JCOM

BIG DATA AND DIGITAL METHODS IN SCIENCE COMMUNICATION RESEARCH: OPPORTUNITIES, CHALLENGES AND LIMITS

A análise de textos auxiliada pelo computador: um laboratório a céu aberto para as ciências sociais

Yurij Castelfranchi

Abstract

Graças, de um lado, à disponibilidade extraordinária de arquivos textuais colossais e, de outro, aos avanços das possibilidades computacionais, hoje o cientista social tem à disposição um extraordinário laboratório, feito de milhões de sujeitos interagindo e bilhões de textos. Uma oportunidade inédita para a ciência, mas também grávida de desafio. Como testar, corroborar modelos¿ Como controlar, interpretar, validar os Big Data¿ Qual o papel da teoria no universo dos padrões e das correlações estatísticas¿ Mostraremos neste comentário algumas características gerais do uso de ferramentas computacionais para a análise de textos, e algumas aplicações nas áreas da comunicação pública da C&T e dos estudos sociais da C&T, mostrando também algumas de suas limitações e armadilhas.

Keywords

Public understanding of science and technology; Representations of science and technology; Science communication: theory and models

O dilúvio informacional chegou às ciências sociais. Governos no mundo inteiro digitalizam e disponibilizam leis e projetos, debates e interrogações parlamentares, indicadores sociais e econômicos, financiamentos de partidos, programas de governo, enquetes e consultas de opinião públicas, mapas, modelos, graças a novas práticas de governança (e à nova retórica) da transparência, *accountability*, participação e deliberação. Dinâmicas dos mercados globalizados e demanda dos clientes obrigam grandes e pequenos veículos de mídia (diários e revistas, Tvs e rádios, infotainment na rede e nas redes sociais) a tornar possível o download de enormes bases de dados, incluindo arquivos daquilo que foi escrito, dito, cantado, ao longo de décadas.

Editoras publicam seus livros em formato digital. Bibliotecas e hemerotecas, bem como corporações como a Google, arquivam milhões de publicações, contendo partes consistentes da literatura e da história mundial. Novas formas de sociabilidade, on e off-line, novos modos, efeitos e afetos do individualismo contemporâneo, levam dezenas de milhões de usuários a ilustrar, narrar, editar, legendar na web 2.0 seus gostos e intimidades, sua cotidianidade e suas memórias, as interações, associações, conflitos, sonhos, rancores e confissões. Estamos, em

suma, ao mesmo tempo exposto a milhões de gigabytes de informação e, ao mesmo tempo, somos incansáveis produtores de ulteriores gigabytes de rastros digitais, sobre nossos movimentos, nosso consumo, nossas invenções, nossa imaginação. Longe de ser um cenário orwelliano, somos vigilantes espontâneos da vida e dos rastros dos outros, e cumplices de nossa vigilância, e auto-vigilância, eletrônica.

Graças ao dilúvio informacional, sonhos e pesadelos dos cientistas sociais se realizam, ao mesmo tempo. De um lado, um inédito, extraordinário, laboratório global de ciências sociais, um colossal experimento a céu aberto, tornou-se possível. Dezenas de milhões de sujeitos voluntários podem ser estudados, ao vivo ou em sua história, em suas interações, seus conflitos, sua sociabilidade, seus rituais, seus processos deliberativos. Modelos podem ser testados, corroborados, modificados, refutados. Novas hipótese podem ser formuladas. De outro lado, esses dados (e os modelos deles derivados) são difíceis de se controlar, analisar, interpretar, validar, refutar, e, o que é pior, segundo alguns, podem tornar as ciências sociais "irrelevantes" e o próprio método científico "obsoleto": teorias e interpretações seriam inúteis, nesta época de correlações e de algoritmos inteligentes, de reconhecimento de padrões? No presente comentário, mostrarei as características gerais e algumas aplicações para os estudos sociais da C&T e a comunicação da ciência do uso de ferramentas computacionais para a análise de textos, mostrando também algumas das promessas, das limitações e dos perigos da euforia computacional ligada aos big data.

A análise de textos auxiliada por computador

Embora a ideia de aplicar técnicas computacionais à produção, tradução, compreensão da linguagem natural fosse uma das primeiras tarefas e promessas para a disciplina da Inteligência Artificial [Castelfranchi e Stock, 2000], e embora desde o pós-guerra existissem programas para análise automatizada de textos, só em anos recentes podemos dizer que, de fato, a computação tornou-se ferramenta crucial para o cientista social, especialmente graças a dois importantes desenvolvimentos [Wiedemann, 2013; Wiedemann, 2015]. De um lado, a análise automática ou semiautomática de textos tornou-se uma opção cada vez mais interessante para o pesquisador graças à facilidade de acesso e ao crescimento imperioso da quantidade de textos em formato digital. O ecossistema dos big data textuais hoje à disposição dos cientistas sociais é extraordinariamente diversificado: entrevistas, discursos públicos, respostas em questionários, forum online, blogs e microblogs, matérias de jornais, comentários dos usuários em matérias ou em posts em redes sociais, leis e projetos, debates parlamentares, pedidos de patentes, transcrições de processos, acordos e contratos, encíclicas, por não falar nas gigantescas bases arquivadas nas bibliotecas e hemerotecas digitais.

De outro lado, a *computer-assisted text analysis* (e, mais em geral, o uso de software para auxiliar a investigação em ciências sociais) teve um impulso extraordinário devido aos avanços recentes, tanto na potência de cálculo das máquinas quanto na sofisticação dos modelos matemáticos e dos algoritmos para coleta, extração, visualização, análise e interpretação dos dados. Nas últimas décadas, novas abordagens estatísticas forneceram ao cientista social ferramentas particularmente adequadas ao estudo da complexidade das variáveis qualitativas e da linguagem humana. Ao mesmo tempo, a potência de cálculo crescente e o custo decrescente dos computadores e de softwares cada vez mais sofisticados, flexíveis e amigáveis, tornaram as máquinas assistentes de pesquisa ideais para muitos pesquisadores.

Hoje os softwares de análise, tanto qualitativa quanto quantitativa, vão muito além de ajudar na marcação de trechos interessantes de textos, ou contar quantas vezes palavras ou conceitos aparecem em um corpus. Permitem explorar não só o texto, mas seu contexto (geográfico, temporal, de meta-dados), e investigar não somente o conteúdo explícito da mensagem, mas também suas dimensões latentes e os aspectos semânticos da linguagem. Ao reconhecer padrões, correlações, identificar conceitos, tópicos, partes da linguagem, atores e suas relações, o software auxilia também análises interpretativas.

Assim, a partir das décadas de 80 e 90, especialmente na área anteriormente menos explorada da análise qualitativa assistida por computador (CAQDA — *computer assisted qualitative data analysis*), se difundiram pacotes software (tais como MAXQDA, ATLAS.ti, NVivo, ALCESTE, etc), que foram se enriquecendo rapidamente com ferramentas também mistas, quali e quantitativas (como nos pacotes QDAMiner e WordStat, Rapidminer, ou em projetos modulares e abertos, de grande interesse, como Knime).

Em menos de vinte anos se multiplicaram redes grupos de pesquisa, cursos especializados e manuais, coletâneas e textos de referências dedicados ao tema como um todo [Popping, 2000], ao uso do computador em abordagens qualitativas [Kuckartz, 2014] ou em áreas mais clássicas como a da análise de conteúdo [Riffe, Lacy e Fico, 2014; Neuendorf, 2016]. Também foi extensa a produção acadêmica dedicada ao território interdisciplinar ("E-humanities, "Digital humanities", etc.), que emergia na interseção entre humanidades, ciências sociais, computação [Schreibman, Siemens e Unsworth, 2016].

Tipologias e processos de computer-assisted text analysis

Há diversas formas de usar auxílios computacionais, dependendo dos dados à disposição, da abordagem teórica, dos objetivos e hipóteses de pesquisa. Mas o uso do computador começa antes de qualquer análise e escolha metodológica. Os algoritmos podem funcionar como baratos ajudantes para a coleta e o processamento dos textos. Para recolher dados, podem ser usados software de "*scraping*" de sites, portais e *feeds* rss, ou robots que efetuam o download de inteiras bases de dados, ou arquivam e organizam os rastros deixados por usuários de redes sociais (ver, por exemplo, Bucchi e Neresini [2011] e Neresini e Lorenzet [2016], bem como a discussão sobre o projeto "TIPS", descrito por Neresini na presente edição da JCOM). Além disso, programas de OCR (*Optical Character Recognition*) são cada vez mais eficazes em transformar imagens digitalizadas de publicações impressas e em textos digitais editáveis e analisáveis, e fornecem ao pesquisador riquíssimo material empírico, quase pronto para o uso.

Posteriormente a esta fase zero da pesquisa, o computador é um poderoso auxílio para o pré-processamento dos dados. Os softwares ajudam a eliminar aquelas preposições, artigos, advérbios, adjetivos que não sejam relevantes para a pesquisa, ou para efetuar procedimentos de "stemming" e "lematização", em que são agregadas palavras diferentes mas que remetem ao mesmo termo (por exemplo, por serem a versão singular, plural, feminina ou masculina, do mesmo vocábulo) ou ao mesmo campo semântico (por exemplo, sinônimos). Além disso, em fase de análise, o computador ajuda na mensuração da confiabilidade da codificação e na validação dos dados e resultados.

Por fim, na análise de textos propriamente dita, os softwares ajudam em ao menos três diferentes maneiras, dependendo da abordagem — dedutiva ou indutiva — e do tipo de operação de codificação do texto, automática ou manual (para uma taxonomia dos tipos de análises auxiliada por computador, ver, por ex., Wiedemann [2013] e Pollach [2012]).

1. Codificação manual, categorias construídas pelos pesquisadores.

Neste caso, o mais tradicional, a pesquisa prevê que os pesquisadores leiam, categorizem, analisem e interpretem o material textual, codificando e classificando as unidades de análise a partir de categorias por eles estabelecidas (por exemplo, com base em hipóteses e teorias prévias, ou com uma abordagem indutiva-dedutiva, como na *Grounded Theory*). Nesta abordagem de tipo dedutivo, com categorias estabelecidas a priori, o papel do pesquisador e o treinamento dos codificadores são cruciais, mas o computador auxilia em organizar, visualizar, recuperar os trechos codificados, e produzir tabelas e dados estruturados a partir do trabalho feito pelos humanos. Tanto em uma análise de conteúdo quantitativa, quanto em abordagens interpretativas, o trabalho de categorizar, anotar, interpretar textos e escrever sobre eles, pode ser auxiliado por software de CAQDA (tais como NVivo, Atlas.Ti, etc.) que constrói catálogos inteligentes, estruturados, do trabalho que o pesquisador humano fez.

2. Codificação automática, categorias construídas por pesquisadores

Mesmo quando as categorias de análise são dadas a priori, ou construídas a partir de uma primeira interpretação do material, algoritmos podem ser usados para auxiliar, especialmente no caso de *corpora* muito grandes, na tarefa de codificar o material. Por exemplo, a forma comum de efetuar análise de conteúdo é a de produzir uma hierarquia de categorias de análises articulada em um livro de códigos (*codebook*) que descreve em detalhes o que codificar, como, quando. O procedimento costuma ser dispendioso e precisa ser refinado com sucessivas rodadas de codificação, alternada por (frustrantes) averiguações da *inter-coder reliability*. Quando o corpus de texto é enorme, a tarefa se torna impossível. Na *Computational Content Analysis*, se tenta resolver o problema fazendo com que as categorias de análises sejam suficientemente simples (ou superficiais) ao ponto que a codificação possa ser feita automaticamente. O pesquisador pode, por exemplo, construir um dicionário, feito de conjuntos de palavras-chave indicativas de uma certa categoria (por exemplo: sentimentos agressivos, jargão científico, etc.).

3. Codificação automática e categorias construídas por via computacional

Um problema da codificação automática é que ela é limitada ao aspecto mais superficial da linguagem (conjuntos de palavras chaves), e ignora em grande medida o contexto, o significado dos textos. O auxílio computacional pode ajudar o pesquisador a dar mais espessura e robustez, mesmo em uma análise automática, no caso de uma abordagem indutiva, em que as categorias de análise não são escolhidas anteriormente, mas a partir de padrões, correlações, temas ou conceitos recorrentes, detectados por via computacional e que não necessariamente estavam presente nas hipóteses da pesquisa.

Exemplos desta construção indutiva de categorias são as abordagens como a análise de discurso automática, ou a lexicometria e a chamada *corpus linguistics*, em que a fase computacional do trabalho antecede a interpretação [Pollach, 2012; Cheng et al., 2008]. Mais em geral, a chamada mineração de textos (*text mining*) é justamente a tentativa de efetuar uma análise semântica dos textos e de sua estrutura [Wiedemann, 2013], extraindo o "sentido" por meio de métodos estatísticos que identificam características intrínsecas do *corpus* junto com aspectos mais interpretativos introduzidos por codificadores humanos. De um lado, então, o software faz emergir, de forma indutiva, padrões e correlações (entre conceitos, objetos, ou temas, por exemplo) presentes nos textos. De outro lado, os pesquisadores podem marcar características importantes do discurso, ou fornecer conjuntos de exemplos para treinar a máquina, e algoritmos de aprendizado podem inferir regras e aprender a codificar os textos.

O software identifica relações quantitativas, estatísticas, entre diferentes partes do texto. Este caminho, oposto ao da análise de conteúdo clássica (em que o pesquisador codifica primeiro, e depois o software calcula quantidades), permite identificar estruturas, recorrências, padrões no texto que não necessariamente haviam sido imaginados a-priori, o que torna este tipo de análise interessante até para pesquisadores tradicionalmente ligados a abordagens puramente interpretativas, como os analistas de discurso e os post-estruturalistas. Software bastante difusos, tais como Alceste, Wordsmith, TextQuest, foram desenvolvidos com este tipo de abordagens em mente [Wiedemann, 2013].

Análises auxiliadas por computador na comunicação pública da ciência e nos estudos sociais da C&T

Em sociologia, antropologia, ciência política, história e, obviamente, na área da comunicação, se multiplicaram, nos últimos anos, pesquisas que aproveitaram a potencialidade das ferramentas computacionais aplicadas à análise de textos. Não podia não acontecer o mesmo também nos estudos sociais da C&T e nas pesquisas sobre comunicação pública da ciência e da tecnologia. Os exemplos já são mais do que poderiam ser mencionados aqui, e vamos apenas mostrar uns poucos exemplos da diversidade de aplicações e abordagens possíveis.

Semino et al. [2005] utilizaram softwares de análise semântica de textos para o estudo do uso de metáforas na comunicação científica. Compararam um corpus de comunicação interna da ciência, entre pares (extraído da revista *Nature Immunology*), com um corpus de comunicação pública (artigos, sobre as mesmas temáticas, de uma importante revista de divulgação científica, o *New Scientist*) e descobriram, de fato, usos e funções diferentes para as metáforas no caso da divulgação e no caso da comunicação especializada. Ciência e tecnologias na mídia também estão sendo estudadas com auxílio de software, inclusive no caso de pesquisas que, classicamente, eram abordadas de forma interpretativa e qualitativa, como a análise de *frames*. Tian e Stewart [2005], por exemplo, abordaram a análise de enquadramento no caso da cobertura da crise da SARS, enquanto Bail [2016] estudou os *frames* no discurso sobre doação de órgãos. Crawley [2007] utilizou auxílio computacional em uma análise qualitativa da cobertura sobre biotecnologias na agricultura em diários comunitários. Castelfranchi, Massarani e Ramalho [2014], identificaram, graças ao uso de dicionários especificamente construídos e outros disponíveis sobre dimensões emocionais e cognitivas da linguagem, que o discurso de divulgação científica em importantes programas da

TV brasileira era marcado por metáforas de guerra, agressão, e fortemente conotado por desigualdade de gênero.

Mas o interesse em estudar grandes *corpora* com o auxílio de softwares não se esgota com o material que circula na mídia. Algumas pesquisas focaram na análise das respostas espontâneas a perguntas abertas em enquetes de opinião, um material empírico frequentemente subutilizado, exatamente pelo elevado custo de análise. Stoneman, Sturgis e Allum [2013] aplicaram técnicas de *clustering* estatístico à análise da descrição espontânea do significado do termo "DNA" fornecida por sujeitos de uma enquete. Já, Tvinnereim e Flø ttum [2015] identificaram temas recorrentes nas opiniões sobre mudanças climáticas, também em respostas a perguntas abertas. Utilizaram uma técnica relativamente recente, e inovadora, de *Structural Topic Modelling*, que permite detetar temas latentes nas declarações. Abordagens de *topic modelling* também foram usados para analisar os temas em grandes corpora de coberturas jornalísticas [Jacobi, Atteveldt e Welbers, 2016].

Sempre no campo dos estudos de caso ligados às mudanças climáticas, Farrell [2016] usou dados textuais e análise de redes sociais para investigar a influência que organizações que recebem financiamento de corporações têm sobre a polarização das opiniões. Veltri e Atanasova [2015], por sua vez, também a respeito de mudanças climáticas, aplicaram uma análise temática automática, em conjunto com análise de redes semânticas, para estudar o compartilhamento de informações através do twitter. Veltri e Suerdem [2013], por sua vez, utilizaram uma abordagem mista, que hibridiza métodos de categorização automática dos textos e codificação por parte de humanos, no estudo do discurso sobre OGM na Turquia. Uma combinação de codificação humana e classificação automática para análise de conteúdo (através do software DiscoverText) também foi intentada para estudar um caso de ativismo online contra a polêmica prática do "fracking" [Hopke e Simis, 2015].

As transcrições de discussões em grupos focais (ferramenta frequentemente utilizada para explorar atitudes e percepções sobre ciência, tecnologia e inovação), também podem ser analisadas com auxílio de ferramentas computacionais, como fizeram Miltgen e Peyrat-Guillard [2014] para investigar influências geracionais e culturais sobre as percepções a respeito da privacidade.

Nossa própria área, de estudos PCST, foi também estudada reflexivamente com técnicas destes tipos, por exemplo analisando a produção em uma das maiores revistas da área, a *Public Understanding of Science* [Bauer e Howard, 2012; Suerdem et al., 2013; Smallman, 2016].

O sucesso e as promessas

As vantagens do auxílio de software para a análise de textos são óbvias, e não apenas para quem está interessado em abordagens clássicas e quantitativas. A análise qualitativa é facilitada e potencializada pelo uso de software de CAQDA. Ferramentas de aprendizado de máquina, detecção de padrões, criação de dicionários, já demonstraram ser valiosas tanto para estudos de tipo interpretativo, quanto para descobertas de aspectos latentes de *corpora*, não sempre fáceis de identificar pela hermenêutica, nem por análises de conteúdo atreladas à mensagem explícita. Mas, para além da ampliação de possibilidades para coleta e análise de dados, a chegada dos algoritmos no gabinete do cientista social pode vir a ser um

novos *húmus*, capaz de estimular inovações teórico-metodológicas, impulsionar a criação de métodos mais ambiciosos e mistos. O computador vem, em certo sentido, como embaixador de uma parcial trégua na polêmica, pouco fecunda, entre defensores do "quali" e do "quanti" na análise dos textos. De um lado, a computação está tornando mais *fuzzy* a fronteira entre as duas, obrigando os pesquisadores a refletirem melhor sobre pesquisas *mixed-methods*, e a investir em equipes interdisciplinares. O *text mining*, embora fundamentado em ferramentas estatísticas e computacionais, não pode ser acusado de ser sinônimo de uma abordagem de cunho reducionista, ou criticado, com base em um *cliché* caricatural, de ser filho de uma "epistemologia positivista", pois auxilia o pesquisador "qualitativo" na extração do sentido e do contexto, e, ao mesmo tempo, na validação e no estudo quantitativo da relevância e confiabilidade da análise efetuada [Wiedemann, 2015]. Muitos hoje hoje na complementaridade de abordagens qualitativas (por exemplo fundamentadas na *Grounded Theory*) e quantitativas (como a análise de conteúdo computacional), e na utilidade de uma aprofundamento teórico e metodológico sobre modelos mistos e triangulação [Kuckartz, 2014]. A capacidade, ainda incipiente, mas crescente, das técnicas de topic modelling, linguística computacional, lexicometria, etc., de investigar o nível semântico, o contexto de enunciação, o lugar de fala, o sentido do texto, permite diminuir a distância entre aquilo que pretende fazer um pesquisador interessado na interpretação de seu objeto de estudo, e aquilo em que uma máquina pode contribuir.

Contudo, ainda está aberto, e acirrado, o debate sobre a influência que o uso de software pode ter sobre o próprio processo de pesquisa, por exemplo devido ao fato de que ele tende a fazer incorporar formas, entidades de conhecimento, unidades analíticas, estilos de pensamento predeterminados, tais como pensar o discurso e a comunicação em termo de unidades codificadas e hierarquias de códigos, de relações predefinidas, formatadas, entre entidades. Corremos o risco de enxergar somente aquilo que já estamos buscando, de conhecer apenas aquilo que o software é programado para detectar [Wiedemann, 2013]. E, mais importante, não podemos esquecer que dispor de novas ferramentas e técnicas não é o mesmo que ter inventado novas metodologias. E acessar mais dados não é sinônimo de ter resolvido problemas teóricos.

Os desafios e as limitações

O arauto da ideia de que o "dilúvio de dados" veio a nós, trazendo uma nova era, foi o editor da revista Wired, Chris Anderson. E a nova era, afirmava Anderson eufórico, é a do "fim da teoria" [Anderson, 2008]. Crianças da "Era do Petabybte", companhias como a Google, dizia o jornalista, não precisam mais de sociólogos, nem de modelos ou hipóteses: indexar, classificar, arquivar dados, e detectar seus padrões, regularidades, dinâmicas, é tudo.

Compreender e interpretar não é preciso. Esqueça a teoria, qualquer que seja a teoria. Não importa porque os seres humanos e os grupos humanos se comportam da forma que se comportam. Só importa rastrear, registrar, mensurar o que eles fazem e, *voilá*, "com suficientes dados, os números falarão por si mesmos". O próprio método científico, a formulação de hipóteses causais, a construção de modelos, os testes, os experimentos, estariam então obsoletos. Saber quais fatos, fenômenos, comportamentos estão correlacionados, seria mais do que suficiente: os algoritmos se encarregarão de encontrar padrões e previsões para o

comportamento humanos, onde teorias e modelos nunca conseguiram. Que a ciência abra alas para a chegada da Google.

Como era previsível, estas e outras afirmações dos entusiastas dos Big Data geraram um oceano de polêmicas, nas quais não pretendemos entrar aqui. Mas vale ressaltar ao menos uma pequena menção aos perigos e as armadilhas da ideia de que os dados, e os padrões, por si, sejam conhecimento. Um problema que há, pelo menos, uma dimensão epistemológica, uma técnica e uma marcadamente política.

Porque, se os dados são importantes, mais ainda o são modelos, hipóteses, interpretações. Para a ciência e para a tomada de decisão. Considerar padrões e previsões o novo sinônimo da palavra verdade, significa fazer uma política, "data driven", isto é, uma política sem política, não mais o lugar da escolha, e do conflito, sobre o viver bem em comum. Respondendo ao Anderson, o escritor James Bridle [2016], diz, por exemplo que a "crença no poder dos dados [...] conduz [...] à crença na verdade de afirmações derivadas de dados. E, se os dados contêm a verdade, então, produzirão, sem necessidade de intervenção moral, os melhores resultados". A segunda dimensão do problema, é técnica. Ao confiar no data mining, corremos o risco de esquecer, ou desconsiderar, que nem sempre sabemos como estimar os erros nos novos modelos estatísticos aplicados à análise de textos (em mais em geral, aos processos sociais). Segundo Grimmer e Stewart [2013], os métodos automáticos podem dar resultados relevantes apenas quando complementados por interpretação e leitura do material por parte dos pesquisadores. E não só: ao comparar diversos métodos, os mesmos autores concluem que tais métodos, no estado da arte atual, sofrem com problemas sérios de validação. Erros e armadilhas no uso de análises auxiliadas por computador são marcantes, e ainda não dispomos de metodologias adequadas para lidar com a automação.

O sociólogo Neal Caren [2015] tem posição parecida: a euforia sobre *Big Data* é acompanhada pelo entusiasmo em incorporar novos métodos estatísticos e computacionais, mas é cedo para avaliar até que ponto tais ferramentas sejam cientificamente robustas e heuristicamente fecundas. A capacidade de um modelo de *machine learning* em fornecer predições não necessariamente significa ter encontrado variáveis causais significativas para explicar um fenômeno social, ou entender realmente seu funcionamento.

Por fim, fora os problemas de validade, robustez, confiabilidade, há algo mais profundo na polêmica. O papel do método, da teoria, e o que entendemos por explicação e por ciência.

Massimo Pigliucci, filósofo da ciência, também critica o triunfalismo simplista de Anderson, e se pergunta: "se deixamos de procurar modelos e hipóteses, estamos ainda realmente fazendo ciência?" [Pigliucci, 2009]. Encontrar padrões é apenas uma parte da prática científica, que se completa ao buscar explicações para os padrões encontrados. Sem modelos, sejam conceituais ou matemáticos, os dados, diz o filósofo, nada são nada senão ruído, a ciência só avança quando consegue fornecer explicações.

Nós concordamos. Não precisamos ser apocalíticos sobre os "perigos" dos *Big Data*, podemos aliás saudar sua chegada como o maior laboratório que as ciências

sociais possam sonhar, mas também não deveríamos ser eufóricos ao ponto de imaginar nos dados a solução para os dilemas fundacionais das ciências sociais, ou para a falta de boas teorias e bons modelos. Não há dúvida: a disponibilidade de enormes *corpora* textuais representa uma oportunidade inédita para os cientistas, um recurso imprescindível. E o auxílio computacional traz extraordinárias, valiosas novidades. Mas nem os dados, nem os algoritmos são, em si, uma nova ciência ou uma resposta às perguntas científicas. Um barômetro, e um ajudante fiel que anote números numa caderneta, não fazem a física e não descobrem como funciona o clima, embora possam identificar padrões e fazer algumas previsões. Do mesmo jeito, regularidades estatísticas e petabytes são mapas, bons mapas, de fenômenos dinâmicos. Mas ter belos mapas coloridos não significa conhecer um país. Fenômenos precisam não apenas ser descritos e retratados, mas também explicados e, como dizia Max Weber muito tempo atrás, compreendidos e interpretados em seu sentido.

Agradecimentos

O autor agradece expressa gratidão pelo apoio do CNPq (Bolsa PQ, Produtividade em Pesquisa), agradece Alisson Soares, Brunah Schall, Gabriela Reznik, do Observatório InCiTe (Inovação, Cidadania, Tecnociência) pelas sugestões e fecundas discussões sobre o texto.

Referências

Anderson, C. (2008). 'The end of theory: The data deluge makes the scientific method obsolete'. *Wired* 16 (7). URL: https://www.wired.com/2008/06/pb-theory/.

Bail, C. A. (2016). 'Cultural carrying capacity: Organ donation advocacy, discursive framing, and social media engagement'. *Social Science & Medicine* 165, pp. 280–288. DOI: 10.1016/j.socscimed.2016.01.049. PMID: 26879407.

Bauer, M. W. e Howard, S. (2012). 'Public Understanding of Science — a peer-review journal for turbulent times'. *Public Understanding of Science* 21 (3), pp. 258–267. DOI: 10.1177/0963662512443407.

Bridle, J. (1 de novembro de 2016). 'What's wrong with big data?' *New Humanist*. URL: https://newhumanist.org.uk/articles/5104/whats-wrong-with-big-data.

Bucchi, M. e Neresini, F. (2011). 'Monitoring Science in the Public Sphere: The Case of Italy'. Em: The Culture of Science. Ed. por M. W. Bauer, R. Shukla e N. Allum. New York, U.S.A.: Routledge.

Caren, N. (2 de abril de 2015). 'The Path to Big Data Sociology Isn't Obvious'. *Mobilizing Ideas*. URL: https://mobilizingideas.wordpress.com/2015/04/02/the-path-to-big-data-sociology-isnt-obvious/.

Castelfranchi, Y., Massarani, L. e Ramalho, M. (2014). 'War, anxiety, optimism and triumph: a study on science in the main Brazilian TV news'. *JCOM* 13 (3), A01. URL: https://jcom.sissa.it/archive/13/03/JCOM_1303_2014_A01.

Castelfranchi, Y. e Stock, O. (2000). Macchine come noi. La scommessa dell'intelligenza artificiale. Bari, Italy: Laterza.

Cheng, A. S., Fleischmann, K. R., Wang, P. e Oard, D. W. (2008). 'Advancing social science research by applying computational linguistics'. *Proceedings of the American Society for Information Science and Technology* 45 (1), pp. 1–12.

Crawley, C. E. (2007). 'Localized Debates of Agricultural Biotechnology in Community Newspapers: A Quantitative Content Analysis of Media Frames and Sources'. *Science Communication* 28 (3), pp. 314–346. DOI: 10.1177/1075547006298253.

Farrell, J. (2016). 'Corporate funding and ideological polarization about climate change'. *Proceedings of the National Academy of Sciences* 113 (1), pp. 92–97. DOI: 10.1073/pnas.1509433112. PMID: 26598653.

Grimmer, J. e Stewart, B. M. (2013). 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts'. *Political Analysis* 21 (3), pp. 267–297.

Hopke, J. E. e Simis, M. (2015). 'Discourse over a contested technology on Twitter: A case study of hydraulic fracturing'. *Public Understanding of Science* 26 (1), pp. 105–120. DOI: 10.1177/0963662515607725.

Jacobi, C., Atteveldt, W. van e Welbers, K. (2016). 'Quantitative analysis of large amounts of journalistic texts using topic modelling'. *Digital Journalism* 4 (1), pp. 89–106. DOI: 10.1080/21670811.2015.1093271.

Kuckartz, U. (2014). Qualitative Text Analysis: A Guide to Methods. Los Angeles, London, New Delhi, Singapore e Washington: SAGE Publications Inc.

Miltgen, C. L. e Peyrat-Guillard, D. (2014). 'Cultural and generational influences on privacy concerns: a qualitative study in seven European countries'. *European Journal of Information Systems* 23 (2), pp. 103–125. DOI: 10.1057/ejis.2013.17.

Neresini, F. e Lorenzet, A. (2016). 'Can media monitoring be a proxy for public opinion about technoscientific controversies? The case of the Italian public debate on nuclear power'. *Public Understanding of Science* 25 (2), pp. 171–185. DOI: 10.1177/0963662514551506.

Neuendorf, K. A. (2016). The Content Analysis Guidebook. Los Angeles, London, New Delhi, Singapore e Washington: SAGE Publications Inc.

Pigliucci, M. (2009). 'The end of theory in science?' *EMBO Reports* 10 (6), p. 534. DOI: 10.1038/embor.2009.111. PMID: 19488038.

Pollach, I. (2012). 'Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis'. *Organizational Research Methods* 15 (2), pp. 263–287. DOI: 10.1177/1094428111417451.

Popping, R. (2000). Computer-Assisted Text Analysis. Los Angeles, London, New Delhi, Singapore e Washington: SAGE Publications Inc.

Riffe, D., Lacy, S. e Fico, F. (2014). Analyzing Media Messages: Using Quantitative Content Analysis in Research. New York, U.S.A. e London, U.K.: Routledge.

Schreibman, S., Siemens, R. e Unsworth, J. (2016). A New Companion to Digital Humanities. Chichester, West Sussex, U.K.: John Wiley & Sons.

Semino, E., Hardie, A., Koller, V. e Rayson, P. (2005). 'A computer-assisted approach to the analysis of metaphor variation across genres'. Em: Corpus-based Approaches to Figurative Language. Ed. por J. Barnden, M. Lee, J. Littlemore, R. Moon, G. Philip e A. Wallington. Birmingham, U.K.: University of Birmingham School of Computer Science, pp. 145–153.

Smallman, M. (2016). '*Public Understanding of Science* in turbulent times III: Deficit to dialogue, champions to critics'. *Public Understanding of Science* 25 (2), pp. 186–197. DOI: 10.1177/0963662514549141.

Stoneman, P., Sturgis, P. e Allum, N. (2013). 'Exploring public discourses about emerging technologies through statistical clustering of open-ended survey questions'. *Public Understanding of Science* 22 (7), pp. 850–868. DOI: 10.1177/0963662512441569. PMID: 23825238.

Suerdem, A., Bauer, M. W., Howard, S. e Ruby, L. (2013). 'PUS in turbulent times II — A shifting vocabulary that brokers inter-disciplinary knowledge'. *Public Understanding of Science* 22, pp. 2–15. DOI: 10.1177/0963662512471911.

Tian, Y. e Stewart, C. M. (2005). 'Framing the SARS Crisis: A Computer-Assisted Text Analysis of CNN and BBC Online News Reports of SARS'. *Asian Journal of Communication* 15 (3), pp. 289–301. DOI: 10.1080/01292980500261605.

Tvinnereim, E. e Flø ttum, K. (2015). 'Explaining topic prevalence in answers to open-ended survey questions about climate change'. *Nature Climate Change* 5 (8), pp. 744–747. DOI: 10.1038/nclimate2663.

Veltri, G. A. e Atanasova, D. (2015). 'Climate change on Twitter: Content, media ecology and information sharing behaviour'. *Public Understanding of Science*. DOI: 10.1177/0963662515613702.

Veltri, G. A. e Suerdem, A. K. (2013). 'Worldviews and discursive construction of GMO-related risk perceptions in Turkey'. *Public Understanding of Science (Bristol, England)* 22 (2), pp. 137–154. DOI: 10.1177/0963662511423334. PMID: 23833021.

Wiedemann, G. (2013). 'Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences'. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 14 (2).

— (16 de novembro de 2015). 'Text Mining #1: Extending the method toolbox: text mining for social science and humanities research'. *Europeana Research*. URL: http://research.europeana.eu/blogpost/extending-the-method-toolbox-text-mining-for-social-science-and-humanities-research.

Autor

Yurij Castelfranchi é formado em física quântica pela Universidade de Roma "La Sapienza", foi jornalista científico e escritor de ciência por cerca de 15 anos. Desde 2002, vive no Brasil, onde fez doutorado em Sociologia na Universidade Estadual de Campinas (UNICAMP). Como escritor, colaborou com diversos jornais diários, revistas, programas de radio e TV, na Itália e no Brasil, e é autor de 6 livros. Foi pesquisador na SISSA (International School for Advanced Studies, Trieste), no Labjor (Laboratório de Estudos Avançados em Jornalismo, Unicamp), e colaborou com a Organização dos Estados Iberoamericanos para a Ciência e a Cultura (OEI). Hoje, é professor da Universidade Federal de Minas Gerais (Brasil), coordenador do Observatório InCiTe (Inovação, Cidadania, Tecnociência), e membro do Instituto Nacional de C&T para a Comunicação Pública da Ciência e da Tecnologia.
E-mail: ycastelfranchi@gmail.com.

How to cite

Castelfranchi, Y. (2017). 'A análise de textos auxiliada pelo computador: um laboratório a céu aberto para as ciências sociais'. *JCOM* 16 (02), C04_pt.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it