

O crescimento do *big data* e as possíveis implicações éticas do seu uso na análise das redes sociais

RAFAEL FERNANDES NOVO
Centro Paula Souza – SP – Brasil
rfnovo@hotmail.com

JOSÉ MANOEL SOUZA DAS NEVES
Centro Paula Souza – SP – Brasil
jmneves.fatec@gmail.com

MARILIA MACORIN DE AZEVEDO
Centro Paula Souza – SP – Brasil
marilia.azevedo@fatec.sp.gov.br

Resumo - O termo *big data* foi cunhado para definir um conjunto de tecnologias e métodos diferenciados para análise de dados, permitindo as empresas evoluírem seu modelo analítico e competirem de forma eficiente. Por possuírem o potencial de coletar e analisar grandes volumes de dados nos mais variados formatos torna-se pertinente a discussão das possíveis implicações éticas do uso destes sistemas. Como resultado, foi possível identificar um aumento substancial no interesse e estudo sobre *big data* em contrapartida a um grupo de usuários do Facebook ainda despreparados para esse novo modelo de análise de dados.

Palavras-chave: Big data; inovação; ética; redes sociais; competitividade.

Abstract - The term Big Data has been coined to define a set of different technologies and methods for data analysis, enabling companies to evolve their analytical model and compete more efficiently. Because they have the potential to collect and analyze large volumes of data in various formats becomes worthwhile discussing the possible ethical implications of using these systems. As a result, it was possible to identify a substantial increase in interest and study on big data versus a group of Facebook users still unprepared for this new model of data analysis.

Keywords: Big data; innovation; ethics; social networks; competitive edge.

1. Introdução

Os sistemas de análise de dados estão presentes em empresas dos mais diversos segmentos, seja no cerne do negócio de multinacionais como Google ou Walmart, seja oferecendo *insights* para áreas específicas como marketing e finanças de empresas de diversos segmentos. A inteligência analítica permite melhorar o poder de competição das empresas, além de ser uma parte importante dos sistemas de BI (*business intelligence* ou inteligência de negócio). (DAVENPORT, 2007).

Nos últimos anos, surgiu o termo *big data* para definir arquiteturas de sistemas capazes de lidar com as novas dimensões dos dados: velocidade, variedade e volume. Brynjolfsson (2012) afirma ainda que soluções de *big data* possuem um potencial ainda maior do que as soluções analíticas tradicionais para trazer benefícios e aumentar a competitividade das empresas e considera que o *big data* trará uma revolução na gestão, afirmando que os executivos das empresas precisam encarar este novo modelo o quanto antes.

Fontes para o *big data* podem englobar tudo, desde dados de voz de *call center* até dados de genoma de pesquisa biológica e medicinal (DAVENPORT, 2012) e podem ser provenientes das mais diversas origens, como uma copiadora ou até mesmo um motor de jato (BROWN, 2011). Nota-se que algumas das fontes de dados de maior importância que podem ser utilizadas são relativamente novas, como as redes sociais, dentre as quais se destacam o Facebook lançado em 2004 e o Twitter em 2006, que aliadas a também recente proliferações de smartphones, tablets, ultrabooks, redes wireless e redes de dados 3G e 4G, permitem às pessoas permanecerem conectadas em praticamente todo o tempo, gerando conteúdo, ideias e opiniões.

Neste novo paradigma, com informações criadas de maneira pulverizada pelos consumidores e com os novos sistemas para armazenamento e análise dos dados, os chamados sistemas de *big data*, surge a necessidade de observar novas disciplinas da informação: privacidade, segurança, retenção, processamento, propriedade e integridade (The Economist, 2010).

2. Referencial Teórico

Diebold (2012) destaca que o termo *big data* tende a evoluir de apenas um fenômeno momentâneo para uma nova disciplina, pois a necessidade de lidar com grandes volumes de dados e, principalmente, extrair informação destes dados vem tornando-se um tema chave para a maioria das ciências.

A necessidade de analisar grandes volumes de dados aumentará nos próximos anos, visto que a quantidade de informações existente, chamado de universo digital, segundo estudo de Gantz (2012), crescerá em até 50 vezes entre os anos de 2010 e 2020 e totalizará um crescimento de até 300 vezes se considerado os anos de 2005 a 2020, partindo de 130 exabytes para aproximadamente 40 zettabytes.

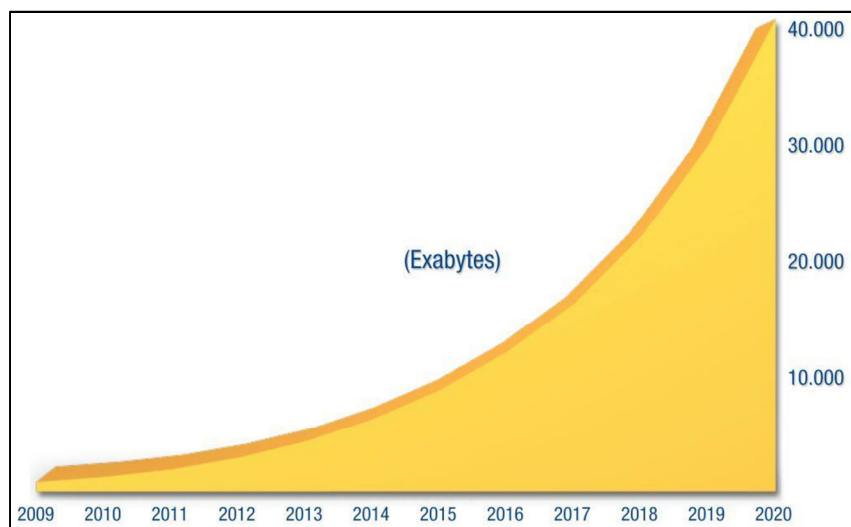


Figura 1 - Crescimento do volume mundial de dados

Fonte: Adaptado de Gantz (2012, p. 3)

Com o uso correto de tecnologias de *big data* e modelos de dados eficientes, aliados as novas fontes de dados, algumas empresas como a Progressive, atuante no ramo de seguros nos Estados Unidos, que como descrito por Davenport (2007) usa informações do escore de crédito do segurado e informações obtidas por rastreadores instalados nos veículos para ajudar a identificar se o segurado respeita as velocidades das vias e pratica condução defensiva, qualificando o condutor de forma a determinar com mais precisão a probabilidade que este se envolve em um acidente. Com o advento do *big data* é possível ampliar ainda mais esta análise lançando mão por exemplo de *tweets* (postagens dos usuários na rede social Twitter), posts no Facebook ou outras informações de fontes diversas espalhadas na internet, que indiquem um bom ou mal comportamento como condutor e consequente grau do risco do segurado.

A grande maioria destes novos dados – cuja análise constitui um diferencial competitivo – são do tipo não estruturado, ou seja, não estão organizados em tabelas e colunas nos molde de um banco de dados tradicional. Por exemplo, o Google processa diariamente aproximadamente 24.000 terabytes de dados e apenas uma ínfima porção destes dados está armazenada em bases de dados tradicionais (DAVENPORT, 2012). Hea (2013) completa que na maioria das empresas 80% dos dados também são de natureza não estruturada, como por exemplo, e-mails, documentos, entre outros.

O poder destas redes sociais merece um destaque especial, pois são uma poderosa fonte de dados a ser utilizada pelas empresas nos seus projetos de *big data*, como mencionado por Brynjolfsson (2012) que cita o uso de mensagens do Twitter que foram analisadas e conseguiram antever em duas semanas os boletins oficiais de propagação da cólera no Haiti após o terremoto em 2010. O autor menciona ainda o estudo realizado por Erik e Lynn Wu que usaram dados obtidos livremente em buscas na internet para prever a variação média do mercado imobiliário dos EUA (Estados Unidos da América), conseguindo obter estimativas mais precisas do que a oficial da Associação Americana dos Corretores de Imóveis, a qual utilizava um modelo complexo, porém alimentado por dados históricos que não conseguiam acompanhar a nova dinâmica do mercado.

No entanto, os exemplos acima trazem à discussão um outro aspecto importante a ser considerado: o lado ético e as possíveis implicações relacionadas a segurança da informação e privacidade no uso destes dados - como por exemplo das próprias redes sociais comentadas aqui – que apesar de serem gerados e disponibilizados pelos próprios usuários que, de certo modo optaram por torná-las públicas, certamente não o fizeram com o intuito de serem usadas por corporações, como no cálculo do seguro do carro.

As redes sociais vem crescendo e hoje possuem elevada significância no cotidiano das pessoas. No ano de 2011 constatou-se que 13% dos norte-americanos adultos usam o Twitter regularmente (SMITH, 2011). Outra evidência da crescente importância não apenas das redes sociais, mas do conteúdo que circula por estas redes, foi a decisão em 2010 da Biblioteca do Congresso Norte Americano em arquivar e disponibilizar todas as postagens dos usuários na rede social Twitter (LOC, 2013).

Esse é um fenômeno global e em franca expansão. O LinkedIn alegou ter em março de 2012 mais de 160 milhões de usuários ativos, crescendo a uma taxa de 2 novos membros por segundo, e este ainda é um número pequeno perto dos 900 milhões de usuários ativos do Facebook (AQUINO, 2012).

Tene (2012) alerta que apesar de existirem políticas de privacidades nas redes sociais, inclusive apresentadas para os usuários no momento de inscrição nos serviços como Facebook ou LinkedIn, estas muitas vezes servem apenas como isenção de responsabilidade para as empresas, ao invés de balizadores das garantias de privacidade para os consumidores. Como agravante as redes sociais deixam disponíveis uma série de

dados pessoais, os quais podem ser acessados por qualquer indivíduo ou corporação com acesso à rede, como demonstrado na Figura 2.

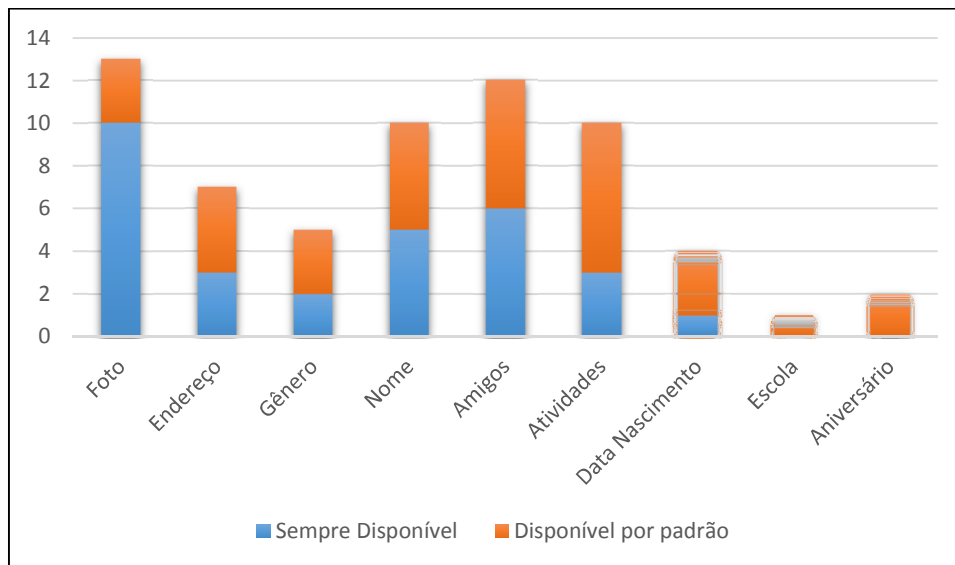


Figura 2 - Informações disponíveis em 13 redes sociais investigadas
Fonte: Adaptado de Krishnamurthy (2010)

Diante da natureza dinâmica da informação digital em um mundo altamente conectado, dar-se-á a necessidade de regulamentações globais que norteiem o uso da internet entre países. Por exemplo, um usuário do Brasil que usa o Facebook tem os seus dados armazenados em servidores localizados nos Estados Unidos e portanto sujeitos a uma série de legislações locais. Uma empresa europeia que utilize estes dados obedece a um outro grupo de normas que podem ser dissonantes das leis brasileiras e estadunidenses. Modelos de regulamentações como os dos oceanos podem servir como base de inspiração para a criação de normas globais para a internet. Além disso é necessário um órgão mundial responsável por este controle, como a Organização Mundial de Comércio (THE ECONOMIST 2010).

Kosinskia (2012) alerta para a grande quantidade de dados disponíveis online oriundos deste novo modelo de interação social e prevê que isto trará discussões no tangente à privacidade e propriedade dos dados, na medida em que torna possível para governos, empresas ou até mesmo outras pessoas inferir informações privadas e não divulgadas dos indivíduos, usando análises de dados.

Atualmente não existe um entendimento consolidado de que os dados gerados por um usuário, como informações postadas em redes sociais, sejam de propriedade do indivíduo. Isso cria abertura para as empresas usarem estes dados de forma ampla e irrestrita. Dentro deste contexto, o The Economist (2010) alerta para o fato de que as pessoas estão disponibilizando informações potencialmente confidenciais nas redes sociais e isso ocasionará uma tensão entre empresas que querem utilizar tais dados em benefícios próprio e os consumidores que querem proteger as suas informações pessoais.

Levará algum tempo para que as organizações entendam completamente o que é viável em relação a capturar, gerenciar, analisar e, acima de tudo, criar inteligência social, a fim de agir sobre o conteúdo gerado pelo usuário (LESLIE, 2012, p. 9).

3. Método

Esta pesquisa objetiva explorar o crescente interesse no *big data*, seu estudo como ferramenta de melhoria da competitividade de empresas e como sua adoção pode gerar implicações éticas no modelo de relacionamento entre corporações e consumidores.

Para tal realizou-se primeiramente uma pesquisa quantitativa, analisando artigos publicados em periódicos registrados no CAPES e contabilizando as citações recebidas por cada artigo e autor. Para a coleta e tratamento dos dados, adotou-se as etapas descritas na Figura 3.

Como a coleta de dados foi realizada ao fim de junho de 2013, o número de publicações referente a este ano foi estimado fazendo uma extrapolação simples do número de artigos publicados até junho. Dividiu-se o este número por 6 e multiplicou-se por 12.

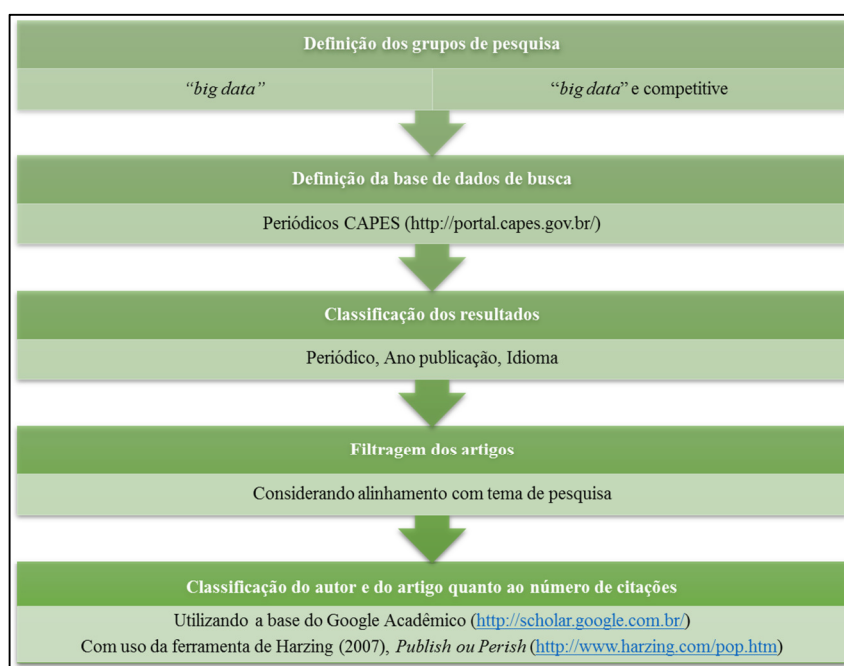


Figura 3 - Etapas da pesquisa bibliométrica.

Completa-se o estudo com uma segunda pesquisa de abordagem qualitativa e quantitativa, por meio de questionário estruturado respondido por usuários do Facebook. Elegeu-se usuários do Facebook, devido a sua maior representatividade entre os usuários brasileiros de redes sociais, com demonstrado por Volgarin (2012)

A coleta de dados ocorreu no período de setembro a novembro de 2013, com quase 50 respostas. O número de respostas representa um percentual ínfimo dos usuários do Facebook e com isso esta pesquisa demonstra apenas tendências de comportamento dos usuários, sem contudo constituir massa crítica válida para um referencial estatístico..

4. Resultados e Discussão

Este é um tema novo, conforme evidenciado na Figura 4, que demonstra um número limitado de artigos publicados sobre *big data* até o ano de 2010. Entretanto o crescimento acelerado no número de publicações nos últimos anos, denota um grande aumento de interesse no tema, que permanece ainda relativamente inexplorado.

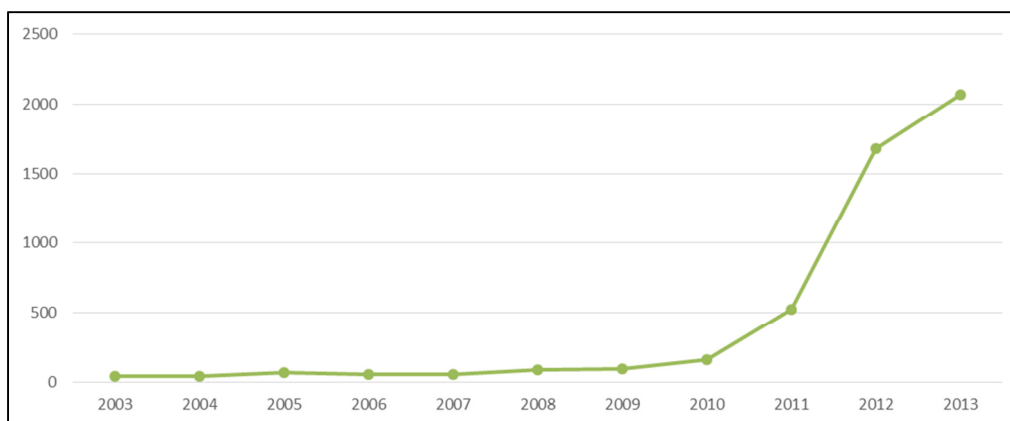


Figura 4 - Resultados por ano de publicação.

Dentre todos os artigos encontrados ao buscar apenas pelo termo *big data*, 99,26% foram escritos na língua inglesa, demonstrando a incipiência do tema no Brasil.

Além do crescente número de artigos, nota-se semelhante aumento no número de citações destes artigos em outras publicações dos mais diversos temas, corroborando com a afirmação de ampla abrangência do *big data* e do incremento de sua relevância como tema de estudo. A busca de referências limitou-se ao Google Acadêmico e objetivou mostrar a tendência de crescimento nas citações, refletida pela reta de tendência linear da **Figura 5**. Não obstante a quantidade de citações de 2012 e 2013 estarem comprometidas pela falta de tempo hábil entre a publicação do artigo e seu uso por outros pesquisadores, a tendência de crescimento permanece positiva.

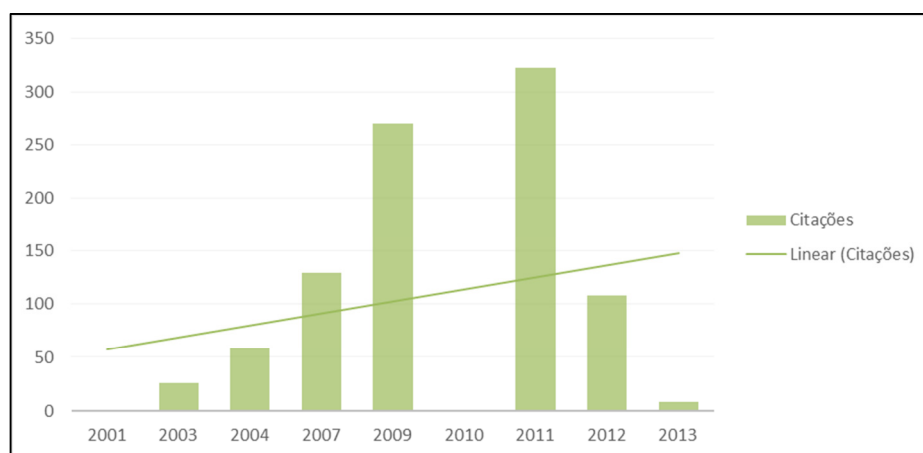


Figura 5 - Citações dos artigos.

Denotado o aumento no interesse sobre o *big data* como ferramenta de competitividade empresarial e tema de estudo acrescenta-se à pesquisa uma visão do uso dos dados de redes sociais e internet.

Dados extraídos de redes sociais como LinkedIn ou Twitter podem trazer metadados que incluam diversas informações adicionais sobre os usuários, como por exemplo a geolocalização dos usuários, como demonstrada em pesquisa representada na Figura 6 que correlaciona a localidade do usuário com o conteúdo postado por este no Twitter pra identificar aonde estão concentrados os usuários que postaram conteúdo com a hashtag (palavra-chave) #bigdata. Esses dados podem ser utilizados como mais um fator de decisão em uma campanha de marketing, permitindo uma abordagem mais

assertiva para as empresas, direcionando a propaganda para o mercado consumidor específico.



Figura 6 - Amostragem de local de origem de tweets com a hashtag #bigdata.

Utilizar as redes sociais como repositório de dados para análise torna-se mais simples a cada dia. Até mesmo pessoas sem experiência ou conhecimentos técnicos avançados podem utilizar ferramentas como o ScraperWiki (disponível em <https://scraperwiki.com/>) para fazer extração de dados de redes sociais como o Twitter e automaticamente realizar processos básicos de *text mining* nesses dados, ou seja, analisar de forma automatizada palavras e frases para extrair correlações e informações relevantes destas, gerando artefatos como a Figura 7, com as principais palavras relacionadas à postagens no Twitter com a hashtag de #bigdata. Na figura, palavras com as fontes maiores significam maior número de ocorrências.



Figura 7 - Análise de palavras relacionadas a big data em postagens no Twitter.

Não se discute aqui a relevância das redes sociais como forma de comunicação entre empresa e cliente, já amplamente utilizadas por várias instituições. Entretanto novas formas complementares de uso, como a análise de dados, abre novas possibilidades como ações de marketing direcionais e previsão comportamental.

Como detalhado na Figura 2, a configuração de segurança e privacidade das redes sociais expõe por padrão várias informações dos usuários. No caso do Facebook é

possível configurar algumas restrições básicas com relação a segurança e privacidade do usuário. Averiguou-se que maioria dos usuários (87%) alega ter realizado ao menos alterações básicas nestas políticas.

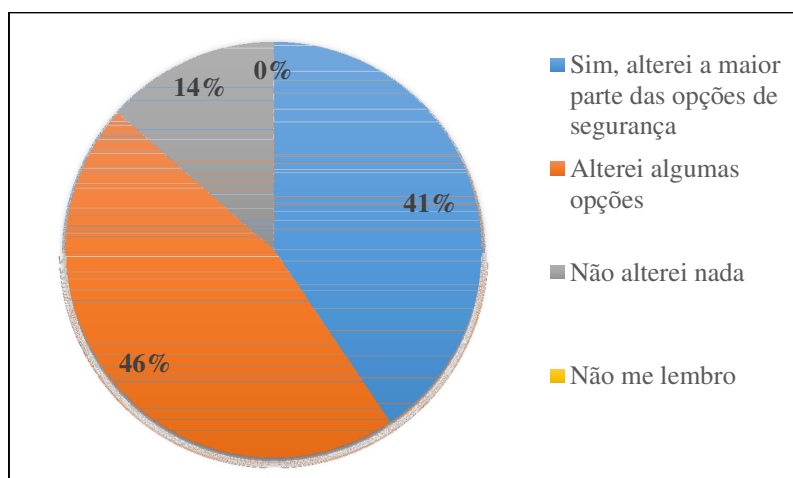


Figura 8 - Usuários e as configurações de segurança e privacidade do Facebook.

Entretanto, ao analisar os perfis dos usuários que responderam a pesquisa, verificando se estes haviam restringido o acesso de desconhecidos a duas informações pessoais (e-mail e postagens), notou-se, como detalhado na Tabela 1, que a maioria dos usuários deixaram expostas informações pessoais para todos os outros usuários da rede. Isso pode representar uma falta de conhecimento técnico dos usuários nas configurações de segurança do Facebook ou uma percepção por parte dos usuários que a exposição pública destas informações não constitui um risco.

Tabela 1 - Exposição dos dados dos usuários

Informações Pessoais	Disponível sem restrições
E-Mail	54%
Postagens	68%

No que tange as postagens no Facebook, 46% dos usuários alegaram que postam qualquer conteúdo que julgaram de interesse dos amigos conectados à rede social. Em contrapartida, apenas 32% afirmaram usar como critério para postagens a cautela quanto a exposição de informações pessoais.

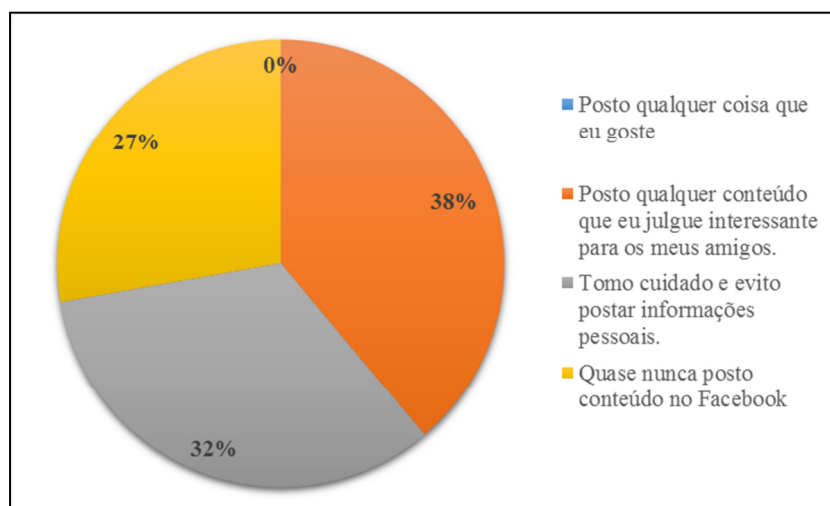


Figura 9 - Critérios ao postar conteúdos no Facebook

Fato relevante que, mesmo entre o grupo de usuários que alegou não ter como principal critério evitar postagens com informações pessoais, a maioria (54%) discorda do uso destes dados por empresas para propagandas ou quaisquer outros casos de uso.

Ao questionar sobre o uso de postagens do Facebook para outros fins, além da propaganda, descobriu-se que apenas 38% são tolerantes a esta prática, sendo que 62% discordam da utilização para outros fins das informações postadas.

Tabela 2 - Aceitação dos usuários no uso das postagens do Facebook

	Concorda	Discorda
Propaganda de produtos e serviços direcionadas a você com base nas suas publicações do Facebook	43%	54%
Uso das informações disponíveis no Facebook para, por exemplo, calcular o seguro do seu carro ou avaliar o seu crédito bancário	38%	62%

5. Considerações finais

Empresas estão nadando em um mar de dados, os quais são muito volumosos ou muito complexos para serem analisados da forma tradicional. E justamente a análise dessa gigantesca massa de dados está rapidamente tornando-se a nova fronteira da competição empresarial.

Dados como cookies ou postagens em redes sociais constitui fonte de informação disponível aos sistemas de *big data* e podem ser utilizados por empresas para que estas obtenham maior conhecimento do mercado alvo. Por sua vez, o consumidor poderá ser beneficiado ao receber ofertas melhor ajustadas às suas necessidades. Nota-se entretanto a desinformação dos usuários das redes sociais sobre os possíveis usos das informações disponibilizadas nestes sítios. Surge a questão ética no uso desses dados que precisa ser debatida com maior profundidade com o objetivo de garantir o respeito ao consumidor. É necessário que o indivíduo tenha maior controle sobre quem pode acessar suas informações e qual o uso que as empresas estão fazendo destes dados, porém o tema ainda carece de regulamentações que estabeleçam padrões de identificação e de segurança dos dados pessoais.

O *big data* se posiciona como um modelo evolutivo da análise de dados tradicionais, suprimindo novas demandas de análise rápida dos dados advindos de diversas fontes e em maior quantidade. Está claro o crescimento pelo interesse e estudo ao redor do tema embora este ainda seja maior na esfera empresarial do que acadêmica.

Referências

- AQUINO, J. Transforming social media data into predictive analytics. **Destination CRM**. 2012
- BROWN, B; CHUI, M; MANYIKA, J. Are you ready for the era of 'big data'?. **McKinsey Global Institute**, 2011
- BRYNJOLFSSON, E; MCAFEE, A. Big Data - A Revolução da Gestão. **Harvard Business Review**, 2012
- DAVENPORT, T; HARRIS, J. **Competição Analítica: Vencendo Através da Nova Ciência**. Campus, 2007.
- DAVENPORT, T; BARTH, P; BEAN, R. How 'big data' is different. **MIT Sloan Management Review**, 2012.
- DIEBOLD, F. **On the origin(s) and development of the term "Big Data"**. Penn Institute For Economic Research, 2012
- GANTZ, J; REINSEL, D. **The digital universe in 2020: Big Data, nigger digital shadows, and biggest growth in the far east**. IDC, 2012
- HEA, H; ZHAB, S; LI, L. Social media competitive analysis and text mining: A case study in the pizza industry. **International Journal of Information Management**, 2013
- KOSINSKIA, M; STILLWELLA, D; GRAEOELB, T. Private traits and attributes are predictable from digital records of human behavior. **PNAS**, 2012
- KRISHNAMURTHY, B; WILLS, C. E. On the leakage of personally identifiable information via online social networks. Proceedings of the 2nd ACM workshop on Online social networks. **ACM**. 2009
- LESLIE, A. Social Analytics & Intelligence: Converting Contextual to Actionable Insights. **Hypatia Research & Advisory**. 2012
- LOC. Library of Congress. Update on the Twitter Archive At the Library of Congress. Janeiro, 2013. Disponível em: <http://www.loc.govtoday/pr/2013/files/twitter_report_2013jan.pdf>. Acesso em 10 outubro de 2013
- NEW rules for *big data*. **The Economist**, 2010
- SMITH, A. 13% of online adults use Twitter. **Pew Research Center**. 2011
- TENE, O; POLONETSKY, J. Privacy in the age of big data: a time for big decisions. **Stanford Law Review Online**. 2012
- VOLGARIN, R; et al. A aplicação de redes sociais no contexto mercadológico. **VIII Workshop de pós-graduação e pesquisa do Centro Paula Souza**. 2013