CENTRO PAULA SOUZA

FACULDADE DE TECNOLOGIA DE AMERICANA

Curso de Segurança da Informação

ANÁLISE DE SENTIMENTOS EM DADOS DESESTRUTURADOS

ALEXANDRA SCHRANKO DE OLIVEIRA

Americana, SP

2014

CENTRO PAULA SOUZA

FACULDADE DE TECNOLOGIA DE AMERICANA

Curso de Segurança da Informação

ALEXANDRA SCHRANKO DE OLIVEIRA
ale.schranko@gmail.com

ANÁLISE DE SENTIMENTOS EM DADOS DESESTRUTURADOS

Trabalho de Conclusão de Curso desenvolvido em cumprimento à exigência curricular do Curso de Segurança da Informação, sob a orientação da Prof.(o) Especialista César Augusto Crócomo.

Área de concentração: Segurança da informação.

Americana, SP

FICHA CATALOGRÁFICA – Biblioteca Fatec Americana - CEETEPS

Dados Internacionais de Catalogação-na-fonte

Oliveira, Alexandra Schranko de

O45a Análise de sentimentos em dados desestruturados / Alexandra Schranko de Oliveira. – Americana: 2014.

42f.

Monografia (Graduação em Tecnologia em Segurança da Informação). - - Faculdade de Tecnologia de Americana – Centro Estadual de Educação Tecnológica Paula Souza.

Orientador: Prof. Esp. César Augusto Crócomo

1. Psicologia social 2. Sociologia da comunicação I. Crócomo, César Augusto II. Centro Estadual de Educação Tecnológica Paula Souza – Faculdade de Tecnologia de Americana.

CDU: 159.922.4

316.77

ALEXANDRA SCHRANKO DE OLIVEIRA

ANÁLISE DE SENTIMENTOS EM DADOS DESESTRUTURADOS

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Segurança da Informação pelo CEETEPS/Faculdade de Tecnologia – FATEC/ Americana.
Área de concentração: Segurança da Informação

Americana, 05 de Dezembro de 2014.

Banca Examinadora:

César Augusto Crócomo (Presidente)
Professor Especialista
Fatec Americana

Edson Roberto Gaseta (Membro)
Professor Especialista
Fatec Americana

Luiz Carlos Degrande Júnior (Membro)
Professor Graduado
Fatec Americana

RESUMO

Vivemos na Era do *Big Data*. A exploração do crescente volume de informação presente na *Web* é enxergada com interesse, tanto pelo mundo dos negócios quanto pelo meio acadêmico. Sabe-se que grande parte desde volume é derivado de Mídias Sociais – onde as pessoas dão vazão a suas opiniões, frustrações e conversam sobre assuntos cotidianos. O material textual produzido através destas interações cria a oportunidade de identificar como um grande número de pessoas se sente em relação a um determinado assunto. A automatização deste processo é chamada de Análise de Sentimentos. Através da análise de uma porção de texto, é possível classificar o sentimento representado como positivo, negativo ou neutro. A utilização deste tipo de tecnologia em Mídias Sociais permite a identificação de tendências de mercado, padrões de consumo e *feedback* instantâneo sobre um determinado produto no mercado. Este trabalho tem como objetivo apresentar a Análise de Sentimentos, através de sua abordagem lexical, e prover uma visão geral de seu funcionamento.

Palavras-chave:
Big Data, Mídias Sociais, Análise de Sentimentos.

ABSTRACT

We live in the age of Big Data. The exploration of the growing volume of information on the Web is attracting great interest from both business and academic fields. Is a well-known fact that a large part of the amount of data available today is generated by Social Media – people ventilate their opinions, frustrations and talk about everyday activities. The textual material produced by these interactions is creating an opportunity to identify how a great amount of people feel about a certain subject. The automatization of this process is called Sentiment Analysis. Through the analysis of a portion of text it is possible to classify it as positive, negative or neutral. This kind of technology allows the identification of market trends, consume patterns and instant feedback of a product in the market. This paper aims to introduce the lexicon-based approach of Sentiment Analysis, providing an overview of how it works.

Palavras-chave:
Big Data, Social Media, Sentiment Analysis.

LISTA DE ABREVIATURAS E SIGLAS

GPS - *Global Positioning System* (Sistema de Posicionamento Global)

IBM - *International Business Machines*

IDC - *International Data Corporation*

NPL - *Natural Language Processing* (Processamento de Linguagem Natural)

PII - *Personally Identifiable Information* (Informações Pessoalmente Identificáveis)

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A Análise de Sentimentos é uma técnica que ganhou visibilidade nos últimos anos devido ao crescimento das mídias sociais. O conceito de identificar e classificar a opinião de um grupo de usuários em relação a um determinado tópico tem despertado o interesse de pesquisadores e se mostrado valioso para a pesquisa de mercado.

Este trabalho é organizado de forma a contextualizar o leitor com o ambiente altamente interativo que é a *Web* atualmente e explicar as motivações e funcionamento básico da Análise de Sentimentos neste contexto.

No Capítulo 2 analisamos o conceito de *Big Data*, com alguns exemplos de sua dimensão e características relacionadas. Tamanho crescimento dos dados presentes na *Web* – tanto em volume quando em variedade – foi considerada uma oportunidade para a produção de conhecimento em diversas áreas.

O Capítulo 3 aborda a Análise de Dados para esta finalidade, assim como o crescimento das Mídias Sociais e seu impacto na produção de material desestruturado, que cria um novo desafio para as ferramentas comuns de análise de dados.

A exploração de conteúdo textual é abordada no Capítulo 4, que apresenta o conceito de Analise de Sentimentos em sua abordagem Lexical – utilizada para a identificação e classificação de opiniões de grupos de usuários como negativa, positiva ou neutra.

No capítulo 5, um breve estudo de caso demonstra a aplicação prática desta tecnologia, identificando menções positivas e negativas sobre um determinado produto por usuários de diversas ferramentas na Internet.

Por fim, o último capítulo considera alguns dos desafios que certamente serão enfrentados por essa tecnologia nos próximos anos.

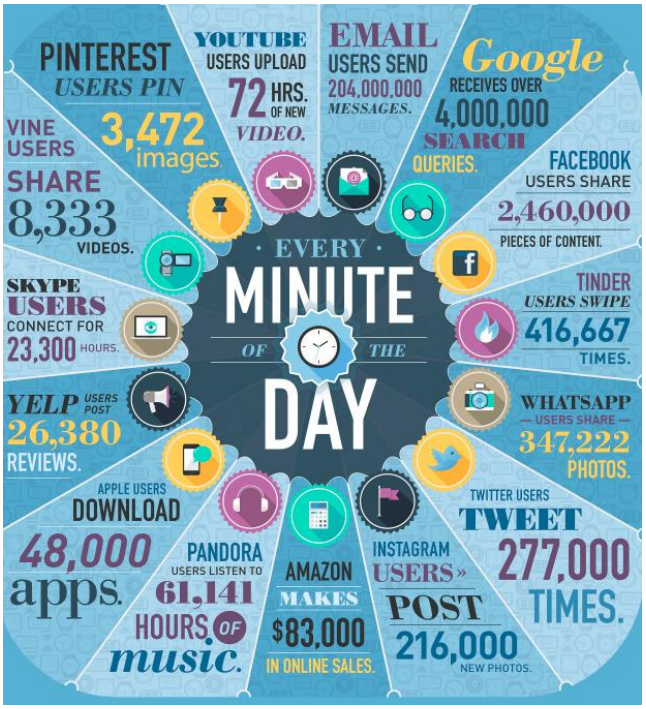
2. BIG DATA

Segundo a IBM (*International Business Machines*), 90% dos dados existentes na atualidade foram criados nos últimos dois anos. Diariamente, cerca de 2.5 quintilhões de *bytes* são gerados por diversas fontes: vídeos, posts em redes sociais, transações comerciais e etc. (IBM, 2014).

De acordo com o *The Economist (*2010) somente o varejista Wal-Mart manipula um volume de aproximadamente 2.5 *petabytes* por hora, oriundos de mais de 1 milhão de transações efetuadas por seus clientes.[1]

Os usuários Instagram[2] por sua vez publicam 216 mil fotos a cada minuto transcorrido, assim como diversos outros usuários de outros serviços, como exemplificado na Figura 1.

Figura 1: Dados produzidos por minuto



Fonte: *Data Never Sleeps* 2.0 – 2014

[1] Essa quantidade de dados é suficiente para ocupar 600 mil DVDs. Fonte: BAYLOR, 2013.

[2] O Instagram é uma Mídia Social focada no compartilhamento de fotos e vídeos. Fonte: INSTAGRAM, 2014.

Big Data refere-se a um conjunto de dados cujas proporções são grandes demais para serem processados por sistemas comuns. Tal definição é intencionalmente imprecisa, pois deixa em aberto a questão do quão grande um conjunto de dados deve ser para se encaixar neste contexto. Assim sendo, também não é abordada a premissa de que *Big Data* deve ser maior do que uma quantidade definida de dados. Assume-se que, com o avanço tecnológico ao longo dos anos, o tamanho de um conjunto de dados qualificado como *Big Data* tende a ser cada vez maior (McKinsey Global Institute, 2011).

Como Boyd e Crawford (2011) observam, a quantidade de dados existentes é indubitavelmente grande, mas esta não é a característica mais relevante deste cenário. O *Big Data* é notável não devido sua dimensão, mas devido ao relacionamento entre os dados. Devido ao esforço de "colecionar" explorar as informações, o *Big Data* é fundamentalmente interconectado. Seu valor surge dos padrões que são encontrados a partir de pequenas porções de informação – sobre um indivíduo, sobre um indivíduo em relação a muitos outros, sobre um grupo de pessoas ou, simplesmente sobre a estrutura da informação em si.

Em 2001, no artigo amplamente citado escrito por Doug Laney (Laney, 2001), são estudadas algumas características do *Big Data*. Estas características mencionadas acabaram por se tornar um padrão no meio acadêmico para o estudo deste tema. São 3 características – ou, como são comumente chamadas, "3Vs" – Volume, Velocidade e Variedade, como veremos a seguir.

Nos últimos anos, o volume de informação coletado e armazenado pela indústria aumentou consideravelmente. Esta tendência é guiada pela redução de custos das soluções de *storage* juntamente com a crescente capacidade de analisar grandes porções de dados. As informações possuem diversas fontes: transações *online, email,* vídeo, imagens, *logs*, interações em redes sociais, bem como recolhidas de sensores acoplados em infra-estrutura, como redes de comunicação, redes de energia, sistemas de posicionamento global (GPS) e etc. (Friedman, Hastie e Tibshirani, 2009)

Este volume tem crescido tão rapidamente que novos termos foram inventados para descrever tais quantias:

Tabela 1: Medida de Dados

Uma quantidade de 1024...	É igual a um...	Comparação
Kilobytes	*Megabyte*	Um CD de músicas comum comporta 600 *Megabytes*.
Megabytes	*Gigabyte*	Um *Gigabyte* pode armazenar o equivalente a quantidade de texto existente em uma fileira de livros com 9 metros de comprimento.
Gigabytes	*Terabyte*	Dez *Terabytes* podem armazenar o equivalente a quantidade de texto presente na Biblioteca do Congresso Americano [3]
Terabytes	*Petabyte*	Um *Petabyte* pode armazenar o conteúdo em texto de cerca de 20 milhões de armários de arquivo com quatro gavetas.
Petabytes	*Exabyte*	Cinco *Exabytes* equivalem a quantidade total de todas as palavras ditas pela humanidade, se convertidas para o formato digital.
Exabytes	*Zettabyte*	Para se fazer o *download* de um arquivo de 1 *Zettabyte* - utilizando uma conexão de alta velocidade - seria necessário cerca de 11 bilhões de anos.
Zettabytes	*Yottabyte*	O conteúdo total da Internet ocupa cerca de 1 Yottabyte.

Fonte: Franks (2012)

Na Tabela 1, algumas unidades de medida utilizadas para volume de dados, assim como alguns exemplos que ajudam a visualizar suas dimensões mais concretamente.

Velocidade refere-se a rapidez com que os dados são criados e analisados. No passado, quando o processamento em lotes era comumente utilizado, era natural receber atualizações todas as noites ou até mesmo uma vez a cada semana. Os computadores e servidores demandavam grande quantidade de tempo para processar estas atualizações. Atualmente, na era do *Big Data*, os dados são criados em praticamente tempo real. Os aparelhos e máquinas conectadas na Internet são capazes de disponibilizar suas informações no momento em que estas são criadas. (Mark van Rijmenam, 2013)

De acordo com Hurwitz, a velocidade na analise de informações maximiza o valor do *Big Data*, criando oportunidades impossíveis de serem aproveitadas há

[3] A Biblioteca do Congresso Americano é uma das maiores do mundo. Ela ocupa um espaço de 4 prédios e possui cerca de 30 milhões de unidades de material escrito. Fonte: OPISHPOSH, 2014.

alguns anos atrás, como por exemplo, o *feedback* instantâneo de um produto no mercado, através do monitoramento de mídias sociais. (Hurwitz et al, 2013)

Como também observado pelo *The Economist* em seu *website (The Economist*, 2010), a decodificação do genoma humano, iniciada em 2003, levou aproximadamente 10 anos para ser concluída. Com a utilização dos recursos computacionais atuais, essa pesquisa poderia ter sido realizada em apenas uma semana.

A noção de variedade de dados incorpora a idéia da utilização de diversos tipos de dados provenientes de diversas fontes. Foram-se os dias em que os dados de uma companhia podiam ser impressos, ordenadamente dispostos em uma mesa e analisados. Em uma pesquisa recente coordenada pela *International Data Corporation* (IDC), comprovou-se que 80% de toda informação digital é composta por dados desestruturados. (IDC, 2014)

Dados estruturados são informações com um elevado grau de organização, que podem ser armazenados em bancos de dados relacionais – ou seja, os dados são guardados em tabelas, com estruturas similares ao observadas em uma planilha. A consulta a um banco de dados relacional é extremamente simples e muito eficiente. (*Brightplanet*, 2012)

Dados desestruturados, por sua vez, são praticamente o caso inverso. Em sua maioria, são textos produzidos por pessoas e em diferentes idiomas. Incluem projetos de pesquisa, publicações internas ou externas, descrições e review de produtos, atividades em redes sociais, emails e etc. (Pornai, 2014).

Muitos autores acreditam que o *Big Data* pode ser mais bem compreendido adicionando-se outros “Vs” em sua definição. Tais “Vs” tem como objetivo explicar aspectos importantes do *Big Data* que as organizações devem ter em mente. Vamos explorar dois desses conceitos: veracidade e valor.

De acordo com Van Rijmenam (2013), possuir um grande volume e variedade de dados fluindo em grande velocidade não significa nada se estes estiverem incorretos. É importante que as organizações assegurem a veracidade dos dados e que a análise sobre eles seja precisa. Informaçoes verídicas criam indicadores confiáveis e margens seguras para projeções de mercado. (Van Rijmenam, 2013).

Desta maneira, as informações disponíveis irão gerar valor para as organizações que estão preparadas para explorá-las. Em 2011, um estudo realizado

pelo McKinsey Global Institute, estimou que as estratégias de *Big Data* gerariam uma economia de 300 bilhões de dólares anuais para o setor de saúde dos Estados Unidos – mais que o dobro gasto anualmente pelo setor de saúde da Espanha. (McKinsey, 2011).

Obviamente, os dados em si não são detentores de valor. Van Rijmenam (2013) ainda afirma que o valor provém da analise realizada nestes dados, na maneira como são transformados em conhecimento e na forma como este conhecimento é utilizado para direcionar a tomada de decisões de uma companhia.

No próximo capítulo iremos explorar o processo de construção de valor sobre os dados coletados do ambiente virtual, utilizando o conceito de Análise de Dados.

3. ANÁLISE DE DADOS

Extrair valor da massa de informações do *Big Data* requer as ferramentas corretas. E, na busca por esse valor, grandes corporações tem investido muitos recursos em ferramentas para a Análise de Dados (do inglês *Data Analysis*).

A Análise de Dados é o processo de examinar uma grande quantidade de dados de diversos tipos e origens, com o objetivo de descobrir padrões outrora ocultos, correlações desconhecidas e outros tipos de informações que auxiliem as organizações a tomar decisões mais inteligentes. (Rouse, 2014).

Porém, os *softwares* de gestão empresarial convencionalmente utilizados apenas "arranham" cerca de 10% a 20% das informações que a maioria das companhias manipulam atualmente: dados estruturados (Deangelis, 2014).

Documentos utilizados no ambiente corporativo são conduzidos pelo seu ciclo de vida por plataformas gerenciadoras de conteúdo. E-mails são igualmente gerenciados, monitorados e armazenados por ferramentas específicas – como o *Outlook* por exemplo. No entanto, tais plataformas estão mais focadas em gerenciamento e retenção de documentos do que em análise de conteúdo. Elas não foram projetadas para prover a análise e exploração dos dados que gerenciam – e nem são capazes disso. (Stewart, 2014)

Além disso, de acordo com uma pesquisa realizada neste ano pelo IDC (2014), 90% dos dados que são relevantes a um negócio provém de fontes externas e são do tipo desestruturado.

Neste contexto, conforme afirmado por Pornai (2014), o potencial para se extrair percepção dos negócios se expande dramaticamente quando a informação empresarial é somada a informação pública. Conteúdos oriundos de redes sociais de toda a sorte podem ser o caminho direto para os corações e mentes dos consumidores. *Blogs, tweets*, comentários e resenhas são o reflexo atual da opinião pública sobre um produto em um determinado momento. Conteúdos mais tradicionais, como artigos jornalísticos e descrições de produtos em *websites* de uma marca tendem a serem moldados como extensão de opiniões públicas.

2.1 – Crescimento das Mídias Sociais

O papel das Mídias Sociais no atual panorama é de grande importância para as empresas que adotam estratégias relacionadas ao *Big Data*. Zabin J. E Jefferies A. (2008) já previam naquela época que:

> Com a explosão da *Web* 2.0, plataformas como *blogs*, fóruns de discussões, redes *peer-to-peer*, e vários outros tipos de mídia social... os consumidores tem a sua disposição um canal de comunicação de alcance e poder sem precedentes, pelo qual podem compartilhar suas experiências e opiniões, positivas ou negativas, sobre qualquer produto ou serviço.
>
> As grandes companhias estão começando a perceber que a voz destes consumidores pode exercer uma enorme influência na formação de opinião de outros consumidores – e, por conseqüência, em sua fidelidade a marcas, em suas decisões de compra e na defesa de suas próprias marcas... As companhias podem responder a estes clientes através de mensagens de *marketing*, posicionamento da marca, desenvolvimento de produtos e outras atividades adequadamente. 4

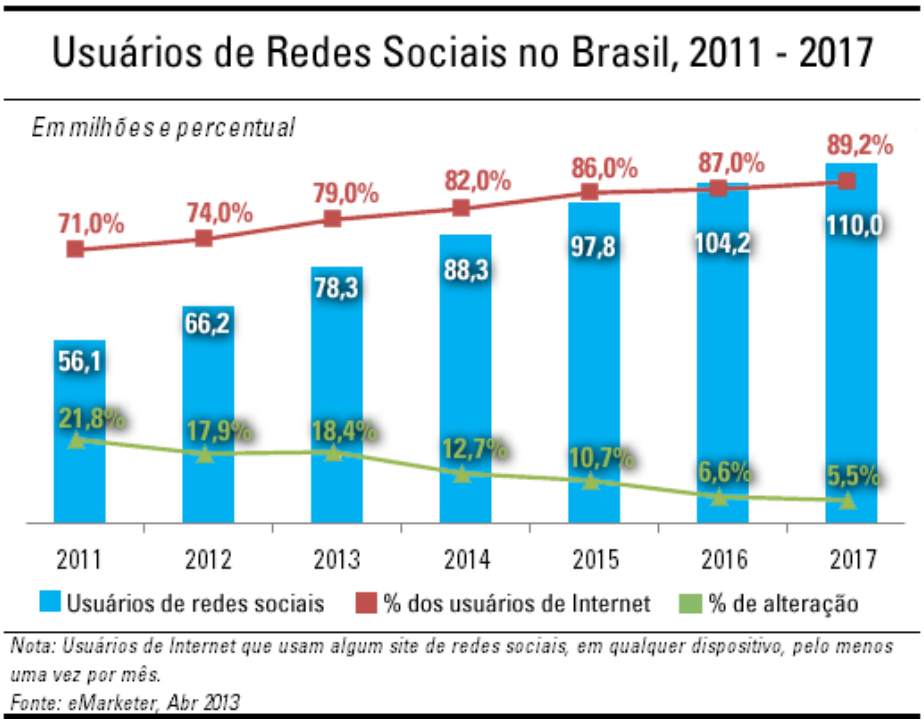
De fato, as Mídias Sociais apresentaram um grande crescimento nos últimos anos e se tornaram parte integrante do comportamento dos usuários na Internet.

Em um estudo realizado em 2013 pela empresa de pesquisa de mercado *eMarketer*, constatou-se que 79% dos internautas brasileiros – cerca de 78,3 milhões de pessoas – são usuários ativos de redes sociais (*eMarketer*, 2013).

---

[4] Traduzido pelo autor: *"With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media . . . consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinions, positive or negative, regarding any product or service. As major companies are increasingly coming to realize, these consumer voices can wield enormous influence in shaping the opinions of other consumers — and, ultimately, their brand loyalties, their purchase decisions, and their own brand advocacy. . . . Companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their marketing messages, brand positioning, product development, and other activities accordingly."*
Fonte: Generating Consumer Insights from Online Conversation Zabin and Jefferies (2008, p. 327)

Figura 2: Usuários de Redes Sociais no Brasil

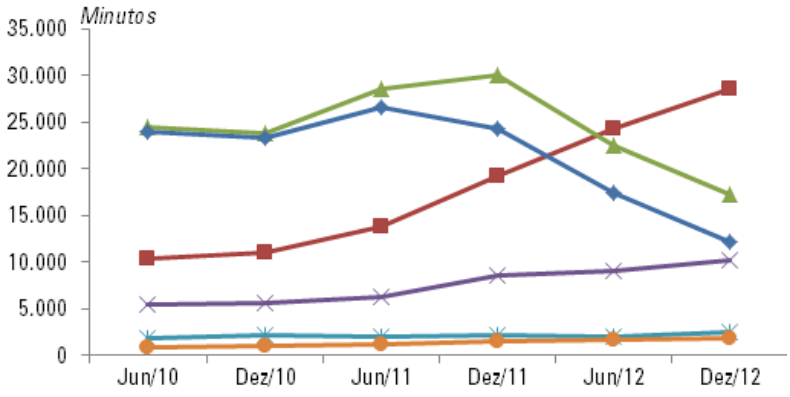


Fonte: Hytrad (2013)

Acompanhando este movimento, um outro estudo realizado pela empresa *comScore* – também no ano de 2013 – constatou que as mídias sociais também apresentaram crescimento em termos de tempo gasto online pelos internautas brasileiros. Em torno de dezembro de 2012, estes internautas estavam gastando, acumuladamente, 158% a mais de tempo com mídias sociais do que gastaram em dezembro de 2010 (*comScore*, 2013).

Figura 3: Tempo gasto pelos internautas brasileiros, por categoria de site

Tempo gasto nos principais sites online, por categoria, pelos internaustas brasileiros



	Jun/10	Dez/10	Jun/11	Dez/11	Jun/12	Dez/12
Mídia Social	10.335	11.071	13.789	19.194	24.299	28.541
Portais	24.466	23.809	28.611	29.951	22.426	17.269
Serviços	24.024	23.220	26.527	24.338	17.435	12.099
Entretenimento	5.397	5.669	6.178	8.534	9.078	10.184
Jogos	1.893	2.146	1.965	2.079	2.017	2.557
Notícias / Informações	813	1.058	1.125	1.546	1.707	1.761

Nota: Idade 6+; Inclui casa e trabalho.
Fonte: comScore Inc., "2013 Brazil Digital Future in Focus," 14/03/2013

Fonte: Hytrad (2013)

Entrelaçando estes dois estudos, podemos enxergar formação de um novo perfil do internauta no Brasil nos últimos anos, em que as pessoas dedicam maior parte de seu tempo online interagindo em mídias sociais, como o *Facebook*[5] e o *Twitter*[6] por exemplo.

As corporações estão cientes desta tendência e reagem intensificando suas estratégias de *marketing* nas mídias sociais. Podemos observar o sucesso deste engajamento observando os indicadores de crescimento do comércio eletrônico – no Brasil e no mundo.

[5] O Facebook é uma mídia social em que os usuários podem compartilhar vídeos, fotos, utilizar serviços de bate-papo, jogos e etc. Fonte: Techtudo, 2014

[6] O Twitter é uma mídia social de comunicação, que oferece um espaço limitado a 140 caracteres por mensagem. Fonte: Twitter, 2014.

2.2 – Expansão do Comércio Eletrônico

Uma pesquisa realizada pela *eMarketer* estima que o comercio digital irá crescer 20.2% até o final de 2014, atingindo a marca de U$ 1.500 trilhões movimentados. Tal crescimento acompanha o aumento do número de usuários conectados a Internet (através de computadores e dispositivos móveis) principalmente nos países de economia emergente, onde cada vez mais pessoas entram para a classe média. (*eMarketer,* 2014).

Tabela 2: Crescimento do Comércio Digital

B2C Ecommerce Sales Growth Worldwide, by Country, 2012-2017
% change

	2012	2013	2014	2015	2016	2017
China*	93.7%	78.5%	63.8%	43.3%	34.4%	29.4%
Indonesia	85.0%	71.3%	45.1%	37.2%	26.0%	22.0%
India**	35.9%	34.9%	31.5%	30.3%	24.5%	20.0%
Argentina	31.1%	6.3%	24.0%	18.0%	12.0%	10.0%
Mexico	55.8%	41.9%	20.0%	14.5%	10.0%	5.0%
Brazil	21.8%	16.5%	19.1%	8.5%	6.9%	6.0%
Russia	34.4%	19.4%	17.1%	10.8%	6.9%	5.2%
Italy	17.0%	16.8%	15.3%	13.5%	12.0%	10.6%
UK	14.5%	16.3%	14.2%	12.2%	9.2%	8.2%
Canada	15.0%	14.2%	14.0%	13.5%	12.5%	11.5%
Spain	10.0%	10.0%	13.8%	11.9%	10.0%	8.0%
Sweden	18.4%	16.2%	13.3%	10.3%	9.0%	8.4%
US***	14.2%	13.4%	11.8%	11.4%	10.9%	10.4%
Norway	14.9%	12.7%	11.0%	10.8%	8.1%	7.2%
Denmark	14.3%	12.4%	10.6%	8.9%	6.5%	5.9%
France	32.3%	10.3%	10.0%	9.8%	7.6%	7.1%
Netherlands	12.7%	11.4%	9.4%	8.4%	6.3%	5.3%
South Korea	12.7%	9.6%	7.4%	4.8%	4.3%	3.6%
Germany	25.6%	5.7%	7.4%	6.9%	6.5%	6.1%
Japan	12.3%	-10.2%	7.1%	6.7%	5.6%	5.0%
Australia	10.5%	6.0%	5.7%	5.1%	5.0%	4.2%
Finland	4.3%	4.4%	3.7%	3.2%	2.7%	2.5%
Worldwide	**22.3%**	**18.3%**	**20.2%**	**17.7%**	**15.9%**	**14.8%**

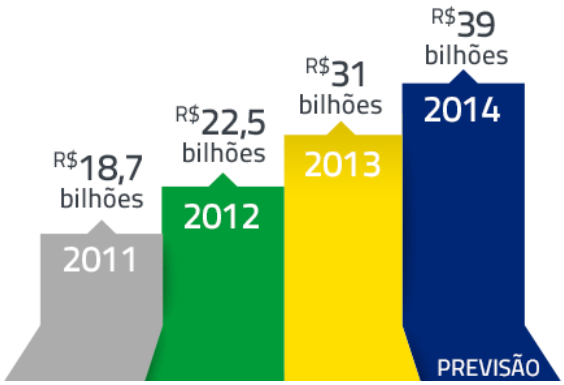
*Note: includes products and services ordered and leisure and unmanaged business travel sales booked using the internet via any device, regardless of the method of payment or fulfillment; *includes sales from businesses that occur over C2C platforms; excludes Hong Kong; **digital travel sales represent roughly 70% of B2C ecommerce sales; ***excludes event tickets*
Source: eMarketer, Jan 2014

167705 www.eMarketer.com

Fonte: eMarketer (2014)

No Brasil, esse valor deverá atingir a marca de R$ 39 bilhões de Reais[7].

Figura 4: Crescimento do comércio eletrônico no Brasil



Fonte: *Webjump* (2014)

O aumento do índice de compras realizadas digitalmente também impulsiona os clientes a se relacionarem com as companhias através das mídias sociais, seja para opinar sobre a qualidade de um produto ou realizar reclamações sobre estes, uma vez que esse tipo de canal de comunicação confere maior visibilidade ao sentimento do consumidor final. (Gesenhues, 2013)

Todos estes dados possuem um enorme potencial para serem convertidos em valor para as organizações, seja em apoio para o processo de tomada de decisões ou para compreensão das necessidades dos consumidores. Infelizmente, essa avalanche de dados produzida através de mídias sociais não é apenas volumosa, mas também é desestruturada.

Os computadores são extremamente úteis no sentido de acumular informação desestruturada, porém ainda deixam a desejar no quesito de analisar estes dados. Isso por que, em sua essência, os computadores são baseados em processos matemáticos, os que o torna melhores em analisar dados estruturados, como listas e planilhas ao invés de som, vídeo, textos e gráficos - que compõem a maioria dos dados existentes na atualidade. (Thornton, 2009)

Capturar, processar e gerenciar esta parcela de dados são tarefas que estão além da escala humana e além da escala de *softwares* comuns. A indústria do *Big*

[7] Cosiderando o valor de R$2,62 do Dólar Comercial em 15/11/2014, esse valor é equivalente a aproximadamente 14 bilhões de dólares.

Data se encaixa neste contexto por ser especializada em manipular conjuntos de arquivos que vão desde alguns *terabytes* até centenas de *petabytes*. Além disso, parte de suas aplicações tem sido projetadas para se extrair sentido das interações em mídias sociais. (Opallios, 2014).

2.3 - Uma nova abordagem para análise de dados

A Internet está evoluindo para uma era em que as opiniões dos usuários estão ganhando visibilidade crescente. A construção de conhecimento a partir da grande massa de dados desestruturados – ou seja, a partir de material textual – é um fator determinante para o sucesso do *marketing* digital, posicionamento de produtos, gerenciamento de reputação de uma marca e muito mais.

Porém, tais dados permanecem difíceis de serem interpretados por computadores, uma vez que são gerados de humanos e para humanos. A automação da análise deste tipo de dados requere aplicações de *Big Data* que compreendam profundamente a linguagem humana e interpretem o significado do material textual produzido – objetivo do qual elas ainda estão muito distantes. (Poria, Cambria, Winterstein e Huang – 2014)

As ferramentas de Análise de Sentimentos tem ganhado destaque no mundo dos negócios – bem como no meio científico – por serem capazes de extrair, com bom grau de precisão, percepções e tendências de mercado a partir de “conversações” em mídias sociais. No próximo capítulo iremos apresentar este conceito, bem como os desafios que está tecnologia trás.

4. ANÁLISE DE SENTIMENTOS

3.1 Origens

A Análise de Sentimentos é considerada um campo de pesquisa dentro da área da Inteligência Artificial chamada Processamento de Linguagem Natural (do inglês *Natural Language Processing*, ou NLP).

O interesse em processamento de linguagem natural surgiu nos meados dos anos 50, quando Alan Turing publicou um artigo intitulado “Computadores e Inteligência” (nome original: *Computing Machinery and Intelligence*), de onde surgiu o chamado Teste de Turing. Neste artigo, o autor afirmou que um computador poderia ser considerado inteligente se fosse capaz de manter uma conversa com um humano sem que este notasse que estava falando com uma máquina.

O objetivo do processamento de linguagem natural é permitir esse tipo de interação para obtenção de informações de sistemas computacionais. Tal tipo de interação foi popularizada em 1968 no filme “2001: Uma Odisséia no Espaço” e na série de TV *Star Trek* (Pang e Lee - 2008).

O crescimento das mídias sociais – e o conseqüente aumento de produção de material textual na *Web* – foi um fator determinante para o desabrochar das pesquisas na área de Análise de Sentimentos, ocorrida em meados de 2001.

3.1 Definição

O papel que as emoções desempenham no mercado não é novidade. As emoções do cliente funcionam como um termômetro de seu comportamento como consumidor. Tais sentimentos tem o poder de moldar a influência de uma marca, modificar suas atitudes, opiniões e percepções.

A Análise de Sentimentos também não é uma novidade neste contexto. As empresas sempre analisaram o sentimento do consumidor em relação a um produto “a moda antiga”, através de caixas de sugestões, pesquisas e entrevistas. Estes métodos tem sido adaptados para se tirar vantagem do ambiente altamente interativo que é a Internet.

A Análise de Sentimentos é o processo de detectar a polaridade de um texto, dentro de uma contextualização específica. Esta técnica é capaz de classificar uma

porção de texto como positiva, negativa ou neutra – atribuindo valores individuais a cada palavra e, posteriormente ao documento completo.

Por exemplo, na frase “Eu amo o verão em Curitiba, mas odeio o inverno”, o algoritmo irá classificar “Eu amo o verão” como positivo e “Odeio o inverno” como negativo. Porém, devida a sua construção, a frase será classificada como neutra, uma vez que a polaridade positiva de uma palavra anula o sentimento negativo da outra.

Segundo Pang e Lee (2008), a detecção e reconhecimento de informação emocional através de recursos computacionais confere mensurabilidade e objetividade para a análise, adicionando confiabilidade, controle de qualidade e verificabilidade, contribuindo positivamente para a credibilidade da pesquisa.

A extração de conteúdo emocional através de processos automatizados tem chamado a atenção de pesquisadores pela velocidade e eficiência no processamento de grandes volumes de informação, que podem ser extraidos em tempo real e prover um precioso e oportuno material para tomada de decisões corporativas .

Devido a capacidade desta técnica de monitorar um assunto específico, muitas empresas a utilizam para monitorar a reputação de seus produtos e serviços. Se a marca for atacada através de conteúdos em mídias sociais, por exemplo, a Análise de Sentimentos irá identificar estas menções e classificá-las como negativas, e a empresa estará ciente e poderá reagir de acordo.

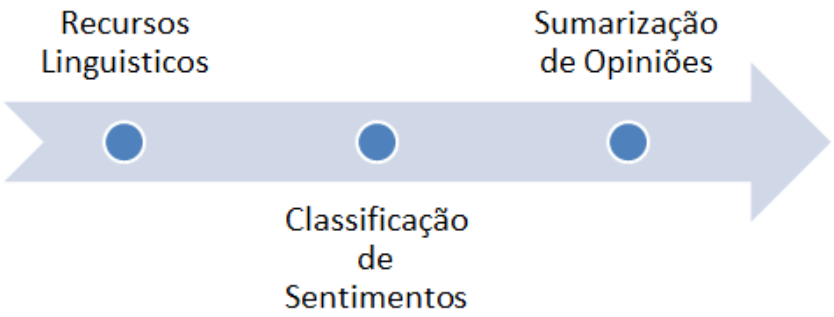
3.2 Metodologia

Existem duas principais metodologias utilizadas para se realizar a Análise de Sentimentos. O primeiro é o método baseado em Aprendizado de Máquina. É um método supervisionado que utiliza uma série de dados como exemplos de situações e, a partir dessa amostragem, é realizada a Análise de Sentimentos no material coletado a partir de comparações com estes exemplos.

Este método é capaz de realizar a extração sentimentos com bastante eficiência, porém, conforme observado por alguns autores, é mais demorado devido ao tempo necessário para se reunir os exemplos que servirão como base. (Augustyniak, Lukasz et al. 2014).

Neste trabalho iremos discorrer sobre outra abordagem – a Lexical – para a apresentação deste tema. Conforme afirmado por Pang e Lee (2008), o processo de análise de sentimentos pode ser dividido em 3 grandes etapas: o desenvolvimento dos recursos lingüísticos, classificação de sentimentos e sumarização. Este capítulo irá conservar esta divisão.

Figura 5: Etapas da análise de sentimentos em um texto



Fonte: Autoria própria

3.2.1 – Obtenção dos Recursos Lingüísticos

Para se realizar a Análise de Sentimento em material textual, é necessário equipar os algoritmos com uma base de dados lexical – ou seja, um grande conjunto de palavras – agrupados em sinônimos que expressam conceitos distintos. (*Wordnet*, 2014).

O léxico pode ser criado manualmente ou ser baseado em compilações já existentes, como o *Genenal Inquirer* da Universidade de Harvard ou o *Wordnet* da Universidade de Priceton – ambos disponibilizados sem custo.

Embora haja alterações na forma como diferentes léxicos são estruturados, de maneira geral são construídos para identificar se uma expressão é subjetiva ou objetiva. Cada palavra é associada a sua definição formal – ou seja, ao seu verbete de dicionário – e a seus respectivos sinônimos.

Adicionalmente, também é criado um parâmetro de categorização humana, agrupando palavras em várias categorias, de acordo com seu significado. A

associação destes dois conjuntos de regras é capaz de distinguir o contexto de uso e polaridade dos termos em uma sentença.

Tabela 3: Exemplo de classificação de adjetivos no *Wordnet*

Adjetivos que expressam:	Positivos	Negativos	Objetivos	Total
Emoção	52	73	3	128
Caráter	355	584	220	1159
Aparência	41	83	46	170
Sabor	40	41	5	86
Outros	29	17	87	133

Fonte: TUVERI, F. ANGIONI, M. 2012

A Tabela 3 indica uma pequena lista de propriedades que podem ser associadas a adjetivos e a quantidade de ocorrências encontradas que expressam valores positivos, negativos ou objetivos[8].

Os advérbios, seus significados, posição e intensidade também são levados em consideração e geralmente possuem orientação objetiva. A seguir, na Tabela 4, é mostrado um exemplo de categorização utilizado pelo *Wordnet*.

Tabela 4: Exemplo de classificação de advérbios no *Wordnet*

Advérbios	Positivos	Negativos	Objetivos	Total
Tempo	0	0	7	7
Modo	184	230	13	427
Lugar	0	0	3	3
Intensificadores	0	0	38	38
Quantidade	0	0	6	6

Fonte: TUVERI, F. ANGIONI, M. 2012

Várias outras categorias de palavras também são classificadas desta mesma forma e são fortes indicadores de sentimentos, como substantivos, palavras de negação, conjunções e verbos. Não é do escopo deste trabalho se aprofundar nas classificações dos grupos de palavras existentes.

[8] Nota do autor: O Wordnet é uma base de dados lexical para palavras em inglês. O número total de adjetivos (e suas ocorrências positivas, negativas ou objetivas) certamente seria maior para uma base de dados em português, visto que este idioma possui mais palavras do que o inglês.

3.2.2 – Análise de Sentimentos

A avaliação do sentimento presente no texto é atingido através da soma dos graus de sentimento expressados nas palavras analisadas. No exemplo a seguir, apresentamos um algoritmo utilizado pela ferramenta Serendio na analise de *tweets* que – além de texto – é capaz de considerar na análise a presença de emoticons.

Figura 6: Algoritmo para cálculo de sentimento - Serendio

```
Algorithm 1: Sentiment Calculation
Data: Preprocessed Twitter data
Result: Output: Positive, Negative, Neutral
Find the list of sentiment words SentiList, its
position in the sentence;
Find the list of sentiment negation words
SentiNegat, its position in the sentence;
Find the list of blind negation words
BlindNegat, its position in the sentence;
if BlindNegat then
 | return negativity;
else
 | if SentiList and SentiNegat then
 | | foreach word in the SentiList do
 | | | if word is atmost the distance of 2
 | | | from SentiNegat then
 | | | | Revert the polarity of the word;
 | | | end
 | | end
 | else
 | | if SentiNegat then
 | | | Add the SentiNegat to the
 | | | negative SentiList;
 | | end
 | end
end
SentiSum=0;
foreach word in the SentiList do
 | SentiSum=SentiSum+sentiment of
 | word;
end
if Hashtag is present then
 | Find all the sentiment words in hash tag
 | using regex matching and add them to
 | SentiList
end
if Emoticon is present then
 | Find sentiment of the emoticon and add
 | emoticon,it's sentiment to SentiList
end
SentiType="neutral";
if SentiSum > 0 then
 | SentiType="positive";
end
if SentiSum < 0 then
 | SentiType="negative";
end
return SentiType;
```

Fonte – PALANISAMY, Prabu, YADAV, Vineet, ELCHURI, Harsha, 2014

O algoritmo demonstra extração das palavras que invariavelmente demonstram sentido negativo (chamadas de '*blind negations*' no inglês – ou seja, 'negação cega') da sentença. A presença de uma *blind negation* indica um sentimento negativo. Se a sentença contém uma *blind negation* os demais passos de análise são pulados e o sentimento é diretamente classificado como negativo.

Em seguida, é feita a extração de palavras que expressam sentimentos e sua comparação com o léxico utilizado pela companhia (SentiList). A polaridade do sentimento pode ser revertida pela ocorrência de uma palavra de negação localizada até duas palavras de distância.

Se a sentença não possuir uma palavra que expresse sentimento, mas possuir uma negação, a própria negação é interpretada como um sentimento negativo. Por exemplo, na expressão "Eu não posso lidar com isso", existe uma negação, mas não existe um sentimento associado.

Emoticons são tabelados pelos nomes dos sentimentos representados, e são analisados em comparação com o léxico utilizado. O texto de uma hash-tag também é extraído e analisado de acordo com o léxico.

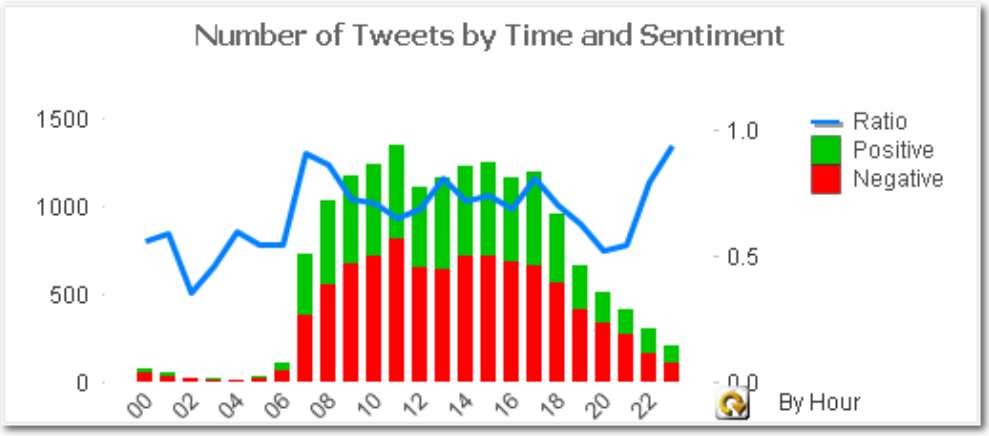
Desta forma, o sentimento representado pelo tweet é a soma dos sentimentos extraídos de todas as entidades presentes. (PALANISAMY, Prabu, YADAV, Vineet, ELCHURI, Harsha, 2014)

3.2.1 – Sumarização de Opiniões

As ferramentas de Análise de Sentimentos operam sobre uma grande quantidade de textos extraídos da Web. A Sumarização de Opiniões nada mais é do que o processo de agrupamento de sentimentos de uma variedade de pessoas em relação ao tema pesquisado – que pode ser um produto, uma situação um evento e etc.

O objetivo é prover uma maneira clara de visualizar os resultados da análise, reduzindo a complexidade e facilitando a identificação de padrões e tendências de mercado. Vemos na figura 7 um exemplo da sumarização de opiniões obtidas a partir do Twitter.

Figura 7: Exemplo de sumarização de sentimentos



Fonte: Qvsource, 2014

A técnica apresentada mostrou-se capaz de processar uma boa quantidade de dados com bastante rapidez e eficiência. Em um pequeno teste recentemente realizado por Kaushik e Mishra (2014), tendo como material de análise *posts* de usuários no Twitter, obteve-se os resultados da tabela 5:

Tabela 5: Performance de algoritmo de classificação

Número de Tweets	Tempo Consumido	Precisão
6,74,412	14.8 segundos	73.5 %

Fonte: KAUSHIK, Chetan, MISHRA. 2014

3.3 Considerações sobre Privacidade

A privacidade é uma das grandes questões levantadas ao se tratar que qualquer assunto relacionado a *Big Data*. As corporações tem armazenado cada vez mais informações de seus clientes, de maneira a utilizá-las em seu processo de análise.

Damos o nome de Informações Pessoalmente Identificáveis (do inglês *Personally Identifiable Information* (PII)) a qualquer informação mantida por uma organização que possa expor a identidade de um cliente. Segundo McCallister (2011), são consideradas PII: Nome, data e local de nascimento, nome da mãe e dados utilizados para identificação biométrica. Qualquer outro tipo de informação

que possa ser diretamente associada a um indivíduo também é considerada PII, como históricos médicos, educacionais, financeiros e etc.

A coleta de dados em mídias sociais deve respeitar esse princípio. Os dados que não são necessários para a Análise de Sentimentos – nome do usuário, por exemplo – devem ser deletados os substituídos nas bases de dados, garantindo um processo de anonimização da informação. As informações sensíveis que por um acaso forem mantidas devem ser protegidas pela companhia.

Ademais, é preciso que as organizações foquem na construção de um relacionamento mais transparente com seus clientes. Políticas de privacidade simples e claramente redigidas, informando-os de como seus dados serão coletados e para qual propósito são um exemplo disso. As companhias também devem trabalhar no sentido de facilitar a experiência do cliente, permitindo-os deletar e/ou editar suas configurações de privacidade sempre que necessário, construindo desta maneira uma relação de confiança com o consumidor e mantendo suas bases de dados com informação atualizada. (Van Rijmenam, 2013).

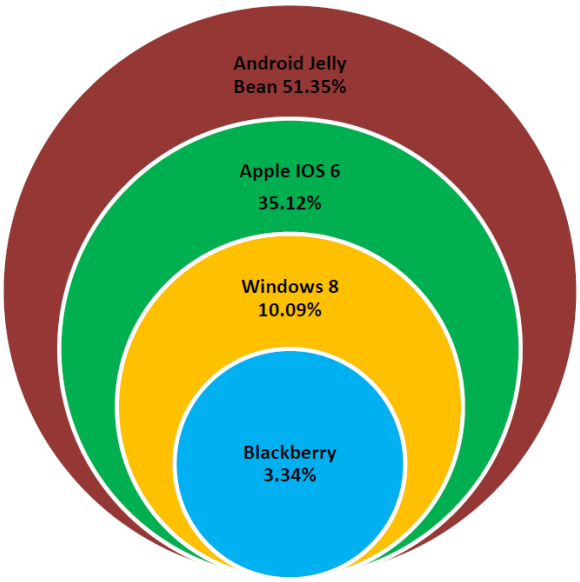
5. ESTUDO DE CASO - Nokia Lumia 920

Um *report* elaborado pela empresa Serendio teve como objetivo estudar os sentimentos dos usuários de mídias sociais e entender o posicionamento no mercado do aparelho Nokia Lumia 920 e seu concorrente, o HTC *Windows* 8X nos Estados Unidos – ambos funcionando com o mesmo sistema operacional, o *Windows* 8.

Para tal, foram analisadas cerca de 75 mil postagens coletadas na *Web*, durante o período de Novembro de 2012 até Janeiro de 2013. A empresa utilizou sua ferramenta proprietária que aplica a técnica de Análise de Sentimentos em sua abordagem lingüística para a realização deste estudo.

Com o objetivo de ter uma melhor compreensão da fatia de mercado ocupada pelo Sistema Operacional *Windows* 8, a empresa realizou um estudo preliminar sobre a popularidade dos principais Sistemas Operacionais, de acordo com a quantidade de menções realizadas na *Web*. (Figura 8)

Figura 8: Popularidade de Sistemas Operacionais de *Smartphones*

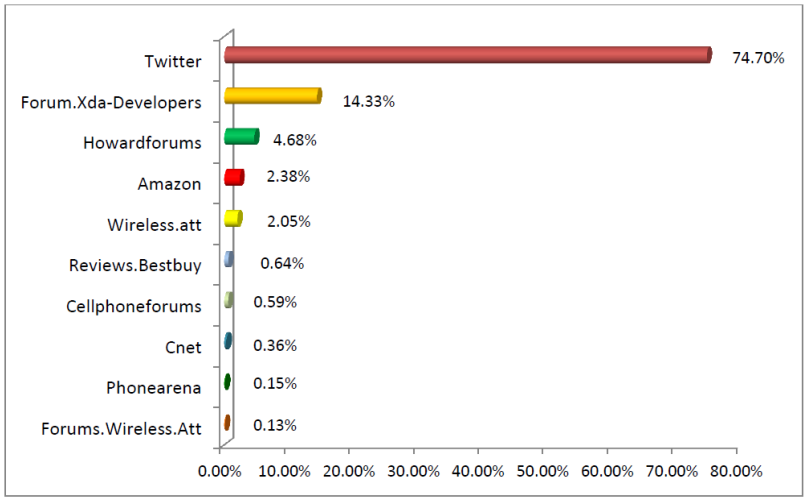


Fonte: Serendio, 2013

Como vemos na Figura 8, a maioria das conversações está relacionada ao Sistema Operacional Android. O *Windows* 8 está em terceiro lugar, figurando em 10,09% das conversações.

Observando os dados coletados durante o período da pesquisa, localizou-se a origem das menções realizadas sobre o Nokia Lumia 920 e o HTC *Windows* 8X, como vemos a seguir:

Figura 9: Menções realizadas - Nokia Lumia 920 e HTC *Windows* 8X

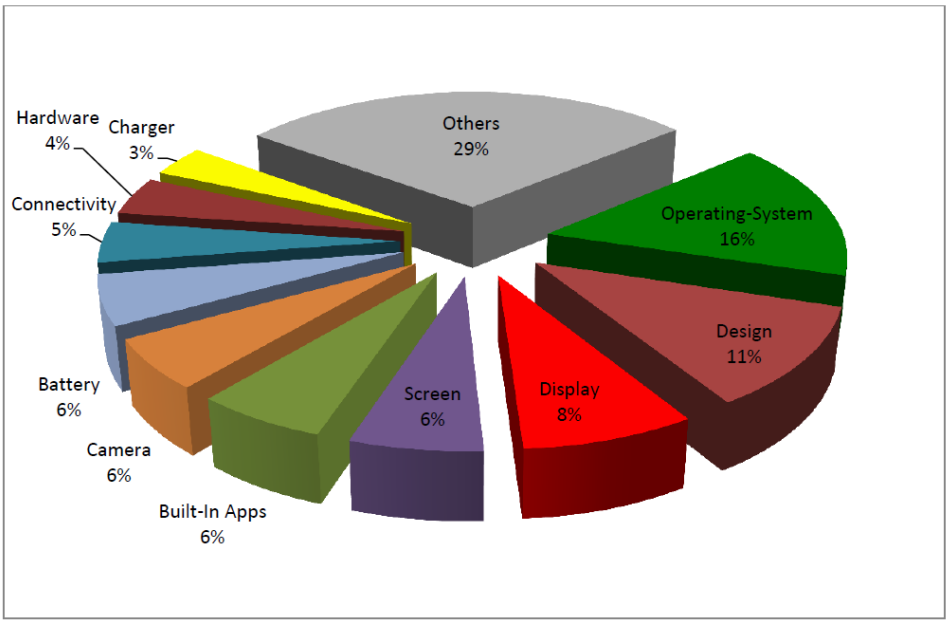


Fonte: Serendio, 2013

A grande maioria das menções foram realizadas na Mídia Social Twitter, seguido de dois fóruns especializados no desenvolvimento de aplicativos para aparelhos móveis. Através desta figura, também é possível notar que poucas referencias foram realizadas em sites em que se é possível fazer avaliação de produtos, como a lojas Amazon e Bestbuy.

Em seguida, na Figura 10, os dados foram filtrados por palavras e volume de ocorrência, com o objetivo de determinar quais características são mais importantes entre os usuários do Nokia Lumia 920 e do HTC *Windows* 8X.

Figura 10: Características consideradas mais importantes pelos usuários



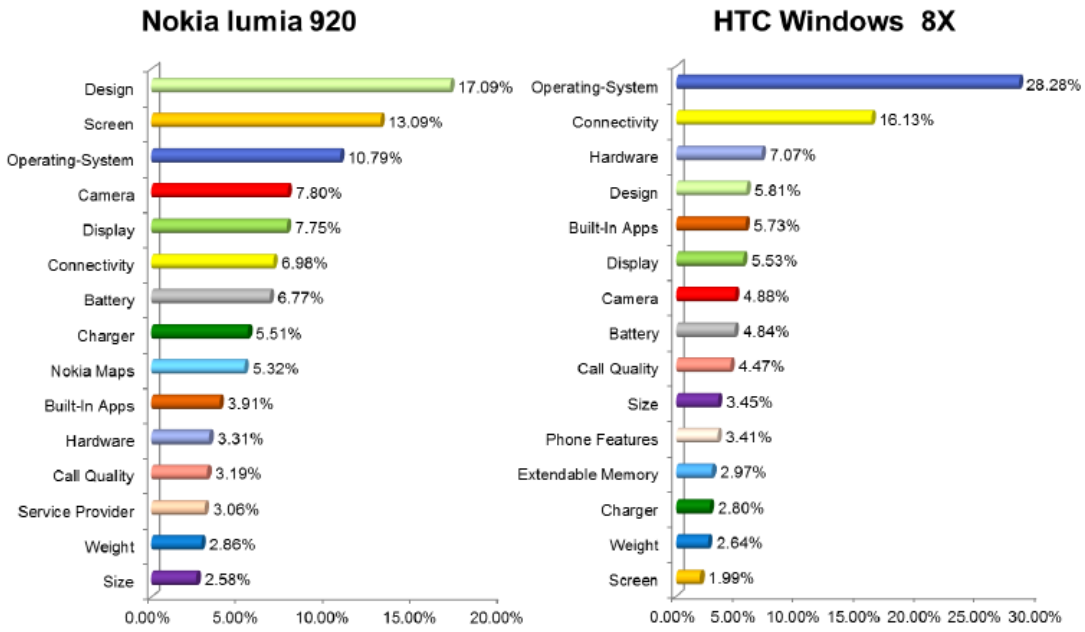
Fonte: Serendio, 2013

A figura nos mostra que Sistema Operacional e *Design* são as características mais populares nas conversações analisadas. É compreensível que o Sistema Operacional seja amplamente citado, uma vez que os dois modelos analisados pertencem a esta categoria.

Design, Display (ou seja, cores e resolução da tela) e *Screen* (considerando aqui o tamanho da tela qualidade do *touch screen*) são outras tópicos de discussão. É possível notar que bateria e câmera geraram pouco volume de discussões

As outras características, sinalizadas na figura com 29% das conversações incluem tópicos como: Email, peso, teclado, rádio e sistema de troca de mensagens. Tais características não foram exploradas mais a fundo pela Serendio pois, no entendimento da empresa, tais características não são determinantes na decisão de comprar ou não um aparelho *Windows* 8.

Figura 11: Características positivamente mencionadas de cada aparelho



Fonte: Serendio, 2013

Na Figura 11 é exibido o volume de menções positivas em relação as características de cada aparelho. Para o Nokia Lumia 920, os pontos mais mencionados foram relativos a seu *Design, Screen*, Sistema Operacional, Câmera e *Display.*

As menções ao aparelho concorrente seguem outra tendência: com exceção do Sistema Operacional, conectividade e *hardware* ganham destaque.

Também é possível verificar que a bateria e câmera do Lumia são considerados melhores do que em seu concorrente. Em contrapartida, O HTC tem mais menções positivas em relacão ao Sistema Operacional, conectividade e aplicativos nativos.

A tabela 6 a seguir lista – em ordem decrescente - as características consideradas mais importantes pelos usuários, em conformidade com a Figura 10. A porcentagem de menções positivas para tais características foram retiradas da figura 11.

Tabela 6: Resultados - Nokia Lumia 920 x HTC Windows 8X

	Menções Positivas	
Características	**Nokia Lumia 920**	**HTC Windows 8X**
Sistema Operacional	10,79%	28.28%
Design	17,09%	5,81%
Display	7,75%	5,53%
Screen	13,09%	1,99%
Aplicativos Nativos	3,91%	5,73%
Camera	7,80%	4,48%
Bateria	6,77%	4,84%
Conectividade	6,98%	16,13%
Hardware	3,31%	7,70%
Carregador	5,51%	2,80%

Fonte: Autoria Própria.

Pode-se observar que o Nokia Lumia 920 possui mais menções positivas em relação as características consideradas importantes pelos usuários. De uma maneira geral, também recebeu um maior número de menções positivas em maior quantidade de categorias. Desta forma, este estudo aponta que o Nokia Lumia 920 é mais bem aceito pelos usuários de *smartphones* com sistema operacional *Windows* do que seu concorrente, o HTC Windows 8X.

A utilização desta ferramenta para o monitoramento de menções apenas requer intervenção humana no momento da escolha de como exibir os resultados obtidos. Seu uso mostrou-se eficiente para determinar o sentimento do produto no mercado através de porções de informação relativamente pequenas (recordando que a maioria das menções foram extraídas do Twitter, que limita cada mensagem a 140 caracteres).

6. CONSIDERAÇÕES FINAIS

Até o ano de 2003, estima-se que a quantidade de dados existentes na Internet era de apenas alguns *exabytes*. Atualmente, essa mesma quantidade de dados é criada em poucos dias. O crescimento das mídias sociais disponibilizou uma maneira de criar conteúdo e compartilhar idéias e opiniões com um número de pessoas sem precedentes.

Porém, esta imensa quantidade de dados é desestruturada – ou seja, produzida por humanos e para humanos, não sendo diretamente processáveis por recursos computacionais.

A Análise de Sentimentos é uma abordagem que pode auxiliar neste contexto, uma vez que é baseada na análise semântica do material textual presente na Internet. As pesquisas existentes na área demonstram que está tecnologia é capaz de analisar uma grande quantidade de dados em poucos segundos e com um grau de precisão muito satisfatório. As ferramentas baseadas nesta tecnologia permitem as corporações a compreenderem seu mercado, possibilitando desde a identificação de padrões e tendências que auxiliam no desenvolvimento de novos produtos até o refinamento de campanhas de *marketing* para atingir públicos específicos.

Neste trabalho, a Análise de Sentimentos foi introduzida como uma opção no processamento da crescente quantidade de material desestruturado. Esta tecnologia ainda não se encontra em um perfeito grau de maturidade, sendo necessário sua evolução no sentido de melhor interpretar a linguagem humana – desde expressões irônicas até gírias e contextos pouco usuais. Além disso, a maior parte das bases de dados lexicais são construídas no idioma Inglês. Criar bases de dados que contemplem outros idiomas mais complexos - como o Português e o Chinês, por exemplo – será certamente um grande desafio a ser superado pela Análise de Sentimentos.

7. SUGESTÕES PARA TRABALHOS FUTUROS

A Análise de Sentimentos é uma tecnologia que se encontra em desenvolvimento, impulsionada pela crescente quantidade de material textual existente na *Web*. A medida que esta técnica evolui e ganha relevância no cenário de *Data Analytics,* é importante salientar a existência de alguns aspectos importantes que poderão ser explorados em trabalhos futuros:

- O conceito de autenticidade é um dos pilares que compõem Segurança da Informação. Com a imensa quantidade de dados existentes na *Web*, provenientes de diversas fontes, é importante determinar a autenticidade da informação coletada para fins de análise. A existência de *Bots* que geram conteúdo – imitando o comportamento humano – pode ser considerado um fator de risco em relação a autenticidade da informação. Por conseqüência, a Análise de Sentimentos realizada sob informações dessa origem pode não refletir uma opinião legítima. Neste contexto, o estudo das técnicas utilizadas para garantir a autenticidade das informações coletadas apresenta-se como um tema de relevância.

- A privacidade é outro aspecto ligado a Segurança da Informação que pode ser explorado. Cada vez mais os usuários da *Web* criam e compartilham conteúdos. E estes, por sua vez, são utilizados pelas corporações como material bruto de análise. Como forma de manter a privacidade durante o processo, existem as técnicas de anonimização de dados, que deletam os substituem as PIIs dos materiais coletados. Porém, esta etapa é vista mais como boa prática do que como exigência Legal. Um estudo que analise a eficiência de tais técnicas, bem como os princípios éticos que permeiam a coleta de dados, também se apresenta como uma possível linha de pesquisa.

REREFÊNCIAS

10 Biggest Libraries in The World. O Pish Posh, 2014.
http://opishposh.com/10-biggest-libraries-in-the-world/ - Acesso em 10/10/2014

2013 Brazil Digital Future in Focus. Comscore, 2013
https://www.comscore.com/por/Insights/Presentations-and-Whitepapers/2013/2013-Brazil-Digital-Future-in-Focus - Acesso em 19/08/2014

A audiência social continua crescendo no Brasil, a medida que novos usuários da web se cadastram nas redes sociais. Hytrade, 2013.
http://www.hytrade.com.br/marketing-de-midia-social/a-audiencia-social-continua-crescendo-no-brasil-a-medida-que-novos-usuarios-da-web-se-cadastram-nas-redes-sociais/ - Acesso em 31/10/2014

About Twitter. Twitter, 2014.
https://about.twitter.com/ - Acesso em 16/11/2014.

Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report, 2011.
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation - Acesso em 25/08/2014

BOYD, Danah, CRAWFORD, Kate. **Six Provocations for Big Data. "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society**", Oxford, 2011.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431 – Acesso em 07/10/2014/

Capture and Share the World's Moments. Instagram, 2014.
http://instagram.com/# - Acesso em 16/11/2014.

Crescimento do E-commerce Brasil. Webjump, 2013.
http://www.webjump.com.br/loja-virtual/crescimento-e-commerce-brasil/ - Acesso em 31/10/2014

Data, data everywhere. The Economist, 2010.
http://www.economist.com/node/15557443 - Acesso em 26/09/2014

Data Never Sleeps 2.0. Domo, 2014.
http://www.domo.com/learn/data-never-sleeps-2 - Acesso em 10/11/2014.

DEANGELIS, Stephen F. **The Growing Importance of Natural Language Processing.** Wired. Fevereiro, 2014.
http://www.wired.com/2014/02/growing-importance-natural-language-processing/ - Acesso em 30/08/2014

Faça download do Facebook e faça parte da maior rede social do planeta! Techtudo, 2014. http://www.techtudo.com.br/tudo-sobre/facebook.html - Acesso em 16/11/2014

FRANKS, Bill. **Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics**. Hoboken: John Wiley & Sons Inc, 2012

FRIEDMAN, Jerome, HASTIE, Trevor, TIBSHIRANI, Robert. **The elements of statistical learning: Data mining, inference and prediction.** Second Edition, Springer, 2009.

GESENHUES, Amy. Survey: **90% Of Customers Say Buying Decisions Are Influenced By Online Reviews.** Marketing Land. Abril, 2013.
http://marketingland.com/survey-customers-more-frustrated-by-how-long-it-takes-to-resolve-a-customer-service-issue-than-the-resolution-38756 - Acesso em 02/09/2014

Global B2C Ecommerce Sales to Hit $1.5 Trillion This Year Driven by Growth in Emerging Markets - Emarketer, 2014.
http://www.emarketer.com/Article/Global-B2C-Ecommerce-Sales-Hit-15-Trillion-This-Year-Driven-by-Growth-Emerging-Markets/1010575 - Acesso em 31/10/2014

HURWITZ, Judith et al. **Big Data for Dummies.** Hoboken: John Wiley & Sons Inc. 2013

Impact of Big Data in Social Media. Opallios, 2014.
http://www.opallios.com/impact-of-big-data-in-social-media/ - Acesso em 26/10/2014

KAUSHIK, Chetan, MISHRA, Atul. **A Scalable, Lexicon Based Technique for Sentiment Analysis.** International Journal in Foundations of Computer Science &

Technology (IJFCST), Vol.4, No.5, September 2014
http://arxiv.org/abs/1410.2265 - Acesso em 01/11/2014

Ken Starr: Spring 2013 Commencement Address. Baylor, 2013.
http://www.baylor.edu/president/news.php?action=story&story=130317 – Acesso em 08/09/2014

LANEY, Doug. **3D Data Management: Controlling Data Volume, Velocity, and Variety.** Gartner Report, 2001
http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf - Acesso em 30/08/2014

LEE, Lilian, PANG, Bo. **Opinion Mining and Sentiment Analysis.** Foundations and Trends in Information Retrieval, Vol. 2, Nos, 2008.
www.cs.cornell.edu/home/llee/omsa/omsa.pdf - Acesso em 18/09/2014

MCCALLISTER, Erika. **Guide to Protecting the Confidentiality of Personally Identifiable Information.** NIST Special Publication 800-122. 2011.
http://books.google.com.br/books?id=tlTrkXB-f3cC&pg=SA4-PA1&dq=guide+ to+protecting+pii&hl=pt-BR&sa=X&ei=BVoVLHaFcuqNtus gIgI&ved =0C B4Q6 AEw AA#v=onepage&q=guide%20to%20protecting%20pii&f=false - Acesso em 01/11/2014

New IDC Study Uncovers Best Practices in Unlocking the Hidden Value of Information. IDC, 2014.
http://www.idc.com/getdoc.jsp?containerId=prUS24993814 – Acesso em 26/09/2014

PALANISAMY, Prabu, YADAV, Vineet, ELCHURI, Harsha. **Serendio: Simple and Practical lexicon based approach to Sentiment Analysis.** Serendio, 2013.
www.aclweb.org/anthology/S13-2091 - Acesso em 30/10/2014

PORNAI, Xavier. **Rewiring to Tackle Unstructured Data.** Wired, Julho, 2014.
http://www.wired.com/2014/07/rewiring-tackle-unstructured-data/ - Acesso em 05/10/2014

RIJIMENAM, Mark van. **Big Data Governance: Controlling and Handling Data.** Big Data Startups. Agosto, 2013. http://www.bigdata-startups.com/big-data-accountability-part-2/ - Acesso em 05/10/2014

RIJIMENAM, Mark van. **Why The 3V's Are Not Sufficient To Describe Big Data.** Big Data Startups. Agosto, 2013. http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/ - Acesso em 05/10/2014

ROUSE, Margareth. **Big Data analytics.** Tech Target, 2012 http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics - Acesso em 15/10/2014

Sentiment Analysis & Text Analytics Connector. QVSource, 2014. http://www.qvsource.com/Connectors-For-QlikView/Text-Analytics-And-Sentiment-Analysis-In-QlikView - Acesso em 25/08/2014

Social Media Report for Nokia Lumia 920 - Nov. 2013 to Jan 11, 2013. Serendio, 2013. http://www.serendio.com/windows-8-smartphones-nokia-lumia-920-vs-htc-windows-phone-8x/ - Acesso em 27/08/2014

STEWART, Darin. **Big Content: The Unstructured Side of Big Data**. Gartner. Maio, 2013. http://blogs.gartner.com/darin-stewart/2013/05/01/big-content-the-unstructured-side-of-big-data/ - Acesso em 02/09/2014

Structured vs. Unstructured Data. Brightplanet, 2012. http://www.brightplanet.com/2012/06/structured-vs-unstructured-data/ - Acesso em 10/10/2014

TUVERI, F. ANGIONI, M. **A Linguistic Approach to Opinion Mining.** Springer Berlin Heidelberg, 2012

http://publications.crs4.it/pubdocs/2012/TA12d - Acesso em 27/10/2014

Unstructured data: A big deal in big data. Digital Reasoning, 2012. http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf - Acesso em 01/08/2014

What is Big Data? IBM, 2014. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html - Acesso em 27/08/2014

What is WordNet? Princeton University, 2014. http://wordnet.princeton.edu/ - Acesso em 27/10/2014

ZABIN, J. JEFFERIES, A. **Social media monitoring and analysis: Generating consumer insights from online conversation.** Aberdeen Group Benchmark Report, January 2008. http://aberdeen.com/research/4587/ra-consumer-insights-online/content.aspx - Acesso em 08/09/2014