Business Intelligence e Big Data: um exemplo prático de aplicação de Text Mining

Célia Satiko Ishikiriyama
csatiko@gmail.com
UFF

Diego Miro
d.miro1089@gmail.com
ENCE

Carlos Francisco Simões Gomes
cfsg1@bol.com.br
UFF

Resumo: No centro das discussões sobre vantagem competitiva e estratégias organizacionais, encontra-se o processo de Business Intelligence. Recentemente, outro conceito surgiu e tornou-se um fenômeno no meio acadêmico e profissional: Big Data. Esse artigo possui como objetivo principal exemplificar a técnica de Text Mining, através de sua aplicação em títulos, palavras-chave e resumo de artigos acadêmicos centradas no tema do objeto de pesquisa: Business Intelligence e Big Data.

Palavras Chave: BI - Big Data - Text Mining - -

1. INTRODUÇÃO

Business Intelligence refere-se ao ato de proporcionar aos negócios o apoio necessário para a tomada de decisão, através do uso de um conjunto de técnicas e ferramentas (MIKROYANNIDIS & THEODOULIDIS, 2010). Para Bologa e Bologa (2011), é possível identificar três principais grupos de atividades para alcançar a inteligência no negócio:

- Acessar, integrar e armazenar dados de diferentes fontes;
- Analisar e transformar dados em informação;
- Apresentar a informação.

Agilidade e capacidade de resposta à mudança oferecem aos negócios a capacidade de competir em uma economia global em mudanças e conhecer o ambiente do negócio tornou-se a chave para manter os negócios rentáveis e competitivos (THOMPSON & VAN DER WALT, 2010).

O BI não só apoia o processo de tomada de decisão, como também permite que as organizações tenham melhores percepções em relação às suas operações através da aplicação de técnicas de análise de dados às suas informações (THOMPSON & VAN DER WALT, 2010). Para reforçar essa ideia, Bologa e Bologa (2011) afirmaram que o uso do BI também torna possível para uma organização incluir comportamento inteligente em suas funções básicas.

A implantação do BI parece simples, mas pode ser complicado e desafiador, dependendo da complexidade do negócio, a quantidade de diferentes sistemas operacionais em uso e da qualidade dos dados disponíveis. A quantidade de informação está crescendo, e, com isso, maiores dificuldades e desafios são colocados ao processo de tomada de decisão (MIKROYANNIDIS & THEODOULIDIS, 2010). A integração do BI com *softwares* pode fornecer soluções para esses problemas (BOLOGA & BOLOGA, 2011).

Dado o cenário, a Tecnologia da Informação (TI) representa um papel fundamental para o sucesso do BI de diversas formas. A oferta de soluções de BI no mercado pode ilustrar a complexidade em lidar com tantos dados provenientes de diversas origens. Essas soluções são consideradas as ferramentas principais para analisar e monitorar o desempenho nas organizações (RUSANEANU, 2013). Os sistemas de informação podem fazer ou facilitar, de forma significativa, a emergência por inovações (SANNER et al., 2014).

Usar um Sistema de BI e aplicações relacionadas pode fornecer apoio à decisão, inteligência competitiva e inteligência operacional (GOMES & RIBEIRO, 2014; SKYRIUS et al., 2013). A implantação de uma solução de BI, contudo, deve levar em consideração planejamento anterior. Um bom plano de arquitetura de BI é fundamental para o sucesso da implantação (ONG et al., 2011). Além disso, o uso da TI tem ajudado as organizações a explorar relacionamentos com clientes como nunca (PHAN & VOGEL, 2010).

Dentro desse contexto, surge o fenômeno *Big Data*, que ainda não possui definição sólida e encontra-se em centro de discussões, tanto no meio acadêmico quanto no mundo dos negócios. Para Tien (2013), *Big Data* é um termo aplicado aos conjuntos de dados, cujo tamanho vai além da habilidade das ferramentas disponíveis para possibilitar o acesso aos dados, análise e/ou aplicação em um tempo razoável. *Big Data* é conhecida por agregar valor ao negócio de várias formas: definição de preço, plano de *marketing*, medicina moderna, dentre outros (TIEN, 2013).

Big Data refere-se a um grande volume de dados, complexo, crescentes conjuntos de dados com fontes múltiplas e autônomas. Com o rápido desenvolvimento de redes, armazenagem de dados e capacidade de coleta de dados, *Big Data* está se expandindo rapidamente em todos os domínios das ciências e engenharia, incluindo ciências físicas, biológicas e biomédicas (WU et al., 2014).

Para Teixeira e Alonso (2014), as organizações precisam ter planejamento adequado e elaborar estratégias que resultem no cumprimento de metas e objetivos estabelecidos. *Big Data* pode também auxiliar as organizações a elaborar planejamento estratégico e alcançar vantagem competitiva.

Para Chen e Zhang (2014), Big Data mudou o modo como fazemos negócios, gestão e pesquisas. Ciências intensivas de dados, especialmente em computação intensiva de dados, estão chegando a um mundo que deseja prover as ferramentas que necessitamos para lidar com problemas de Big Data (CHEN & ZHANG, 2014).

Business Intelligence e *Big Data* possuem muitos pontos em comum, já que ambos buscam o autoconhecimento, vantagem competitiva, melhoria de desempenho e *insights* para melhor posicionamento no mercado, além de possuir a tecnologia como pilar fundamental.

O objetivo deste estudo é exemplificar uma aplicação da técnica de *Text Mining*, para identificar conceitos relacionados ao *Business Intelligence* e *Big Data* de forma integrada, para viabilizar a construção da definição, dos processos e das tecnologias integrantes dessa nova forma de buscar autoconhecimento e diferenciais no mercado competitivo.

2. METODOLOGIA

A pesquisa do presente estudo pode ser classificada como pesquisa exploratória. Para Gil (2010), a pesquisa exploratória tem como objetivo proporcionar maior familiaridade com o objeto de estudo ou problema.

Este estudo dividiu-se em duas etapas principais: pesquisa na literatura e análise exploratória dos dados, através do uso da técnica Text Mining. Para Radovanović & Ivanović, o campo do *text mining* objetiva a extração de informação útil de dados não estruturados, através da identificação e exploração de padrões interessantes.

Text mining refere-se à extração de informação de dados não estruturados e um "saco de palavras", que é uma representação textual em forma de vetor espacial (RADOVANOVIĆ & IVANOVIĆ, 2008), com o objetivo de descobrir novos conhecimentos (STAVRIANOU et al., 2007).

Na primeira etapa, foram realizadas duas pesquisas em abril do presente ano, em duas bases de busca: Scopus e Web of Science. Os critérios utilizados para a pesquisa em ambas as bases foram os seguintes:

- Palavras pesquisadas: "*Business Intelligence* AND *Big Data*"
- Tipo de documento: artigos
- Demais campos: sem restrições

As pesquisas resultaram em 96 artigos, sendo 70 provenientes do Scopus e 26 artigos, da Web of Science. Dos 96 artigos, seis foram encontrados em ambas as bases, o que resultou em 90 artigos para a aplicação da técnica de análise. O perfil destes artigos foi traçado quanto ao ano de publicação, países dos autores e área de pesquisa.

Os artigos decorrentes da pesquisa também foram utilizados na segunda etapa: a análise exploratória. Para esta etapa, foram criadas três bases de dados. A primeira contendo

os títulos dos artigos; a segunda, as palavras-chave e a terceira, os resumos. A técnica *Text Mining* foi aplicada, em cada uma das bases, a fim de descobrir conceitos correlatos.

A aplicação da técnica mencionada ocorreu através do software R-project, que é um ambiente de software para estatística computacional e elaboração de gráficos. A versão do programa utilizada foi a 3.1.2, em conjunto com dois pacotes do mesmo programa: "tm" para o *text mining* e "*wordcloud*" para a geração de nuvens de palavras.

3. ANALISE DOS RESULTADOS

Os resultados que serão apresentados nesta seção delimitam-se à pesquisa explicitada na seção anterior. Os resultados serão expostos em duas subseções: perfil dos artigos e análise exploratória que, por sua vez, será subdivida em três partes: análise dos títulos, análise das palavras-chave e análise do resumo.

3.1. PERFIL DOS ARTIGOS

Como se observa na Figura 1, a maioria dos artigos encontrados foi publicada de 2012 em diante (79%), quando começou a apresentar comportamento crescente, atingindo o seu pico em 2014. Como as pesquisas para este estudo foram realizadas em abril de 2015, o comportamento da curva pode sugerir que a quantidade de publicações em 2015 ultrapasse o resultado obtido em 2014.

Figura 1: Quantidade de artigos por ano de publicação.

A Figura 2 ilustra a quantidade de autores por país. O país com o maior número de autores envolvidos foi Estados Unidos, com 30 autores, representando 33% do total. Em segundo lugar, aparece China, com 9% do total, representando 8 autores.

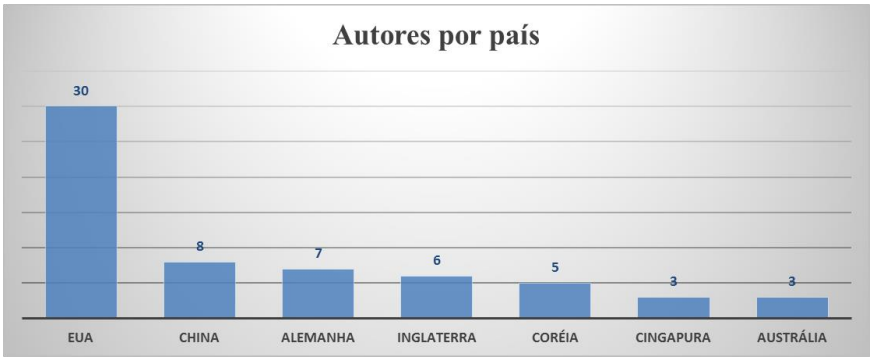

Figura 2: Autores por país.

Segundo área de pesquisa, os resultados podem ser observados na Tabela 1. A área de pesquisa mais recorrente foi Ciência da Computação (39%), seguida de Engenharia (14%) e Negócios, Gestão e Contabilidade (14%). As áreas de pesquisa mais citadas indicam as áreas com aplicação mais pesquisadas.

Tabela 1: Artigos por área de pesquisa.

ÁREA PESQUISA	*Artigos*	*%*
Ciência da Computação	60	39%
Engenharia	21	14%
Negócios, Gestão e Contabilidade	21	14%
Ciências Sociais	10	6%
Ciências de Decisão	8	5%
Matemática	5	3%
Ciências Biológicas	5	3%
Ciências de Materiais	4	3%
Medicina	4	3%
Arte e Humanidades	4	3%
Psicologia	3	2%
Biologias	2	1%
Economia e Finanças	2	1%
Engenharia Química	2	1%
Física e Astronomia	1	1%
Química	1	1%
Ciências Ambientais	1	1%
Ciências Planetárias e da Terra	1	1%

3.2. ANÁLISE EXPLORATÓRIA

A primeira etapa da análise exploratória usando text mining foi realizada com os dados não estruturados dos 90 artigos selecionados na etapa da pesquisa. Os textos foram organizados em três bases de dados separadas em formato csv (*comma-separated values*) e posteriormente importadas para o R.

A fim de manter a coerência na análise de palavras, "*big*" seguido imediatamente por "*data*" foi interpretado como uma única palavra "*big data*". A mesma lógica foi aplicada para "*business*" e "*intelligence*". Assim como as palavras que aparecem tanto no singular como no plural, as quais tenham o mesmo sentido no texto.

Após todos os tratamentos nos dados, o passo seguinte foi elaborar uma tabela de frequência das palavras. Em seguida, gerar a nuvem de palavras em dois níveis, sendo o primeiro com todas as palavras e o segundo excluindo as palavras da pesquisa: *Business Intelligence* e *Big Data*, uma vez que se espera que estas apareçam em maior frequência. A nuvem de palavras é uma forma gráfica que permite a percepção imediata das palavras mais frequentes pelo tamanho e esquema de cores. Por fim foi analisada a semântica dos resultados obtidos no passo anterior.

Esse processo foi realizado para as três partes dos artigos analisadas (título, palavras-chave e resumo).

3.2.1. ANÁLISE DOS TÍTULOS

Nesta subseção, serão expostos os resultados referentes à análise da base de títulos dos artigos selecionados.

A Figura 3 ilustra as nuvens de palavras geradas pelos títulos.

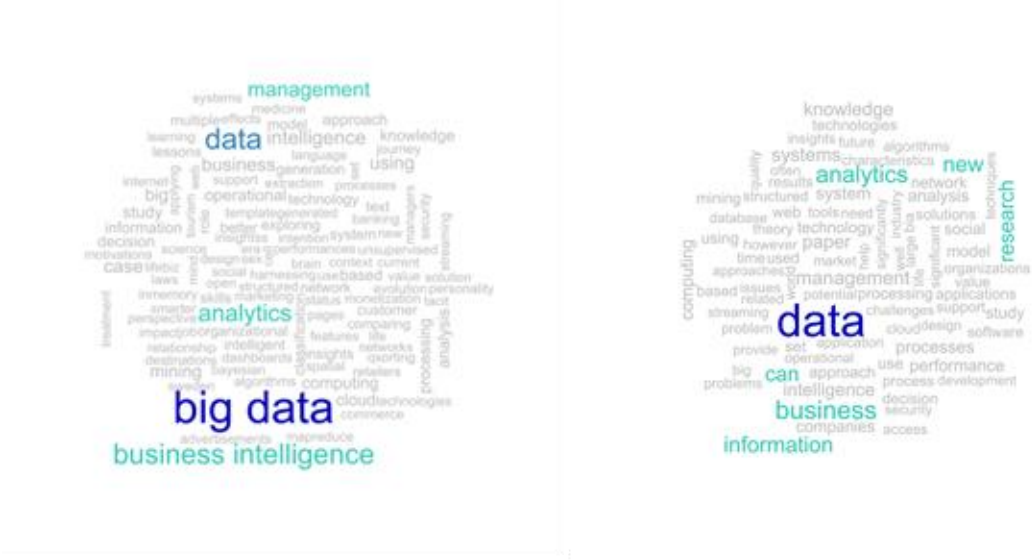

Figura 3: Nuvens de palavras dos títulos em dois níveis

As palavras mais frequentes dos títulos, com exceção de "*big data*" e "*business intelligence*" foram: dados (3%), analítico (2%), inteligência (1%) e gestão (1%). Na Tabela 2, estão descritas as palavras com ocorrência maior ou igual a quatro. Estas palavras representam 28% do total de 623 palavras encontradas.

Tabela 2: Tabela de frequência de palavras dos títulos

Palavra	*Frequência*	*%*
Big data	32	5%
Data	17	3%
Business intelligence	16	3%
Analytics	12	2%
Management	9	1%
Intelligence	8	1%
Case	7	1%
Business	7	1%
Using	7	1%
Big	6	1%
Mining	6	1%
Analysis	5	1%
Cloud	5	1%
Operational	5	1%
Computing	5	1%
Study	5	1%
Decision	4	1%
Processing	4	1%
Knowledge	4	1%
Approach	4	1%
Information	4	1%
Based	4	1%

3.2.2. ANÁLISE DAS PALAVRAS-CHAVE

Nesta subseção, serão apresentados os resultados referentes à análise da base de palavras-chave dos artigos selecionados na Tabela 3.

A Figura 4 ilustra as nuvens de palavras geradas pelas palavras-chave.

As palavras mais frequentes dos títulos, com exceção de "*big data*" e "*business intelligence*" foram: dado (3%), analítico (3%), negócio (3%), gestão (3%), conhecimento (2%) e informação (2%). Na Tabela 3, estão descritas as palavras com ocorrência maior ou igual a quatro. Estas palavras representam 34% do total de 517 palavras encontradas.

Figura 4: Nuvens de palavras das palavras-chave em dois níveis

Tabela 3: Tabela de frequência de palavras das palavras-chave

Palavra	*Frequência*	*%*
Data	15	3%
Analytics	14	3%
Management	13	3%
Business	13	3%
Knowledge	11	2%
Big data	10	2%
Information	9	2%
Computing	7	1%
Business intelligence	7	1%
Analysis	6	1%
Systems	6	1%
Intelligence	6	1%
Support	5	1%
System	5	1%
Value	4	1%
Semantic	4	1%
Processing	4	1%
Customer	4	1%
Social	4	1%
Decision	4	1%
Technology	4	1%
Networks	4	1%

Theory	4	1%
Performance	4	1%
Internet	4	1%
Web	4	1%

3.2.3. ANÁLISE DOS RESUMOS

Nesta subseção, serão expostos os resultados referentes à análise da base dos resumos os artigos selecionados.

A Figura 5 ilustra as nuvens de palavras geradas pelos resumos e a Tabela 4 descreve as frequências das palavras.

Figura 5: Nuvens de palavras dos resumos em dois níveis

Tabela 4: Tabela de frequência de palavras dos resumos

Palavra	*Frequência*	*%*
Data	224	2%
Big data	139	1%
Business	93	1%
Analytics	80	1%
New	71	1%
Information	70	1%
Business intelligence	68	1%
Research	64	1%
Can	63	1%
Management	53	1%
Systems	49	0%
Paper	49	0%
Knowledge	48	0%
Intelligence	45	0%
Computing	40	0%
System	39	0%
Processes	38	0%
Performance	36	0%
Analysis	36	0%
Companies	32	0%

Study	31	0%
Model	30	0%

As palavras mais frequentes dos resumos, com exceção de *big*, *data*, *big data* e *business intelligence* foram: analítico, novo, informação, pesquisa, gestão e sistemas. As palavras com frequência igual ou superior a trinta representam 14% do total de palavras analisadas (10.019).

4. CONCLUSÃO E CONSIDERAÇÕES FINAIS

A partir das análises expostas na seção anterior, percebe-se o interesse crescente sobre o objeto de pesquisa. As áreas de pesquisa com mais aplicação são ciência da computação, engenharia e negócios. No entanto, há aplicação nas mais diversas áreas, passando pelas áreas humanas e estendendo-se até a medicina e pesquisas da biologia.

Como pode ser observado na Tabela 5, o resumo comparativo entre as 12 palavras mais encontradas em cada subseção da seção 3 demonstra a frequência do conceito "*Big Data*" é superior que a de "*Business Intelligence*", o que leva indica que o termo *Big Data* está em plena ascensão e representa um papel de maior expressividade nos artigos.

Tabela 5: Tabela resumo de frequências das palavras

Título	*Freq.*	*%*	*Palavra-chave*	*Freq.*	*%*	*Resumo*	*Freq.*	*%*
Big data	32	5%	*Data*	15	3%	*Data*	224	2%
Data	17	3%	*Analytics*	14	3%	*Big data*	139	1%
Business intelligence	16	3%	*Management*	13	3%	*Business*	93	1%
Analytics	12	2%	*Business*	13	3%	*Analytics*	80	1%
Management	9	1%	*Knowledge*	11	2%	*New*	71	1%
Intelligence	8	1%	*Big data*	10	2%	*Information*	70	1%
Case	7	1%	*Information*	9	2%	*Business intelligence*	68	1%
Business	7	1%	*Computing*	7	1%	*Research*	64	1%
Using	7	1%	*Business intelligence*	7	1%	*Can*	63	1%
Big	6	1%	*Analysis*	6	1%	*Management*	53	1%
Mining	6	1%	*Systems*	6	1%	*Systems*	49	0%
Analysis	5	1%	*Intelligence*	6	1%	*Paper*	49	0%

Ao analisar as 12 palavras mais frequentes de forma conjunta e excluindo todas as palavras referentes ao objeto de pesquisa ("*big data*", "*big*", "*data*", "*business Intelligence*", "*business*" e "*intelligence*") e também palavras "*can*", "*study*" e "*paper*", foram encontradas as palavras mais relacionadas ao tema. A Tabela 6 ilustra a frequência destas palavras.

Tabela 6: Tabela resumo de frequências das palavras

Palavras Top 12	*Freq.*	*%*
Analytics	106	1%
Systems	104	1%
Information	83	1%
Management	75	1%
New	73	1%
Research	65	1%
Knowledge	63	1%
Computing	52	0%
Analysis	47	0%
Performance	41	0%

Processes	41	0%
Technology	36	0%

Das 12 palavras da Tabela 6, quatro estão associadas à tecnologia: sistemas, informação, computação e a própria tecnologia. Cinco palavras estão associadas a processos de gestão e análise (analítico, análise, gestão, processos e pesquisa), um adjetivo (novo) e, finalmente, três palavras são substantivos relacionados ao objeto de pesquisa: conhecimento, e desempenho.

5. REFERÊNCIAS

BOLOGA, A & BOLOGA, R. Business Intelligence using Software Agents. Database Systems Journal, v. 2, 2011, pp. 31-42.

Chen, C.L.P.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, v. 275, 2014, pp. 314-347.

GIL, A. C. Como elaborar projetos de pesquisa. São Paulo: Atlas, 2010.

GOMES, C. F. C. & RIBEIRO, P. C. C. Gestão da Cadeia de Suprimentos integrada à Tecnologia da Informação. Rio de janeiro: Editora SENAC Rio de Janeiro, 2013.

MIKROYANNIDIS, A. & THEODOULIDIS, B. Ontology management and evolution for business intelligence. International Journal of Information Management, v. 30, 2000, pp. 559–566.

Ong, I.; Siew, P.; Wong, S. A Five-Layered Business Intelligence Architecture. IBIMA Publishing, v. 2011, 2011.

PHAN, D. D. & VOGEL, D. R. A model of customer relationship management and business intelligence systems for catalogue and online retailers. Information & Management, v. 47, 2010, pp. 69-77

R CORE TEAM (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

RADOVANOVIĆ, M. & IVANOVIĆ M. Text Mining: Approaches and Applications. Novi Sad J. Math, v. 38, 2008, pp. 227-234.

RUSANEANU, A. Comparative Analysis of the Main Business Intelligence Solutions. Informatica Economica, v. 17, 2013, pp. 148-156.

SANNER, T.A.; MANDA, T.D.; NIELSEN P. Grafting: Balancing Control and Cultivation in Information Infrastructure Innovation. Journal of the Association for Information Systems, v. 15, 2014, pp. 220-243.

SKYRIUS, R.; KAZAKEVIČIENĖ, G.; BUJAUSKAS, V. From Management Information Systems to Business Intelligence: The Development of Management Information Needs. International Journal of Artificial Intelligence and Interactive Multimedia, v.2, 2013, pp. 31-37.

STAVRIANOU, A.; ANDRITSOS, P.; NICOLOYANNIS, N. Overview and Semantic Issues of Text Mining. SIGMOD Record, v. 36, 2007, pp. 23-34.

Teixeira , C. A. C. & ALONSO, V. L. C. A Importância do Planejamento Estratégico para as Pequenas Empresas, SEGET, Rio de Janeiro, 2014.

THOMPSON, W. J. J. & VAN DER WALT, J. S. Business intelligence in the cloud. SA Journal of Information Management, v.2, 2010.

TIEN, J.M. Big Data: Unleashing Information. Journal of Systems Science and Systems Engineering, v.22, 2013, pp. 127-151.

Wu, X.; Zhu, X.; Wu, G.Q.; Ding, W. Data Mining with Big Data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, v. 26, 2014.