YOUTUBE, FACEBOOK, TWITCASTING E GOOGLE NEWS: USO DE MÉTODOS DIGITAIS E BIG DATA NA PESQUISA EM COMUNICAÇÃO

TANIA LUCÍA COBOS
UNIVERSIDAD TECNOLÓGICA DE BOLÍVAR
CARTAGENA DAS ÍNDIAS, BOLÍVAR, COLÔMBIA
TCOBOS@UTB.EDU.CO

ANA LÚCIA NUNES DE SOUSA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
RIO DE JANEIRO, RIO DE JANEIRO, BRASIL
ANALUCIA@NUTES.UFRJ.BR

YOUTUBE, FACEBOOK, TWITCASTING E GOOGLE NEWS: USO DE MÉTODOS DIGITAIS E BIG DATA NA PESQUISA EM COMUNICAÇÃO

Resumo: Este trabalho analisa dois casos de coleta de dados utilizando métodos digitais. A primeira se centra no uso de Netvizz, Nvivo y Gephi para trabalhar com dados de Facebook, YouTube e TwitCasting; e a segunda, na utilização de um scraper bot, Microsoft Excel e Tag Cloud Generator para explorar dados da plataforma Google News. As experiências demonstram que a utilização de métodos digitais na pesquisa em comunicação apresentam grande riqueza informativa e possibilidades de aprofundamento analítico que justificam seu uso, pese às debilidades apresentadas e cuidados necessários à coleta e análise.
Palavras-chave: Métodos computacionais; metodologia; pesquisa; dados massivos; comunicação e jornalismo.

YOUTUBE, FACEBOOK, TWITCASTING Y GOOGLE NEWS: USO DE MÉTODOS COMPUTACIONALES Y BIG DATA EN LA INVESTIGACIÓN EN COMUNICACIÓN

Resumen: Este trabajo analiza dos casos de recolección de datos utilizando métodos digitales. El primero se centra en utilizar Netvizz, Nvivo y Gephi para trabajar con datos de Facebook, YouTube y TwitCasting; y el segundo, usando un scraper bot, Microsoft Excel y Tag Cloud Generator para explorar datos de la plataforma Google News. Las experiencias muestran que el uso de métodos digitales en la investigación en comunicación presenta una gran riqueza informativa y posibilidades de profundización analítica que justifican su uso, a pesar de las debilidades presentadas y el cuidado necesario para la recopilación y análisis.
Palabras clave: Métodos digitales, metodología, investigación, datos masivos; comunicación y periodismo.

***YOUTUBE, FACEBOOK, TWITCASTING* AND *GOOGLE NEWS*: USE OF COMPUTATIONAL METHODS AND BIG DATA IN COMMUNICATION RESEARCH**

Abstract: This work analyzes two cases of data collection using digital methods. The first focuses on using Netvizz, Nvivo and Gephi to work with data from Facebook, YouTube and TwitCasting; and the second, using a scraper bot, Microsoft Excel and Tag Cloud Generator to explore data from the Google News platform. The experiences demonstrate that the use of digital methods in research in communication presents great informational wealth and possibilities of analytical deepening that justify its use, despite the weaknesses presented and the necessary care for collection and analy-

sis.
Keywords: Digital methods, methodology, research, big data; mass communication and journalism.

1 INTRODUÇÃO

A quinta revolução tecnológica, a era da informática e das telecomunicações (PÉREZ, 2004, p. 44) – dominada por companhias tecnológicas estadunidenses como Google, Facebook, Twitter, entre outras, é caracterizada pelo uso de algoritmos[1], que favoreceram a captura de grandes quantidades de dados (estruturados, semi estruturados e não estruturados), armazenados em bases de dados públicas e privadas, genericamente chamadas de big data ou dados massivos.

Uma das funções que vêm sendo empregadas a estes dados é a pesquisa científica. Para Hilbert (em HOPENHAYN, 2017), as Ciências Sociais se transformaram em uma das áreas mais ricas no que se refere ao acesso a dados científicos. Atualmente, as pessoas parecem levar um sensor individual 24h por dia: sabemos onde estão, o que compram, quando dormem, quem são suas amizades, ideias políticas, vida social, etc. Toda esta informação, gerada no mundo digital, é armazenada em bases de dados massivos que podem ser exploradas através de análises complexas de fenômenos sócio-culturais abordados, entre outros, a partir da perspectiva comunicacional.

O termo *big data*, de acordo a Hadi et al. (2015, p.16), foi introduzido no âmbito computacional por Roger Magoulas, da Agência O'Reilly b vbMedia, em 2005, para se referir a uma grande quantidade de dados que as técnicas tradicionais de gestão de dados não podiam administrar e processar devido à complexidade e tamanho. Os dados massivos estão compostos por um grande número de peças de informação que podem ser cruzadas, comparadas, agregadas e desagregadas em grande profundidade. Ainda que não exista uma definição rigorosa, Mayer-Schönberger e Cukier (2013, p.17) sugerem que os dados massivos:

> Se referem a coisas que podem ser feitas em grande escala, mas não numa escala inferior, para extrair novas percepções ou criar novas formas de valor, de modo que transformam aos mercados, as organizações, as relações entre cidadãos e os governos, etc.

1 "Conjunto de regras que, aplicadas sistematicamente a dados de entrada apropriados resolvem um problema em um número finito de passos elementares" (Peña Marí em FANJUL, 2018).

Hadi et al. (2015, p. 20) e Marr (2016) identificam cinco grandes características do *big data*: volume (*volume*), variedade (*variety*), velocidade (*velocity*), veracidade ou validade (*veracity or validity*) e valor (*value*). Volume se refere ao grande tamanho; variedade à diversidade de tipologia e fontes de dados; velocidade à rapidez com que estes são gerados; veracidade e validade à garantia de qualidade dos dados ou à sua autenticidade e credibilidade; e valor à utilidade ou benefício que seus proprietários podem obter ao explorá-los.

No início, a criação destas bases de dados massivas respondeu a interesses comerciais e de mercado das empresas de tecnologia. Hoje, entretanto, é inegável que a captura, armazenamento, compartilhamento, análise e visualização em busca de padrões que permitam delinear correlações está presente em praticamente todas as facetas da vida humana. Assim, se fazem presentes nas estratégias de marketing, comércio eletrônico, nas telecomunicações, governo eletrônico, processos eleitorais, saúde pública, vigilância; e em outros campos, como o científico e, no caso específico que tratamento neste texto: a comunicação. Atualmente, também não se pode esquecer que o *big data* enfrenta grandes desafios, tais como: a ética na captura, manejo da privacidade, atualização e reforço das desigualdades e preconceitos, entre outros.

Neste artigo, nos propomos a analisar criticamente a utilização de dados massivos como metodologia de pesquisa na comunicação, e a sua captura e tratamento através de métodos computacionais e/ou métodos digitais. Nos interessa reflexionar em que medida o *big data* pode gerar mais conhecimento e se a implementação dos métodos digitais é apropriada para o campo da comunicação. Para tal, partimos de uma discussão teórica e aterrisamos em dois exemplos de pesquisa na qual esta combinação foi utilizada. A primeira, em relação a dados gerados por seres humanos em Facebook, YouTube e TwitCasting; e a segunda, se refere a dados gerados em Google News por meios jornalísticos.

2 UM OLHAR CRÍTICO AO USO DE DADOS MASSIVOS

Os defensores de dados massivos argumentam que é necessário mudar o paradigma científico utilizado até o momento, já que sua utilização só tem sentido se também se aceita a imprecisão da metodologia, a necessidade de confiar em correlações e – mais importante – que "os dados massivos tratam do quê, e não do porquê. Nem sempre precisamos conhecer as causas do

fenômeno, preferencialmente, podemos deixar que os dados falem por si mesmos" (MAYER-SCHÖNBERGER e CUKIER, 2013, p. 26-27).

É possível capturar, estruturar e armazenar os grandes acontecimentos mundiais em bases de dados igualmente grandes. Os dados massivos representam um avanço no que se refere à análises macro, mas são uma ferramenta pouco útil quando a intenção é analisar um fenômeno em suas singularidades. Neste sentido, a necessidade de conhecer um fenômeno detalhadamente é considerada inútil pelos defensores dos dados massivos, para os quais é suficiente conhecer a tendência geral. É questionada, inclusive, a necessidade de fazer amostras e ter hipóteses de pesquisa, argumentam que "agora temos tantos dados a nossa disposição, e tanta capacidade de processamento, que já não precisamos escolher com dificuldade uma aproximação ou um pequeno punhado delas e examiná-las uma a uma" (MAYER-SCHÖNBERGER e CUKIER, 2013, p.75).

É preciso considerar que os dados massivos fornecem uma quantidade assombrosa de informação e possibilidades à ciência e à sociedade, mas mesmo assim não escapa aos críticos e céticos quanto a seu verdadeiro papel e potencial. Muitos dos defensores dos dados massivos trabalham com a crença em sua completa objetividade. Assim, para eles, é suficiente "lançar os números dentro dos maiores *clusters* de computadores do mundo e deixar que os algoritmos estatísticos encontrem os padrões que a ciência não encontrou" (ANDERSON, 2008). Mas recoletar e datificar – transformar em dados – uma quantidade tão grande de informação pode resultar em um trabalho muito complexo. O pesquisador precisa conhecer profundamente os softwares que auxiliam no processo. Também é possível que haja confusão na combinação entre diferentes tipos de informação provenientes de fontes distintas e vários tipos de erros, transformando a análise em um procedimento de alto risco (MAYER-SCHÖNBERGER e CUKIER, 2013; MAHRT e SCHARKOW, 2013; ROGERS, 2013).

Alguns pesquisadores apontam que a análise de dados massivos pode mostrar o que fazem os usuários, mas não porque fazem (MAYER-SCHÖNBERGER e CUKIER, 2013). Também costumam revelar informação superficial e pouca sensibilidade ao contexto em que os dados foram gerados (MANOVICH, 2012; MAHRT e SCHARKOW, 2013; BOYD E CRAWFORD, 2012). Outro problema, segundo Andersen (em BOLLIER, 2010, p. 12), é o risco de tirar conclusões a partir de um único conjunto de dados. Desta forma, aponta que seria mais seguro usar sets de dados provenientes de múltiplas fontes

e que: "sempre que estatísticas são feitas, encontramos correlações ruins e laços que proximidade que, na verdade, não existem". Cassin (2008, p.113) argumenta que a automatização não garante objetividade. Igualmente Andersen (em BOLLIER, 2010, p.13) questiona a suposta objetividade dos dados, uma vez que devem ser "limpados" e isto afetaria a objetividade, já que é um processo subjetivo realizado pela pessoa que investiga, informando quais variáveis importam e quais não.

Mahrt e Scharkow (2013, p. 21) questionam a validez dos dados massivos em casos onde o pesquisador "deixa que os dados falem por si mesmos", contrário ao que defendem Mayer-Schönberger e Cukier (2013). Nestes casos, os pesquisadores costumam utilizar quaisquer dados disponíveis e, logo, construir uma justificativa teórica para sua utilização. Mahrt e Scharkow (2013, p. 25) asseveram que esta estratégia é totalmente contrária à teoria tradicional e atenta contra a validez e alcance dos resultados.

Por estes motivos, não são poucos os pesquisadores que questionam se, de fato, mais dados significam mais conhecimento. Em muitos contextos, uma pequena mostra pode dizer mais e responder melhor às inquietudes de uma pesquisa que milhares de dados (BOLLIER, 2010; MAHRT e SCHARKOW, 2013; KING e LOWE, 2003; SCHRODT, 2010; KRIPPENDORFF, 2004).

Mas as críticas ao *big data* não se limitam ao campo científico. No século XXI, como já mencionamos, os dados são a alma dos negócios. Em geral, os usuários não têm consciência de que suas marcas digitais formarão parte de uma pesquisa, seja comercial, policial ou científico-acadêmica. Parte-se do pressuposto que as pessoas concordam automaticamente com a utilização de suas publicações, fotos, vídeos, e demais interações, mas há questões pendentes envolvendo o direito à privacidade e os direitos de autoria. Todos os rastros gerados pelos internautas ou em qualquer tipo de ferramenta de comunicação estão datificados e podem ser transformados em mercadorias de alto valor e interesse para corporações econômicas e governos (MAYER-SCHÖNBERGER e CUKIER, 2013, p. 51). Os usuários, em sua maior parte, têm pouco ou nenhum conhecimento de que tudo que é realizado online está sendo transformado em mercadoria, implicando, algumas vezes, em violações à privacidade, liberdade civil e consentimento (BOLLIER, 2010).

3 OS MÉTODOS COMPUTACIONAIS COMO OPÇÃO METODOLÓGICA

Como já foi mencionado, a análise de dados massivos requer o domínio de utilização de determinados *softwares* ou programas informáticos que

permitam processar e visualizar estes enormes conjuntos, uma vez que a capacidade humana para fazer uma análise manual é reduzida. Rieder (2013) afirma que, há mais de uma década, programas informáticos são utilizados para capturar, produzir ou utilizar de outra maneira os dados massivos, objetivando investigar diferentes aspectos da internet. Os métodos digitais e/ou computacionais possuem uma série de vantagens comparados aos métodos tradicionais, tais como: relativas ao custo, velocidade, exaustividade, detalhe, entre outros, mas também relacionadas à rica contextualização que pode ser proporcionada pela estreita relação entre os dados e as propriedades do meio (entendida como tecnologias, plataformas, ferramentas, sítios web, etc). Para Rogers (2015), os métodos digitais são técnicas para o estudo das mudanças sociais e das condições culturais usando dados oriundos da Internet.

Esta metodologia faz uso de conjuntos de dados massivos armazenados, por exemplo, *hiperlinks*, etiquetas, marcadores temporais, interações de qualquer natureza nas redes sociais digitais como os *likes*, elementos compartilhados, *retweets*, comentários, entre outros, e busca compreender como são tratados pelos métodos incorporados nas plataformas digitais dominantes. Os métodos digitais se esforçam para reorientar a finalidade dos métodos e serviços *online* ao ponto de vista de pesquisa social, e como uma prática de investigação, formando parte do giro computacional das humanidades e das ciências sociais, e dentro desta última, a comunicação.

Como metodologia, objetiva orientar a finalidade dos dados massivos armazenados na internet em diferentes plataformas online (ex: Facebook, Twitter, Google, etc.) para a pesquisa social, utilizando-se de métodos e ferramentas informáticas cuja implementação dependerá de qual plataforma terá os dados extraídos, de como os dados deverão ser estruturados e visualizados. Neste sentido, como já foi mencionado, o pesquisador deve dar-se a tarefa prévia de conhecer o manejo e dominar os programas informáticos que serão utilizados para estas tarefas. É importante mencionar que estes métodos computacionais, são, além disso "experimentais e situacionais" (ROGERS, 2015, p. 9), já que são construídas, em algumas ocasiões, sobre plataformas que podem deixar de funcionar ou simplesmente desaparecer, como páginas *web* ou determinadas determinadas funcionalidades das redes sociais online e outros serviços conexos.

Os métodos computacionais facilitam a automatização, mas não substituem totalmente o critério interpretativo do pesquisador. Os dados falam

e as correlações são mostradas, mas o que significam, implicam, sugerem, o que é deduzido ou inferido disso é tarefa da equipe investigadora. Esta, por sua vez, deve ser consciente das limitações técnicas: a temporalidade dos serviços *web*, a instabilidade dos fluxos de dados que pode ocorrer por reconfigurações das API (*application programming interface*), a qualidade dos dados capturados; as limitações, instabilidades e imprecisões dos algoritmos e o viés que ocasiona a "limpeza" ou curadoria dos dados para seu processamento. É necessário ter presente, assim mesmo, que os métodos digitais não só permitem determinar tendências gerais em meio à massividade, mas também aprofundar em detalhe a "letra pequena" do fenômeno, e o variado leque de programas informáticos permite fazer leituras simultâneas dos dados.

É importante ter em conta que, pese à agitação em torno das possibilidades abertas pelas técnicas computacionais e seus *softwares* e programas de análise de dados, este ainda é um campo com muitas dificuldades e riscos. Manovich (2012, p. 9-10) sugere que os dados massivos devem ser utilizados em combinação com outras técnicas, incluindo as clássicas:

> Idealmente, queremos combinar a habilidade humana para compreender e interpretar – coisa que os computadores ainda não podem fazer – com a capacidade das máquinas para analisar grandes conjuntos de dados, utilizando os algoritmos que criamos para tal.

Nuttall et al (2011) apontam na mesma direção, sugerindo uma abordagem científica que possa combinar os métodos que trabalham com dados e a etnografia. Finalmente, Rogers (2013) assevera que os métodos digitais precisam de um longo tempo de dedicação, além de um olhar crítico para a análise dos dados, pois somente assim poderá produzir resultados satisfatórios.

Por último, há muitos desafios em relação a "que objetos considerar, como criar uma mostra, como analisar, como interpretar, como chegar aos resultados" (ROGERS, 2013, p.85). Todos estes desafios foram constantes nos exemplos que serão expostos a seguir. Em primeiro lugar, apresentaremos um estudo sobre o vídeo ativismo no Brasil, durante o Mundial de Futebol da FIFA, em 2014. Neste caso, a obtenção dos dados foi gerada por usuários nas plataformas sociais Facebook, YouTube e TwitCasting. No segundo caso, temos uma pesquisa sobre o tratamento de fontes jornalísticas de quatro edições ibero-americanas de Google News, em 2015, a partir da

realização de um *scraping* ou "raspagem" de dados gerados por algoritmos, sem intervenção humana.

4 PRIMEIRO CASO: IMPLEMENTANDO MÉTODOS DIGITAIS PARA A PESQUISA NO FACEBOOK, YOUTUBE E TWITCASTING

Nosso primeiro caso se refere a uma pesquisa de doutorado intitulada "*De la calle a la red: videoactivismo en el contexto de las protestas en contra del Mundial de Fútbol en Río de Janeiro (2014)*" (SOUSA, 2017). Nesta pesquisa foram implementados métodos digitais, investigação participativa e entrevistas semi-estruturadas, propondo um olhar amplo e profundo sobre o vídeo ativismo desenvolvido na cidade do Rio de Janeiro (Brasil) durante o Mundial de Futebol da FIFA, em 2014. No âmbito deste trabalho, vamos a referir-nos apenas aos métodos digitais utilizados, ainda que mencionemos as outras técnicas empregadas aos leitores, principalmente pesquisadores em formação, para que possam ter nitidez de todo o processo metodológico envolvido na realização da pesquisa e como estas diversas técnicas se relacionam e se complementam.

Em termos práticos, os métodos digitais aplicados aos meios sociais possibilitam que os dados sejam coletados de forma automatizada a partir das plataformas, visualizados e, posteriormente, analisados. Os dados podem ser capturados através de um *scraping* ou através da utilização de APIs. O Laboratório Digital Methods Initiative, vinculado à Universidade de Amsterdam, lista vários *softwares* ou programas de extração de dados (alguns desenvolvidos por eles), baseados nas APIs específicas de cada plataforma, que facilitam este trabalho de coleta. No caso desta pesquisa, optamos pela utilização destas ferramentas e também, em alguns momentos, pela coleta manual. Como já foi mencionado, as metodologias digitais são compostas por diversas técnicas, que fazem uso de diferentes *softwares* diversos para a captura, visualização e análise de dados. Este processo exige muito esforço, dedicação e pode levar bastante tempo, tanto a aprendizagem das ferramentas, como testes e comprovações com as bases de dados geradas pelos sistemas.

No caso específico desta pesquisa, não se havia pensado na utilização destas técnicas até o início do trabalho de campo, em junho de 2014. Foi neste momento que o potencial de potencial de Facebook, YouTube e TwitCasting foi descoberto, deixando evidente que as dinâmicas desenvolvidas nestes ambientes virtuais eram fundamentais para compreender o vídeo ati-

vismo como um processo comunicativo de forma completa. A partir deste momento, começamos a estudar estas ferramentas e a provar vários *softwares*. Após selecionar as principais ferramentas disponíveis, optamos por utilizar, inicialmente, Netvizz para a captura dos dados em Facebook; Nvivo para gerar nuvens de etiquetas e categorizar os dadas para a análise e Gephi para a visualização.

Quando o trabalho de campo foi iniciado, a coleta de dados – ou primeira captura – era realizada diariamente utilizando a *app* Netvizz, que coletava as publicações, comentários e demais ações dos usuários no Facebook. No decorrer do processo, a app foi atualizada, permitindo que dados de dias anteriores pudessem ser capturados a qualquer momento. A partir deste momento já não era mais necessária a coleta diária e optamos por suspender este procedimento, direcionando os esforços para a investigação participativa e entrevistas, que eram realizadas ao mesmo tempo, durante o Mundial de Futebol da FIFA, em junho e julho de 2014.

Esta decisão se tornou um problema posteriormente, já que quando reiniciamos a coleta de dados e tentamos corroborá-los numa segunda captura notamos várias incongruências na base de dados. Na segunda captura, Netvizz estava limitado devido a restrições impostas por Facebook. Além disso, algumas publicações haviam sido apagadas da plataforma, e também houve mudanças na própria API do Facebook. Os principais problemas encontrados na utilização de Netvizz foram: 1) a eliminação da *fanpage* de uns dos meios estudados, o Jornal A Nova Democracia, durante o desenvolvimento da pesquisa; 2) os dados são maleáveis, ou seja, a data da captura pode determinar que um conteúdo específico fosse ou não coletado, uma vez que usuários e páginas modificam suas configurações de privacidade, alterando, assim, os dados possíveis de ser coletados.

Netvizz gera apenas folhas de cálculo com os dados capturados, assim decidiu-se recorrer ao *software* Gephi para a visualização. Entretanto, a utilização do programa se mostrou complexa, mesmo dedicando bastante tempo à tarefa. Como os dados necessários à pesquisa eram simples, exploramos apenas algumas das funcionalidades do *software*, para gerar a visualização da rede e a conexão dos ativistas no entorno das plataformas de meios sociais, neste caso específico, o Facebook. Ao final, mesmo utilizando apenas os recursos básicos e simples de Netvizz e Gephi, conseguimos concluir que ambas são ferramentas poderosas para explorar os dados digitais.

Os dados do YouTube foram coletados, inicialmente, de forma manual.

Construímos uma base de dados com todos os vídeos do período da mostra, totalizando 173 vídeos. Para analisar as ações em torno da narrativa audiovisual de forma mais profunda, observar as interações e comentários da audiência, foi realizada uma segunda coleta de dados, com a utilização do *software* YouTube Data Tools. Esta ferramenta, por sua vez, permitiu visualizar os seguintes dados: informações do canal, lista de vídeos, informações e conteúdo de cada um dos vídeos, comentários e outras interações da audiência, etc.

Já em relação ao TwitCasting, uma plataforma usada para vídeo *streaming* através de telefones celulares, os dados foram capturados de forma manual, uma vez que não havia, até aquele momento, uma ferramenta específica para a captura de dados nesta plataforma. Esta é outra limitação que precisa ser considerada. Com a variedade de plataformas de mídias sociais utilizadas atualmente, muitas delas ainda não dispõe de ferramentas digitais adequadas para a coleta, visualização e análise dos dados. Nestes casos, a coleta deve ser manual ou a pesquisa precisará providenciar o desenvolvimento da ferramenta.

Em resumo, os dados foram coletados utilizado Netvizz para Facebook, YouTube Data Tools para YouTube, e manualmente no caso de TwitCasting. No total, foram coletados dados de 173 vídeos, nas três plataformas. Entretanto, optamos por utilizar apenas os 10 vídeos mais visualizados de cada plataforma, totalizando uma análise de: 1) em Facebook foram 20 mensagens/*posts*, que acompanhavam a publicação dos vídeos, que concentraram 4.455 *likes*, 6.211 comentários e foram compartilhados 6.555 vezes; 2) no YouTube foram 20 vídeos, 1.886.143 visualizações, 4.523 comentários, 1.427 *dislikes*, e 3.462 "compartilhar"; 3) no TwitCasting, 82 vídeos e 20.500 comentários. Todos estos dados, posteriormente, foram analisados utilizando o software Nvivo. A análise realizada buscou revelar as tendências de conteúdo das mensagens, ou seja, realizamos uma análise conteúdo das mesmas. Esta análise permitiu avaliar o discurso dos atores envolvidos no processo comunicativo, dando a conhecer as ações de cada um dos atores e o papel que desempenharam na narrativa vídeo ativista.

5 SEGUNDO CASO: IMPLEMENTANDO MÉTODOS COMPUTACIONALES PARA INVESTIGAR EN GOOGLE NEWS

A pesquisa de doutorado *"Medios de comunicación iberoamericanos y agregadores de noticias: análisis a las ediciones de Google News Brasil, Co-*

lombia, España, México y Portugal" (COBOS, 2017) adotou uma metodologia mista, combinando métodos computacionais, consulta documental e entrevistas (tanto presenciais como virtuais). A triangulação destes métodos permitiu realizar uma análise dos meios de comunicação jornalísticos, com ênfase nos ibero-americanos, indexados na edições Google News dos países mencionados, em aspectos como sua identificação, localização geográfica, quotas de agregação de notícias, empresa proprietária, e as percepções e experiências dos editores-chefe, diretores ou proprietários dos meios sobre o agregador de notícias. No caso da Espanha, como a edição do Google News estava interrompida, somente foram aplicados os métodos tradicionais de coleta de dados (entrevistas pessoais e telefônicas).

Novamente, no âmbito deste trabalho, nos referimos somente aos métodos computacionais utilizados neste trabalho, mas fazemos menção a outras técnicas empregadas para que possa ser evidente como os vários métodos utilizados se complementam. Inicialmente, quando a pesquisa foi iniciada, em 2014, não se vislumbrava a utilização de métodos computacionais. A ideia surgiu após ler o artigo, publicado em um blog, "Lista de fuentes de Google News España" (DANS, 2005). O texto mencionava o uso de um *script* em PHP para listar estas fontes. Além disso, a participação em uma conferência de Bernhard Rieder — professor da Universidad de Ámsterdam e membro de Digital Methods Initiative —, realizada na Universitat Autònoma de Barcelona, na qual foram apresentadas várias ferramentas computacionais que poderiam ser utilizadas nas pesquisas em comunicação, entre as quais Google News Scraper confirmou o interesse em utilizar tais técnicas.

Dado o objetivo geral do projeto era necessária a captura de notícias das edições mencionadas de Google News, e ao ver que era possível fazê-lo de forma massiva e automatizada, utilizando um *scraper* ou *scraper bot* (raspador), o que brindaria uma maior e melhor aproximação ao fenômeno, optou-se por estudar em profundidade em que consistia o *web scraping*. Posteriormente, determinou-se as variáveis que o *scraper bot* deveria capturar e ao compreender que a ferramenta Google News Scraper não era a mais adequada para realizar a tarefa, um desenvolvedor de *software* foi integrado à equipe para criar um *scraper bot* en PHP capaz de capturar e armazenar, em uma base dados MySQL, as nove variáveis estipuladas para cada notícia.

Aqui é importante destacar que a pesquisadora possuía conhecimentos prévios do jargão informático e uma compreensão inicial de como funcionava o *web scraping*, elementos que facilitaram muito a comunicação com

o desenvolvedor no processo de construção, prova, ajustes e funcionamento do scraper bot e o armazenamento dos dados, assim como a posterior exportação a folhas de cálculo de Microsoft Excel para o processamento e análise. É importante mencionar que, como toda técnica informática, que a mesma não está isenta de erros, e que isto faz parte das limitações do projeto. Por exemplo, lentidão no processamento dos dados devido à saturação da memória do computador, eventuais quedas de serviço em Google News, em algum momento, etc.

No total, foram coletados 5.048.150 milhões de notícias, oriundas de 2.378 meios jornalísticos iberoamericanos. Uma vez finalizado o *scraping* e a organização das informações em tabelas de Microsoft Excel, foi realizada uma revisão manual, na qual percebeu-se a necessidade de realizar uma limpeza manual aos dados para sanar as imprecisões detectadas no funcionamento de StoryRank (o algoritmo que opera em Google News) ou produto de falhas humanas na inserção dos dados (de identificacao do meio jornalístico). Assim, houve a necessidade de corrigir manualmente a erros presentes na origem dos dados para, depois de todo este processo que levou meses, conseguir analisar os dados. Assim, é importante atentar para o fato de que a automatização na coleta de dados não implica necessariamente que os mesmos estejam corretos na origem, pode haver anomalias que vão determinar a necessidade de um trabalho manual antes da análise, como no caso mencionado. Voltando ao nosso exemplo, uma vez que os dados foram revisados, procedeu-se à correlacioná-los, utilizando filtros, ordenações, eliminação de duplicidades e tabelas dinâmicas de Microsoft Excel. Além disso, foram gerados gráficos e visualizações de dados utilizando o mesmo programa.

Outra ferramenta de dados computacionais utilizadas foi Tag Cloud Generator. Esta ferramenta foi utilizada para, a partir dos títulos das notícias, criar nuvens de etiquetas que permitissem identificar os termos mais frequentes e, assim, ter uma aproximação aos temas selecionados por Google News para criar sua agenda, nos diferentes canais do serviço, tanto por cada edição, como em um panorama geral. Cabe mencionar que, uma vez geradas as diferentes nuvens de etiquetas, os artigos e conectores foram eliminados manualmente (Ex: como, a, o, os, este, etc.), para deixar apenas as palavras com significados temáticos (Ex: nome de personagens, lugares, fatos etc.).

À guisa de conclusão, as bases de dados geradas com as notícias cap-

turadas em cada edição do agregador ou *datasets* (*Noticias de Google News ediciones Brasil, Colombia, México y Portugal, 2015*), foram liberadas com licenças Creative Commons no Dipòsit Digital de Documents da Universitat Autònoma de Barcelona para que possam ser utilizadas em outras pesquisas científicas. Assim, tratamos de contribuir com o movimento de Dados Abertos ou *Open Data*, que tem por objetivo oferecer a pesquisadores conjuntos de dados para que possam ser explorados livremente, de forma individual ou combinada em outros projetos de pesquisa.

6 CONCLUSÕES

Chegados a este ponto, é evidente que, na segunda década do século XXI, as ciências sociais se converteram agora em uma das áreas mais ricas em dados a partir do boom tecnológico que vivemos. Assim, isto se converte em grandes oportunidades de pesquisa, mas também tem seu lado escuro. Na discussão teórica é possível apreciar que o uso de dados massivos e a implementação de métodos computacionais para seu processamento é, ainda, um campo contraditório, de experimentação, com suas potencialidades e riscos, com seus partidários e detratores.

Este cenário também resulta desafiador a nível técnico para os cientistas sociais, os quais devem aprender técnicas informáticas, dominar a terminologia básica e aprender o manejo de programas para a coleta, processamento e visualização. Tudo isto implica uma curva de aprendizagem, além de não substituir, de nenhuma forma, a análise e raciocínio de quem investiga. Além disso, em alguns casos, é necessário interactuar com desenvolvedores de *software* e "traduzir-lhes" as necessidades da pesquisa. Também é importante reconhecer que o desenvolvimento do código informático não é uma "varinha mágica" que faz com que, automaticamente, apareçam os dados, ou seja, compreender que todas as ferramentas tem suas limitações. Por outro lado, o processamento de dados massivos demanda a compra de licença de alguns programas (nem todos são *freeware*), além de disponibilidade de computadores com alta capacidade de armazenamento, necessária para processar os dados com celeridade. Isto implica, portanto, a execução de um orçamento na atividade investigativa.

Finalmente, em relação aos dois casos apresentados, distintos entre si, se observa que os dados massivos tratados através dos métodos digitais trouxeram riqueza informativa para: no primeiro caso, identificar as ações dos vídeo ativistas durante os protestos realizados no contexto da FIFA

World Cup 2014; e no segundo, registrar o comportamento de um algoritmo em relação às notícias que foram coletadas e hierarquizadas, nas referidas edições do agregador de notícias. É importante considerar que, neste último caso, houve a necessidade de agregar trabalho manual, particularmente na limpeza dos dados antes de que fossem analisados. Nos dois casos também é menester considerar a combinação dos métodos digitais com técnicas tradicionais de pesquisa, particularmente as qualitativas, para obter resultados complementares entre si. Os métodos computacionais são, de fato, uma alternativa metodológica para os pesquisadores em comunicação, mas sua utilização requer a compreensão profunda de suas limitações.

REFERÊNCIAS

ANDERSON, Chris. The end of theory. **Wired**. Publicado em 23/06/2008. Disponível

em <https://www.wired.com/2008/06/pb-theory/>. Acesso em 15 set. 2020.

BOLLIER, David. **The Promise and Peril of Big Data**. Washington: The Aspen Institute, 2010. Disponível em <https://assets.aspeninstitute.org/content/uploads/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf>. Acesso em 15 set. 2020.

BOYD, Danah e CRAWFORD, Kate. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. **Information, Communication & Society**, v. 15, n. 5, p. 662-679. Taylor & Francis, 2012.

CASSIN, Barbara. **Googléame, la segunda misión de los Estados Unidos**. México DF: Fondo de Cultura Económica, 2008.

COBOS, Tania Lucía. **Medios de comunicación iberoamericanos y agregadores de noticias: análisis a las ediciones de Google News Brasil, Colombia, España, México y Portugal**. Barcelona: Universitat Autònoma de Barcelona, 2017. Disponível em <https://ddd.uab.cat/record/188096> Acesso em 15 set. 2020.

DANS, Enrique. **Lista de fuentes de Google News España**. Enrique Dans. Publicado em 29.03.2005. Disponível em <https://www.enriquedans.com/2005/03/lista-de-fuentes-de-google-news-espana.html> Acesso em 15 set. 2020.

FANJUL, Sergio. Matemáticas. En realidad, ¿qué [...] es exactamente un algoritmo?. **El País.** Publicado em 24.03.2018. Disponível em <https://retina.elpais.com/retina/2018/03/22/tendencias/1521745909_941081.html> Acesso em 15 set. 2020.

HADI, Hiba Hassin, SHNAIN, Ammar Hameed, HADISHAHEED, Sarah e AHMAD, Azizahbt Haji. Big data and five V's characteristics. **International Journal of Advances in Electronics and Computer Science**, v. 2, n. 1, p. 16-23. IRAJ, 2015. Disponível em <http://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf>Acesso em 15 set. 2020.

HOPENHAYN, Daniel. Martin Hilbert, experto en redes digitales: "Obama y Trump usaron el Big Data para lavar cerebros". **The Clinic**. Publicado em 19/01/2017. Disponível em <https://www.theclinic.cl/2017/01/19/martin-hilbert-experto-redes-digitales-obama-trump-usaron-big-data-lavar-cerebros/> Acesso em 15 set. 2020.

KING, Gary e LOWE, Will. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. **International Organization**, v. 57, n. 3, p. 617-642. Cambridge Core, 2003.

KRIPPENDORFF, Klaus. **Content analysis. An introduction to its methodology**. 2a ed. Thousand Oaks, CA: Sage, 2004.

MAHRT, Merja e SCHARKOW, Michael. The Value of Big Data in Digital Media Research. **Journal of Broadcasting & Electronic Media**, v. 57, n. 1, p. 20-33. Taylor & Francis, 2013.

MANOVICH, Lev. Trending: The Promises and the Challenges of Big Social Data. Em GOLD, Matthew (ed.) **Debates in the Digital Humanities**. Minnesota: The University of Minnesota Press, 2012.

MARR, Bernard. **Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results**. Nueva York, NY: Wiley, 2016.

MAYER-SCHÖNBERGER, Viktor e CUKIER, Kenneth. **Big data: la revolución de los datos masivos**. Madrid: Turner Publicaciones, 2013.

SOUSA, Ana Lúcia. **De la calle a la red: videoactivismo en el contexto de las protestas en contra del Mundial de Fútbol en Río de Janeiro (2014)**. Barcelona: Universitat Autònoma de Barcelona, 2017. Disponível em <https://ddd.uab.cat/record/188119> Acesso em 15 set. 2020.

NUTALL, Peter, SHANKAR, Avi, BEVERLAND, Michael e STALLWOTRTH, Cheryl. Mapping the Unarticulated Potential of Qualitative Research: Stepping out from the Shadow of Quantitative Studies. **Journal of Advertising Research**, v. 51, n. 1, p. 153-166. WARC, 2011.

PÉREZ, Carlota. **Revoluciones tecnológicas y capital financiero: la dinámica de las grandes burbujas financieras y las épocas de bonanza**. México DF: Siglo XXI Editores S.A de C.V., 2004.

RIEDER, Bernhard. Studying Facebook via data extraction: the Netvizz application. **Annual ACM Web Science Conference**, París, maio 2-4 2013, p. 346-355. New York: ACM.

ROGERS, Richard. **Digital Methods**. Cambridge, MA: MIT Press, 2013.

ROGERS, Richard. **Digital methods for Web research**. Em SCOTT, Robert, BUCHMANN, Marlis e KOSSLYN, Stephen (eds.) Emerging trends in the social and behavioral sciences: An Interdisciplinary, Searchable, and Linkable Resource. Nueva York, NY: Wiley, 2015.

SCHRODT, Philip. Automated production of high-volume, near-real-time political event data. **American Political Science Association 2010 Annual Meeting**, Washington, setembro 2-5 2010. Disponível em <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1643761##> Acesso em 15 set. 2020.

Tania Lucía Cobos

Doutora em Comunicação e Jornalismo pela Universitat Autònoma de Barcelona (Espanha). Professora e pesquisadora do Curso de Comunicação Social da Universidad Tecnológica de Bolívar (Cartagena de Indias, Colômbia).
E-mail: tcobos@utb.edu.co

Ana Lúcia Nunes de Sousa

Doutora em Comunicação e Jornalismo pela Universitat Autònoma de Barcelona (Espanha) Professora e pesquisadora do Instituto Nutes de Educação em Ciências e Saúde e do Programa de Pós-graduação em Educação em Ciências e Saúde da Universidade Federal do Rio de Janeiro.
E-mail: analucia@nutes.ufrj.br