Semiótica e big data: o valor da 'textualização' na lógica capitalista da cultura dataficada

Letícia Moraes
Universidade Federal da Paraíba, Brasil

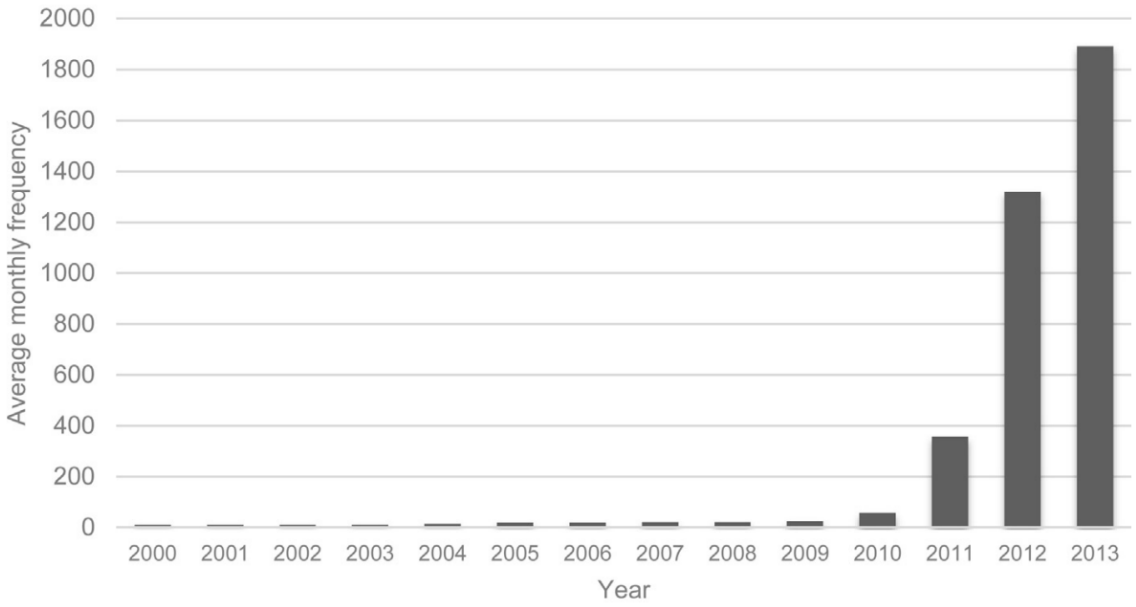
Introdução

O termo *big data* tem cada vez mais se tornado popular, frequentemente encontramo-lo sendo usado em textos jornalísticos e até mesmo em rodas de conversas entre amigos; engana-se, no entanto, quem pensa tratar-se de um neologismo, pois seu emprego para referir-se a uma grande e complexa coleção de dados remonta ao século XX. É difícil, no entanto, precisar quem foi o primeiro a cunhá-lo, o economista Francis Diebold (2012, p.3), ao estudar sua origem, atribui o uso, em meados dos anos 1990, à *Silicon Graphics Inc.* e a John Mashey e aponta que a expressão se popularizou somente após 2011, tendo como mola propulsora a ascendência das ações da IBM (*International Business Machines Corporation*)[1] e de outras companhias de tecnologia, cujo principal (ou um dos principais) foco é a análise de dados. Vale ressaltar que a estreita relação entre o *big data* e o capital não é gratuita, como será visto ao longo da discussão.

No gráfico a seguir (ver Figura 1), é possível cotejar a sua popularização em documentos coletados na *ProQuest Research Library,* entre os anos de 2000 e 2013; comparado aos anos de 2009 e 2010, percebe-se um grande aumento na quantidade de vezes em que a expressão foi utilizada em 2011. Observe que o salto se repete nos anos seguintes, isso é, em 2012 e em 2013.

[1] A IBM é uma empresa de tecnologia dos Estados Unidos ativa desde o século XIX, atualmente está presente em mais de 175 países e é considerada a maior empresa de TI (tecnologia de informação) no mundo.

Figura 1. O uso do termo "big data" na *ProQuest Research Library*.



Fonte: Gadomi e Haider (2015, p. 139).

Embora o gráfico acima mensure somente os primeiros anos do uso da expressão, foi na década seguinte que ela ganhou ainda mais popularidade, ao fazer parte das discussões acadêmicas, das notícias televisas, de temas de documentários, dentre outras práticas que permeiam o cotidiano das pessoas. A frequência de seu uso acompanhou a gradativa popularidade da inteligência artificial (doravante IA)[2] no século XXI; sabe-se, por exemplo, que muitas empresas das áreas de informática e tecnologia que apostaram na extração e na produção de dados ganharam repercussão e grandes aportes de investimento.

Diante do cenário exposto, nossa investigação tem como principal objetivo compreender como o *big data* pode ser semioticamente explicado e de que maneira o fenômeno da textualização tornou-se central para a sua existência. Do ponto de vista do quadro teórico que embasa a discussão, é a semiótica discursiva, isso é, da *Escola de Paris*, erigida por Algirdas Julien Greimas (1966; 2008[1979]), a partir dos anos 1960, que traz a base necessária para a compreensão das grandes coleções de conjuntos de dados como resultantes do fenômeno da textualização.

[2] A IA pode ser entendida como o uso de tecnologias para a resolução de problemas e tomadas de decisões com pretensões de simular o raciocínio humano.

O trabalho é dividido em três principais partes: na primeira, discorremos sobre o *big data*, contextualizando-o dentro do quadro teórico da semiótica discursiva; em seguida, trazemos o conceito de "textualização" da semiótica com o objetivo de cotejar os meandros responsáveis pela geração das grandes coleções de dados na cultura dataficada; e, por fim, ainda que brevemente, realizamos uma discussão sobre como o acúmulo dessas coleções transformou-se em uma espécie de capital desejável para a sustentação do poder e das formas de manipulação simbólica.

1 Em torno do *big data*

Apesar da popularidade da expressão nos tempos atuais, o termo *data* existe desde os gregos, com significado que remete a algo "dado, um presente". De acordo com a sua etimologia, o dado (*data*) provém do adjetivo latino *datus-a-um* e significa "aquilo que foi entregue, presenteado" (Santos Saraiva, 2006). Em sua forma verbal, o *dare*, cujo sentido é "dar ou presentear", mantém uma relação de antonímia com o verbo *capere*, cuja acepção é "tomar ou receber"; se considerarmos, de maneira literal, a definição etimológica, um dado é algo que foi oferecido por alguém como um presente; ora, sabe-se que se uma pessoa doa algo, outra está apta para recebê-lo e, do ponto de vista de quem recebe o presente, a ação tem um sentido maior de passividade.

Recuperamos o passado etimológico porque a expressão *big data* beneficia-se dele; durante muito tempo, havia uma crença de que os dados já estavam prontos no mundo digital para serem recebidos de maneira passiva por alguém ou, melhor, por alguma empresa. Nos primeiros anos das redes sociais, acreditava-se que as empresas (em especial, as mídias digitais, como o "Facebook") apenas recebiam dados pessoais de seus usuários e esses eram fornecidos, como um presente, pelo serviço "gratuito" prestado.

Alguns anos depois, o termo "minerar" popularizou-se, junto com a expressão *data mining,* tornando mais evidente que, embora os dados possam ser produzidos pelos usuários, eles não são, em muitas situações, disponibilizados pelas pessoas de maneira consciente. Ao contrário, os dados mais valiosos - do ponto de vista do capital- precisam ser minerados – ação que envolve dois grandes processos que antecedem a análise do material: (i) a

extração de informações em uma grande fonte de dados e (ii) a limpeza do arquivo, retirando tudo aquilo que é considerado impureza (ou seja, o que não se considera útil para os fins desejados)[3]. Somente após concluída a etapa de triagem é que os analistas estão aptos a colocarem todos os elementos em relação para que os padrões almejados possam ser construídos ou encontrados.

Diante da repercussão dessa informação, não demorou para que a prática do uso e da venda dos dados de usuários de mídias digitais se tornasse assunto comum em manchetes jornalísticas (também em filmes, documentários e outras produções midiáticas de grande alcance); inúmeros casos de empresas que mineravam e vendiam as bases de dados de seus usuários, muitas vezes sem que os mesmos tivessem consciência, foram expostos. Nesse ponto da discussão, vale retomarmos à discussão etimológica do termo ("data"): ironicamente o *big data*, do ponto de vista das práticas que envolvem sua produção, extração e armazenamento, está mais próximo, metaforicamente, da acepção de "tomar" do verbo *capere* do que do sentido empregado no verbo *dare*.

Apesar dos inúmeros escândalos e processos judiciais movidos nos últimos anos contra o uso indevido dessas informações sensíveis, os dados são, atualmente, considerados centrais na dinâmica da nossa sociedade, gerando capital econômico, cultural e social (Diebold, 2012). Entre as maiores empresas globais atuantes nos dias de hoje, aquelas que nasceram e expandiram-se pelo ambiente digital figuram na lista das principais no mercado financeiro; o *Facebook* é um exemplo: além de rede social, é considerado, também, uma mídia (*new media*) (Manovich, 2001), um sistema projetado para possibilitar a interação das pessoas e o compartilhamento e a criação de textos, sejam eles manifestados como frases, imagens, vídeos, áudios, *emoticons*, etc. Apenas essa mídia contava, em 2017, com dois bilhões de perfis ativos; são milhões de

[3] Cf MANDELLI, Mariana, O dilema das redes se resolve com educação midiática. **Folha de São Paulo**, São Paulo, 24 de setembro de 2020. Opinião. Disponível em: https://www1.folha.uol.com.br/educacao/2020/09/o-dilema-das-redes-se-resolve-com-educacao-midiatica.shtml. Acesso 10 mai. 2024.

milhões de usuários interagindo, criando, compartilhando e consumindo conteúdo em uma escala cada vez maior[4].

Há uma outra característica comum nesse tipo de corporação atuante como mídia social, trata-se da manipulação dos mecanismos de distribuição das informações para a base de usuários[5]. O gerenciamento dos dados pessoais e a existência de regras (conhecidas como algoritmos) que distribuem o que cada indivíduo consumirá, embora muito comuns nesse modelo de negócio, são características pouco conhecidas pelo público em geral que utiliza as plataformas e até mesmo por pesquisadores da área, uma vez que esses códigos costumam ser secretos; em razão disso, a expressão "black box" tem sido usada com frequência para referir-se às questões éticas relacionadas ao uso da IA em diferentes aplicações. A título de exemplificação, citamos a pesquisa "Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical A.I.", conduzida pelos pesquisadores Juan Manuel Durán (TU Delft) e Karin Jongsma (UMC Utrecht), publicada em 2021, que debate os enviesamentos e as consequências da "black box" nos algoritmos usados na área da medicina.

No que tange ao uso do big data, sabemos que, de um lado, há a extração de informações sobre práticas sociais de indivíduos e grupos inteiros e, de outro, o uso desses dados para a geração infinita de novas informações; esse processo de retroalimentação é conhecido, nas áreas das humanidades digitais e da computação social, como dataficação, tradução do inglês *datafication*, termo foi cunhado, em 2013, no âmbito acadêmico, por duas figuras militantes na área da governança e regulação da internet, o professor Viktor Mayer-Schoenberger (University of Oxford) e o jornalista americano Kenneth Cukier.

Nessa perspectiva dos autores, a dataficação corresponde ao fenômeno que transforma ações em dados quantificáveis e, posteriormente atua na alteração dos comportamentos, das ações e dos conhecimentos das pessoas, com base em algoritmos presentes em um sistema de inteligência artificial; podendo ser, grosso modo, resumida como algumas sequências finitas de

[4] O número foi estimado pelo próprio *Facebook* e desconsidera os perfis *fakes*. A notícia pode ser conferida em https://www.terra.com.br/noticias/facebook-atinge-marca-de-2-bilhoes-de-usuarios,b647b4af6f82ca8b483f01a143c1ea9c2nf55uu3.html.

[5] Cf. Documentários "Privacidade hackeada" (2019), dirigido por Karim Amer e Jehane Noujaim, e "O dilema das redes sociais" (2020), dirigido por Jeff Orlowski, ambos produzidos e distribuídos pela *Netflix*.

ações executáveis, chamadas de algoritmos, que avaliam a performance dos dados previamente coletados e tomam decisões a partir de regras preestabelecidas (Mayer-Schoenberger; Cukier, 2013, p. 28). Nota-se, no entanto, que essa explicação técnica, popular nas pesquisas das áreas da computação, esconde uma característica muito importante e cara a nós, pesquisadores das humanidades: ao postular uma certa neutralidade, com o uso da explicação estritamente técnica, tudo se passa como se as tecnologias fossem isentas de quaisquer tipos de vieses e crenças humanas.

Sabemos, no entanto, que o uso da inteligência artificial é assentado a partir de decisões humanas, resguardando vieses políticos, raciais, sociais, históricos, ideológicos e culturais. Longe de ser um procedimento autônomo e não humano, o *big data* e a inteligência artificial são perpassados, em seus mais diferentes processos, por decisões tomadas por pessoas (e corporações dirigidas por pessoas) que carregam com si suas experiências, crenças, histórias de vida, ideologias, valores culturais e familiares.

Se a dataficação constitui um novo paradigma na ciência e na sociedade, ela deve ser também compreendida como um processo transpassado pelas axiologias culturais, como lembra-nos José van Dijck (2017), professora na área de mídia e sociedade digital na *Universidade de Utrecht*. Van Dijck é autora de diversos trabalhos que versam sobre os impactos sociais da tecnologia; para ela, vivemos em uma sociedade da plataforma que não se separa das instituições tradicionais e amalgama as plataformas digitais às estruturas sociais. Dessa maneira, a plataformização não é uma "coisa", mas um processo em que todos nós estamos inseridos, sendo sustentada pela relação de confiança erigida entre o indivíduo que "permite" o uso de seus dados em troca das facilidades apresentadas pelas plataformas (que contemplam desde o uso de um servidor *on-line* ou de aplicativos no celular até resultados clínicos baseados em análises algorítmicas de dados).

É interessante notar que, em termos semióticos, estabelece-se um contrato fiduciário entre o usuário e o processo de plataformização. Os semioticistas Algirdas Julien Greimas e Joseph Courtés (2008[1979]) postulam que esse contrato

> apresenta-se então como um contrato enunciativo [...]. O contrato fiduciário que assim se instaura, pode repousar em uma evidência (isto é, numa certeza imediata) ou então ser precedido de um fazer persuasivo (de um fazer-crer) do enunciador, ao qual corresponde um fazer interpretativo (um crer) da parte do enunciatário (Greimas; Courtés, 2008[1979], p. 86).

Para uma melhor compreensão entre a relação do processo de plataformização e os dados, pensemos no seguinte exemplo: em uma rede farmacêutica, o cliente pode ser levado a crer que ganhará uma vantagem imperdível, como um desconto no preço de um determinado medicamento se ele informar seus dados pessoais no ato da compra. Observe que o enunciador é bastante eficaz em sua jornada do fazer persuasivo ao agir por meio de um "fazer-parecer-verdadeiro" que cria um "efeito de verdade". Esse efeito de verdade não é necessariamente a realidade, pois aos preços dos produtos é adicionado um valor extra e somente os clientes que fornecem os dados pagam o valor original. Ou seja, as empresas criam artificialmente um desconto para convencer os clientes a concederem seus dados privados[6]. No entanto, por conta do efeito de verdade e do contrato fiduciário estabelecido, ainda que se trate de um desconto fictício, o enunciatário é levado a crer que está em conjunção com uma vantagem financeira.

Essa estratégia deu tão certo que se tornou corriqueira em muitos setores que comercializam produtos básicos e necessários para a sobrevivência, como nos casos dos supermercados e das farmácias. Algumas supostas promoções são apenas desbloqueadas quando o consumidor informa o número de um documento pessoal que imediatamente permite o sistema ter acesso ao banco de dados do cliente e computar aquele item comprado em seu histórico. Cabe-nos, aqui, fazer um pequeno parêntese: essa estratégia não age apenas no "crer", mas também na modalidade virtualizante do "dever-fazer" (Greimas; Courtés, 2008[1979], p. 135), que possibilita, junto ao "querer-fazer", as condições mínimas para o fazer, uma vez que podemos também compreendê-la no âmbito da ordem da prescrição; o enunciatário é, de certa maneira,

[6] A *Drogaria Araújo S/A* foi condenada, pelo Procon-MG, a uma pena de multa de mais de R$7 milhões de reais por condicionar tais descontos mediante o fornecimento de dados sensíveis do consumidor, sem que haja uma transparência sobre o destino desses dados. (G1. Drogaria Araújo é multada em mais de R$7 milhões por condicionar descontos a fornecimento de CPF. **G1**, São Paulo, 05 de dezembro de 2018, Minas Gerais. Disponível em: https://g1.globo.com/mg/minas-gerais/noticia/2018/12/05/drogaria-araujo-e-multada-em-mais-de-r7-milhoes-por-condicionar-descontos-a-fornecimento-de-cpf.ghtml. Acesso 10 mai. 2024).

obrigado a informar os seus dados para não pagar a mais por um produto. A eficácia da estratégia é atestada uma vez que grande parte dos clientes não percebem que foram obrigados a fazer isso; eles são convencidos de que ofereceram o seu número de identificação pessoal por livre escolha e, por isso, foram agraciados com uma vantagem imperdível.

A semioticista brasileira Diana Luz Pessoa de Barros (2001[1990], p. 50) lembra que as modalidades (o querer, o dever, o poder e o saber) são resultantes da conversão da categoria tímico-fórica fundamental e agem nas relações dos sujeitos com os valores. Dentro do contexto de uma sociedade regida pelo sistema econômico capitalista, há uma pressão para que os sujeitos ganhem vantagens financeiras ou, por outra perspectiva, economizem o máximo de dinheiro possível na compra de um produto e, assim, estejam aptos a entrar em conjunção com outros objetos também tidos como necessários ou desejáveis. Por conseguinte, ainda que os usuários tenham (ou venham a ter) consciência das práticas de venda de dados realizada por tais empresas, há de se considerar que, diante da escassez financeira enfrentada pela maioria das famílias brasileiras, aliada aos altos custos dos medicamentos comercializados e as sucessivas crises econômicas vivenciadas nas últimas décadas em todo o mundo, para muitas pessoas, especialmente para aquelas pertencentes aos grupos minorizados, não é uma alternativa pagar a mais em um item na farmácia "apenas" para a proteção de seus dados pessoais. Trata-se, portanto, de uma não-escolha, uma vez que o contexto em que essa pessoa está inserida a obriga a fazer isso, sendo modalizadas por um tipo de "dever".

Todo *marketing* em volta dessa estratégia usada por grandes empresas privadas revela que o interesse das corporações por dados pessoais de sua base de clientes é cada vez maior; para elas, há inúmeras vantagens, inclusive econômicas, pois usam esses bancos de dados (o *big data*) para publicidade e/ou vendem essas informações sigilosas para outras corporações interessadas em construir padrões de determinados grupos sociais; a partir, por exemplo, de um histórico de compras de remédios em uma drogaria, pode-se ter uma ideia dos problemas de saúde do consumidor, seus hábitos de compras, poder aquisitivo, dentre outras informações.

Os próprios dados que alimentam esse processo na cultura dataficada são constituintes de extensas coleções de artefatos provenientes das mais diversas

fontes, de mídias sociais (textos verbais e não-verbais, curtidas, data de nascimento, rede de amigos, nome dos familiares, indicações de estados de humor, etc.), de empresas financeiras de cartão de crédito (produtos comprados, padrão de gastos e de consumo, marcas e lojas favoritas, etc.), de cadastros geridos pelos governos (identidade, tipo sanguíneo, estado civil, histórico de doenças e transtornos psicológicos, renda mensal, profissão, bens próprios, etc.), para citar alguns. São, portanto, gerados a partir das práticas sociais humanas.

O semioticistas francês Jacques Fontanille, em seu livro *Pratiques sémiotiques* (2008, p. 34), esquematiza os níveis de pertinência semiótica para análise e concebe as cenas práticas como comportamentos e ações humanas, que podem ser compreendidas como "semióticas" por portarem um plano da expressão e um plano do conteúdo. Mas diferente de um objeto semiótico gerado em um suporte estável, como uma pintura, elas produzem sentido ao vivo, isso é, ao mesmo tempo em que estão em curso, visto que seu próprio movimento é gerador de sentido e sua significação é dinâmica.

A partir dessa formulação fontanilliana, somos levados a crer que o *big data* pode ser compreendido semioticamente como um grande conjunto de coleções de objetos (multi)semióticos gerados especialmente pelo (e no) desencadeamento de práticas humanas. Alguns desses objetos semióticos têm seu sentido produzido dentro da própria prática, ao vivo, no momento em que ela se desenrola; esse é o caso, por exemplo, de uma transação monetária realizada por pix no ato da compra de um produto no supermercado. Embora a ação em si do pagamento possa ser compreendida como uma prática não estabilizada em um suporte estável, dela decorrem objetos semióticos que são estabilizados e passam a fazer parte de uma grande coleção com outros artefatos similares (em termos de conteúdo e/ou expressão). Diante do exposto, temos como intenção jogar luz em uma das principais características do *big data*, a de permitir e gerar objetos semióticos em suportes estáveis a partir de ações e comportamentos humanos em curso.

Nessa mesma direção, as sociólogas Danah Boyd e Kate Crawford (2012) argumentam que o *big data* é composto por centenas de milhares de objetos culturais (para nós, objetos semióticos), que caracterizam traços de uma cultura e podem ser usados para a análise de comportamentos individuais e de um

grupo inteiro. Mas isso, a nosso ver, só é possível, ou melhor dizendo, mais facilmente realizável, porque objetos semióticos estabilizados foram gerados a partir e no desenrolar dessas cenas práticas. Salienta-se, ainda, que a extração, a produção e o uso das grandes coleções de dados não funcionam apenas como uma observação do mundo e das pessoas, visto que elas, de acordo com as autoras, também constroem o mundo que conhecemos hoje e terão cada vez mais influência nas escolhas e nas ações dos sujeitos no decorrer deste século.

Na perspectiva dos estudos dos níveis de pertinência semiótica (Fontanille, 2008), esses objetos semióticos são usados, dentro do sistema capitalista dataficado, para não apenas compreender o comportamento e as ações humanas, mas também para remodelá-los, agindo sobre o nível das práticas e – também – das formas de vida. Essas, por suas vezes, são depreendidas como constituintes imediatos da semiosfera, estruturadas a partir da repetição e da regularidade das estratégias adotadas nas articulações das práticas. Assim, ao agir sobre as práticas continuamente, o processo desencadeado pelo *big data* também altera nossas formas de vida, a maneira como nos organizamos dentro da cultura, transfigurando a percepção sobre a realidade e alterando as interações estabelecidas entre os sujeitos e destes com o mundo em volta.

Em perspectiva mais crítica, a idealizadora do *Center for Critical Internet Inquiry*, na UCLA (*University of California*, Los Angeles), Safyia Umoa Noble (2022[2018]), em sua obra *Algoritmos da opressão*, aponta que o uso dos sistemas de inteligência artificial, junto aos algoritmos e ao *big data*, é acompanhado de diversos debates sobre enviesamento de dados, racismo algorítmico e outras formas de preconceitos, cujas consequências fortalecem ainda mais a exclusão de determinados grupos marginalizados. Em um experimento realizado, ao digitar "meninas negras" (no original, "black girls") no buscador da *Google*, a autora se deparou com diversos conteúdos pornográficos como resultados da pesquisa; o episódio foi responsável pelo início da jornada de Noble na compreensão de como os vieses por trás dos algoritmos e dos bancos de dados reforçam a exclusão e a marginalização de grupos minorizados.

À vista disso, somos levados a crer que o *big data* – sua produção, uso e acesso - deve fazer parte das discussões dos humanistas/semioticistas, pois aquilo que convencionalmente chamamos de dados são, em outros termos, objetos semióticos gerados a partir de práticas sociais em nosso dia a dia. Sendo assim, com o intuito de dar seguimento à discussão, no próximo tópico, examinaremos a geração do *big data* a partir da perspectiva do conceito de textualização, proveniente dos estudos da semiótica discursiva, da *Escola de Paris*, erigida por Algirdas Julien Greimas e seus(suas) colaboradores(as), na década de 1960, com desenvolvimentos até os dias atuais.

2 O fenômeno da textualização e a geração do *big data*

No quadro dos estudos da semiótica, o *big data* pode ser entendido epistemologicamente como uma extensa e dinâmica coleção de infinitas coleções de objetos (multi)semióticos, que são permeados por axiologias e gerados a partir de diferentes práticas sociais em curso no interior de uma cultura. É importante notar que essa definição não concerne a um conjunto de artefatos em específico, mas ao conceito geral de *big data*. Assim, a dataficação (Mayer-Schonberger; Cukier, 2013), como explicitado anteriormente, a geração de objetos semióticos a partir de ações humanas e a sua intrínseca relação com as práticas e as formas de vidas, permite a transformação dos atos em traços digitais[7]. Diferentes ações e comportamentos (atos) são textualizados e geram objetos semióticos digitais impressos em um suporte (mais ou menos) estável.

O próprio suporte, como destacam os semioticistas Maria Giulia Dondero e Everaldo Reyes-García (2016), tem função na significação; ele age como um mediador da prática e do objeto semiótico estável (chamado por eles de "texto"). Nesse caso, é o suporte que torna a dataficação legível para ser considerada um tipo de textualização, essa definida por Greimas e Courtés (2008[1979], p. 504) como um "conjunto dos procedimentos – chamados a se organizarem numa sintaxe textual – que visam à constituição de um contínuo

[7] *Datafication* é definido, pelos cientistas da computação, como a transformação de "diversos aspectos" da nossa vida em dados digitais. Estes são transformados em informação e entendidos como uma nova forma de valor (Biltgen; Ryan, 2016, p. 151).

discursivo, anteriormente à manifestação do discurso nesta ou naquela semiótica [...]".

As expressões "sintaxe textual" e "contínuo discursivo", retiradas da citação acima, revelam o papel da textualização como uma instância geral e atuante no processo geracional do objeto semiótico, como procuramos demonstrar em Moraes (2021). Enquanto a textualidade está para a natureza do objeto, a textualização dá conta do percurso de geração do texto, uma espécie de *continuum*, e tem o seu fim com a manifestação do objeto semiótico. Esse, por sua vez, pode ser constituído, na instância da manifestação, por qualquer linguagem (ou código), seja ela visual, matemática, algorítmica, *3d*, gestual, etc. e ainda pela combinação de duas ou várias delas, como são os ditos objetos multissemióticos, a exemplo da HQ, do ballet, do filme, da peça teatral, etc.

Essa compreensão da textualização é importante no âmbito da disciplina semiótica, pois ela permite a conceptualização da noção de texto enquanto uma grandeza geral, um todo organizado de sentido. Na perspectiva de texto como uma grandeza, a textualização designa um percurso, uma espécie de continuum, que leva à manifestação do texto-objeto (do objeto semiótico manifestado); ela é uma instância produtora de sentido e permite a junção do plano do conteúdo com o plano da expressão.

Para que essas grandezas semióticas sejam um todo organizado de sentido é preciso que haja uma dependência em sua estrutura, isso acontece quando uma parte depende do sentido das outras partes, em uma situação de correlação (Hjelmslev, 2010[1975], p. 3). Ao estudar a obra do linguista Louis Trolle Hjelmslev, Sémir Badir (2014, p. 144) destaca a sua contribuição para a concepção de texto por ser o estudioso dinamarquês o primeiro (ou um dos primeiros) a perceber que a frase não tem existência autônoma, o seu sentido é constituído enquanto tal porque ela é compreendida como um objeto textual. A título de exemplificação, várias frases aleatórias que não se relacionam umas com as outras (ou com o todo) formam apenas um amontoado de frases, mas se há uma relação de dependência de sentido entre elas, então há um texto. O mesmo ocorre também com outras linguagens; uma obra de arte pintada a óleo não é apenas um conjunto aleatório de quaisquer pinceladas, há uma relação estabelecida entre a direção, a pressão, o formato

das pinceladas com as cores, as texturas etc., compondo um todo coerente, dotado de intencionalidade e de uma direção significante.

Para melhor compreender esse processo aplicado à dataficação, retomamos o trabalho de Jathan Sadowski (2019), pesquisador do *Emerging Technologies Research Lab*, na Monash University, Austrália. Sua pesquisa incide sobre a tecnopolítica e defende que os dados se tornaram uma espécie de capital; de acordo com ele, os dados são transformados em *bits* de informações discretos, cuja gravação digital é, *a posteriori*, processada por uma máquina, mas seu valor fundamenta-se na compreensão de que são um tipo de "capital humano". Tudo se passa, no âmbito da semiótica, como se as ações e os comportamentos humanos engendrassem uma espécie de textualização de certas práticas e ações sociais, cujos objetos semióticos resultantes sejam os portadores de sentidos próprios da vida humana[8].

Essas textualizações podem ser produzidas a partir de diferentes práticas e manifestadas em suportes digitais diversos. Por exemplo: elas podem ser geradas no interior de uma prática semiótica específica, como é a condição dos dados fabricados em uma compra no supermercado que foi paga com o cartão de crédito. Observe que esses constituintes, por exemplo, a fatura emitida no ato da compra, revelam informações sobre o valor gasto na compra, o padrão de consumo alimentar, as marcas preferidas pelo usuário do cartão, etc. e tornam-se disponíveis graças à textualização digital engendrada. Uma vez gerados esses objetos semióticos, eles passam a fazer parte de uma grande coleção de artefatos semióticos que serão usados em análises, cujos resultados produzem novos objetos semióticos que podem ser usados para modificar e gerar novas práticas sociais e, por conseguinte, formas de vida no mundo natural.

Em razão disso, sustentamos a tese de que uma das características mais importantes do *big data* seja a sua constituição enquanto um tipo de textualização das ações e dos comportamentos humanos no mundo, cuja manifestação resulta objetos semióticos variados: fotografias (exemplo: *selfies* postadas nas redes sociais, fotografias capturadas por satélites), vídeos

[8] Nem toda coleção de dados do *big data* tem, a priori, um valor cultural. Neste texto, no entanto, estamos lidando especificamente com os objetos que permitem a produção de sentido sobre os humanos e os seus comportamentos.

(provenientes de câmeras de segurança em locais públicos e privados, postados nas redes sociais e outras plataformas, como o *YouTube*), textos escritos (*e-mails*, textos compartilhados em redes sociais, pesquisas realizadas nos buscadores, documentos oficiais de governos, documentos pessoais de identificação), áudios (telefonemas, gravação de voz pelo celular), etc.

A dataficação não pode ser compreendida, dessa maneira, como uma mera tradução do analógico para o digital, pois no contínuo discursivo da textualização novas significações são geradas a depender, por exemplo, de como esses artefatos são agrupados e em que linguagem são textualizados (numérica, visual, verbal escrita, etc.). No processo da dataficação, o texto-objeto, ao final do percurso, não é semelhante ao primeiro objeto que lhe deu origem, pois ela é, retomando Fontanille (2008), gerada a partir de uma cena prática e concerne ao processo de textualização de uma ação (uma semiose em curso) em um texto-objeto (de múltiplas linguagens e códigos, incluindo algoritmos, linguagem numérica, etc.) fixo a um suporte mais estável que o primeiro.

Assim, para a dataficação, o que interessa, por exemplo, na leitura de um texto digital, é a quantidade de tempo gasto na passagem dos olhos em cada parte da tela, os movimentos realizados, as ações que o leitor executou durante a leitura, etc., isso é, os gestos e as ações que compõem a cena prática da leitura; algumas das ações e dos comportamentos do leitor podem ser textualizados, usando as linguagens topológica e numérica, em um suporte mais estável (ex. uma tabela contendo dados temporais).

O processo da dataficação que se inicia com a produção, a extração e o armazenamento dos dados do dia a dia das pessoas são práticas já consideradas corriqueiras, pois, como apontam Kambatla *et al*. (2014, p. 2562), há um conjunto enorme de artefatos já armazenados e todos os dias essa quantidade aumenta[9]: cada vez que clicamos em um *website*, fazemos uma chamada usando o celular, pagamos uma compra com o cartão de crédito, usamos o *gps*, nos comunicamos por meio de uma rede social, baixamos um aplicativo, somos filmados por uma câmera com detector facial, um (ou vários) registro(s) digital(is) do comportamento humano individual ou coletivo é(são)

[9] Em 2008, estimava-se que havia 9.57×1021 bytes de dados. Naquele momento, já era esperado que esse número dobrasse de tamanho a cada dois anos.

criado(s) e armazenado(s) de maneira cumulativa por alguns seletos grupos e/ou incorporações. Interessa-nos, agora, melhor compreender a relação estabelecida entre o processo da dataficação e o capital dentro do sistema econômico atual.

3 O valor capital do *big data*

O *big data* não funciona apenas como uma observação do mundo e das pessoas; ele constrói, em alguma medida, o mundo em que vivemos (Boyd; Crawford, 2012), uma vez que o seu uso é capaz de modificar a nossa percepção sobre o meio ao redor, transformando a nossa maneira de interagir com os outros sujeitos e com o mundo. Tendo isso em consideração, alguns pesquisadores atuantes nas áreas computação social e da ética aplicada à inteligência artificial alertam para o fato de que os dados dirigem a lógica da nossa sociedade (por exemplo: as eleições, as políticas públicas, as relações entre os consumidores e as empresas e vice-versa), influenciando, em algum nível, o comportamento humano.

Essas grandes coleções de artefatos digitais são constituídas especialmente por objetos culturais (Boyd; Crawford, 2012), que caracterizam traços de uma cultura e podem, sob certas circunstâncias, permitir a análise dos comportamentos individuais e/ou de um grupo. Entre os diversos conteúdos manifestados nos textos-objetos presentes nessas coleções de artefatos, encontram-se fotografias (de *selfies*, pratos de comida, lugares, etc.), vídeos, localizações geográficas, textos verbais escritos, registros de compras realizadas (em supermercados, lojas de eletrodomésticos, lojas virtuais, etc.), buscas feitas na internet, *hashtags,* rotas e trajetos percorridos por carros, transações financeiras, programas de *tv* assistidos, dados pessoais (números de identidade, peso, estado civil, salário, tipo de sangue, gênero, idade, etc.), sites e conteúdos acessados, histórico de doenças, redes de relações familiares (mãe, pai, irmãos, primos, tios, etc.) e de amizade (pessoas que tem um contato mais próximo ou mais distante), etc.

Quase todas as nossas ações deixam, em maior ou menor medida, um rastro digital, visto que a sociedade em que vivemos é cada vez mais dataficada. A internet é considerada o maior sistema de engenharia já criado pela

humanidade, são bilhões de computadores conectados (Kurose; Ross, 2013, p. 10), desde máquinas de lavar roupas *smart* até as câmeras instaladas nas ruas, nos estabelecimentos comerciais e nos prédios residenciais. Esse grande conjunto de coleções de dados diz respeito ao nosso comportamento, aos nossos padrões sociais e à nossa forma de conceber o mundo, produzindo sentido sobre nós mesmos. Cabe salientar que quase todo esse conhecimento é produzido, manuseado e estocado por organizações privadas, tais como bancos e empresas da tecnologia, e governos.

A partir da análise de tais conjuntos de coleções, é possível predizer certos comportamentos sobre uma pessoa ou um grupo, trazendo respostas para perguntas como: qual a probabilidade dessa pessoa pagar o empréstimo que está solicitando ao banco? Que tipo de coisa uma pessoa ou um grupo de pessoas consome? O que faria essas pessoas consumirem mais? Quais os seus hobbies? Qual o padrão estabelecido nos relacionamentos desse grupo? Em qual o candidato provavelmente esse indivíduo ou grupo de pessoas votará? O que faria ele mudar ou decidir o voto? Como visto, esse conhecimento incide sobre as práticas humanas e as formas de vida (Fontanille, 2008) e desvelam muito os gestos e comportamentos nas ações individuais e coletivas, na maneira como nos organizamos enquanto grupos dentro de uma sociedade.

Com as novas formas de organização do capitalismo na era da dataficação, o capital percebeu que os dados estão além do valor de mercadoria; eles são convertidos em capital econômico. Sadowski (2019) explica que, assim como o capital social e o cultural, descrito por Bordieu (1974), o "data capital" pode ser, sob certas circunstâncias, convertido em capital econômico. E a maior parte desse "data capital", que tem valor econômico extraído no mundo, é sobre as pessoas, seus comportamentos, suas crenças, suas práticas individuais ou coletivas:

> O capital de dados é mais do que conhecimento sobre o mundo, são bits discretos de informação registrados digitalmente, processáveis por máquina, facilmente aglomerados e altamente móveis. Assim como o capital social e cultural, o capital de dados é conversível, em determinadas condições, em capital econômico. [...] No capitalismo digital, os dados não substituem o dinheiro, mas são elevados e colocados "no mesmo nível do capital financeiro", como afirma um

> relatório da *Oracle* e da *MIT Technology Review Custom* (2016, p. 2) (Sadowski, 2019, p. 4)[10].

A lógica do sistema capitalista na era do *big data* continua sendo a do acúmulo: coletar e circular dados o tempo todo, produzindo novas mercadorias que gerarão mais dados e edificando infraestruturas para gerenciar todo esse capital produzido em um sistema contínuo e ininterrupto. A importância desse apontamento para o nosso estudo reside justamente na transformação de um objeto digital com valor cultural em valor econômico, em uma relação de capital, em dinâmica parecida com aquela descrita pelo filósofo Karl Marx (1998[1867]), em sua obra *O Capital*, *vol. 1*, e sintetizada na fórmula D – M – D (D = dinheiro; M= mercadoria). Ou seja, quando o dinheiro é usado para comprar uma mercadoria que é então vendida por mais dinheiro; na cultura dataficada, quanto mais dados, mais novos dados poderão ser gerados e, assim como o dinheiro, eles têm a capacidade de gerar valor porque são valor – e não apenas mercadorias como se supunha anteriormente.

Junto com Jean-Claude Passeron, Bordieu (2018[1964]), na obra *Os herdeiros: os estudantes e a cultura*, expande a teoria de Marx (1998[1867]), acrescentando novas formas de capital que não são convertidas diretamente em dinheiro, mas que podem também ser em certas condições; as novas formas de capital são descritas como capital cultural e capital social. O capital cultural, por exemplo, que acrescenta no *status* de uma pessoa e contribui para o seu sucesso, é uma representação de classe e é, geralmente, transmitido no interior de uma família. O capital social, por sua vez, é relacionado ao *network* de um indivíduo, o que possibilita certos privilégios para pessoas que possuem um bom estoque de capital social.

O capital cultural e o capital social configuram-se como formas de dominação e de poder em um espaço de lutas, em que estruturas simbólicas permitem que um grupo esteja acima do outro. Cotejando a teoria de Bordieu (1974), podemos pensar na dinâmica de extração e de produção dos dados no cenário capitalista atual; é preciso enfatizar que o valor gerado pelos dados não

[10] No original: "Data capital is more than knowledge about the world, it is discrete bits of information that are digitally recorded, machine processable, easily agglomerated, and highly mobile. Like social and cultural capital, data capital is convertible, in certain conditions, to economic capital. [...] In digital capitalism, data is not a substitute for money, but is rather elevated and put 'on the same level as financial capital,'as a report by Oracle and MIT Technology Review Custom (2016: 2) states".

é sempre monetário, embora também o seja em certos momentos. Por trás deste movimento, há uma dinâmica de acumulação, que faz parte da lógica do capitalismo, por exemplo: uma empresa pode até mesmo nunca usar parte dos dados que foram produzidos para um fim específico ou então não fazer dinheiro diretamente com o uso das gigantes coleções armazenadas por ela, mas ainda assim continuará em busca de capital de dados, pela razão que

> o capitalista não está preocupado com a utilização imediata dos dados ou com uma coleção específica, mas sim com o fluxo incessante de criação de dados. Este ponto é ilustrado pelo fato de os dados serem muitas vezes recolhidos sem uma utilização específica em mente. De fato, a prática de recolher dados primeiro e de os descobrir depois é cada vez mais uma parte essencial do funcionamento das empresas e dos organismos governamentais. (Sadowski, 2019, p. 5)[11].

O acúmulo dos dados gera poder, dominação e exploração. Por isso, a posse desse capital é um privilégio para poucos; as pessoas e grupos sociais que produzem os dados não têm acesso a eles, muitas vezes nem mesmo para fins acadêmicos. Há ainda de se considerar a perpetuação da dinâmica de dominação exercida pelo Norte no Sul global; as grandes empresas de dados estão, em sua grande maioria, localizadas nas regiões do Norte, com destaque especial para os Estados Unidos, e atuam em quase todo o globo.

No entanto, ainda que não haja um uso imediato de tais dados, não se pode assegurar o mesmo para o futuro, pois muitas empresas que não utilizavam esses objetos semióticos uma década atrás, hoje lucram com a venda de informações para outras empresas do sistema capitalista; isso porque quanto mais dados uma empresa, governo ou organização possui, melhor poderá, por exemplo, traçar perfis dos clientes, cidadãos do país, grupos sociais, etc. A análise de perfis torna possível dizer quem são as pessoas ou grupos mais suscetíveis a compreenderem uma mensagem como verídica e até mesmo a compartilhar para outros membros de sua família, de seu grupo social, de sua rede de contatos.

[11] No original: "The capitalist is not concerned with the immediate use of a data point or with any single collection, but rather the unceasing flow of data-creating. This point is illustrated by the fact that data is very often collected without specific uses in mind. Indeed, the practice of collecting data first and figuring it out later is increasingly a core part of how businesses and government bodies operate".

A dataficação, por meio da textualização das ações e dos comportamentos humanos em objetos semióticos mais estáveis, age em dupla direção: de um lado, um objeto semiótico é produzido a partir de (ou dentro de) uma prática e, de outro, ele (como integrante de uma grande coleção) influencia na geração de novas práticas em devir, alterando, como pontuou Sadowski (2009), a lógica da nossa sociedade. Esse processo, realizado pelo uso das inteligências artificiais, influencia o comportamento humano em diversos âmbitos, desde o gosto estético até as eleições nacionais.

À vista disso, esses objetos semióticos, conhecidos como *big data*, podem ser compreendidos como uma nova ferramenta de acumulação, que geram mais poder para aqueles que têm condições de concentrar maiores quantidades, sendo importante salientar que esse ciclo não tem fim: quanto mais dados acumulados maior é o poder de acumular, e, assim, de agir, por meio das tecnologias de inteligência artificial, sobre as práticas e as formas de vida que estruturam uma cultura dataficada. Eles são, pois, uma nova espécie de capital, isso é, um capital semiótico.

Considerações finais

No processo da geração do *big data*, apercebe-se que alguns esboços das ações cotidianas das pessoas são transformadas em traços digitais; elas passam por uma espécie de textualização para uma linguagem digital e multisemiótica e são armazenadas em grandes grupos de coleções. No que se refere à textualização, a sua importância reside no fato de que a inteligência artificial não tem acesso direto aos humanos e ao mundo natural; por consequência, a relação entre a inteligência artificial e o mundo só pode ocorrer via dataficação, ou seja, por meio de um processo desencadeado pela textualização dos gestos, das ações e dos comportamentos humanos, que desvelam sentidos da vida humana e de suas organizações e interações.

É justamente o processo da textualização o responsável por fazer com que o *big data* seja, além de um objeto cultural, como foi descrito por Boyd e Crawford (2012), também uma forma de capital, característica apontada por Sadowski (2019), mas diferente de outras formas de capital, acreditamos que o *big data* seja um "capital semiótico", estando a textualização no cerne da

questão da dataficação e da geração de novas formas de vidas nas culturas dataficadas.

Por fim, cremos ser oportuno salientar que o uso da inteligência artificial vem acompanhado de um número considerável de discussões sobre o enviesamento de dados, a presença de racismo algorítmico e o recrudescimento de outras formas de preconceitos, cujas consequências fortalecem ainda mais a exclusão dos grupos desfavorecidos e minorizados (Noble, 2022). Uma maior presença de humanistas e semioticistas neste debate faz-se urgente e necessária, uma vez que o *big data* não pode ser depreendido apenas como um compilado de informações sobre as pessoas. Como objetos (multi)semióticos variados, eles contam uma parte da narrativa de nossa vida e as formas de vida de uma cultura dataficada; e são usados para fins, muitas vezes, nefastos pelos detentores do capital.

Referências

BADIR, Sémir. **Épistémologie sémiotique**. La théorie du langage de Louis Hjelmslev. Patis: Honoré Champion, 2014.

BARROS, Diana Luz Pessoa de. **Teoria do discurso:** fundamentos semióticos. 3ª ed. São Paulo: Humanitas FFLCH/USP, 200 [1990].

BILTGEN, Patrick; RYAN, Stephen. **Activity-based intelligence:** principles and applications. Boston: The Artech House Electronic Warfare, 2016.

BOURDIEU, Pierre. **A economia das trocas simbólicas.** Trad. Sérgio Micelli. São Paulo: Perspectiva, 1974.

BOURDIEU, Pierre. PASSERON, Jean-Claude. **Os herdeiros**: os estudantes e a cultura. Trad. Nilton Valle. Santa Catarina: EdUFSC, 2018[1964].

BOYD, Danah; CRAWFORD, Kate. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. **Information, Communication & Society,** v. 15, n. 5, p. 662-279, 2012. DOI: https://doi.org/10.1080/1369118X.2012.678878.

DIEBOLD, Francis. A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline. **PIER Working Paper**, v. 2, (12-003), p. 1-8, 2012. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843. Acesso em: 20 jan. 2024.

DONDERO, Maria Giulia; REYES-GARCÍA, Everaldo. Les supports des images: photographie et images numériques. **Revue Française des Sciences de l'information et de la communication**, v. 9, 2016. DOI: https://doi.org/10.4000/rfsic.2124.

DURÁN, Juan Manuel; JONGSMA, Karin Rolanda. Who is afraid of Black Box algorithms? On the epistemological and ethical basis of trust in medical A.I. **Journal of medical ethics**, v. 47, p. 329-335, 2021. DOI: http://dx.doi.org/10.1136/medethics-2020-106820.

FONTANILLE, Jacques. **Pratiques sémiotiques.** Paris: PUF, 2008.

GANDOMI, Ana; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. **International Journal of Information Management,** v. 35, n. 2, p. 137-144, 2015. DOI: https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

GREIMAS, Argildas. Julien; COURTÉS, Joséph. **Dicionário de Semiótica**. Trad. Alceu Dias Lima, Diana Luz Pessoa de Barros, Edward Lopes e Ignácio Assis da Silva. São Paulo: Contexto, 2008[1979].

HJELMSLEV, Louis. **Résumé d'une théorie du langage.** Trad. e edição eletrônica Alain Herreman, 2010[1975].

KAMBATLA, Karthik. et al. Trends in big analytics. **Journal of Parallel and Distributed Computing,** v. 74, n. 7, p. 2561-2573, 2014. DOI: https://doi.org/10.1016/j.jpdc.2014.01.003.

KUROSE, James; ROSS, Keith. **Redes de computadores e a internet**. Uma abordagem top-down. 6 ed. Trad. Daniel Vieira. São Paulo: Pearson Education do Brasil, 2013.

MANOVICH, Lev. **The language of new media**. Cambridge MA: MIT Press, 2001.

MARX, Karl. **O Capital:** Livro 1. O processo de produção do capital. Trad. Reginaldo Anna. São Paulo: Civilização Brasileira, 1998[1867].

MAYER-SCHONBERGER, Viktor; CUKIER, Kennethy. **Big data:** a Revolution that will transform how we live, work, and think. New York: Houghton Mifflin Harcourt, 2013.

MORAES, Leticia. **A noção de texto na semiótica**. 2021. Tese (Doutorado em Semiótica e Lingüística Geral) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2021. DOI: https://doi.org/10.11606/T.8.2021.tde-29112021-185047.

NOBLE, Safya Umoja. **Algoritmos da opressão.** Trad. Felipe Damorim. São Paulo: Ed. Rua do Sabão, 2022[2018].

SADOWSKI, Jathan. When data is capital: Datafication, accumulation, and extraction. **Big Data & Society**, v. 6, n. 1, p. 1-12, 2019. DOI: https://doi.org/10.1177/2053951718820549.

SANTOS SARAIVA, Francisco Rodrigues. **Dicionário latino-português.** Rio de Janeiro: Livraria Garnier, 2006.

van DIJCK, José. Confiamos nos dados? As implicações da dataficação para o monitoramento social. ***MATRIZes,*** v. 11, n. 1, p. 39-59, 2017. DOI: https://doi.org/10.11606/issn.1982-8160.v11i1p39-59.