave, questionando-a.

Utilizar o método "Permissions"

O método "Permissions" apenas está disponível para a loja Google Play Store. Seja no momento da instalação, como no decorrer do uso das *apps*, o utilizador coerciva ou voluntariamente confere à *app* acesso a um conjunto de dados através de permissões que estas vão solicitando. Essas permissões podem dar acesso a diversos dados através de funcionalidades do *smartphone* (exemplo: leitura ou edição de ficheiros armazenados, localização, acesso a contactos, à câmara ou ao microfone) que, nem sempre, podem parecer fazer sentido no funcionamento da *app*. Algumas dessas permissões podem ser classificadas como perigosas devido à natureza dos dados aos quais têm acesso.[9]

Tendo em conta o exposto, o método "Permissions" pode ser o apropriado para explorar ou responder a questões no âmbito do estudo de *apps*.

Recolher dados

— Ir a https://tools.digitalmethods.net/app-scrapers/;
— selecionar Google Play Store;
— clicar na opção "Permissions" de entre os métodos disponíveis;
— clicar em "Create a new query";
— preencher os diferentes campos.

---

6 https://appnee.com/imagesorter/
7 Consulte o procedimento "Explorar contas associadas à/s *hashtag*/s" no capítulo dedicado ao Instagram para relembrar como usar a função "COUNTIF".
8 https://medialab.iscte-iul.pt/conhecer-pessoas-namoro-e-sexo-as-respostas-da-app-store/
9 Pode explorar a tipologia de permissões neste artigo https://www.geeksforgeeks.org/what-are-the-different-protection-levels-in-android-permission

Quadro 11.5 Resumo dos campos e *inputs* para pesquisa através do método "Permissions"

Campo	*Input*
name	Dar um nome à pesquisa. Este será o nome do ficheiro de resultados. Convém ser um nome de que se lembre e que situe a pesquisa Sugestão: Nome da loja_AppID_Data. Neste caso atribuímos: Google Play_calm_20022024.
appId	Indicar o Google Play ID da app da qual se pretende saber as permissões. Pesquisámos pelo id ""(*).
short	No caso de ser selecionado, os nomes das permissões serão indicados. Indicámos "full detail".

(*) Para encontrar o ID na Google Play da *app*, basta procurar essa informação no URL da *app* em questão. No exemplo que estamos a usar: https://play.google.com/store/apps/details?id=com.calm.android.
Na App Store, o ID da *apps* é dado por um número que também se encontra no correspondente URL: https://apps.apple.com/us/app/calm/id571800810

Fonte: elaboração própria da autora.

Vamos, para este método, realizar uma pesquisa através da *app* Calm.[10] Esta *app*, disponível desde 2012, está classificada na categoria "Health e Fitness" da Google Play Store. Na sua descrição apresenta-se como uma *app* para sono, meditação e relaxamento "recomendada pelos melhores psicólogos, terapeutas e especialistas em saúde mental". Cem milhões de pessoas já fizeram o *download* da mesma, a qual acumula cinco milhões de subscritores, tendo faturado, em 2022, cerca de 330 milhões de euros (Curry, 2024).

Tendo em conta a sua área de atuação, tempo no mercado e número de utilizadores, pareceu-nos interessante e importante explorar as permissões às quais esta *app* tem acesso. Pode usar esta ou outra *app* que considere interessante para seguir a próxima receita. No quadro 11.5 resumem-se os campos, os *inputs* e os parâmetros utilizados para a pesquisa através do método "Permissions".

— Clicar no botão "Create". Os dados serão recolhidos. Poderá confirmar através da mensagem "Processing:1 of 1" que aparecerá.
— Uma vez terminada a recolha, realizar o *download* do ficheiro no formato CSV.

Explorar os dados

Abra o ficheiro com o Google Sheets. Siga o procedimento correspondente indicado na receita Utilizar o método "Search" neste capítulo. Redimensione as linhas, caso seja necessário.

10 https://www.calm.com/pt

Algumas possibilidades de investigação

Lembre-se de trabalhar numa cópia da folha de resultados, mantendo a versão original da recolha inalterada.

— A quantas permissões a *app* tem acesso?
— Que tipo de permissões são essas?
— Outras *apps*, semelhantes à investigada, também têm acesso a essas mesmas permissões?
Pode realizar várias recolhas, com outras *apps* ID, e combiná-las entre si.
O artigo "*Online dating* e privacidade: Quanto está disposto a revelar por um *match*?" (Sepúlveda e Crespo, 2022) ilustra como as permissões conferidas às *apps* podem constituir um objeto de estudo. Foca-se na questão da tipologia de dados às quais as *apps* têm acesso, mesmo quando as pessoas não as estão a utilizar, mas as têm instaladas.

Utilizar o método "Similar"

A pesquisa pelo método "similar", disponível para as lojas Google Play e Apple App Store, apresenta, como resultados, um conjunto de *apps* semelhantes àquela através da qual se está a pesquisar. Assim, o *input* de pesquisa será o ID da *app*. A sua lógica de recolha assenta na lógica de sugestão e recomendação das lojas de *apps*. Este método pode ser interessante não só para estudar as lógicas de sugestão e recomendação das lojas, como para construir uma base de dados.

Recolher dados

— Ir a https://tools.digitalmethods.net/app-scrapers/;
— selecionar Google Play Store ou Apple Play. Selecionámos o segundo;
— clicar na opção "Similar" de entre os métodos disponíveis;
— clicar em "Create a new query";
— preencher os diferentes campos.

Dentro do tema das pesquisas anteriores, continuaremos, para este método, com a *app* Calm como objeto de estudo. No quadro 11.6 resumem-se os campos, *inputs* e parâmetros de pesquisa.

— Clicar no botão "Create". Os dados serão recolhidos. Poderá confirmar através da mensagem "Processing:1 of 1" que aparecerá;
— uma vez terminada a recolha, realizar o *download* do ficheiro no formato CSV.

Quadro 11.6 Campos, *inputs* e parâmetros de pesquisa utilizados no método "Similar"

Campo	*Input*
name	Dar um nome à pesquisa. Este será o nome do ficheiro de resultados. Convém ser um nome de que se lembre e que situe a pesquisa. Sugestão: Nome da loja_AppID_Data Neste caso, atribuímos: iTunes App Store_571800810_20022024
appId	Indicar o App Store ID da qual se pretende saber as permissões. Pesquisámos pelo ID "571800810".
lang (optional)	Indicar o código, de duas letras, do idioma relativo aos resultados. Deixámos o predefinido "en".
country (optional)	Indicar o código, de duas letras, do país de onde se pretende obter similar apps. Deixámos o predefinido "us".
fullDetail (optional)	Se selecionado, apresenta os detalhes das apps que fazem parte dos resultados (exemplo: nome, categoria, descrição, número de downloads). Indicámos "full detail".

Fonte: elaboração própria da autora.

Explorar os dados

Abra o ficheiro com o Google Sheets. Siga o procedimento correspondente indicado na receita Utilizar o método "Search". Redimensione as linhas, caso seja necessário.

Algumas possibilidades de investigação

Lembre-se de trabalhar numa cópia da folha de resultados, mantendo a versão original da recolha inalterada.

— Que *apps* são sugeridas?
— A que categorias pertencem?
— Como se caracterizam?
— Ao realizar a pesquisa pelo mesmo *app* ID, mas alterando o parâmetro "country", quais são os resultados?
— Ao realizar a pesquisa pelo correspondente *app* ID, na loja Google Play, quais são os resultados?
— Ao criar uma rede de resultados em função da loja de *apps*, a que conclusões consegue chegar? Pode explorar a ferramenta table 2 Net[11] e o *software* Gephi.[12]

11 https://medialab.github.io/table2net/
12 https://gephi.org/

Referências bibliográficas

Bandinelli, C., e A. Gandini (2022), "Dating apps: the uncertainty of marketised love", *Cultural Sociology*, 16 (3), pp. 423-441, doi:10.1177/17499755211051559.

Bucher, T. e A. Helmond (2018), "The affordances of social media platforms", em Jean Burgess, Thomas Poell, Alice Marwick, *The SAGE Handbook of Social Media*, Sage.

Ceci, L. (2023a), "Hours spent on mobile apps 2019-2022, by country", *Statista*, disponível em https://www.statista.com/statistics/1269704/time-spent-mobile-apps-worldwide/.

Ceci, L. (2023b), "Most popular Apple App Store categories as of 3rd quarter 2022, by share of available apps", *Statista*, disponível em https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/.

Ceci, L. (2023c), "Most popular Google Play app categories as of 3rd quarter 2022, by share of available apps", *Statista*, disponível em https://www.statista.com/statistics/279286/google-play-android-app-categories/.

Ceci, L. (2023d), "Monthly global unique WhatsApp users 2020-2023", *Statista*, disponível em https://www.statista.com/statistics/1306022/whatsapp-global-unique-users/.

Curry, D. (2024), "Calm revenue and usage statistics", *Business of Apps*, disponível em https://www.businessofapps.com/data/calm-statistics

Dieter, M., C. Gerlitz, A. Helmond, N. Tkacz, F. van der Vlist, e E. Weltevrede (2019), "Multi-situated app studies: methods and propositions, *Social Media + Society*, 5 (2), pp. 1-15, doi:10.1177/2056305119846486.

Dieter, M., A. Helmond, N. Tkacz, F. van der Vlist, e E. Weltevrede (2021), "Pandemic platform governance: mapping the global ecosystem of COVID-19 response apps", *Internet Policy Review*, 10 (3), disponível em https://policyreview.info/articles/analysis/pandemic-platform-governance-mapping-global-ecosystem-covid-19-response-apps.

Gerlitz, C., A. Helmond, D. Nieborg, e F. van der Vlist (2019), "Apps and infrastructures — a research agenda", *Computational Culture*, 7.

Illouz, E. (2018), *Emotions as Commodities: Capitalism, Consumption and Authenticity*, Routledge.

Morris, J. W., e E. Elkins (2015), "There's a history for that: apps and mundane software as commodity", *The Fibreculture Journal*, 25, pp. 63-88, doi:10.15307/fcj.25.181.2015.

Sepúlveda, R. (2024), "Fostering intimacy in a digital environment: couples, mobile apps and romantic relationships", em I. Amaral, R. B. de Simões, e A. M. M. Flores (eds.), *Young Adulthood Across Digital Platforms*, Leeds, Emerald Publishing Limited, pp. 93-109, https://doi.org/10.1108/978-1-83753-524-820241006.

Sepúlveda, R. e M. Crespo (2022), "*Online* dating e privacidade: quanto está disposto a revelar por um match?", *MediaLab*, disponível em https://medialab.iscte-iul.pt/online-dating-e-privacidade/.

Schrock, A. (2015), "Communicative affordances of mobile media: portability, availability, locatability, and multimediality", *International Journal of Communication*, 9, pp. 1229-1246.

Zhang, S. (2017), "The awful pinkness of period apps", *The Atlantic*, disponível em https://www.theatlantic.com/health/archive/2017/05/period-apps-pink/525207/.

Capítulo 12

Métodos de inquirição *online*

Transcendendo os limites tradicionais na pesquisa social

Tiago Lapa
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Introdução

No mundo digital de hoje, a relevância dos métodos de investigação *online* aumentou, transformando fundamentalmente os cenários de investigação. À medida que as tecnologias digitais continuam a evoluir, permitem abordagens mais sofisticadas, diversificadas e eficientes à recolha de dados. Esta mudança não é apenas uma questão de conveniência, mas uma transformação estratégica que se alinha com as tendências transversais de digitalização em todos os setores. Para os investigadores, a capacidade de realizar inquéritos, entrevistas ou até mesmo realizar estudos quase-experimentais *online* alarga drasticamente o âmbito de potenciais participantes e diversifica o conjunto de dados além das restrições geográficas, aumentando potencialmente a extensão e a relevância dos estudos.

Os métodos de inquirição *online* abrangem qualquer forma de recolha de material empírico ou estratégia de investigação que utilize ferramentas e plataformas digitais para facilitar a recolha e análise de dados. Isto pode incluir inquéritos por questionário ou entrevistas *online*, grupos focais digitais com recurso a um leque amplo de ferramentas disponíveis — desde programas de videoconferência a fóruns de discussão e plataformas como Reddit ou Quora —, experiências baseadas na *web* e etnografia digital, entre outros. No entanto, podemos perguntar em que medida os métodos de inquirição *online* são uma transposição de métodos tradicionais e presenciais bem estabelecidos — como o questionário ou a entrevista — ou se apresentam características, procedimentos e desafios próprios. Ou seja, em que medida o contexto digital pode ser considerado apenas uma circunstância mediadora ou se estamos perante algo transformativo para os procedimentos de pesquisa.

Neste capítulo é discutido como a ascensão das tecnologias digitais transformou os métodos de pesquisa, alinhando-se com a digitalização em vários setores. Procede-se à apresentação dos métodos de inquirição *online*, incluindo as suas formas e as ferramentas utilizadas, como questionários *online*, grupos focais digitais, entrevistas *online* ou a etnografia digital, como à análise das inovações trazidas pelos métodos de inquirição *online*. Neste sentido, avaliamos criticamente a transposição de métodos tradicionais para o *online* e a forma como a tecnologia pode fazer

a diferença. Isso implica uma discussão sobre novas oportunidades metodológicas que vão além dos métodos convencionais e a exploração do impacto das tecnologias emergentes na recolha e análise de dados.

Procedemos depois no presente capítulo para a discussão das principais vantagens dos métodos de inquirição *online*, como acessibilidade ampliada, custo-efetividade e rapidez na recolha de dados, mas sem descurar a análise crítica das limitações e desafios, incluindo questões de amostragem, validade e controlo de contexto de pesquisa, além dos desafios éticos e relacionados com a segurança e privacidade dos dados.

Na parte conclusiva é feita uma revisão das principais ideias desenvolvidas ao longo do capítulo e como elas contribuem para o entendimento dos métodos de inquirição *online*, além de uma reflexão sobre o futuro desses métodos e o seu potencial para continuar a evoluir e influenciar a pesquisa social.

A novidade do ponto de vista metodológico

Uma questão central que tem merecido a atenção dos metodólogos que se debruçam sobre a inquirição *online* é como a tecnologia pode fazer a diferença e que novas oportunidades metodológicas além do convencional podemos explorar. O nosso posicionamento é que o domínio digital não confere apenas uma mediação, nem é uma mera transposição dos métodos tradicionais *offline*, embora a metodologia associada às pesquisas *online* seja construída com clara referência às metodologias aplicadas de modo presencial, postal ou telefónico, e com elas seja comparada. Daí que o campo dos métodos de inquirição *online* se tenha destacado e evoluído como um subcampo dos métodos de pesquisa social, com os seus próprios problemas e questões, por direito próprio. Se, por um lado, partilha elementos com os métodos tradicionais em termos de boas práticas na construção e implementação dos instrumentos de recolha de dados, por outro lado, apresenta enquadramentos metodológicos específicos, com as suas características e procedimentos próprios, que estão em fluxo constante devido à inovação tecno-social.

Os métodos de inquirição *online* têm introduzido várias inovações metodológicas significativas que transformam a forma como os dados são recolhidos, analisados e interpretados em pesquisas sociais e de mercado. Essas inovações são em grande parte impulsionadas pelo avanço tecnológico, que não apenas facilita novas formas de interação com os participantes (Couper e Miller, 2008), mas também oferece ferramentas para uma análise mais profunda e automatizada dos dados recolhidos. Questões como o recrutamento, a verificação de identidade e a ausência de sinais e pistas visuais e sociais influenciaram o desenvolvimento de uma variedade de métodos de pesquisa social à medida que foram sendo transferidos e adaptados para os ambientes *online*. A atual era digital assistiu a um aumento dramático no volume de dados disponíveis, bem como na velocidade com que podem ser recolhidos e processados. Isto abriu novos caminhos para os investigadores, permitindo explorações mais dinâmicas e imediatas de questões de pesquisa e de identificação de tendências sociais complexas. Ou seja, as possibilidades

tecnológicas do espaço *online* também têm exercido a sua influência sobre o desenvolvimento metodológico. Esta evolução sugere a necessidade de se desenvolverem novos instrumentos que permitam uma recolha de dados eficaz e adaptada aos novos meios na *internet*.

Um segundo ponto é que, embora os métodos de inquirição *online* possam constituir ferramentas poderosas de suporte à investigação social permitindo hoje o que no período pré-digital era entendido como impossível ou impraticável — como, por exemplo, proceder a uma pesquisa comparativa entre países sem a necessidade de viajar —, não devemos olhar para e usar esses métodos de forma acrítica. Não podemos encarar, portanto, os instrumentos de inquirição *online* de modo fetichista, isto é, venerados e preferidos por defeito no desenho da pesquisa por serem entendidos como convenientes, eficazes, na vanguarda ou simplesmente "melhores". Na verdade, a mudança não é necessariamente sempre progressiva ou evolutiva como caucionam Madge e O'Connor (2005), havendo circunstâncias em que os métodos *online* poderão ser vistos como o substituto pobre da "excelência" dos métodos *offline*. Claro que podemos contrapor que a inquirição *online* tem virtudes próprias, pelo que os seus métodos não podem ser vistos como parentes pobres na família dos métodos de pesquisa.

Comparando métodos tradicionais aos métodos de inquirição *online*, há vantagens, limitações e desafios de parte a parte. Se, por um lado, a inquirição *online* proporciona vantagens cruciais na recolha, análise e divulgação de informação e resultados de pesquisa, por outro, apresenta limitações e desafios em termos de amostragem, taxas de resposta, validade, controlo dos contextos de pesquisa, éticos e relativos à segurança dos dados, à exclusão digital, entre outros. Por exemplo, a dependência de ferramentas digitais exige medidas robustas de cibersegurança para proteger informações sensíveis recolhidas *online* e uma atualização constante de equipamento, licenças e sobretudo de competências e literacia por parte dos investigadores. Ou com a persistência de segmentos infoexcluídos na população, em particular os seniores no contexto português, a aplicação dos métodos *online* torna-se inviável quando os consideramos como objeto de estudo. A confiabilidade dos dados coletados *online* depende de um controlo rigoroso das condições em que a pesquisa é realizada, além de uma análise cuidadosa das características tecnológicas dos participantes. Logo, há a necessidade de os investigadores *online* praticarem o seu ofício com reflexividade, como assinalam Madge e O'Connor (2005), o que também é válido para a aplicação dos métodos tradicionais, e constituindo este um alerta central do nosso texto.

Terceiro, é importante fazer uma distinção entre a investigação que examina a *internet*, e a pesquisa que utiliza a *internet* como meio ou ferramenta para realizar uma pesquisa. Portanto, é de esclarecer que a inquirição *online* refere-se especificamente à pesquisa com ferramentas digitais e que não se deve confundir métodos com objetos de estudo. Claro que os métodos de pesquisa *online* são úteis de uma forma mais óbvia quando o fenómeno a ser investigado está fortemente ligado à *internet*, embora nem sempre. Se um investigador quiser estudar a infoexclusão, ou seja, os fatores que impedem a utilização das tecnologias da informação e comunicação (TIC) por parte de um segmento da população, não faz sentido enveredar

pela pesquisa *online*. Por outro lado, muitos investigadores usam métodos *online* para pesquisar problemas que não estão relacionados ou estão vagamente relacionadas com a *internet*.

A quarta consideração é que os métodos de pesquisa social, tanto *online* quanto *offline* ou tradicionais, não são desenvolvidos ou aplicados num vácuo. Eles são profundamente influenciados e moldados por uma variedade de fatores contextuais, incluindo sociais, políticos e intelectuais. A forma como os investigadores abordam as suas investigações está entrelaçada com o mundo social em que vivem, e com os temas e tópicos que eles estão a pesquisar.

As características demográficas e culturais de uma sociedade podem determinar quais questões são consideradas importantes e merecem ser investigadas. Por exemplo, uma sociedade com uma população crescentemente envelhecida pode focar-se mais em pesquisas relacionadas com a literacia digital (ou falta dela) entre os seniores.

Ademais, as normas sociais influenciam o que é considerado ético na condução de pesquisas. Isso pode afetar desde a escolha dos temas de pesquisa até os métodos de recolha de dados, respeitando o que é socialmente aceitável e ético.

As políticas públicas também podem incentivar ou restringir certos tipos de pesquisa, influenciando a alocação de fundos e recursos.

O contexto social e político da investigação pode ter impactos na capacidade dos pesquisadores de conduzir estudos no terreno. Neste sentido, pode ser mais seguro e viável realizar pesquisas *online*, por exemplo, quando se estudam contextos marcados pelo extremismo político.

Além disso, as tendências e debates intelectuais atuais influenciam quais teorias e métodos são privilegiados na pesquisa social. Isso pode ser visto na forma como certos paradigmas ganham popularidade e moldam as abordagens de pesquisa. O avanço da tecnologia também pode ser considerado um fator intelectual, pois novas tecnologias podem criar métodos de pesquisa ou aperfeiçoar os existentes, especialmente no contexto *online*.

Esses fatores são interdependentes e sobrepõem-se, criando um ambiente dinâmico em que os métodos de pesquisa são continuamente adaptados e refinados. Por exemplo, uma mudança política que afeta a liberdade académica pode levar a novas abordagens intelectuais e metodológicas como resposta. Da mesma forma, um avanço tecnológico pode abrir novas possibilidades para a pesquisa social, mas as suas implicações éticas e sociais precisarão de ser consideradas de forma reflexiva e cuidada. Entender estes fatores é crucial para os investigadores ao desenvolverem, escolherem e aplicarem métodos de inquirição *online*, assegurando que os seus estudos sejam relevantes, éticos e eficazes dentro do contexto em que operam.

Uma quinta consideração remete para a multidisciplinaridade dos métodos de inquirição *online*, pois estes integram diferentes abordagens teóricas e metodológicas de diversas áreas do conhecimento para realizar pesquisas e recolher dados através da *internet*, adaptadas às necessidades emergentes de novos objetos de pesquisa social. Os meios *online* alcançam uma variedade muito ampla de indivíduos abrangendo diferentes culturas, países e contextos socioeconómicos. Cada um

desses grupos pode requerer abordagens distintas que são informadas por diversas disciplinas.

Constituindo as tecnologias digitais a base dos métodos de inquirição *online*, aspetos como o *design* de interface (Dillman, Smyth e Christian, 2014), usabilidade, segurança de dados e processamento de informações são fundamentais. Aliás, como os dados recolhidos *online* podem ser quantitativos e qualitativos, abrangendo de respostas a questionários até análises de comportamento em *websites* (por exemplo, rastreamento de cliques), este campo metodológico mobiliza um conjunto vasto de métodos e técnicas, incluindo técnicas computacionais para examinar dados ou realizar análises de conteúdo, e contributos do conjunto das ciências humanas, incluindo as ciências cognitivas, para dotar de sentido essas mesmas análises.

Muitas questões de pesquisa que são exploradas através de métodos *online* cruzam fronteiras disciplinares, exigindo uma integração de conhecimento que abarca várias áreas, da ética em pesquisa até metodologias de análise de dados complexos. Um exemplo é a análise do discurso de ódio *online* (Lapa e di Fátima, 2023), que remete para a colaboração de várias áreas do conhecimento para entender e identificar esse fenómeno. No estudo das motivações por trás do discurso de ódio e os seus impactos nas vítimas e na sociedade, a psicologia pode explorar os efeitos emocionais do discurso de ódio, enquanto a sociologia analisa como ele se dissemina dentro de grupos sociais e as ciências da comunicação analisam como o discurso de ódio é disseminado através de diferentes *media* sociais, incluindo a dinâmica de como as informações são compartilhadas e se tornam virais. A multidisciplinaridade permite uma abordagem mais holística e eficaz no estudo do discurso de ódio nas redes sociais *online*, proporcionando uma compreensão mais profunda dos mecanismos subjacentes.

Os métodos de inquirição *online* podem ser adaptados para atender a necessidades específicas de pesquisa em diversos campos, tornando-os extremamente versáteis e aplicáveis a uma vasta gama de questões de pesquisa. Acrescente-se que a colaboração entre disciplinas frequentemente gera inovações metodológicas. A junção de métodos quantitativos e qualitativos de diferentes campos pode levar a abordagens mais ricas e holísticas. Por exemplo, a combinação de análises estatísticas avançadas (próprias da estatística ou ciência da computação) com contributos teóricos e conceptuais da psicologia ou sociologia enriquece a interpretação dos dados. A combinação de teorias psicológicas com tecnologias de rastreamento de dados pode levar a novas maneiras de entender o comportamento dos utilizadores *online*, assim como o *software* de mineração de texto ou análise de sentimentos emerge muitas vezes da colaboração de linguistas com cientistas da computação.

Tais colaborações também podem resultar na melhoria da qualidade e precisão dos dados. Colaborações entre especialistas em *media* digitais e ciências sociais podem melhorar as estratégias de amostragem e alcance para as pesquisas *online*, garantindo que as amostras sejam mais heterogéneas e menos enviesadas. Neste sentido, o trabalho conjunto entre psicólogos, cientistas sociais e estatísticos pode ajudar no desenvolvimento e na validação de instrumentos de pesquisa que são

culturalmente sensíveis e metodologicamente sólidos. A colaboração interdisciplinar pode contribuir ainda para a integração eficaz entre teoria e empiria. Por exemplo, teorias de comportamento humano (psicologia) podem ser combinadas com modelos de difusão de informações (ciências da comunicação) para entender como as pessoas respondem a pesquisas *online* e como disseminam informações. Em suma, a multidisciplinaridade dos métodos de inquirição *online* não só enriquece a pesquisa, proporcionando-lhe maior profundidade e abrangência, mas também a torna mais relevante e aplicável num espectro mais amplo de contextos e problemas contemporâneos.

Como sexta consideração, na esteira de Mann e Stewart (2000), temos as formas como a escolha da estratégia metodológica pode influenciar significativamente o tipo de resultados obtidos, bem como as conclusões que podem ser generalizadas para contextos mais amplos. A metodologia determina o que é medido, como é medido, e o contexto no qual as medições são feitas. O desenho de uma pesquisa *online* deve ter estas considerações em conta, fazendo parte da competência do investigador antever o tipo de resultados a que pode chegar, mediante as escolhas metodológicas que faz. Por exemplo, especificamente em ambientes *online*, a generalização dos resultados pode enfrentar limitações devido à natureza dos métodos utilizados. Métodos qualitativos, como entrevistas ou grupos focais, oferecem resultados profundos e uma compreensão detalhada sobre o comportamento e as perceções dos participantes, mas podem ser mais difíceis de generalizar devido ao carácter frequentemente não estruturado dos dados.

Tendo em conta o que foi dito até agora, como última consideração temos um aspeto crucial dos métodos de inquirição *online* destacado por Best e Krueger (2004, p. 85): embora iniciar um projeto de recolha de dados *online* possa parecer um processo menos complicado e mais acessível quando comparado aos métodos tradicionais, executá-lo com sucesso é substancialmente mais complexo e exige uma abordagem meticulosa e refletida. As ferramentas tecnológicas disponíveis hoje permitem aos pesquisadores configurar e lançar uma pesquisa com relativa rapidez e a um custo reduzido. No entanto, a facilidade de configuração não é sinónimo de qualidade na recolha de dados. Para garantir que esta é bem-sucedida, é fundamental planear meticulosamente cada etapa do processo de inquirição *online*, desde a definição clara dos objetivos até a escolha dos instrumentos e métodos de análise mais adequados para atingir esses objetivos. Além disso, os investigadores devem estar atentos às questões éticas, como o consentimento informado e a privacidade dos dados, garantindo que estas questões sejam tratadas com a devida seriedade. É igualmente essencial considerar os desafios específicos associados à inquirição *online*, como a representatividade da amostra, o envolvimento e motivação dos participantes e a precisão das respostas. Assim, apesar das vantagens aparentes dos métodos *online* em termos de custo e eficiência, eles requerem uma consideração cuidadosa e estratégica para superar essas limitações. E é sobre isso que nos debruçaremos na parte seguinte do texto.

Vantagens, limitações e desafios

Os métodos de inquirição *online* oferecem várias vantagens significativas que transformam a pesquisa e a recolha de dados. Primeiro, a capacidade de romper barreiras geográficas, acedendo a uma grande quantidade de participantes simultaneamente e em diferentes geografias, mas igualmente diversificados do ponto de vista sociodemográfico. Isso é especialmente importante em estudos comparativos, globais ou quando se procura estudar nichos específicos da população que seriam difíceis de alcançar fisicamente ou algo que seria logisticamente complexo e custoso com o recurso a métodos tradicionais. Portanto, os métodos *online* podem apresentar melhores rácios custo-benefício em comparação aos tradicionais, que geralmente envolvem interações presenciais, viagens e envio de materiais físicos, reduzindo assim significativamente os custos económicos, de tempo e operacionais.

A rapidez é outra vantagem notável dos métodos *online*, pois a automatização das ferramentas digitais permite uma recolha e análise de dados mais rápidas, essenciais para estudos que necessitam de agilidade, como aqueles que monitorizam tendências de mercado ou comportamentos eleitorais de forma sincrónica ou em tempo real.

A acessibilidade e conveniência são inegáveis, com a *internet* a permitir que os participantes dos estudos se envolvam em pesquisas de qualquer parte do mundo, o que alarga o âmbito das oportunidades de investigação. A escalabilidade é outra característica dos métodos *online*, permitindo dimensionar estudos para acomodar um número maior de inquiridos sem um aumento proporcional no esforço ou custo. A capacidade de recolher e analisar dados em tempo real permite aos investigadores tomar decisões informadas rapidamente e obter *feedback* instantâneo.

Os métodos de inquirição *online* também permitem a recolha e o processamento de uma ampla variedade de tipos de dados, como interações multimédia, que podem aumentar o envolvimento dos inquiridos, ou dados gerados pelos utilizadores nos *media* sociais e outras pegadas digitais, fornecendo uma rica fonte de informações que podem ser analisadas para descobrir padrões de comportamento e preferências. Além de recolher respostas ativas dos participantes (como em questionários), os métodos *online* também podem capturar dados passivos, como padrões de navegação na *web*, interações nos *media* sociais e até mesmo dados biométricos em dispositivos conectados. Ferramentas como aplicações móveis e *software* de pesquisa baseado na *web* podem enviar notificações e promover interações imediatas, capturando dados de forma sincronizada com eventos em andamento, o que é particularmente útil em estudos longitudinais e diários.

A tecnologia digital permite ainda uma adaptabilidade significativa no *design* dos questionários. Estes podem agora ser elaborados de forma ramificada, permitindo que a apresentação das questões seja flexibilizada com base nas respostas anteriores, que a pesquisa seja mais dinâmica e personalizada, o que, por sua vez, pode levar a medições e recolhas mais precisas e aprofundadas.

Os investigadores podem ainda optar por métodos sincrónicos, como entrevistas via videoconferência, que simulam a interação face a face, ou métodos assíncronos, como fóruns *online* e *emails*, que permitem aos participantes responderem

no seu próprio ritmo, aumentando a conveniência para ambos, os pesquisadores e os inquiridos.

Contudo, esses métodos não estão isentos de desafios e limitações que discutimos de seguida de forma não exaustiva. Um dos maiores desafios enfrentados pelos métodos de inquirição *online* é garantir que a amostra seja representativa da população-alvo. Muitas pesquisas *online* dependem de voluntários que escolhem participar, o que pode levar a um viés de seleção, em que os indivíduos que são mais propensos a usar a *internet* e interessados no tema da pesquisa estão sobrerrepresentados. Além disso, certos grupos demográficos, como seniores ou indivíduos de estatuto socioeconómico despreviligiado, podem ter acesso limitado à *internet*, resultando na sua sub-representação. A dependência de tecnologia fiável e o tipo de conectividade à *internet* também pode excluir segmentos da população em regiões com infraestruturas tecnológicas limitadas. A diversidade tecnológica e as competências ou literacia digital dos potenciais participantes ou público-alvo podem ingulmente influenciar o grau de facilidade de uso e da acessibilidade dos métodos de inquirição *online*. Acrescente-se que problemas técnicos, como falhas num *website* ou serviço, problemas de compatibilidade com diferentes dispositivos ou navegadores e interrupções de conexão à *internet* podem prejudicar a experiência dos utilizadores e afetar o processo de recolha de dados.

A autenticidade das respostas é outra preocupação significativa. Sem interações face a face, é mais difícil verificar a identidade dos respondentes e garantir que as respostas são genuínas. Além disso, a falta de controlo sobre o ambiente em que a pesquisa é realizada *online* pode levar a variações nos dados que são difíceis de quantificar ou controlar, o que tem impacto na validade e fiabilidade dos resultados.

As taxas de resposta para inquéritos *online* podem ser mais baixas do que aquelas obtidas por métodos tradicionais. A natureza impessoal e a sobrecarga e fadiga respeitante às informações *online* podem levar a um menor envolvimento, interesse e motivação dos participantes. Isso é agravado pela facilidade com que os inquiridos podem abandonar um estudo *online*, muitas vezes devido ao comprimento do questionário ou da entrevista ou ao *design* pouco atraente.

Nas pesquisas que recorrem às entrevistas ou aos grupos focais digitais, entre outras técnicas, a recolha de dados *online* muitas vezes não permite a observação de pistas não-verbais, como linguagem corporal e expressões faciais, que são valiosas em muitos campos de pesquisa. Isso pode limitar a profundidade e o contexto dos dados coletados, afetando a interpretação dos resultados. A falta de interação pessoal pode, portanto, afetar a qualidade dos dados recolhidos, especialmente em pesquisas que envolvem comportamentos e emoções humanas complexas.

Os métodos de inquirição *online* apresentam desafios éticos únicos que precisam de ser considerados cuidadosamente pelos investigadores para garantir a integridade dos seus estudos. Um dos principais pilares éticos da pesquisa é o consentimento informado. A garantia que todos os participantes compreenderam plenamente o propósito da pesquisa, os procedimentos envolvidos, os potenciais riscos e benefícios e os direitos de saída a qualquer momento pode ser mais difícil de alcançar nas pesquisas *online* devido à distância física e à falta de interação direta. Por exemplo, a facilidade de carregar em "concordo" em formulários digitais

pode não refletir um verdadeiro entendimento do consentimento por parte do respondente.

Determinados grupos podem ser mais vulneráveis em contextos de pesquisa *online*, incluindo menores de idade, pessoas com limitações cognitivas, ou indivíduos em situações de risco (como vítimas de abuso ou pessoas com condições de saúde mental). A pesquisa *online* deve ser projetada para assegurar que esses grupos são tratados com dignidade e que a sua participação seja verdadeiramente voluntária e informada.

Também é de considerar que a forma como as perguntas são formuladas e as pesquisas são projetadas *online* podem influenciar as respostas dos participantes. Isto é ao mesmo tempo um desafio metodológico que diz respeito à fiabilidade das medições, mas igualmente um problema ético, pois existe o risco de manipulação, intencional ou não, que pode levar a dados enviesados. Os investigadores devem ser meticulosos no desenho da pesquisa para evitar a introdução de vieses.

Outro aspeto ético importante é a gestão do ciclo de vida dos dados recolhidos, incluindo a sua retenção e destruição apropriada após o término da pesquisa. Deve-se, portanto, definir claramente e cumprir os prazos de retenção de dados, garantindo que sejam destruídos de forma segura para evitar qualquer uso indevido subsequente.

Além disso, as questões de segurança e privacidade de dados são preocupações críticas, especialmente quando as pesquisas tratam de temas sensíveis, já que a recolha de dados *online* pode ser suscetível a violações de segurança e a requisitos acrescidos de anonimização dos dados. A identidade dos participantes pode ser mais difícil de proteger devido à rastreabilidade digital e ao armazenamento de metadados. Assegurar o anonimato e a confidencialidade dos dados em ambientes em que as informações podem ser facilmente copiadas, armazenadas e potencialmente acedidas por terceiros não autorizados apresenta desafios significativos. Tal exige que os investigadores sigam regulamentos rigorosos de proteção de dados e a observância de normas legais como o RGPD (Regulamento Geral sobre a Proteção de Dados da União Europeia). A falha em proteger os dados pode ter implicações legais e danificar a confiança pública nas pesquisas académicas ou de mercado.

A própria natureza dos dados recolhidos *online* ou das próprias ambições de pesquisa, que podem prever a captura de grandes volumes de dados não estruturados, conteúdos multimédia e dados de monitorização comportamental, apresenta desafios significativos para análise. A necessidade de ferramentas avançadas de análise de dados e competências específicas pode constituir um obstáculo, especialmente para pesquisadores que não têm formação em estatística ou ciência de dados. Neste sentido, o investigador tem de ser realista e avaliar as competências próprias ou da sua equipa quando define os objetivos da pesquisa *online*.

Portanto, enquanto os métodos de inquirição *online* oferecem numerosas vantagens e revolucionam a pesquisa moderna, também apresentam desafios significativos que necessitam ser abordados para maximizar a sua eficácia e garantir a integridade e relevância dos dados recolhidos.

Conclusão

Como vimos, os métodos de inquirição *online* incluem uma variedade de técnicas, como questionários, trabalho etnográfico, entrevistas e grupos focais que são realizados utilizando ferramentas *online* e plataformas digitais. Este capítulo teve como propósito esclarecer como esses métodos se adaptam e se integram com as tecnologias digitais, permitindo uma recolha de dados mais dinâmica e diversificada. Os métodos de inquirição *online* oferecem uma abordagem flexível, económica e inovadora à investigação. Permitem a recolha de diversos tipos de dados e facilitam uma ampla participação para além de fronteiras geográficas.

No entanto, estes métodos também apresentam desafios significativos, incluindo questões relacionadas com a ética da pesquisa, a dependência tecnológica, segurança de dados e envolvimento dos participantes. Foram destacadas as complexidades em garantir a privacidade dos participantes e obter um consentimento verdadeiramente informado num ambiente *online*, ressaltando a necessidade de práticas rigorosas e transparentes. Foram igualmente abordadas questões de representatividade e autenticidade, problemas que advêm da falta de interação pessoal e o possível viés de autosseleção dos inquiridos, entre outras, que podem ter impacto na qualidade e na confiabilidade dos dados recolhidos. Para que os métodos de inquirição *online* sejam utilizados de forma eficaz, é crucial abordar estas limitações e, ao mesmo tempo, aproveitar as vantagens para melhorar a qualidade e a amplitude da investigação e da aprendizagem.

Apesar destes desafios, o potencial dos métodos de investigação *online* para revolucionar a investigação é inegável. Fornecem plataformas poderosas para recolha e análise de dados, permitem a rápida disseminação de resultados e oferecem oportunidades de recolha flexíveis e inclusivas. O desenvolvimento e aperfeiçoamento contínuos destes métodos continuarão, sem dúvida, a moldar os cenários da investigação, impulsionando a inovação e a criação de conhecimento num mundo cada vez mais digital. Ao entender as dinâmicas descritas, os investigadores podem aproveitar melhor as vantagens dos métodos *online*, enquanto mitigam as suas limitações, contribuindo assim para o avanço do conhecimento de maneira responsável e inovadora.

Não será arriscado dizer que o futuro dos métodos de inquirição *online* é promissor. À medida que as tecnologias digitais avançam, esses métodos tornam-se mais sofisticados, acessíveis e integrados nos padrões de pesquisa em diversas disciplinas. Um exemplo é a recente democratização da utilização da inteligência artificial (IA) generativa. A integração da IA nos métodos de inquirição *online* poderá suportar análises mais profundas e precisas de grandes conjuntos de dados. Isso inclui a capacidade de identificar padrões e tendências complexas, realizar análises preditivas e automatizar a categorização e codificação de respostas qualitativas. À medida que mais dispositivos se tornam conectados, a capacidade de recolher dados de comportamento em tempo real através de dispositivos IdC (*internet* das coisas) oferecerá novas oportunidades para pesquisas comportamentais e ambientais, proporcionando uma compreensão mais dinâmica e contextualizada dos padrões de vida dos participantes.

Por outro lado, o uso destas novas ferramentas poderá introduzir desafios, metodológicos e éticos, e limitações próprias. O futuro dos métodos de inquirição *online* poderá ser fortemente influenciado por mudanças nas leis de privacidade e proteção de dados. A adaptação a essas mudanças será fundamental para manter a confiança das pessoas e a legitimidade dos processos de pesquisa.

Também será relevante acompanhar a forma como as ferramentas de inquirição *online* continuarão a tornar-se mais acessíveis e se se traduzem numa maior democratização da investigação em diversos sectores além do académico, permitindo que uma gama mais ampla de investigadores, incluindo aqueles em regiões com menos recursos, conduza estudos sofisticados. Há ainda quem aponte que as plataformas *online* podem capacitar comunidades e indivíduos ao permitirem que estes participem ativamente nos processos de pesquisa, não apenas como sujeitos, mas como colaboradores ativos (Markham, 2008). Finalmente, há ainda a esperança de que o desenvolvimento de melhores ferramentas para verificar a autenticidade e a precisão dos dados recolhidos *online* ajudará a melhorar a confiança nos resultados das pesquisas.

Referências bibliográficas

Best, S. J., e B. S. Krueger (2004), *Internet Data Collection*, 141, Sage.

Couper, M. P., e P. V. Miller (2008), "Web survey methods: introduction", *Public Opinion Quarterly*, 72 (5), pp. 831-835.

Dillman, D. A., J. D.Smyth, e L. M. Christian (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, John Wiley e Sons.

Eynon, R., J. Fry, e R. Schroeder (2017), "The ethics of online research", *The SAGE Handbook of Online Research Methods*, 2, pp. 19-37.

Lapa, T., e B. Di Fátima (2023), "Hate speech among security forces in Portugal", *Hate Speech on Social Media*, 277-293.

Madge, C., e H. O'Connor (2005), "Mothers in the making? Exploring liminality in cyber/space", *Transactions of the Institute of British Geographers*, 30 (1), pp. 83-97.

Mann, C., e F. Stewart (2000), *Internet Communication and Qualitative Research: A Handbook for Researching Online*, Londres, Sage.

Markham, A. (2008), "The Internet in qualitative research", em L. Givens (ed.), *The Sage Encyclopedia of Qualitative Research Methods,* Thousand Oaks, CA, Sage, pp. 454-458.

Parte 4 | Análise de dados e apresentação de resultados

Capítulo 13

A análise de dados

Propostas, exemplos e sugestões

Sofia Ferro-Santos
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia(CIES-Iscte), Lisboa, Portugal

Rita Sepúlveda
ICNOVA — Instituto de Comunicação da Nova, Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisboa, Portugal

Inês Narciso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Se um dos grandes desafios relacionados com a realização de pesquisa no âmbito dos métodos digitais é relativo à recolha dos dados, como abordar o fenómeno, questionar a plataforma, decidir que ferramentas utilizar e ter competências para dominá-las; outro prende-se com a análise dos dados obtidos. A norma é que a base de dados seja facultada em ficheiros, por exemplo, no formato CSV, JSON, GFD. Assim, antes de proceder à análise propriamente dita, a nossa orientação é que leve o seu tempo para explorar a base de dados. Considere este passo como intermédio e com o propósito de conhecer e familiarizar-se com a tipologia de dados e os dados propriamente ditos que são apresentados no *corpus*. Pode explorar se e como estes se relacionam entre si, assim como poderá trabalhá-los ou analisá-los.

Acreditamos que, à medida que vai explorando a base de dados obtida, terá necessidade de a organizar, talvez trocar a posição de algumas colunas, juntar colunas entre si, destacá-las através da atribuição de alguma cor ou até eliminar algumas das que contêm dados. Faz tudo parte do processo! Não obstante, lembre-se, antes de começar esta análise exploratória, de realizar uma cópia da primeira base de dados. Assim garante que terá sempre o resultado original da recolha de dados que efetuou.

No decorrer dessa análise exploratória, é natural que comece a tentar responder às perguntas de partida que definiu anteriormente e como parte do desenho da pesquisa, guiando-se pelas mesmas. Seja através de dados quantitativos, qualitativos ou, claro, uma combinação de ambos. Recordemos que a abordagem metodológica dos métodos digitais se posiciona mais como qualiquantitativa, do que uma divisão entre ambas. Adicionalmente, toda a análise deve ter em conta a plataforma na qual o fenómeno está a ser estudado, pois esta, como ambiente, tem impacto no mesmo e encerra em si uma capacidade explicativa do próprio fenómeno.

Existem variadas abordagens para proceder à análise dos dados. A resposta sobre qual será a mais adequada está na pergunta de partida que definiu. Abordagens qualitativas envolvem explorar o objeto de estudo com maior profundidade por meio de dados não numéricos. Assim o texto, as imagens, as *hashtags* poderão ser o seu alvo de análise com o objetivo de descobrir significados, usos, reconhecer

padrões, analisar relações e compreender contextos sociais. Métodos como a análise temática, visual e de redes podem fazer sentido.

Já a abordagem quantitativa envolve a análise sistemática de dados numéricos. Métricas como gostos, visualizações ou comentários, entre outras possibilidades, podem servir para testar hipóteses, identificar padrões ou relações. Métodos com recurso a análises estatísticas ou matemáticas e análise de dados podem auxiliar na resposta às perguntas definidas.

Como será expectável, as abordagens apresentam em si vantagens e desvantagens. Não obstante, as desvantagens podem ser colmatadas através da combinação das abordagens, oferecendo assim *insights* distintos. Novamente, a resposta às perguntas sobre quais são as técnicas e os métodos mais adequados e o caminho a seguir deverá estar refletida no desenho da pesquisa e determinada de acordo com a pergunta de investigação.

A nossa proposta neste capítulo passa por ilustrar como procedemos à análise de dados mediante diferentes desafios de investigação. Assim, mais do que facultar receitas sobre como analisar dados, exemplificamos como o fizemos. Acreditamos que a partilha destes procedimentos pode servir como inspiração e ponto de partida para o desenvolvimento de abordagens próprias. Note que estes exemplos que partilhamos não devem ser vistos como os únicos possíveis. Eles são apenas exemplos.

Analisar a oferta da App Store quando se pesquisa pela expressão *couple*[1]

Contexto

A oferta nas *apps stores* é variada incluindo-se as aplicações (*app*) para casais, isto é, aquelas que prometem promover a comunicação entre os membros de um relacionamento amoroso. A investigação propõe oferecer um mapeamento da oferta da App Store e a sua caracterização.

Pergunta de partida

Como é que as *couple apps* se apresentam?

Recolha de dados

Através da ferramenta DMI iTunes App Store Scraper, módulo "search" que recolhe uma lista de *apps* e os seus detalhes através da pesquisa por uma determinada

1 Este procedimento faz parte da investigação que resultou no capítulo Sepúlveda, R. (2024), "Fostering Intimacy in a Digital Environment: Couples, Mobile Apps and Romantic Relationships", Amaral, I., de Simões, R.B. e Flores, A.M.M. (eds.) *Young Adulthood Across Digital Platforms*, Emerald Publishing Limited, Leeds, pp. 93-109. https://doi.org/10.1108/978-1-83753-524-820241006

expressão. Os parâmetros de pesquisa definidos foram: expressão de pesquisa "couple", número de *apps* n=1000, idioma= en e país=us.

Análise de dados

A recolha de dados originou uma base de dados com 507 *apps*. Com o objetivo de avaliar o posicionamento das *apps*, analisou-se como estas se apresentavam, explorando o ícone que as representa e a descrição das *apps*.

Para tal, foi realizada uma análise temática (Braun e Clarke, 2006). Como método qualitativo, envolve a leitura de dados e a identificação de padrões nos mesmos, os quais darão origem a temas. A análise temática compreendeu os seguintes passos: 1. familiarização com os dados, isto é, por um lado, ler as descrições das *apps* para identificar como estas se posicionavam, a quem se dirigiam ou quais as promessas, e, por outro, ver como eram os ícones dessas mesmas *apps*, os seus símbolos ou cores; 2. codificação dos dados, este passo traduziu-se em destacar no texto da descrição das *apps* frases ou partes de frases e criar "códigos" que tivessem a capacidade de descrever esse conteúdo. A mesma lógica foi aplicada para os ícones, procedendo-se à anotação dos mesmos; 3. identificação de padrões entre os códigos criados e geração de temas. Estes tinham como objetivo serem mais abrangentes que os códigos, conseguindo até agrupar em si vários códigos; 4. os temas foram revistos de forma que sejam o mais precisos possível e 5. foi atribuído um nome.

A abordagem da temática qualitativa foi indutiva, isto é, os dados recolhidos determinaram os temas (Fereday e Muir-Cochrane, 2006). Porém, em outros casos, essa abordagem poderia ser de carácter dedutivo, isto é, partindo de alguns temas pré-estabelecidos com origem em teoria, conceitos ou resultados de estudos já existentes.

Num segundo momento de análise pretendia-se olhar se e como as *apps* formataram possibilidades na criação do perfil no que dizia respeito ao género, orientação sexual e tipologia de relacionamento. Assim, foram selecionadas, de entre as 507 aplicações da amostra, as dez com a classificação, atribuída pelos utilizadores, mais elevada. A análise realizada a essas aplicações foi feita com recurso ao método *walkthrough* que é considerado uma abordagem qualitativa. Este método é definido como "uma forma de interagir diretamente com a interface de uma aplicação para analisar os seus mecanismos tecnológicos e referências culturais incorporadas para entender como ela orienta os utilizadores e molda a experiência" (Light *et al.*, 2018: 882).

Os procedimentos consistiram em realizar o *download* e a instalação das dez *apps* selecionadas e criar o perfil em cada uma destas. Todos os passos da criação do perfil foram registados e anotados, através da construção de uma grelha de observação, permitindo concluir especificamente sobre os três tópicos que se pretendia investigar: género, orientação sexual e tipo de relacionamento dos utilizadores.

Compreender lógicas de *cross platform* com o #EstudoemCasa como objeto de estudo

Contexto

A pandemia provocada pela covid-19 e o conjunto de restrições impostas geraram novos ritmos educativos nos quais as salas de aula físicas foram substituídas por ecrãs. No contexto português, o programa "Estudo em Casa",[2] lançado pelo Ministério da Educação, foi criado para mitigar os efeitos das interrupções letivas. Estilizado #EstudoEmCasa, tratava-se de um programa televisivo exibido na RTP que tomou emprestada a linguagem de um objeto nativo digital: a *hashtag* (Liu, 2009; Rogers, 2013).

Durante períodos de confinamento, o digital foi utilizado não só como meio de acesso à educação, como também as redes sociais *online* serviram de meio através do qual os utilizadores partilhavam experiências e se manifestavam no âmbito da educação.

Pergunta de partida

Como estudar manifestações da educação em contexto digital? O objetivo central da investigação foi identificar como os utilizadores no Instagram, Twitter/X e YouTube apropriaram o #EstudoEmCasa.

Recolha de dados

— Instagram | #EstudoEmCasa e @EstudoEmCasa2020
A recolha de dados realizou-se através do Phantombuster. Respetivamente através do módulo "Instagram Hashtag Collector" (n=10 000), para recolher conteúdo identificado com #EstudoEmCasa, e do módulo "Instagram Posts Extractor" para recolher as publicações do perfil @EstudoEmCasa2020 (n=313 posts).
— Twitter | #EstudoEmCasa
Com o intuito de analisar o conteúdo publicado no Twitter/X com o #EstudoEmCasa realizou-se a recolha de dados através do Phantombuster e do módulo "Twitter Hashtag Collector"[3]
— YouTube | #EstudoEmCasa e EstudoEmCasa
Para a recolha de vídeos no YouTube recorreu-se à ferramenta YouTube Data Tools (Rieder, 2015), especificamente ao módulo "Video List".[4] Realizaram-se duas recolhas: 1) através da *query* "estudoemcasa" e 2) através da *query* "#estudoemcasa". Cada recolha originou uma lista com 250 vídeos e,

2 Pode consultar o estudo completo em Flores, A.M. e Sepúlveda, R. (2021), "Método digitais e educação: uma proposta de investigação", em Ana Nobre, Ana Mouraz, e Marina Duarte (eds.), *Portas Que o Digital Abriu na Investigação em Educação,* Universidade Aberta, pp.226-255, 10.34627/uab.edel.15.11.
3 Este módulo não está atualmente disponível.
4 https://ytdt.digitalmethods.net/

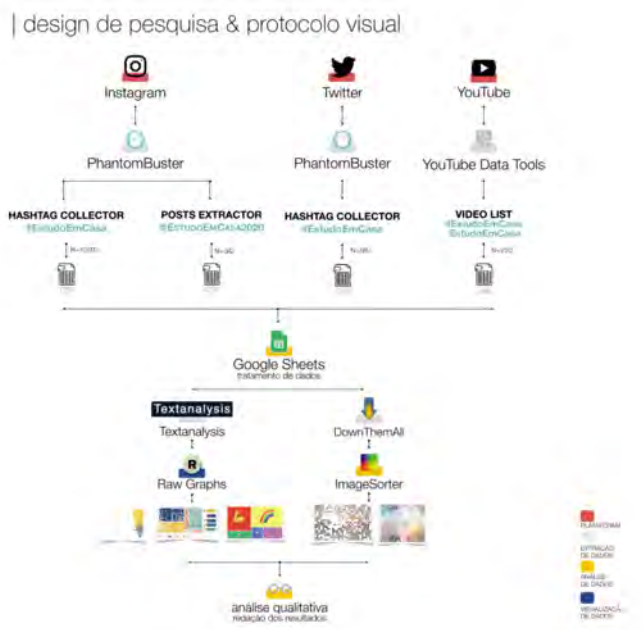

Figura 13.1 *Design* de pesquisa e protocolo visual seguido para o objeto empírico #EstudoEmCasa
Fonte: Flores A.M. e Sepúlveda, R. (2021).

depois de eliminados os repetidos, obteve-se uma base de dados final (n=320).

Na figura 13.1 é resumido o desenho da pesquisa contemplando as ferramentas para recolha de dados e como foi realizada a análise dos mesmos.

Análise de dados

— Instagram | #EstudoEmCasa
A base de dados foi explorada em termos de datas das publicações (temporalmente situadas entre os anos 2017 e 2020), o formato das publicações (986 vídeos e 9014 fotografias). As legendas foram analisadas para contextualizar temporalmente as publicações, uma vez que parte delas era anterior ao programa #EstudoEmCasa. Tal permitiu concluir que a *hashtag* tinha sido alvo de apropriação noutros contextos que não os do âmbito do programa.
As imagens também foram exploradas tematicamente e seis dimensões foram identificadas. Cinco dessas dimensões estavam relacionadas com o contexto propriamente dito do estudo (dicas, explicações, materiais) e uma dimensão era específica sobre o programa português #EstudoEmCasa.

— Instagram | @EstudoEmCasa2020
A análise realizada ao conteúdo proveniente do perfil EstudoEmCasa2020 foi semelhante à do #EstudoEmCasa. A análise detalhada das publicações permitiu identificar duas grandes dimensões: 1) realização de tarefas escolares e 2) acompanhamento à distância das aulas. Possibilitou também identificar os *posts* que acumulavam mais *engagement* e concluir sobre os contributos do perfil no contexto da questão.
Relativamente às legendas dos *posts* associados ao #EstudoEmCasa e ao perfil @EstudoEmCasa2020, procedeu-se a uma análise emocional incidindo sobre a presença de *emojis* e a sua tipologia. Para tal, recorreu-se à ferramenta Text-Analysis.[5] A análise permitiu concluir quanto à tipologia de *emojis*, aos *emojis* mais ou menos utilizados, aqueles que automaticamente remeteram para o objetivo de estudo (através do *emoji* televisão) e para o contexto no qual se realizava o estudo (pela presença do *emoji* arco-íris).
— Twitter/X | #EstudoEmCasa
O conteúdo publicado no Twitter/X e associado ao #EstudoEmCasa foi analisado, permitindo classificar as publicações em cinco categorias. Uma vez que a *hashtag* também era usada noutros países, o país de origem também foi classificado. Através do Twitter/X, os utilizadores expressavam considerações sobre o programa. A análise do número de publicações em função do tempo permitiu concluir sobre dinâmicas *offline* que se refletiam no *online*.
— YouTube | "Estudo em Casa" e "#EstudoEmCasa"
Os vídeos publicados no YouTube (n=320) foram analisados com a pretensão de identificar quais os temas de conteúdo que tinham maior *engagement*. Esses foram avaliados em função da classificação nativa da própria plataforma e tendo em conta o somatório de visualizações, gosto, não gosto, favoritos e comentários. A lógica de uso mostrou como o conteúdo da categoria comédia se sobrepôs ao conteúdo da categoria educação. Os resultados permitiam discutir sobre o sistema de recomendação do YouTube que tende a apresentar ao utilizador conteúdo mais visionado dentro de determinada categoria (Rieder, Matamoros-Fernández e Coromina, 2018).

Analisar como é que os atores políticos em Portugal, Espanha, Itália e Grécia se destacam ao falar da Europa nas plataformas de redes sociais

Contexto

No âmbito de um projeto europeu — EUMEPLAT, *European Media Platforms: assessing positive and negative externalities for European Culture* — desenvolveu-se uma pesquisa sobre como diferentes agentes se referem à Europa e à cultura europeia em várias plataformas de redes sociais. Foram desenvolvidas várias pesquisas e

5 https://labs.polsys.net/tools/textanalysis/

análises no âmbito do projeto pelos investigadores que o integravam, sendo uma delas a forma como os atores políticos de quatro países mediterrâneos (Portugal, Espanha, Itália e Grécia) se destacam a falar da Europa em três redes sociais distintas: Facebook, Twitter/X e YouTube.[6]

Recolha de dados

A definição dos países em análise não se deveu apenas ao facto de serem países do sul da Europa, mas particularmente devido à história recente partilhada com a crise económica da década de 2010 e a sua relação com as instituições europeias e internacionais. As três plataformas escolhidas (Facebook, Twitter/X e YouTube) são de forma consistente as mais utilizadas pelos cidadãos dos países em análise (Newman *et al.*, 2021). Após a definição dos países e das plataformas em estudo, foi preciso definir as dimensões e o período de análise. Em relação às dimensões de análise, além do tema geral da Europa, foram também acrescentadas como alvo de objeto de estudo outros temas como a saúde, clima e economia — as áreas de maior relevância para os cidadãos europeus segundo o Eurobarometer 2020 (European Commission, 2020). Foram, assim, criadas quatro dimensões de análise: Europa; Saúde e Europa; Clima e Europa; Economia e Europa. Para estudar estas quatro dimensões, foram definidas palavras-chave nas línguas dos países em análise para criar as *query*. O período de análise foi composto por três momentos diferentes: em setembro de 2021; em outubro de 2021; em novembro de 2021. O objetivo do estudo não consistia numa análise ao longo do tempo, mas sim numa análise sincrónica. A recolha de dados foi feita com o recurso a três ferramentas: CrowdTangle (para o Facebook), Brandwatch (para o Twitter/X) e YouTube Data Tools (para o YouTube). Estas três ferramentas utilizam as API (*Application Programming Interface*) das plataformas, o que garante que os dados recolhidos estão de acordo com os termos e condições das mesmas.

Análise de dados

Uma forma de analisar o conteúdo de publicações é codificá-las, ou seja, definir categorias (códigos) que vamos atribuir às publicações. Dependendo do que se está a estudar, será possível definir que uma publicação pode pertencer a apenas uma categoria ou a várias. Podemos ainda criar subcategorias, no sentido em que uma publicação pode pertencer a diferentes categorias, mas dentro de cada categoria só podemos escolher uma das subcategorias (são mutuamente exclusivas). A operacionalização desta codificação pode ser feita em Excel, em que cada linha é uma publicação e, além das colunas com a publicação e a sua metainformação (os *outputs* gerados pelo CrowdTangle, por exemplo), acrescentam-se as colunas relativas às diferentes categorias. Nesse caso, para cada coluna com uma categoria diferente codifica-se cada linha com "0" (não pertence a essa categoria) ou "1" (pertence a essa categoria).

---

6 Pode ler o estudo completo em: Moreno, J. C., S. Ferro Santos, e R. Sepúlveda (2024), "Taking Europe home: how political agents stand out in their approach to Europe on social media", *Observatorio (OBS*)*, 17 (5), https://obs.obercom.pt/index.php/obs/article/view/2419

No caso deste estudo, foi necessário fazer um passo intermédio — ou uma primeira categoria — que consistia em identificar as publicações que estavam dentro e as que estavam fora de âmbito. Apesar de se ter usado uma *query* com palavras-chave para identificar publicações com conteúdo dentro do âmbito pretendido de análise, podem ter sido recolhidas algumas publicações que usam as palavras-chave de formas não previstas e, por isso, que estejam fora do âmbito (p. ex., a palavra "euro" como referência ao preço de algo ou à competição de futebol, e não em referência à zona euro ou outras questões dentro do âmbito de estudo). O objetivo deste passo era identificar qual era o *top* 10 de publicações de cada país/âmbito/plataforma/mês que estivessem dentro do âmbito. O *top* 10 referia-se às publicações com mais interações (soma de reações, comentários e partilhas) no Facebook, com mais *reach* no Twitter/X e com mais relevância segundo o algoritmo de pesquisa do YouTube.

Após este passo intermédio, foi realizada a codificação das categorias e subcategorias para cada publicação no *top* 10 de cada país/âmbito/plataforma/mês, com base num *codebook* que foi desenvolvido pela equipa de investigação. O *codebook* foi criado com base na revisão da literatura das dimensões em análise (em particular na Europa, cultura europeia e europeização) e com base nas perguntas de investigação que tinham sido desenvolvidas. O objetivo era identificar padrões e criar categorias com base numa análise temática (Braun e Clarke, 2006). O processo de criar o *codebook* foi bastante iterativo, com várias fases em que se incluíram, se excluíram, se organizaram e se redefiniram categorias e subcategorias. Para garantir que todos os membros da equipa de codificação tinham o mesmo entendimento em relação às categorias, foram feitos testes com amostras da base de dados que foram codificadas por todos os investigadores de forma independente e cujos resultados foram comparados.

Exemplos de categorias (com escolhas múltiplas) que foram codificadas neste projeto são: tipo de agente que publicou o conteúdo (sendo que em fases posteriores só foram utilizados os agentes políticos); posição política do agente que fez a publicação; âmbito geográfico da publicação; etc.

Para dar maior credibilidade e rigor científico à análise foi calculado o *Intercoder Reliability* (ICR) para uma amostra de 20% do primeiro mês da amostra para cada país e para cada dimensão. Esta medida permite perceber se os dois ou mais codificadores estão a interpretar da mesma forma as publicações e as categorias, dando solidez à análise feita porque lhe retira subjetividade.

Após a codificação dos agentes que fizeram as publicações no *top* 10 foi possível alcançar a amostra final pretendida para a investigação (que consistia em 306 publicações), identificando, assim, as publicações mais relevantes sobre a Europa por parte dos agentes políticos nos quatro países em estudo. A análise de dados realizada sobre esta amostra final permitiu responder à questão de partida: como é que os atores políticos em Portugal, Espanha, Itália e Grécia se destacam a falar da Europa nas plataformas de redes sociais? O resultado desta investigação foi publicado por Moreno *et al.* (2024).

Desinformação em formato áudio no WhatsApp durante a pandemia de covid-19

Contexto

Durante a pandemia de covid-19,[7] houve uma proliferação significativa de conteúdo falso e/ou enganador disseminado através do WhatsApp. Em março de 2020, desenvolvemos uma investigação que visava explorar a disseminação de desinformação na plataforma de mensagens, nos estágios iniciais da pandemia em Portugal.

Pergunta de partida

Qual foi o papel da desinformação sobre a covid-19, nomeadamente em formato de áudio, que circulou, em Portugal, no WhatsApp, no início da pandemia?

Recolha de dados

Para a recolha de dados, foi criada uma conta nova no WhatsApp, especificamente para este projeto de investigação. Esta conta recebeu 988 unidades de conteúdo relacionadas com a covid-19, entre os dias 12 e 15 de março de 2020. Os dados foram enviados na sequência de um apelo público nas redes sociais do Iscte-IUL e num meio de comunicação social, o jornal *Diário de Notícias*, incentivando os utilizadores a enviar informações sobre a covid-19 recebidas pelo WhatsApp. Esta metodologia de apelar aos utilizadores que encaminhem conteúdo, informando-os em resposta de todos os dados do projeto e garantias de proteção de dados, é uma boa forma de contornar as limitações inerentes ao acesso de plataformas mais restritas, como as plataformas de mensagens (Piaia *et al.*, 2022).

Preparar os dados

O conteúdo de cada mensagem recebida foi manualmente descarregado, sem incluir dados identificativos, como o telefone ou o nome do remetente, e salvo na pasta correspondente à data de receção. As mensagens de texto foram copiadas e salvas em arquivos de texto. Todo o conteúdo duplicado foi identificado, contabilizado e catalogado. As mensagens foram classificadas como originais ou duplicadas, e os duplicados foram removidos para evitar redundâncias na análise. Uma das técnicas utilizadas que facilitou a definição de duplicados foi a identificação de ficheiros áudio/imagem e vídeo com a mesma dimensão (em termos de kb/mb), porque o conteúdo multimédia reencaminhado no WhatsApp mantém a dimensão. Esta estratégia de identificação de duplicados consistiu na visualização, por

7 Pode ler o estudo completo em: Cardoso, G., R. Sepúlveda, e I. Narciso (2022), "WhatsApp and audio misinformation during the Covid-19 pandemic", *Profesional de la Información*, 31 (3), e310321, https://doi.org/10.3145/epi.2022.may.21.

pasta, dos ficheiros por ordem de tamanho. Foi criada uma grelha de registo, com uma linha para cada unidade de conteúdo recebido, com *links* para o conteúdo multimédia correspondente, quando aplicável. O conteúdo exclusivamente em texto era copiado para a grelha de registo diretamente.

Análise de dados

Foi efetuada uma análise quantitativa e qualitativa sobre os dados. Em termos quantitativos, procuraram-se quantificar variáveis como o tipo de formato e as características do conteúdo. Os formatos de conteúdo recebidos incluíam texto, imagem, áudio e vídeo. As mensagens também foram codificadas com base em variáveis descritivas, como formato, a quem se endereçavam, se a comunicação estava na primeira ou terceira pessoa e se o conteúdo era, ou não, desinformativo. Foi criado um livro de códigos, baseado em estudos preliminares sobre a covid-19 e desinformação e alguns dados do próprio conteúdo. Este processo de codificação e organização de dados foi muito semelhante ao descrito no procedimento relativo ao estudo "Analisar como é que os atores políticos em Portugal, Espanha, Itália e Grécia se destacam ao falar da Europa nas plataformas de redes sociais". A análise destes dados teve uma componente de verificação de factos, que obrigou ao alinhamento com padrões internacionais da *International Fact-Checking Network*, nomeadamente a análise de fontes oficiais sobre as diversas alegações presentes no conteúdo. O conteúdo foi classificado em três categorias: verdadeiro, desinformação (conteúdo impreciso) e desinformação (conteúdo incorreto). Toda a amostra foi codificada, separadamente por dois codificadores, e o coeficiente Kappa de Cohen (0.81) foi calculado. Esta medida afere a concordância entre dois avaliadores na classificação de dados. O valor de 0.81 indica uma concordância substancial entre os avaliadores, demonstrando a confiabilidade de codificação dos dados.

A análise qualitativa focou-se nos áudios, e na identificação das narrativas desinformativas e das estratégias utilizadas na produção desse conteúdo. Foram observados temas como críticas à resposta das autoridades, capacidade dos serviços de saúde e teorias da conspiração. As estratégias identificadas incluíram a personificação, criando uma conexão emocional com o ouvinte, e a alegação de acesso privilegiado a informações governamentais. Este processo envolveu ouvir várias vezes os áudios, para identificar padrões e gerar temas e categorias. Muitas vezes nestes processos de categorização é necessária uma familiarização com os dados, antes de avançar para a criação de categorias.

A metodologia inovadora na recolha de dados, que envolveu a criação de uma conta específica no WhatsApp para receber conteúdo diretamente dos utilizadores, e a categorização do conteúdo recebido em diferentes camadas, permitiram uma compreensão aprofundada do contexto em que circularam os áudios desinformativos no início da pandemia de covid-19 em Portugal.

Analisar as redes de interação dos deputados da Assembleia da República no Twitter/X

Contexto

Têm sido realizados alguns estudos sobre os fenómenos de *filter bubble* (Pariser, 2011) e de *echo chamber* (Sunstein, 2006) nas redes sociais, nomeadamente no Twitter/X.[8] Estes fenómenos são referidos como sendo causas para a diminuição da qualidade deliberativa do debate público (Habermas, 2022). No entanto, há críticos desta visão, como Bruns (2021), que referem que a existência destes fenómenos tem tido resultados mistos nos estudos empíricos e que a base teórica assenta em determinismo tecnológico, escondendo o problema de fundo: que é social e político e não algorítmico. A maior parte dos estudos empíricos destes fenómenos em círculos políticos têm sido realizados em países com sistemas políticos diferentes do português. Em Portugal, o recrutamento legislativo acontece, na sua maioria, dentro do partido (Teixeira *et al.*, 2012), e a taxa de adoção do Twitter/X é bastante reduzida (Haman e Skolnik, 2021), o que pode alterar as motivações para os agentes políticos interagirem em plataformas de redes sociais.

Recolha de dados

O primeiro passo para realizar a recolha de dados foi identificar que deputados tinham conta no Twitter/X. Utilizando o *website* do Parlamento, foram identificados os 230 deputados em 2 de abril de 2022. Desses, foi possível identificar conta de Twitter/X com acesso público para 128. O segundo passo foi delimitar o período de recolha de dados, tendo sido definido o período de uma semana em quatro meses diferentes (abril, maio, junho e julho de 2022). Não se pretendia estudar o fenómeno de campanha eleitoral, e ao escolher quatro períodos distintos e espaçados no tempo foi possível analisar diferentes momentos da vida política, como a aprovação do Orçamento do Estado. Para recolher as publicações e metainformação das publicações de Twitter/X dos deputados, foi utilizada uma API do Twitter/X com acesso livre para trabalhos académicos, mas que já não se encontra em funcionamento. O resultado da recolha de dados originou um total de 2192 publicações por 69 deputados (de todos os partidos políticos).

Peparação dos dados

Antes de realizar a análise e visualização da rede de interação foi preciso realizar dois passos intermédios. Primeiramente foi necessário identificar das 2192 publicações quais delas tinham alguma forma de interação. Foram codificados manualmente quatro formatos de publicação: *tweet* (44%), *retweet* (25%), *reply* (17%),

8 Pode ler o estudo completo em: Ferro-Santos, S., G. Cardoso, e S. Santos (2024), "Para além da bolha (de filtro): interações dos deputados no Twitter", *Media & Jornalismo*, 24 (44), e4403, https://doi.org/10.14195/2183-5462_44_3.

quote-tweet (8%) e *tweet-mention* (6%). Destes, só os últimos quatro são de interação. Alguns tipos de interação — em particular os *quote-tweets* — só são possíveis de identificar através de codificação manual porque, na realidade, são *links* (para outro *tweet*). Caso se quisesse fazer redes de interação apenas com *retweet* e *reply*, não seria preciso fazer codificação manual.

O segundo passo de preparação consistiu na identificação de todas as contas com as quais os deputados interagiram, tendo sido identificadas e codificadas 757 contas diferentes. As contas foram categorizadas através do seu perfil (bio), *tweets* recentes e informação pública disponível no Google como sendo contas de esquerda ou direita, sempre que era possível fazer essa distinção, ou de *media*. No fim destes dois passos intermédios já tínhamos a informação de que precisávamos para fazer as redes de interação dos deputados: as contas com as quais os deputados interagiram (e a sua inclinação política) e as diferentes formas de interação.

A análise da rede de interação e a sua visualização foram feitas com o *software* Gephi.[9] Para usar o Gephi na construção de uma rede de interação é preciso criar dois ficheiros de Excel diferentes: um para os *nodes* (neste caso as contas de Twitter/X) e outro para as *edges* (neste caso a interação entre as contas). As *edges* podem ser unidirecionais — a conta A interage com a conta B — ou multidirecionais — a conta A e a conta B interagem mutuamente. No ficheiro dos *nodes* (as contas), foi acrescentada uma coluna para identificar se a conta era de *media*, e, não sendo, se era de esquerda, de direita ou se não era possível identificar a inclinação política. Foram feitos dois pares de folhas de Excel *nodes*/*edges*: uma para os *retweets* e outra para os *reply*. As outras duas formas de interação só foram analisadas estatisticamente e não com visualização porque não tinham tanta expressão.

Explorar os dados

Importar os dados para o Gephi:

— abrir o Gephi;
— ir ao menu "File" e selecionar "Import Spreadsheet";
— selecionar o ficheiro de Excel que foi criado com os "nodes" e na opção "import as" escolher "nodes tables";
— confirmar no menu "Data Laboratory" que os ID e as *label* (nomes das contas) estão em conformidade com o ficheiro Excel;
— ir novamente ao menu "File" e selecionar "Import Spreadsheet";
— selecionar o ficheiro de Excel que foi criado com os "edges" e na opção "import as" escolher "edges tables" — este ficheiro só será aceite se o nome das duas colunas for "source" e "target" e se não houver ID que não estão presentes já no *Data Laboratory* (que foram importados no ficheiro "nodes");
— confirmar que no "import report" o "graph type" é "directed" e escolher a opção "append to existing workspace".

9 https://gephi.org/

Visualizar a rede no Gephi

— Ir ao menu "Overview", onde já estará uma primeira representação visual;
— na parte inferior do painel, clicar no "T" para que a "label" (o nome das contas) fique visível. Nesse menu, também é possível alterar o tamanho e a cor da "label";
— no menu à direita fazer "run" do "average degree" (guardar o relatório se quisermos os dados);
— no menu "appearance", à esquerda, escolher a imagem da tela de pintura e a opção "partitition". Neste exemplo, nas opções, poderíamos escolher "inclinação política" como o critério para alocar uma cor diferente aos *nodes*. Selecionar "apply";
— no menu "appearance" à esquerda, escolher a imagem seguinte (bolas com diferentes tamanhos) e selecionar "ranking". Na escolha do "attribute", podemos escolher se queremos que os nós sejam maiores com base no *out-degree* (uma conta que respondeu/fez *retweet* a muitas outras seria maior) ou no *in-degree* (uma conta que tem muitas respostas/*retweets* seria maior). Depois de escolher o atributo, alteramos o valor mínimo e máximo de cada nó. Selecionar "apply";
— no menu "layout", à esquerda, mas em baixo, escolher a opção "Force Atlas 2". Esta opção organiza a visualização por "comunidades" aproximando os nós com mais ligações em conjunto. Selecionar "run" e "stop";
— após a aplicação do "Force Atlas 2", no mesmo menu, selecionar a opção "Noverla" e correr essa opção para separar os nós. Selecionar "run". Se os nós continuarem demasiado juntos e os quisermos separar ainda mais para garantir uma melhor leitura, selecionar, no mesmo menu, a opção "Expansion". Fazer "run" desta opção até os nós estarem com a visualização pretendida.

Apesar de não ter sido abordado no exemplo apresentado, é possível também atribuir características aos *edges* e alterar as cores com base nessas características. Por exemplo, se em vez de querermos ter duas visualizações diferentes para as *replys* e para os *retweets,* quiséssemos ter todas as interações no mesmo mapa, podíamos acrescentar uma coluna ao ficheiro de *edges* com a codificação de cada tipo de interação e definir a cor dos *edges* com base nessa característica.

Guardar imagem da rede

— Selecionar o menu "Preview" no canto superior esquerdo;
— fazer "Refresh" no menu, na parte inferior;
— adicionar "labels", se necessário, na parte "Node Labels" e "Show labels". Fazer "Refresh";
— quando a imagem estiver pronta para exportação, selecionar no canto inferior "Export SVG/PDF/PNG" e guardar a imagem no formato pretendido.

A figura 13.2 é um exemplo de uma imagem de rede de interação que foi exportada do Gephi. Neste exemplo do trabalho de Ferro-Santos *et al.* (2024) — da rede de contas a quem os deputados da AR fizeram *retweet* durante quatro semanas entre

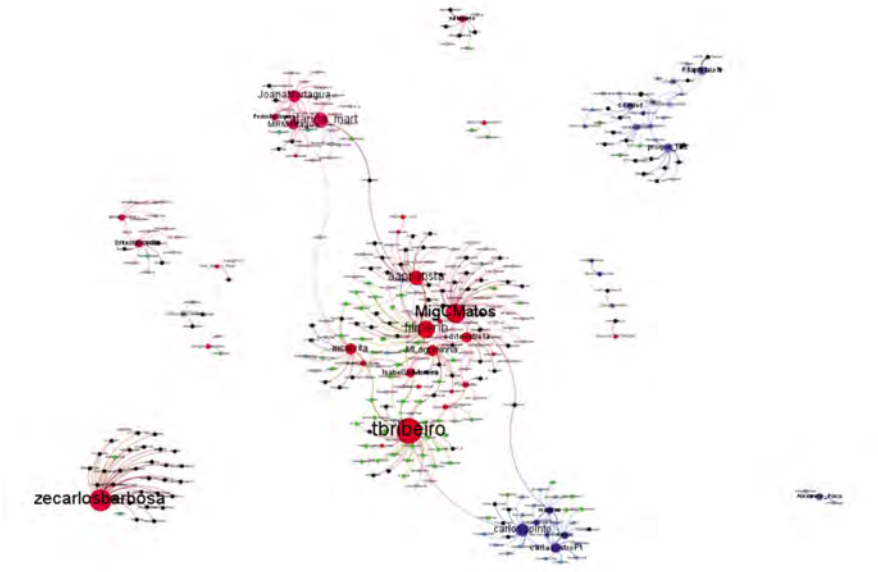

Figura 13.2 *Retweet network* dos deputados da Assembleia da República durante quatro semanas entre abril e julho de 2022

Fonte: Ferro-Santos *et al.* (2024).

abril e julho de 2022 — o tamanho dos nós foi baseado no *out-degree*, e as cores baseadas na inclinação política: azul — direita, vermelho — esquerda, preto — não identificável, e os *media* (verde).

Análise de dados com recurso a ferramentas de inteligência artificial

As ferramentas de inteligência artificial (IA), como o ChatGPT, têm demonstrado um potencial significativo, entre outras possibilidades, no auxílio na análise de dados provenientes do digital, especialmente de redes sociais *online* (Costa, 2023). A sua utilização começa a ser introduzida gradualmente na academia, e é importante conhecer as possibilidades que oferece, desde que considerando sempre também as suas limitações.

Análise de texto

Os modelos de linguagem generativa, como o ChatGPT, e outras ferramentas de NLP (Natural Language Processing) podem ser usados para análise de texto,

nomeadamente para categorizar automaticamente grandes volumes de texto, identificando temas e tópicos principais. São também extremamente úteis para fazer uma primeira análise rápida dos dados e identificar categorias ou temas que permitam a criação de um livro de códigos com base nos dados recolhidos. Estas ferramentas de NLP podem igualmente ajudar na análise de sentimentos, e na identificação de emoções predominantes, categorizando, por exemplo, excertos de texto como positivos, negativos ou neutros.

Exemplo: Festival da Eurovisão

- Objetivo: compreender o peso geopolítico do Festival da Eurovisão, usando como estudo de caso a participação de Israel em 2024;
- recolha de dados: utilize uma das várias ferramentas mencionadas neste manual para extrair comentários e *posts*, por exemplo do Facebook, sobre notícias do Festival da Eurovisão de 2024;
- configuração do ChatGPT: inicie sessão com o ChatGPT;[10]
- análise preliminar: prepare um *prompt* detalhado para a análise dos comentários. Um *prompt* é uma instrução ou comando fornecido a uma ferramenta de inteligência artificial, com o intuito de gerar uma resposta ou de guiar a ferramenta para executar uma ação específica. Os *prompts* a ferramentas de NLP são mais eficazes se identificarmos como o modelo deve agir, qual a tarefa e qual o tipo de *output* que pretendemos.
  - Exemplo de *prompt*: "Aje como investigador académico na área da geopolítica da cultura. Nos próximos *prompts* vou partilhar um conjunto de comentários extraídos das redes sociais a publicações sobre o Festival da Eurovisão.";
- identificação de categorias: Solicite ao ChatGPT para ler os comentários e identificar categorias principais.
  - Exemplo de *prompt*: "Analisa os comentários e procura identificar categorias para os agrupar de acordo com o pendor político dos mesmos, se não têm conotação política, e se têm, se existem subcategorias entre os comentários políticos. Identifique as categorias e subcategorias numa lista com a respetiva descrição de cada uma";
- codificação: solicite ao ChatGPT que identifique comentários relativos à participação de Israel no Festival da Eurovisão e que forneça os dados em formato CSV para futura revisão e análise.
  - Exemplo de *prompt*: "Coloque todos os comentários numa lista em formato CSV. Em frente à coluna de texto do comentário, identifique na segunda coluna com um SIM, todos aqueles que se referem à participação de Israel no festival.";
- análise de sentimento: solicite ao ChatGPT que avalie o sentimento relativo à participação de Israel no Festival da Eurovisão.

10 Pode, se desejar, usar outras ferramentas de NLP, como o CoPilot ou o Hugging Face.

    - Exemplo de *prompt*: "Classifique os seguintes comentários como positivos, negativos ou neutros: [*copiar e colar todos os comentários identificados*]";
- revisão e ajustes: reveja as categorias, comentários e sentimentos identificados, ajustando conforme necessário para garantir precisão.

Análise de dados em formato tabela

As ferramentas de IA podem ajudar logo na fase de preparação dos dados, na sua uniformização e otimização, na identificação de anomalias e na limpeza de bases de dados, bem como na sua conversão em formatos mais úteis. Após a sua transposição para um formato CSV ou semelhante, ferramentas como o CoPilot 365 ou o GPTExcel conseguem realizar análises estatísticas e gerar *insights* sobre os dados, criando relatórios e gráficos detalhados e identificando ligações entre variáveis, que, em grandes volumes de dados, podem não ser percetíveis.

Exemplo: análise de publicações sobre a medida do Autovoucher

- Contexto: o Autovoucher foi um apoio dado pelo XXIII Governo de Portugal, para combater o aumento dos preços dos combustíveis, que vigorou entre novembro de 2021 e março de 2022;
- objetivo: extrair, uniformizar e analisar dados de publicações em redes sociais e de pesquisas no Google entre outubro de 2021 e abril de 2022 sobre a medida do Autovoucher. O objetivo era compreender o impacto da medida nas redes e as reações, e identificar variações durante o período analisado;
- recolha de dados: utilizar ferramentas de extração de dados, como o ZeeSchuimer ou o SentiOne, já exploradas no presente manual, para extrair em formato tabela as publicações sobre a medida do Autovoucher do TikTok, Facebook e Twitter/X para o período analisado. Recolher também, para o período analisado, as tendências de pesquisa do Google sobre o termo, em formato tabela;
- uniformização dos dados das redes sociais *online*:
    - abrir o CoPilot 365, uma ferramenta de inteligência artificial da Microsoft que permite a realização de *prompts* diretos sobre ficheiros Office, nomeadamente Excel. Para aceder ao CoPilot 365 é preciso ter uma licença de Office e de CoPilot365. O custo atual da licença de CoPilot 365 é de 29 euros por mês, por utilizador. Pode aceder depois ao CoPilot 365 no Excel, no canto superior direito do ecrã. Não se esqueça de identificar sempre as células/colunas/linhas correspondentes nos *prompts;*.
    - uniformização de datas e horas: selecione a coluna de datas. Peça ao CoPilot 365 para padronizar o formato das datas. Exemplo de *prompt*: "Uniformize as datas na coluna A para o formato YYYY-MM-DD.";
    - uniformização dos tipos de conteúdo: selecione a coluna do tipo de conteúdo. Peça ao CoPilot 365 para categorizar o conteúdo como "texto", "foto", "vídeo".
      Exemplo de comando: "Considera que todos os conteúdos do Twitter/X e do Facebook têm o tipo de conteúdo identificado em inglês. Traduz

para português. Identifica todos os conteúdos de origem do TikTok como conteúdos em 'vídeo'";

- análise de interações antes, durante e depois da medida ser aplicada:
  - agrupamento mensal. Crie uma nova coluna para o mês. Peça ao CoPilot 365 para extrair o mês da coluna de datas. Exemplo de comando: "Cria uma nova coluna com o mês extraído da coluna de datas.";
  - análise de padrões: solicite ao CoPilot 365 que agrupe e conte as publicações em quatro fases: anúncio da medida, implementação da medida, decorrer da medida e fim da medida, e analise as interações. Exemplo de comando: "Agrupa as publicações por quatro períodos e calcula a média de interações para cada mês.";
  - análise de ligações: solicite ao CoPilot 365 que procure ligações entre dois conjuntos de dados, como a covariância, a correlação ou até a criação de gráficos de dispersão. Vamos supor que queremos avaliar a correlação entre o número de interações e o número de pesquisas no Google.
    Exemplo de comando: "Considera a média de interações e avalia se existe correlação com as variações nas pesquisas Google";
- análise por origem das publicações:
  - classificação: peça ao CoPilot 365 para classificar as publicações de acordo com a origem (cidadãos, organizações civis, organizações políticas, instituições públicas, notícias). Exemplo de comando: "Classifica as publicações na coluna [*indique nome ou número da coluna*] conforme a origem: cidadãos, organizações civis, organizações políticas, instituições públicas, notícias.";
  - análise de interações: solicite ao CoPilot 365 que compare o número de interações ou as plataformas mais usadas pelos diferentes grupos. Exemplo de comando: "Compara o número médio de interações entre diferentes tipos de origem;";
- procura de tendências:
  - peça ao CoPilot para analisar qual a característica que tem mais impacto no número de interações: a altura em que surge ou a origem da publicação?
    Exemplo de comando: "Considerando o número de interações com mais impacto, que variável tem mais impacto: a altura da publicação, a origem ou o tipo de conteúdo?"

Análise de dados com recurso à visão computacional

Ferramentas de inteligência artificial (IA) com capacidades de visão computacional podem interpretar e classificar imagens e vídeos, identificar objetos e até mesmo explorar sentimentos a partir de expressões faciais. São capacidades que não só permitem analisar grandes volumes de dados, como também restringir, de entre um grande lote de imagens, aquelas que, por exemplo, desejamos analisar em maior detalhe. Igualmente importante é o facto de as ferramentas de visão computacional conseguirem, caso as imagens e os vídeos contenham texto, extrair o mesmo. Esta é uma

vantagem operativa, uma vez que tal texto poderá ser mais relevante e central para a análise do que a própria imagem em si ou, no caso de dados recolhidos de redes sociais, ser mais relevante do que o texto que acompanha a publicação. Note que a colocação de texto na imagem é uma das estratégias de produtores de conteúdo que está a violar as normas das plataformas (como discurso de ódio, desinformação, apelo à violência, etc.) para evitar a deteção e potencial remoção da mesma (Dan *et al.*, 2021).

Entre as ferramentas disponíveis, o Memespector-GUI(Chao, 2021), de código aberto e sem necessidade de programação em Python, permite realizar análises de imagens usando API de visão computacional como Google Vision API, Microsoft Azure Cognitive Services e Clarifai.[11] Trata-se de uma ferramenta de uso gratuito que opera com recurso à visão computacional de plataformas comerciais para analisar imagens. As API a que recorre também costumam oferecer créditos gratuitos, que permitem processar milhares de imagens antes de exigir um pagamento. Embora seja particularmente útil para análises de grandes volumes de dados, também permite analisar pequenos volumes de dados.

Exemplo: estratégias de redução do desperdício no Instagram

- Objetivo: compreender que tipo de estratégias de redução de desperdício são utilizadas pelos utilizadores de Instagram, através da análise de publicações com a *hashtag* #zerowaste; recolha de imagens: utilize o PhantomBuster, cuja utilização foi detalhadamente explorada neste manual. Armazene essas imagens numa pasta do seu computador;
- importação para Memespector: aceda ao Memespector-GUI e selecione a opção de analisar uma pasta de imagens. Selecione a pasta com as imagens recolhidas;
- configuração do Memespector-GUI: configure a API de visão computacional de sua escolha (Google Vision API, Microsoft Azure Cognitive Services ou Clarifai):
  - adicione a sua chave de API no Memespector-GUI;[12]
  - selecione os parâmetros de análise, como deteção de etiquetas, objetos e texto;
- defina parâmetros: defina categorias e parâmetros de análise para aplicar no Memespector-GUI que lhe permitam compreender que tipo de estratégias de redução de desperdício são apresentadas pelos utilizadores do Instagram:
  - produtos sustentáveis:
    - produtos reutilizáveis: identificar garrafas de água, sacos de compras e outros produtos reutilizáveis;
    - produtos biodegradáveis: detetar utensílios de cozinha, escovas de dentes e outros itens biodegradáveis;
    - produtos a granel: analisar alimentos e produtos de higiene vendidos a granel;

11 https://publicdatalab.org/2021/10/27/memespector-gui/

12 Uma explicação mais detalhada de todo o processo para instalação e obtenção de chaves de API para uso no Memespector-GUI pode ser encontrada na sua página de GitHub: https://github.com/jason-chao/memespector-gui.

- práticas de redução de desperdício:
  - reciclagem e compostagem: identificar imagens mostrando práticas de reciclagem e compostagem;
  - DIY (*Do It Yourself*) e reaproveitamento: detetar projetos DIY e reaproveitamento de materiais;
  - lojas de 2.ª mão: identificar exemplos de compra ou venda em lojas ou plataformas *online* de venda em 2.ª mão;
- soluções sustentáveis:
  - hortas caseiras e urbanas: identificar iniciativas de hortas caseiras ou urbanas;
  - uso de energia renovável: detetar o uso de painéis solares ou outras fontes de energia renovável;
  - iniciativas comunitárias: analisar bancos de alimentos não utilizados, trocas de roupas e outras iniciativas comunitárias ligadas à economia circular;

— análise de conteúdo visual: use o Memespector-GUI para processar as imagens e identificar diferentes tipos de categorias:
- deteção de objetos:
  - objetos específicos: utilizar o Memespector-GUI para detetar objetos como garrafas reutilizáveis, sacos de compras, centros de compostagem, painéis solares, etc.;
  - objetos mais presentes: contar a frequência com que surgem os diferentes tipos de objetos em todas as imagens analisadas;
- classificação de práticas: classificar as imagens em categorias como reciclagem, compostagem, DIY, etc.;
- reconhecimento de texto:
  - texto em imagens: utilizar OCR (Optical Character Recognition) para extrair e analisar textos presentes nas imagens, como dicas, receitas e instruções de práticas sustentáveis. Ao usar uma ferramenta de NLP para agrupar estes textos extraídos de acordo com as diferentes práticas identificadas, cria mais uma camada de dados para análise;
- análise de cor:
  - cores predominantes: analisar as cores predominantes nas imagens para entender a estética visual associada ao movimento *zero waste;*
- análise de sentimento:
  - Expressões faciais e ambiente: analisar expressões faciais e ambientes das imagens para inferir sentimentos e emoções, como felicidade, tranquilidade e satisfação, associados às práticas *zero waste*;

— *outputs*: o MemeSpector-GUI tanto permite reunir as imagens de cada grupo numa subpasta, como permite a exportação numa tabela CSV que identifica as imagens e as etiquetas associadas;

— análise dos resultados: os resultados recolhidos permitem responder ao objetivo tratado, nomeadamente através de:
- identificação de padrões:
  - tipos de produtos mais utilizados: identificar os produtos sustentáveis mais comuns nas publicações;

  - práticas de redução de desperdício mais comuns: determinar quais as práticas que são mais frequentemente promovidas pelos utilizadores;
- análise temporal e geográfica:
  - temporal: estudar como as estratégias de redução de desperdício evoluem ao longo do tempo. Analisar tendências sazonais e eventos específicos que influenciam as práticas;
  - geográfica: comparar práticas de diferentes regiões geográficas para identificar variações culturais e regionais nas práticas de redução de desperdício.

Limitações e desafios

Embora as ferramentas de IA, em rápido desenvolvimento, ofereçam inúmeras vantagens na preparação e análise de dados, apresentam também limitações que devem ser consideradas (Karjus, 2023). Os modelos de linguagem e de visão computacional podem gerar informações imprecisas ou erros factuais. A precisão depende, entre outros, da qualidade e do volume dos dados com que o modelo treinou e da capacidade do utilizador em dar *prompts* (comandos) precisos. Pode ser necessário escrever vários *prompts* e ir afinando-os, até a ferramenta compreender na totalidade o objetivo do utilizador. Em tópicos mais representados há menos probabilidade de os erros ocorrerem, mas, mesmo assim, as análises realizadas por IA devem ser sempre complementadas com verificações humanas. Também é importante considerar que os modelos de IA podem refletir vieses e preconceitos presentes nos dados de treino, o que pode conduzir a resultados tendenciosos ou incorretos (Omena *et al*., 2023). É essencial identificar esses vieses durante a fase de desenvolvimento e utilizar técnicas de mitigação.

As conclusões e análises da ferramenta devem ser revistas e integradas no contexto, num texto da autoria do investigador. É importante, quando se utilizam dados não públicos, como, por exemplo, conteúdo de um grupo de WhatsApp, incluir, no formulário de consentimento facultado aos participantes, a indicação de que pondera utilizar uma ferramenta de IA para a análise de dados.

Relativamente à referência, em trabalho de investigação, do uso deste tipo de ferramenta para análise de dados, algumas revistas científicas (Grove, 2023; Nature, 2023) destacam, nas suas políticas de publicação, que a sua utilização deve ser divulgada na secção relativa à metodologia ou numa secção específica do artigo. Esta divulgação deve incluir o nome da ferramenta de IA, a versão e a forma como foi utilizada. Por se tratar de um tema em profunda mutação e desenvolvimento, é importante que os investigadores consultem regularmente o comité de ética da sua instituição e as políticas específicas das revistas onde pretendem publicar.

A importância e a necessidade da visualização de dados

Contexto

A visualização de dados é a representação gráfica de dados e informação por meio de elementos visuais (Botelho *et al.*, 2017). O seu objetivo principal é, ao facilitar a compreensão de valores, padrões, tendências, comunicar informações de forma clara e eficaz.

No contexto dos métodos digitais, a visualização de dados é particularmente relevante. Tal deve-se não só aos métodos digitais envolverem o uso de técnicas computacionais para recolher e analisar volumes de dados, muitas vezes de grande dimensão, e que são difíceis de interpretar através dos meios tradicionais, mas também porque esses dados provêm de plataformas digitais com *affordances* próprias e meios de expressão particulares. Representar dados que provêm de *hashtags* poderá requerer outras necessidades do que representar visualmente o uso de *emojis* como forma de expressão ou as visualidades associadas a um determinado tópico ou perfil no Instagram.

Consideramos então que uma vantagem significativa da visualização de dados em métodos digitais é não só a sua capacidade de simplificar informações complexas, já que grandes conjuntos de dados podem ser complicados e difíceis de entender quando apresentados em texto ou tabelas, mas também poder adequar dessa visualização em função da proveniência e tipologia dos dados.

Adicionalmente, a visualização de dados oferece suporte à análise exploratória dos mesmos. Apresentar os dados visualmente permite aos investigadores explorar padrões e relações que, de outra forma, seria mais desafiante ou menos evidente identificar, podendo levar a novas hipóteses e *insights*.

Além disso, a visualização de dados pode melhorar a qualidade narrativa da análise de dados. Contar histórias com dados é uma forma eficaz de transmitir informações complexas de maneira atrativa e acessível. Ao integrar elementos visuais nas narrativas, os investigadores podem criar histórias cujo impacto no público seja maior, tornando os resultados das suas investigações mais memoráveis e impactantes.

Pergunta de partida

Quais as visualidades associadas à hashtag #25deabrilsempre no Instagram em tempos de confinamento?[13]

Recolha de dados

É importante mencionar que a recolha de dados deste estudo foi realizada em 2020, através de uma ferramenta que atualmente já não está a funcionar. Como referimos ao longo deste manual, este é um dos desafios de fazer investigação no contexto

13 Pode consultar o estudo completo em: https://medialab.iscte-iul.pt/25-de-abril-no-instagram-a-celebracao-da-liberdade-em-tempos-de-confinamento/.

das plataformas sociais digitais. Porém, pode recorrer ao PhantomBuster. Para reproduzir este estudo ou outro semelhante cujo ponto de partida seja uma *hashtag*, siga a receita "Ponto de partida: *hashtag*" apresentada no capítulo deste manual sobre o Instagram. Quando reproduzida, explore o passo a passo "Obter imagens dos *posts*" também presente no mesmo capítulo.

Análise de dados

Uma vez que tenha obtido uma pasta com todas as imagens associadas à *hashtag* da sua recolha, vamos abri-la no programa ImageSorter.[14] Este *software* permite organizar conjuntos de imagens, ajudando a encontrar imagens semelhantes, sendo útil para identificar e explorar padrões.

Uma vez instalado e aberto o ImageSorter, terá de indicar onde é que, no seu computador, está a pasta das imagens que pretende explorar. Assim, no menu "Explorer" (lado esquerdo do ecrã), indique o caminho através da pasta "Users". Uma vez localizada a pasta, clique na mesma e, automaticamente, as imagens começarão a aparecer na parte central. Poderá ser necessário dar algum tipo de autorização para que tal aconteça. Na figura 13.3 está representado o resultado das imagens recolhidas associadas à *hashtag* #25deabrilsempre e visualizadas através do ImageSorter. Cada uma das imagens, na forma quadrada que compõe a figura 13.3, corresponde a uma das imagens dos *posts* cujos utilizadores do Instagram publicaram e, na legenda do mesmo, colocaram #25deabrilsempre.

Esta visão geral permite explorar padrões entre as imagens. Para facilitar esse processo, clique na opção "Sort by color" no ImageSorter. Esta encontra-se entre o menu "Explorer" e a parte central. É representada por um quadrado de várias cores e uma seta. Ao clicar nesse ícone, o *software* organiza as imagens por semelhança de cor. Essa organização e consequente exploração permite não só ter uma visão geral das imagens associadas, neste caso à *hashtag*, como também a grupos específicos de imagens. Este passo permitiu-nos, no estudo em questão (Sepúlveda e Crespo, 2020), identificar tipologias de expressões visuais no contexto do 25 de Abril e associadas a #25deabrilsempre e que podem ser observadas na figura 13.4.

Além das imagens dos posts procedeu-se à análise dos *emojis* presentes nas legendas dos posts como elemento visual através do qual os utilizadores comunicavam sobre o 25 de Abril. Para obter os *emojis*, por favor, siga as indicações do "Recolher *emojis* dos comentários" apresentadas no capítulo deste manual sobre o Instagram. Como estamos a recolher *emojis* de legendas de posts, e não de comentários, como é exemplificado no procedimento referido, deverá focar-se na coluna "Description" e não na coluna "Comment". Todos os demais passos mantêm-se.

O propósito é gerar a base de dados de *emojis* presentes nas legendas dos *posts* associados à *hashtag* que está a estudar, e que conterá os campos "emoji", "alias" e "frequency".

---

14 https://visual-computing.com/project/imagesorter.

Figura 13.3 Representação coletiva das imagens recolhidas associadas a #25deabrilsempre gerada pelo ImageSorter

Fonte: Sepúlveda e Crespo (2020).

Figura 13.4 Visualizações particulares das imagens recolhidas associadas a #25deabrilsempre

Fonte: Sepúlveda e Crespo (2020).

Para visualizar os *emojis* associados a tal *hashtag* em função da sua frequência, vamos usar o *software* RAWGraphs.[15] Esta ferramenta *online*, gratuita e de código aberto é uma aliada na visualização de dados. Note que, dependendo dos dados e objetivos, terá de preparar os mesmos antes de tentar gerar uma visualização.

Passo a passo:

a. Abrir https://www.rawgraphs.io/
b. Clicar em "Use it now".
   A geração de visualizações através do RAWGraphs está dividida em cinco campos: 1. "Load your data", 2. "Choose a chart", 3. "Mapping", 4. "Customize" e 5. "Export".
c. "Load your data"
   Neste primeiro campo, terá de colocar os dados que pretende visualizar. Pode fazê-lo através de diferentes formas (exemplos: copiar da sua base de dados e colar no RAWGraphs, fazer *upload* da sua base de dados ou através de um URL).
   Uma vez carregados os dados, terá a oportunidade de verificar se os mesmo estão em conformidade para os visualizar. Inconformidades com os dados costumam ser assinaladas a amarelo pelo RAWGraphs. Essas inconformidades podem ser resultantes da inapropriada preparação dos dados, pelo que deverá ter atenção à mesma. Verifique, também no RAWGraphs, se a tipologia de dados está correta (*string* representada por AA, *number* representada pelo símbolo "#" e *date* representada pelo símbolo relógio). Deverá fazê-lo na coluna de dados correspondente, na tabela gerada pelo RAWGraphs após ter carregado os dados.
   Neste passo, copiámos os dados provenientes da nossa base de dados de *emojis* presentes nas legendas dos *posts* (figura 13.5), após se ter seguido a lógica do procedimento "Recolher *emojis* dos comentários", e colámos no RAWGraphs;.
d. "Choose a chart"
   No segundo campo, vai poder eleger um gráfico entre os vários possíveis. Para entender as tipologias, diferenças e possibilidades dos gráficos, aconselhamos a que os explore um a um. Isto porque existirão gráficos mais adequados para determinado tipo de dados do que outros. Cada gráfico vem acompanhado com um tutorial sobre como usá-lo, o que ajuda bastante na construção do mesmo.
   Neste caso elegemos o Treemap. Tal como é indicado, o mesmo exibe dados estruturados hierarquicamente em função de uma dimensão quantitativa. É composto por retângulos cujo tamanho depende da dimensão quantitativa.

15 https://www.rawgraphs.io/

Quadro 13.1 Correspondência entre "dimensions" e "chart variables" do exemplo em questão

Dimensions	Chart variables
Frequency	Hierarchy
Frequency	Size
Emoji	Color
Alias	Label

Fonte: elaboração própria da autora.

e. "Mapping"
Neste terceiro campo, irá passar à construção propriamente dita do gráfico. Para tal terá de indicar que "dimensions" da nossa base de dados ("emoji", "alias" e "frequency"), que são indicadas no lado esquerdo do ecrã, correspondem a cada "chart variables" ("hierarchy", "size", "color" e "label"), campos indicados na parte central do ecrã. Para tal, terá de arrastar as dimensões da nossa base de dados ("emoji", "alias" e "frequency") para as correspondentes variáveis ("hierarchy", "size", "color" e "label").
Caso esteja a indicar alguma dimensão que não seja suportada pela variável em questão, o RAWGraphs irá assinalar.
No exemplo concreto com o qual estamos a trabalhar, o respetivo mapeamento está representado na quadro 13.1.
À medida que se vão atribuindo dimensões às variáveis, o RAWGraphs vai automaticamente gerando o gráfico. É uma funcionalidade útil para se ir verificando como este está a resultar, como os dados estão a ser distribuídos e ir modificando, caso se deseje ou seja necessário.

f. "Customize"
Neste campo, tal e como o nome indica, terá a oportunidade de customizar o seu gráfico. No parâmetro "artboard", poderá definir o tamanho do gráfico, das margens, a cor de fundo, se a legenda está presente ou não, e a dimensão desta. No parâmetro "chart", poderá determinar o tipo de retângulo através do qual o gráfico será construído. No parâmetro "colors", poderá indicar quais as cores associadas a cada um dos retângulos/dado e, por fim, no parâmetro "labels", poderá definir detalhes de como estas aparecem.

g. "Export"
No último campo, poderá exportar a sua visualização. Terá de atribuir uma nomenclatura à sua visualização (por defeito aparecerá "viz") e depois selecionar o formato no menu *dropdown*. Note que o formato SVG lhe permitirá editar a visualização em programas de edição de imagem.
Na figura 13.5, é possível observar a visualização final. Esta foi gerada no RAWGraphs e, posteriormente, editada para que os *emojis* tivessem uma dimensão maior. Algo que não é possível fazer no RAWGraphs.

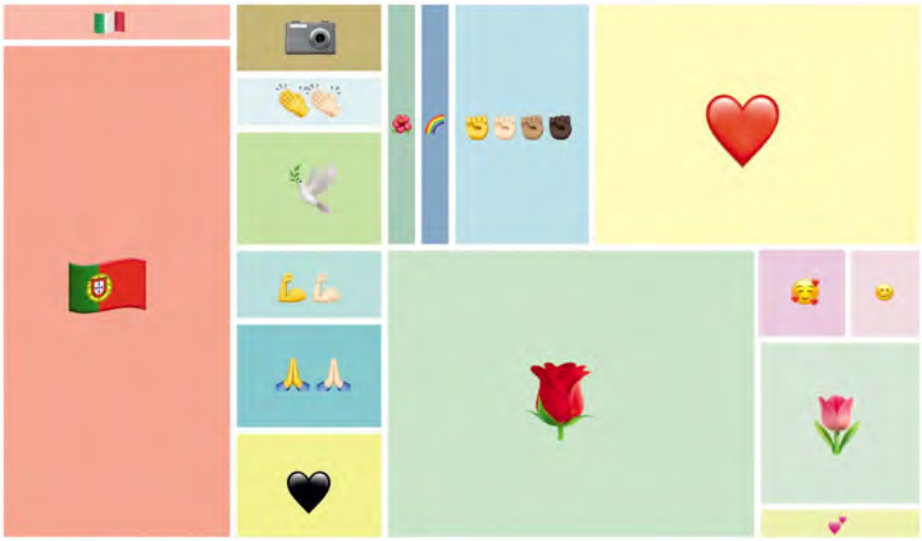

Figura 13.5 Resultado da visualização no RAWGraphs dos *emojis* mais frequentes entre as publicações associadas a #25deabrilsempre

Fonte: Sepúlveda e Crespo (2020).

Referências bibliográficas

Botelho, M. C., E.T. Vilar, E. Cardoso, A.A. Silva, P.D. Almeida, L. Rodrigues, e S. Rodrigues (2017), "The four faces of information visualization: a conceptual framework for a postgraduate program", em L.P. Reis, Á. Rocha, B. Alturas, C. Costa, e M. P. Cota (eds.), *2017 — 12th Iberian Conference on Information Systems and Technologies (CISTI)*, Lisboa, Portugal, IEEE.

Braun, V. e V. Clarke (2006), "Using thematic analysis in psychology", *Qualitative Research in Psychology*, 3 (2), pp. 77-101, https://doi:10.1191/1478088706qp063oa

Bruns, A. (2021), "Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problema", em M. Peréz-Escolar, e J. M. Noguera-Vivo, *Hate Speech and Polarization in Participatory Society*, Routledge, pp. 33-48.

Cardoso, G., R. Sepúlveda, e I. Narciso (2022), "WhatsApp and audio misinformation during the Covid-19 pandemic", *Profesional de la Información*, 31 (3), https://doi.org/10.3145/epi.2022.may.21

Chao, T. H. J. (2021), "Memespector GUI: graphical user interface client for computer vision APIs (version 0.2) [*software*], disponível em https://github.com/jason-chao/memespector-gui.

Costa, A. P. (2023), "Qualitative research methods: do digital tools open promising trends?", *Revista Lusófona de Educação*, 59 (59), https://doi.org/10.24140/issn.1645-7250.rle59.04.

Dan, V., B. Paris, J. Donovan, M. Hameleers, J. Roozenbeek, S. van der Linden, e C. von Sikorski (2021), "Visual mis- and disinformation, social media, and democracy", *Journalism & Mass Communication Quarterly*, 98 (3), pp. 641-664, https://doi.org/10.1177/10776990211035395.

European Commission (2020), *Standard Eurobarometer 93 — Summer*, disponível em https://europa.eu/eurobarometer/surveys/detail/2262.

Fereday, J., e E. Muir-Cochrane (2006), "Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development", *International Journal of Qualitative Methods*, 5 (1), pp. 80-92, https://doi.org/10.1177/160940690600500107.

Ferro-Santos, S., G. Cardoso, e S. Santos (2024). "Para além da bolha (de filtro): interações dos deputados no Twitter", *Media & Jornalismo*, 24 (44), e4403, https://doi.org/10.14195/2183-5462_44_3.

Grove, J. (2023), "First AI ethics policy unveiled by Cambridge University Press", *Times Higher Education*, disponível em https://www.timeshighereducation.com/news/first-ai-ethics-policy-unveiled-cambridge-university-press.

Habermas, J. (2022), "Reflections and hypotheses on a further structural transformation of the political public sphere", *Theory, Culture & Society*, 39 (4), pp. 145-171, http://doi.org/10.1177/02632764221112341.

Haman, M., e M. Skolnik (2021), "Politicians on social media. The online database of members of national parliaments on Twitter", *Profesional de la Información*, 30 (2), http://doi.org/10.3145/epi.2021.mar.17

Karjus, A. (2023), "Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence", https://doi.org/10.48550/arXiv.2309.14379

Light, B., Burgess, J., e Duguay, S. (2018), "The walkthrough method: an approach to the study of apps", *New Media & Society*, *20* (3), 881-900, https://doi.org/10.1177/1461444816675438

Liu, A. (2009), "Digital humanities and academic change", *English Language Note*s, 47, pp. 17-35.

Moreno, J. C., S. Ferro Santos, e R. Sepúlveda (2024), "Taking Europe home: how political agents stand out in their approach to Europe on social media", *Observatório (OBS)**, 17 (5), disponível em https://obs.obercom.pt/index.php/obs/article/view/2419.

Nature (2023), "Tools such as ChatGPT threaten transparent science; here are our ground rules for their use", *Springer Nature*, 623, https://doi.org/10.1038/d41586-023-00191-1.

Newman, N., R. Fletcher, A. Schulz, S. Andi, C. T. Robertson, e R. K. Nielsen (2021), *Reuters Institute Digital News Report 2021*, Oxford, Reuters Institute for the Study of Journalism, disponível em https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.

Omena, J.J., E. Bitencourt, J. Chao, A.M. Flores, R. Sepúlveda, L. Draisci, e Q. Du (2023), "Cross vision-API studies. Digital methodologies for understanding computer vision", *Digital Methods Initiative*, disponível em https://www.digitalmethods.net/Dmi/WinterSchool2023CrossVisionApiStudies.

Pariser, E. (2011), *The Filter Bubble: What the Internet Is Hiding from You*, Penguin.

Piaia, V., E. Matos, T. Dourado, P. Barboza, e S. Almeida (2022), "Ethical issues in WhatsApp research: notes on political communication studies in Brazil", *Revue*

Française des Sciences de l'Information et de la Communication, 25, https://doi.org/10.4000/rfsic.13328.

Rieder, B., A. Matamoros-Fernandez, e Ò Coromina (2018), "From ranking algorithms to 'ranking cultures': investigating the modulation of visibility in YouTube search results", *Convergence*, 24 (1), pp. 50-68, https://doi.org/10.1177/1354856517736982.

Rogers, R. (2013), *Digital Methods*, MIT Press.

Sepúlveda, R. e Crespo, M. (2020), "25 de Abril no Instagram: a celebração da liberdade em tempos de confinamento #25deabrilsempre", MediaLab Iscte, disponível em https://medialab.iscte-iul.pt/25-de-abril-no-instagram-a-celebracao-da-liberdade-em-tempos-de-confinamento/.

Sunstein, C. R. (2006), *Infotopia — How Many Minds Produce Knowledge*, Oxford University Press.

Teixeira, C. P., A. Freire, e A.M. Belchior (2012), "Parliamentary representation in Portugal: deputies' focus and style of representation", *Portuguese Journal of Social Science*, pp. 99-117, http://doi.org/10.1386/pjss.11.2.99_1.

Capítulo 14

Lista de ferramentas

Ao longo deste manual fomos referindo várias ferramentas e *softwares* que são de extrema utilidade no âmbito da pesquisa em contexto digital, sejam essas ferramentas e *softwares* indicados para a recolha de dados, outras que permitem a análise e aquelas através das quais é possível gerar visualizações. No quadro seguinte reunimos essas, e outras, ferramentas. Não obstante, como referimos ao longo deste manual é natural que tais ferramentas possam ficar desatualizadas, obsoletas ou deixem de funcionar. É um dos desafios inerentes a realizar investigar no contexto digital.

Ferramenta	Descrição	Acesso	Custo	Plataforma
4CAT Capture and Analysis Toolkit	Ferramenta para a captura de conteúdos das redes sociais	https://github.com/digitalmethodsinitiative/4cat	Gratuito	Telegram, Tumblr, Instagram, TikTok, 9gag, Imgur, Linkedin, X, Douyin, YouTube, Weibo
Brand24	*Software* de monitorização de redes sociais e *websites*	https://brand24.com/	Pago	X, Facebook, Instagram, YouTube, Linkedin, Reddit, TikTok, websites
Brandwatch	*Software* de monitorização de redes sociais e *websites*	https://www.brandwatch.com/	Pago	Facebook, Instagram, WhatsApp, Messenger, YouTube, TikTok, Linkedin, X, Reddit, Tumblr, websites
ChatGPT	Ferramenta de inteligência artificial.	https://chatgpt.com/	Freemium	N/A
CoPilot 365	Ferramenta de inteligência artificial de apoio nas ferramentas Office 365.	https://www.microsoft.com/en-us/microsoft-365/	Pago	N/A
DataTab	Plataforma de cálculos estatísticos	https://datatab.net/statistics-calculator/descriptive-statistics	Pago	N/A
Datawrapper	*Software* de visualização de dados	https://www.datawrapper.de/	Freemium	N/A
Epoch Converter	*Software* para converter datas	https://epochconverter.com	Gratuito	N/A
Export comments	Ferramenta para exportar comentários do Facebook	https://exportcomments.com/	Freemium	Facebook
Facepager	Ferramenta para extrair dados de redes sociais	https://github.com/strohne/Facepager	Gratuito	YouTube, X, Facebook, Amazon, websites
Find Facebook ID	Ferramenta para localizar o ID de uma página de Facebook	https://findidfb.com	Gratuito	Facebook
Flourish	*Software* de visualização de dados	https://flourish.studio	Freemium	N/A
Gephi	*Software* para visualização e análise de redes	https://gephi.org	Gratuito	N/A

Ferramenta	Descrição	Acesso	Custo	Plataforma
IBM Watson NLU	Ferramenta de processamento de linguagem natural	https://www.ibm.com/products/natural-language-understanding	Freemium	N/A
Image sorter	*Software* para visualizar e organizar imagens	https://visual-computing.com/project/imagesorter/	Gratuito	N/A
Infogram	*Software* de visualização de dados	https://infogram.com	Freemium	N/A
Mediacloud	Plataforma para recolha de dados de *websites* de notícias	https://www.mediacloud.org	Gratuito	Notícias
Meltwater	*Software* de monitorização de redes sociais e *websites*	https://www.meltwater.com	Pago	X, Facebook, Instagram, YouTube, Reddit, Twitch, Pinterest, websites
Memespector-GUI	Ferramenta que permite o uso de API de visão computacional	https://github.com/jason-chao/memespector-gui	Gratuito	N/A
Newswhip	*Software* de monitorização de redes sociais e *websites*	https://www.newswhip.com	Pago	Facebook, Instagram, Reddit, YouTube, TikTok, Linkedin, X, websites
PhantomBuster	Ferramenta para recolher dados de contas de redes sociais	https://phantombuster.com	Freemium	Instagram, Linkedin, Facebook, Twitter
Postman	Plataforma de desenvolvimento de API	https://www.postman.com/	Freemium	N/A
RawGraphs	Software de visualização de dados	www.rawgraphs.io	Gratuito	N/A
SentiOne Listen	*Software* de monitorização de redes sociais e *websites*	https://sentione.com/features/listen	Pago	Facebook, Instagram, X, YouTube, TikTok, Reddit, websites
StorySaver.net	*Software* para recolher Instagram *stories*, destaques e vídeos	https://www.storysaver.net/	Gratuito	Instagram
Table to net	*Software* para criar uma rede através de uma tabela	https://medialab.github.io/table2net/	Gratuito	N/A
Talkwalker	*Software* de monitorização de redes sociais e *websites*	https://www.talkwalker.com	Pago	Facebook, Instagram, X, TikTok
Telepathy DB	Ferramenta para pesquisa e recolha de dados do Telegram.	https://telepathydb.com/	Freemium	Telegram
Web Scraper	Ferramenta para fazer *scraping* de páginas *web*	https://webscraper.io/	Freemium	Websites
YouTube Data Tools	Ferramenta para extrair dados do YouTube.	https://ytdt.digitalmethods.net/	Gratuito	YouTube
Zeehaven	Ferramenta que converte ficheiros JSON em CSV	https://publicdatalab.github.io/zeehaven/	Gratuito	N/A
Zeeschuimer	Extensão do *browser* para fazer *scraping* de dados de redes sociais	https://github.com/digitalmethodsinitiative/zeeschuimer	Gratuito	TikTok, Instagram, X, LinkedIn, 9gag, Imgur, Douyin, Gab

Reflexões finais
Prospetiva e pesquisa digital: futuros, desafios, oportunidades e tendências

Gustavo Cardoso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

Inês Narciso
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

José Moreno
Instituto Universitário de Lisboa (Iscte-IUL), Centro de Investigação e Estudos de Sociologia (CIES-Iscte), Lisboa, Portugal

A generalização da comunicação em rede (Cardoso, 2023) torna os métodos digitais cada vez mais relevantes e necessários na investigação académica. Esta relevância decorre da capacidade dos métodos digitais em capturar a complexidade e a dinâmica do nosso mundo imerso na mediação, e da centralidade do meio digital como espaço de expressão e comunicação da atualidade (Cardoso, 2023; Kemp, 2024; Newman *et al.*, 2023).

Os métodos digitais permitem explorar fenómenos sociais em tempo real e de forma abrangente, utilizando técnicas integradas no meio que estudam, como *scraping*, análise de redes e visualizações de dados. Segundo Rogers (2024), os métodos digitais não são meramente adaptações de técnicas tradicionais, mas sim novas abordagens que aproveitam as características inerentes do mundo digital para a recolha e análise de dados.

Pela especificidade do meio em que são desenvolvidos, os métodos digitais apresentam desafios e oportunidades específicos, sendo que a constante mutabilidade e adaptação é uma das características que pode representar esta dualidade. Um dos objetivos deste manual, e de toda a oferta formativa desenvolvida em redor dos métodos digitais, é dotar o investigador de um conhecimento que lhe permita transformar a instabilidade dos métodos digitais em flexibilidade. Assim, encaram-se os métodos digitais como uma disciplina em atualização permanente, na procura de novas formas para recolha e análise de dados de um meio digital em constante evolução.

O acesso estável e constante aos dados é o principal desafio de quem recorre às metodologias digitais, sobretudo na recolha de dados das redes sociais. A frequente reconfiguração ou descontinuação de API e as dificuldades no acesso às mesmas são agravadas pela qualidade instável dos dados *online*, sobretudo para estudos longitudinais, devido a alterações na política de dados disponíveis ou até à eliminação de contas e páginas, como aconteceu durante a análise de campanhas de desinformação russa nas eleições presidenciais dos EUA (Albright, 2017). Adicionalmente, a política de eliminação de contas também apresenta desafios, como a maior percentagem de

contas pró-Brexit eliminadas após o referendo, comparativamente às contas que faziam campanha para permanecer na União Europeia (Bastos, 2021).

Este manual reconhece que o meio digital é uma fonte vital de dados que requer abordagens inovadoras e adaptativas. Neste capítulo final, procuramos traçar os atuais desafios e tendências na pesquisa digital. Em termos de desafios, explorar o atual panorama do acesso aos dados, as alterações recentes nas políticas de dados das plataformas, como o encerramento de API e de ferramentas de apoio como o CrowdTangle, focando também a legislação pertinente, como a Lei dos Serviços Digitais (DSA). Abordamos também as limitações e contingências que os investigadores deverão continuar a enfrentar, apesar da legislação. Estruturalmente, por um lado, a quantidade de dados que é possível obter é uma oportunidade relevante, por outro lado, o facto de os dados continuarem a ser controlados e condicionados pelas plataformas e pelos seus algoritmos é um desafio permanente à investigação.

Contextualmente, importa prestar atenção à generalização do crescimento da inteligência artificial (IA), da cultura de *celebridades* (Cardoso, 2023) e de uma oferta de recursos, cada vez mais próxima das redes e *media* sociais como o Facebook ou Instagram, por parte das plataformas de mensagens como o WhatsApp.

Desafios e oportunidades no acesso aos dados

A evolução das plataformas de redes e *media* sociais teve um impacto significativo nas metodologias utilizadas na investigação em ciências sociais. Inicialmente, as plataformas ofereciam um acesso relativamente aberto aos dados através das suas API, permitindo aos investigadores recolher e analisar grandes conjuntos de dados. No entanto, nos últimos anos, e particularmente em resposta ao caso Cambridge Analytica, tem-se observado uma posição mais restritiva nas políticas de acesso aos dados.[1] Este fenómeno, nomeado por alguns investigadores de métodos digitais como APIcalypse (Bruns, 2019), caracterizou-se pelo endurecimento gradual das políticas de acesso aos dados por parte de plataformas importantes como a Meta e o X (antigo Twitter).

O termo "APIcalypse" capta a tendência crescente das plataformas de redes sociais de restringirem o acesso às suas API, o que obrigou os investigadores a usar as ferramentas disponibilizadas pelas próprias plataformas ou a procurarem métodos alternativos de recolha de dados, como o *scraping* e a doação direta de dados pelos utilizadores (Bruns, 2019; Perriam *et al*., 2020). A pressão por parte da comunidade académica e da sociedade civil junto das instituições reguladoras, sobretudo perante o crescimento de fenómenos como a desinformação e o discurso de ódio *online*, contribuiu para a inclusão de legislação da União Europeia que torna obrigatória a partilha de dados das plataformas com os investigadores, nomeadamente o artigo 40.12 da Lei de Serviços Digitais (DSA).

---

1 Sobre o caso Cambridge Analytica: https://tek.sapo.pt/noticias/internet/artigos/as-perguntas-e-respostas-mais-importantes-sobre-o-caso-cambridge-analyticafacebook

A recente decisão da Meta de encerrar o CrowdTangle, uma ferramenta muito utilizada para a recolha de dados públicos do Facebook e Instagram, e a decisão do X (ex-Twitter) de fechar o acesso à sua API por trás de uma *paywall* com preços inacessíveis à comunidade académica exemplificam a continuação de uma cultura de restrição aos dados por parte das plataformas.

O DSA representa um esforço regulamentar significativo destinado a aumentar a transparência e a responsabilidade das plataformas digitais. O artigo 40.12 exige especificamente que as plataformas digitais de grande dimensão (VLOP) forneçam aos investigadores acesso, em tempo real, a dados públicos. Apesar do potencial do artigo 40.12, a sua implementação tem sido repleta de desafios, sobretudo com o acesso ao X (Silverman, 2024a).

Esta mudança de API abertas para modelos de acesso mais controlados tem implicações significativas para as metodologias utilizadas na investigação em ciências sociais, o que terá sido evidente ao longo de todo este manual. O acesso tradicional aos dados através de API permitia uma recolha de dados mais abrangente e sistemática, o que é crucial para estudos longitudinais e análises em grande escala. Em contrapartida, métodos alternativos como ferramentas de terceiros e *scraping* de dados são frequentemente menos transparentes na forma como selecionam e recolhem esses dados, e podem colocar desafios éticos e legais.

O novo modelo de acesso aos dados da Meta, que envolve a Biblioteca de Conteúdos, exemplifica as atuais limitações que os investigadores de métodos sociais enfrentam.[2] Embora a Biblioteca de Conteúdos ofereça algum nível de acesso a dados públicos, é muito mais restritiva do que o CrowdTangle (Silverman, 2024b). Os investigadores deparam com processos de candidatura rigorosos, portabilidade de dados limitada e requisitos de atualização de dados frequentes, o que prejudica a robustez e a flexibilidade da investigação sobre ou no âmbito das redes sociais.

O aviso, com apenas seis meses de antecedência, do encerramento da ferramenta CrowdTangle também exemplifica a importância de construir projetos de investigação que considerem os riscos inerentes à constante mutação de políticas de dados das plataformas, projetando, na sua origem, planos alternativos para resposta à pergunta de partida. Este panorama traduz também como é imperativo os investigadores explorarem novas estratégias de recolha de dados (Trans *et al.* 2024).

Além do acesso aos dados, a própria forma como estes são disponibilizados e o modo como expressam o comportamento dos utilizadores é um desafio para os investigadores. No primeiro caso porque os dados estão disponíveis apenas quando e como a plataforma decide e, em segundo lugar, porque são condicionados pelas regras de utilização e pelas gramáticas e *affordances* dessa mesma plataforma. Ao contrário do que normalmente se pode considerar, o investigador que consegue aceder aos dados dos utilizadores de uma rede social não acede aos dados em bruto, mas sim aos dados filtrados pela plataforma (ela disponibiliza apenas uma pequena parte dos dados coletados) e condicionados pela forma como os utilizadores podem comunicar nessa

2 Meta Content Library: https://transparency.meta.com/researchtools/meta-content-library/

plataforma. Uma publicação na plataforma X, limitada a 280 caracteres, é necessariamente diferente de um vídeo na plataforma TikTok, embora possam ser ambos a expressão de um conjunto de participantes nessas plataformas.

Por outro lado, os dados são igualmente condicionados pelos algoritmos das plataformas. Ou seja, os conteúdos que são mostrados a cada participante — com os quais este interage e aos quais reage, muitas vezes publicando novos conteúdos — são determinados por um algoritmo que o investigador não conhece e ao qual não tem acesso. Por isso, ao contrário do que possa parecer, os comportamentos de sociabilidade e de comunicação em plataformas *online* não são inteiramente espontâneos, mas antes suscitados pelas características e funcionamento da própria plataforma. O que isto significa, no fundo, é que a objetividade algorítmica é um mito (Gillespie, 2019). As plataformas digitais produzem um volume inédito de dados, com os quais o investigador pode trabalhar. Mas esses dados não garantem objetividade. Contextualizar a relevância da investigação digital dentro desta limitação é um dos desafios que se colocam ao investigador.

Acrescente-se que o que se passa no mundo mediado algoritmicamente (Cardoso, 2023) e o que ocorre sem mediação fora dele são dois fenómenos que estão intimamente interligados, mas não é possível extrapolar a partir de qualquer uma das dimensões para o todo. Embora o investigador, munido de grandes quantidades de dados sobre a comunicação que ocorre na dimensão mediada do nosso quotidiano, pudesse ser tentado a fazer extrapolações. Na realidade, tal não deve ser feito, pois os dados oriundos da mediação são moldados a partir das características e funcionalidades da plataforma digital em que ocorrem. Aquilo que for concluído a partir da análise da mediação apenas deve vincular esse contexto observacional.

No entanto, aos desafios colocados pela investigação a partir de dados digitais, também se juntam algumas oportunidades. A primeira é a enorme quantidade de dados disponíveis. Nos métodos de investigação tradicionais uma dos condicionantes é justamente, muitas vezes, a escassez de dados. No ambiente digital, pelo contrário, os dados podem ser abundantes e essa é uma oportunidade que se coloca aos investigadores. Além disso, os dados serem gerados sem a interferência do investigador, espontaneamente (tendo em conta as características e funcionalidades da plataforma já mencionadas), constitui uma vantagem adicional. No entanto, também constitui um desafio integrar e trabalhar essa grande quantidade de dados, tendo em consideração essa mesma espontaneidade, considerando que os dados não foram produzidos especificamente para a investigação. O desenho de pesquisa deve ter isso em conta e a recolha e análise dos dados deve refleti-lo.

Por fim, é também importante afirmar que a abundância e facilidade na interligação e relacionamento dos dados digitais constitui igualmente uma oportunidade para a investigação, uma vez que torna possível explorar simultaneamente várias hipóteses de uma forma diferenciada à dos métodos de investigação tradicionais (Salganik, 2019).

As tendências de contexto

Ao longo do presente manual, procurámos apresentar técnicas e ferramentas que permitam abordar de forma mais eficaz os desafios e oportunidades, considerando algumas das tendências que contextualizam o atual espaço comunicativo mediado. Entre elas, destacam-se três com influência direta nas abordagens metodológicas dos investigadores: 1. a consolidação de uma cultura da celebrização, que tem nos *influencers* a sua face mais visível; 2. integração da IA e as questões de autenticidade e originalidade que o seu uso generalizado implica, e 3. a transição para espaços sociais *online* menos abertos, nomeadamente as plataformas de mensagens como o WhatsApp e Telegram.

A consolidação de uma cultura da celebrização (Cardoso, 2023) é, simultaneamente, causa e reflexo de uma transformação significativa na forma como os participantes interagem com as plataformas. O fenómeno surge inicialmente ligado ao YouTube, mas tem vindo a ganhar transversalidade, com o crescimento do TikTok e dos Instagram *reels*. Neste contexto, alguns participantes transformam-se em criadores, que aproveitam as plataformas digitais para rentabilizar a sua produção criativa (Florida, 2022). Esta mudança cria uma divisão crescente entre a maioria dos participantes, que escolhem essencialmente interagir e partilhar, e os que buscam a celebrização, os influenciadores. Estes últimos procuram dominar e definir a agenda nas redes e *media* sociais. Os influenciadores buscam o envolvimento e procuram moldar o discurso público, muitas vezes motivados por oportunidades de monetização, tais como parcerias com marcas e financiamento direto das plataformas, mas outras vezes buscam apenas o reconhecimento *interparticipantes*.

Os influenciadores e criadores representam um estágio de evolução da alteração da paisagem mediática iniciada pelo *star system* do cinema, expandida pelo jornalismo cor-de-rosa, aprofundada pelos *reality shows* e generalizada pelas redes e *media* sociais (Cardoso, 2023). Tirando partido das características de funcionalidades das plataformas digitais, aqueles que buscam ser influenciadores e criadores de conteúdos tornam-se parte da desintermediação que as tecnologias digitais proporcionam (Robles-Morales e Córdoba-Hernández, 2019). Tratando-se de um processo social que acompanhou o século da mediatização, é de esperar que a reconfiguração continue, suscitando novos objetos e possibilidades de investigação.

No entanto, para os investigadores das redes e *media* sociais, é importante também considerar as limitações existentes numa paisagem povoada pela busca da celebrização. Se, por um lado, os influenciadores fornecem uma grande quantidade de dados sobre métricas de envolvimento, tendências de conteúdo e dados demográficos do público, por outro lado, a predominância de conteúdos gerados por influenciadores pode obscurecer tendências, preocupações e reações autênticas dos participantes em geral, tornando difícil captar uma verdadeira representação do sentimento da esfera pública na sua versão contemporânea. Nesse sentido, é importante desenvolver metodologias que diferenciem o conteúdo criado por influenciadores e o conteúdo criado pelos restantes participantes, garantindo uma compreensão equilibrada dos temas estudados. Os comentários e demais interações, como expressão mais comum dos participantes em geral em detrimento das

publicações autónomas dos influenciadores, ganham relevância como objeto de estudo, nomeadamente em plataformas como o TikTok e o Instagram.

Por sua vez, a integração da IA na criação de conteúdo revolucionou as redes e *media* sociais, permitindo a geração de conteúdo de forma rápida e pouco dispendiosa. Ferramentas de IA como o ChatGPT ou como o Midjourney, podem produzir texto, imagens e vídeos realistas, que parecem orgânicos, o que leva a preocupações sobre a autenticidade das interações e conteúdo que recolhemos das redes e *media* sociais, (Mishra e Awasthi, 2023).[3] À medida que o conteúdo gerado por IA se torna mais prevalente, a distinção entre o orgânico e o não orgânico torna-se mais difícil, devendo ser considerada pelos investigadores, uma vez que a categorização da validade dos dados pode depender também do tipo de autenticidade do conteúdo analisado.

Nos conteúdos gerados por IA, à questão da autenticidade individualizada e da convivência entre uma autenticidade oriunda do jornalismo e ciência e outra negociada caso a caso pelos participantes vem-se juntar, ainda, a categorização da origem do conteúdo face à sua autenticidade.

Além disso, a conjugação das várias tecnologias de IA permite automatizar processos de modo que seja montada mais facilmente uma rede de produção e disseminação de conteúdos, nomeadamente criando perfis falsos e gerando comentários e interações não orgânicos para promover determinadas narrativas, perfis ou conteúdos. A redução do custo deste tipo de infraestruturas de amplificação não orgânica, aproveitando-se do conhecimento adquirido sobre os algoritmos e as funcionalidades de recomendação de conteúdos, torna-as acessíveis à população em geral e o seu uso, potencialmente, generalizável ou, pelo menos, em maior escala.

No entanto, o incremento na capacidade da IA de produzir conteúdo realista e automatizar processos de amplificação é acompanhado de melhorias na deteção de conteúdos gerados por IA. Face ao ritmo de desenvolvimento da tecnologia é ainda cedo para avaliar o peso que esta questão terá no futuro, destacando-se a importância de acompanhar este desenvolvimento e o seu impacto nas metodologias digitais.

Do mesmo modo, são hoje em dia numerosos os sistemas de IA que proporcionam utilizações académicas muito significativas para a investigação. Tendo em especial consideração a oportunidade de recolher grandes volumes de dados através dos métodos digitais e as inúmeras possibilidade de relacionamento e associação entre esses dados, a inteligência artificial pode ter um papel muito relevante no apoio ao trabalho do investigador, não só no sentido de facilitar certas tarefas, como de permitir linhas de análise que de outra forma não seriam possíveis. Algumas dessas novas possibilidades na formatação e análise de dados, são exploradas no presente manual.

Em 2024, as aplicações de mensagens estão a ultrapassar as plataformas de redes sociais em termos de utilizadores mensais ativos em 20% (Curry, 2024; Kemp, 2024). Os dados sobre o número de horas passadas nas diversas aplicações apontam

3 https://openai.com/index/chatgpt/, https://www.midjourney.com/home

também para uma migração de participantes das redes e *media* sociais abertos para as plataformas de mensagens privadas, numa tendência que pode ser denominada como de adoção de "cortinas sociais". Pois, tal como as cortinas na janela não impedem o vislumbrar do interior de uma casa, apenas tornam a imagem mais difusa, também as plataformas de mensagens não estão a mudar as tendências já criadas pela comunicação em rede, mas estão a moldá-las de um modo diferente.

Várias variáveis podem explicar esta migração. Algumas, de caráter mais especulativo, como o conceito de *enshittification*, cunhado por Cory Doctorow, defendem que as plataformas, no seu lançamento, apresentavam serviços de alta qualidade para atrair novos participantes, mas que, gradualmente, degradaram esses serviços para dar prioridade aos anunciantes e ao lucro, levando ao afastamento dos seus participantes iniciais (Doctorow, 2024). No entanto, ainda que o grau de uso diminua, o abandono total das redes e *media* sociais não ocorre, o que nos permite colocar a hipótese de que os participantes na comunicação em rede, perante uma oferta de ferramentas diversificadas, estão apenas a assumir um comportamento racional de gestão da sua reserva pessoal perante as plataformas (Cardoso, 2023). Utilizando-as de forma diferente para fins diferentes, os quais requerem diferentes tipos de aproximação à reserva individual. Este fenómeno parece evidente, quando os participantes passam a usar mais as plataformas como o WhatsApp, que oferecem maior privacidade. Para as próprias plataformas, pressionadas por iniciativas regulatórias como a Lei dos Serviços Digitais (na Europa, mas também noutros continentes), essa deslocação dos conteúdos de plataformas abertas e visíveis para outra fechadas (nalguns casos encriptadas) e menos visíveis tem sido vista como uma forma potencial de escapar a essa pressão pública. Ou seja, podemos colocar a hipótese de ser do interesse estratégico das plataformas que os participantes interajam em círculos mais pequenos e menos visíveis, desde que o modelo de negócio publicitário continue a ser rentável dentro do ecossistema das suas marcas. Tal estratégia tende a ser mais resiliente para as plataformas que possuem a propriedade simultânea de redes e *media* sociais e, também, de mensagens, como é o caso da Meta, com o Facebook, Instagram, WhatsApp e Facebook Messenger.

Para os investigadores das redes e *media* sociais, esta migração coloca desafios significativos. Os dados destes ambientes fechados são menos acessíveis, necessitando de novas metodologias que respeitem a privacidade dos utilizadores e, ao mesmo tempo, permitam acesso aos dados.[4]

O desafio metodológico da mudança

Os desafios e tendências apresentados destacam a complexidade e a dinâmica da comunicação em rede e refletem a importância de acompanhar essas mudanças e

4 Algumas dessas estratégias foram exploradas no presente manual no capítulo que aborda as plataformas de mensagens.

adaptar as metodologias utilizadas. Flexibilidade necessária pela dificuldade no acesso aos dados devido ao encerramento de API e ferramentas de apoio, como o CrowdTangle, que a legislação emergente, como a Lei dos Serviços Digitais, ainda não veio colmatar. Por outro lado, mesmo no quadro da implementação dessa legislação e considerando que ela prevê o acesso dos investigadores aos dados, esse acesso continuará a ser condicionado e controlado pelas plataformas e os resultados continuarão a ser influenciados pelos algoritmos, algo que os investigadores continuarão a ter de enfrentar como um desafio. Paralelamente, a ascensão dos influenciadores cria novas dinâmicas de representatividade e comportamento nas redes, enquanto a integração da inteligência artificial levanta questões sobre a autenticidade dos dados. Por fim, a migração para plataformas de mensagens converge com os desafios já identificados sobre a dificuldade de acesso aos dados.

No entanto, a maior dificuldade não advém da mudança de um qualquer objeto das ciências da comunicação. Pois, no quadro científico aqui analisado, o das ciências da comunicação, coexiste na sua matriz fundacional de investigação a assunção prévia da instabilidade do objeto "comunicação".

Se os métodos digitais estudam a comunicação das pessoas nas plataformas, então, independentemente do uso dos dados recolhidos se destinar a uma abordagem de *business analytics*, sociologia, economia, psicologia social, antropologia social, história contemporânea, ciências da comunicação, serviço social ou urbanismo, este é e sempre será um estudo da imprevisibilidade. Pois estudam-se comportamentos sociais baseados na recolha de dados, essencialmente, oriundos da comunicação.

No entanto, esses dados não são todos inteligíveis para o investigador. Apenas os dados passíveis de serem tratados pelos métodos digitais são inteligíveis. Os dados inteligíveis para o investigador são os mesmos que fazem sentido para o emissor e recetor que os comunicam. Depois há os outros dados, os ininteligíveis. Isto é, os dados que apenas fazem sentido para os detentores do código em que são recolhidos, guardados e analisados. Isto é, os dados que as plataformas obtêm a partir da dataficação da comunicação (Cardoso, 2023). Estes últimos estão na maioria dos casos fora do alcance do investigador e dos métodos digitais, pois são a "alma do negócio" das plataformas. Este é o limite dos métodos digitais, aquele que é estrutural e não contextual, como os aqui atrás abordados.

Na realidade não existe um desafio metodológico da mudança, o desafio está na génese dos métodos digitais, está na sua flexibilidade para lidar com a mudança. Quanto à resposta à pergunta "que mudança poderá ser essa?", este manual não será, de certeza, o espaço para o debater, pois aqui procurámos fazer prospetiva, pois somos cientistas, e não futurologia, pois não somos futurólogos.

Referências bibliográficas

Albright, J. (2017), "Welcome to the era of fake news", *Media and Communication*, 5 (2), pp. 87-89, https://doi.org/10.17645/mac.v5i2.977.

Bastos, M. (2021), "This account doesn't exist: tweet decay and the politics of deletion in the Brexit debate", *American Behavioral Scientist*, 65 (5), pp. 757-773, https://doi.org/10.1177/0002764221989772.

Bruns, A. (2019), "After the 'APIcalypse': social media platforms and their fight against critical scholarly research", *Information, Communication & Society*, 22 (11), pp. 1544-1566, https://doi.org/10.1080/1369118X.2019.1637447.

Cardoso, G. (2023), *A Comunicação da Comunicação. As Pessoas São a Mensagem*, Lisboa, Mundos Sociais.

Curry, D. (2024) "Messaging app revenue and usage statistics", *Business of Apps*, disponível em https://www.businessofapps.com/data/messaging-app-market/.

Doctorow, C. (2024), "'Enshittification' is coming for absolutely everything", *Financial Times*, disponível em https://www.ft.com/content/6fb1602d-a08b-4a8c-bac0-047b7d64aba5.

Florida, R. (2022), "The rise of the creator economy", *The Creative Class Group*, disponível em https://creativeclass.com/reports/The_Rise_of_the_Creator_Economy.pdf.

Gillespie, T. (2018), *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*, Yale University Press, https://doi.org/10.12987/9780300235029.

Kemp, S. (2024), "Digital 2024: global overview report", *Data Reportal*, disponível em https://datareportal.com/reports/digital-2024-global-overview-report.

Luca, S., P. Schlesinger, A. Iramina, e A. McCluskey (2023) "Policy futures for the digital creative economy: proceedings of the University of Glasgow/University of Sydney Symposium", *CREATe Working Paper 2023/2*, University of Glasgow.

Mishra, A., e S. Awasthi (2023), "Chat GPT: revolutionizing communication or threatening authenticity?", *Management Dynamics*, 23 (1), pp. 165-168, disponível em https://doi.org/10.57198/2583-4932.1321.

Newman, N., R. Fletcher, K. Eddy, C.T. Robertson, e R.K. Nielsen (2023), "Digital news report 2023", *Reuters Institute for the Study of Journalism*, disponível em https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023.

Perriam, J., A. Birkbak, e A. Freeman (2020), "Digital methods in a post-API environment", *International Journal of Social Research Methodology*, 23 (3), pp. 277-290, https://doi.org/10.1080/13645579.2019.1682840.

Robles-Morales, J. M., e A.M. Córdoba-Hernández (2019), *Digital Political Participation, Social Networks and Big Data: Disintermediation in the Era of Web 2.0*, Palgrave Macmillan Cham, https://doi.org/10.1007/978-3-030-27757-4.

Salganik, M. J. (2019), *Bit by Bit: Social Research in the Digital Age*, Princeton University Press.

Silverman, B. (2024a), "European regulators are right to be concerned about the state of X's transparency", Mozilla Foundation, disponível em https://foundation.mozilla.org/en/blog/EU-Digital-Services-Act-and-The-State-of-X-Transparency/

Silverman, B. (2024b), "CrowdTangle is dead, long live CrowdTangle! Substack some good trouble", disponível em https://brandonsilverman.substack.com/p/crowdtangle-is-dead-long-live-crowdtangle/comments

Trans, M., D. Beraldo, L. Draisci, L. Afsahi, M. Brennan, V. Goldschmidt, e H. Xu (2024), "APIcalypse now: redefining data access regimes in the face of the Digital Services Act", *Digital Methods Initiative*, disponível em https://www.digitalmethods.net/Dmi/WinterSchool2024APIcalypse.

White, J. (2024), "See how easily A.I. chatbots can be taught to spew disinformation", *New York Times*, disponível em https://www.nytimes.com/interactive/2024/05/19/technology/biased-ai-chatbots.html.

Manual de Métodos para Pesquisa Digital

Este manual reúne um conjunto de propostas práticas sobre como realizar investigação no âmbito das plataformas de redes e media sociais digitais e na pesquisa digital. Contempla também uma reflexão teórica sobre o que implica realizar investigação na qual o objeto de estudo são posts, vídeos, perfis ou hashtags em ambientes online variados e com características próprias. Trata-se de um manual inovador pensado para investigadores, estudantes, docentes e todos os profissionais que, embora não pertencendo ao universo académico, desejam realizar análises no âmbito de plataformas digitais e/ou sobre as expressões de dimensão social, económica, financeira, cultural e política que com elas se relacionam.

Gustavo Cardoso
Professor catedrático de Ciências da Comunicação no departamento de Sociologia do Iscte-IUL. Dirige, na mesma instituição, o Doutoramento em Ciências da Comunicação e o MediaLab CIES. Coordena o OberCom (Observatório da Comunicação) e a participação portuguesa no IBERIFIER – Iberian Digital Media Observatory. É membro da Academia de Ciências de Lisboa e foi agraciado com o grau de Grande-Oficial da Ordem do Infante D. Henrique.

Rita Sepúlveda
Investigadora no ICNOVA. O seu trabalho tem-se centrado na transformação da intimidade no contexto da apropriação de plataformas digitais. É docente convidada no Iscte-IUL onde leciona sobre comunicação e metodologias de investigação. É autora de vários artigos científicos, capítulos de livros e do livro *Swipe, Match, Date*.

IBERIFIER
Iberian Digital Media Observatory

editora MUNDOS SOCIAIS

cies _ iscte
Centro de Investigação e Estudos de Sociologia

iscte
INSTITUTO UNIVERSITÁRIO DE LISBOA

ISBN 978-989-8536-94-5