Text Mining em Redes Sociais para Análise de Marketing

Neander de Abreu Prates[1], Sylvio Barbon Junior[1].

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

neanderthalabreu@hotmail.com, @uel.com, jomi@inf.furb.br

Abstract. *The internet has grown in the last twenty years, so social networking has had a huge growth, especially in the last ten years, companies seek information on social networks to improve sales and acceptance of its products through analysis of consumer profiles that post it on his own, so have researched and developed tools in order to analyze the data for better analysis using data mining and text mining, so this work will be surveyed and compared tools and text mining techniques to acquire meaningful content that can be used in marketing and advertising.*

Resumo. *A internet cresceu muito nos últimos vinte anos, assim também as redes sociais tem tido um grande crescimento, principalmente nos últimos dez anos, empresas buscam nas redes sociais informações para melhorar a venda e aceitação de seus produtos através de análises de perfis dos consumidores que as postam nela por vontade própria, sendo assim têm se pesquisado e desenvolvido ferramentas de forma a analisar os dados para melhores análises utilizando data mining e text mining, portanto neste trabalho irão ser pesquisadas e comparadas ferramentas e técnicas de text mining para adquirir conteúdos significativos que possa ser utilizadas na área de marketing e publicidade.*

1. Introdução

Houve um grande crescimento no número de usuários de internet de forma muito rápida, nesses últimos vinte anos, de forma a mudar muitos costumes e padrões de vida de uma boa parte da população global. Tendo em vista isso, o comércio busca se adaptar a essas novas tendências, sendo uma das principais tendências importantes para o comércio a popularização das redes sociais.

A grande quantidade de dados na internet levou a necessidade de técnicas de data mining, para as mais diversas áreas, destacando-se junto a elas as técnicas de text mining, na qual são usadas em diversas áreas, porém com um foco bem relevante junto às redes sociais para análise de reviews de produtos, agrupamento e análises de perfis, etc.

2. Fundamentação Teórico-Metodológica e Estado da Arte

2.1. Crescimento da internet

A internet está cada vez mais presente no cotidiano das pessoas em diversas áreas das vidas delas. Desde os primeiros navegadores populares no início da década de 90,

Mosaic (em 1993) e Netscape (em 1994), o número de pessoas com internet cresceu de uma forma impressionante, como se pode observar nas tabelas 1 e 2 retiradas do site "Internet World Stats"[1], mesmo o crescimento com relação à porcentagem de pessoas com acesso a internet é algo de se admirar como se pode ver na tabela 1, pois mais de cem por cento de pessoas com acesso a internet não passará, logo era de se esperar que a penetração da internet fosse diminuir gradualmente, apesar que isso ocorre nos últimos anos nos países mais desenvolvidos, porém os países em desenvolvimento tiveram um grande crescimento nesses mesmos anos na qual fez com que o crescimento continuasse de forma elevada como se pode perceber na tabela 2.

Tabela 1. History and Growth of the Internet from 1995 till Today[2].

DATE	NUMBER OF USERS	% WORLD POPULATION	INFORMATION SOURCE
December, 1995	16 millions	0.4 %	IDC
December, 1996	36 millions	0.9 %	IDC
December, 1997	70 millions	1.7 %	IDC
December, 1998	147 millions	3.6 %	C.I. Almanac
December, 1999	248 millions	4.1 %	Nua Ltd.
March, 2000	304 millions	5.0 %	Nua Ltd.
July, 2000	359 millions	5.9 %	Nua Ltd.
December, 2000	361 millions	5.8 %	Internet World Stats
March, 2001	458 millions	7.6 %	Nua Ltd.
June, 2001	479 millions	7.9 %	Nua Ltd.
August, 2001	513 millions	8.6 %	Nua Ltd.
April, 2002	558 millions	8.6 %	Internet World Stats
July, 2002	569 millions	9.1 %	Internet World Stats
September, 2002	587 millions	9.4 %	Internet World Stats
March, 2003	608 millions	9.7 %	Internet World Stats
September, 2003	677 millions	10.6 %	Internet World Stats
October, 2003	682 millions	10.7 %	Internet World Stats
December, 2003	719 millions	11.1 %	Internet World Stats
February, 2004	745 millions	11.5 %	Internet World Stats
May, 2004	757 millions	11.7 %	Internet World Stats
October, 2004	812 millions	12.7 %	Internet World Stats
December, 2004	817 millions	12.7 %	Internet World Stats

[1] http://www.internetworldstats.com/

[2] http://www.internetworldstats.com/emarketing.htm

March, 2005	888 millions	13.9 %	Internet World Stats
June, 2005	938 millions	14.6 %	Internet World Stats
September, 2005	957 millions	14.9 %	Internet World Stats
November, 2005	972 millions	15.2 %	Internet World Stats
December, 2005	1,018 millions	15.7 %	Internet World Stats
March, 2006	1,023 millions	15.7 %	Internet World Stats
June, 2006	1,043 millions	16.0 %	Internet World Stats
Sept, 2006	1,086 millions	16.7 %	Internet World Stats
Dec, 2006	1,093 millions	16.7 %	Internet World Stats
Mar, 2007	1,129 millions	17.2 %	Internet World Stats
June, 2007	1,173 millions	17.8 %	Internet World Stats
Sept, 2007	1,245 millions	18.9 %	Internet World Stats
Dec, 2007	1,319 millions	20.0 %	Internet World Stats
Mar, 2008	1,407 millions	21.1 %	Internet World Stats
June, 2008	1,463 millions	21.9 %	Internet World Stats
Sept, 2008	1,504 millions	22.5 %	Internet World Stats
Dec, 2008	1,574 millions	23.5 %	Internet World Stats
Mar, 2009	1,596 millions	23.8 %	Internet World Stats
June, 2009	1,669 millions	24.7 %	Internet World Stats
Sept, 2009	1,734 millions	25.6 %	Internet World Stats
Dec, 2009	1,802 millions	26.6 %	Internet World Stats
June, 2010	1,966 millions	28.7 %	Internet World Stats
Sept, 2010	1,971 millions	28.8 %	Internet World Stats
Mar, 2011	2,095 millions	30.2 %	Internet World Stats
Jun, 2011	2,110 millions	30.4 %	Internet World Stats
Sept, 2011	2,180 millions	31.5 %	Internet World Stats
Dec, 2011	2,267 millions	32.7 %	Internet World Stats
Mar, 2012	2,336 millions	33.3 %	Internet World Stats
June, 2012	**2,405 millions**	**34.3 %**	Internet World Stats

Tabela 2. WORLD INTERNET USAGE AND POPULATION STATISTICS

June 30, 2012[1].

[1] http://www.internetworldstats.com/stats.htm

WORLD INTERNET USAGE AND POPULATION STATISTICS June 30, 2012						
World Regions	Population (2012 Est.)	Internet Users Dec. 31, 2000	Internet Users Latest Data	Penetration (% Population)	Growth 2000-2012	Users % of Table
Africa	1,073,380,925	4,514,400	**167,335,676**	15.6 %	3,606.7 %	7.0 %
Asia	3,922,066,987	114,304,000	**1,076,681,059**	27.5 %	841.9 %	44.8 %
Europe	820,918,446	105,096,093	**518,512,109**	63.2 %	393.4 %	21.5 %
Middle East	223,608,203	3,284,800	**90,000,455**	40.2 %	2,639.9 %	3.7 %
North America	348,280,154	108,096,800	**273,785,413**	78.6 %	153.3 %	11.4 %
Latin America / Caribbean	593,688,638	18,068,919	**254,915,745**	42.9 %	1,310.8 %	10.6 %
Oceania / Australia	35,903,569	7,620,480	**24,287,919**	67.6 %	218.7 %	1.0 %
WORLD TOTAL	**7,017,846,922**	**360,985,492**	**2,405,518,376**	**34.3 %**	**566.4 %**	**100.0 %**

2.2. Redes Sociais Onlines

A ideia geral de redes sociais não é algo novo, pois o já dizia Aristóteles há 400 anos antes de Cristo que o ser humano necessita procurar e criar comunidades, contudo com a popularização da World Wide Web e o desenvolvimento de tecnologias da informação, as redes sociais atingiram uma nova dimensão com o surgimento das redes sociais onlines [1].

Pode-se dizer que o surgimento das redes sociais online se deu em 1997, desde então passaram por alguns ciclos e são esses: O começo das redes sociais onlines (1997-2002); O crescimento das redes sociais onlines e o aumento de sua popularidade (2003 – 2009); Um fenômeno global (2010 – até o presente). [1][13].

Ao observar tanto a história quanto crescimento da internet e das redes sociais, principalmente com relação à proporção de pessoas com essas tecnologias ao comparar com a totalidade da população mundial, se pode inferir a ideia de que se está passando por uma revolução tecnológica/social, pois a popularização de ambas não chega há 20 anos, sendo assim ao levar em conta que as duas tecnologias juntas colaboram com a globalização cultural tem-se portanto uma nova mentalidade dos indivíduos dessa geração sendo necessárias novas abordagens nas diversas áreas de suas vidas como: social, comercial, educacional, etc.

Consumidores estão cada vez mais utilizando tecnologias como ferramentas para compras, de forma que as redes sociais se tornaram um lugar de virtual para consumidores compartilharem as informações de forma voluntária e pessoal possuindo assim agora os meios para comunicar suas opiniões sobre produtos e empresas para outros consumidores como eles próprios em um ponto crítico no ciclo de vida, o inicial [4]. Com relação as pessoas que viajam, não só em agências de viagens, as redes sociais onlines estão colaborando positivamente como nas trocas de informações, compartilhamento de experiências, comentários, opiniões, análises de hotéis, sugestões

de férias e pacotes de viagens, sendo que essas informações, análises e compartilhamento de experiências, influenciam em mais de 10 bilhões de dólares em compras online neste setor de viagens todos os anos [14]. Ao se tratar de produtos de supermercados, por exemplo, as redes sociais são uma poderosa ferramenta que pode ser usada como forma de análise para prevenção de uma visão ruim para o produto por parte dos consumidores [2].

Os profissionais de marketing buscam cada vez mais aprender sobre comunidades virtuais para assim influenciar as escolhas dos membros, disseminar rapidamente conhecimento e percepções sobre seus produtos, oportunidades para se relacionar melhor com seus clientes, etc [3].

As principais redes sociais onlines hoje de acordo com o ranking do site Alexa[1] são: Facebook; Twitter; Linked In; Google+. As redes sociais onlines se diversificam por focos normalmente, no caso dessas redes mais populares se tem como slogan: Facebook - No Facebook você pode se conectar e compartilhar o que quiser com quem é importante em sua vida[2]; Twitter - Descubra o que está acontecendo, agora mesmo, com as pessoas e organizações que lhe interessam[3]; Linked In - Mais de 200 milhões de profissionais utilizam o LinkedIn para compartilhar informações, ideias e oportunidades[4]; Google+ - Compartilhe apenas com as pessoas certas, Converse cara a cara a cara, Não perca nenhuma foto[5].

2.3. Data mining

Em virtude da quantidade e qualidade de dados hoje disponíveis no mercado, departamentos de data mining ganharam importância para tarefas como predição de churn e gestão de campanha para aumentar a conscientização sobre produto ou marca [5].

Pode-se dizer que data mining é um exercício interdisciplinar, na qual atua nas áreas de estatística, tecnologias de banco de dados, aprendizagem de máquina, reconhecimento de padrões, inteligência artificial, visualização, entre outras áreas [15]. Data mining é processo de descoberta de padrões de dados de forma automática, de tal forma que os padrões devem levar a alguma vantagem relevante, normalmente uma vantagem econômica [6].

É muitas vezes definido, o data mining, no contexto mais amplo de descoberta de conhecimento em bancos de dados, conhecido como KDD (do inglês *knowledge discovery in databases*), as fases do processo do KDD são: seleção de dados; pré-processamento; transformação quando necessária; data mining para extrair relacionamentos e padrões; interpretação e avaliação das estruturas descobertas [15].

---

1 http://www.alexa.com/topsites/category/Computers/Internet/On_the_Web/Online_Communities/Social_Networking

2 https://www.facebook.com/

3 https://twitter.com/

4 http://www.linkedin.com/

5 www.google.com/+

Onde o processo de procura de relacionamentos envolve passos como: determinação da natureza e estrutura da representação a ser usada; decisão de quantificar e comparar o quão bem se encaixa nos dados diferentes representações escolhendo uma função "score"; escolher um algoritmo de processo para aperfeiçoar a função "score"; decidir quais princípios de gerenciamento de dados são requeridos para programar algoritmos eficientemente [15].

Algumas das principais técnicas de data mining são: modelagem preditiva; modelagem descritiva; mineração de padrões; detecção de anomalias [16]. Modelagem preditiva é o processo de modelagem na qual o modelo tenta prever a melhor probabilidade de um resultado, como exemplos têm: naive Bayes; árvores de decisão; regressão logística. A modelagem descritiva é o processo de modelagem na qual se descreve o corrente estado, seus objetos e suas relações, como exemplos: k-means; redes neurais; spectral clustering. Mineração de padrões é um método de achar padrões existentes nos dados, como exemplos: regras de associação; graph mining; detecção de anomalias. Detecção de anomalias serve para detectar padrões que se destacam muito dos outros parecendo que não pertence ao mesmo conjunto de dados, como exemplos de técnicas populares para detecção de anomalias se têm: kNN(k-nearest neighbor); análise de cluster baseado em outlier detection; máquina de vetores de suporte de uma classe; replicador de redes neurais.

2.4. Text mining

O text mining assim como o data mining procura extrair informações uteis das fontes de dados através de exploração e identificação de dados interessantes, porém as fontes de dados do text mining são conjuntos de documentos e os padrões de interesse são encontrados em textos não estruturados ao invés de registros formalizados nos banco de dados, mesmo assim o text mining normalmente possui muitas semelhanças em termos de arquitetura de alto nível [7].

Ao levar em conta que o text mining trabalha com textos de forma não estruturada, pode se usar processamento de linguagem natural, pois o mesmo tenta, de forma grosseira, descobrir quem fez o quê a quem, quando, como, onde e por quê, utilizando de conceitos linguísticos, como parte da fala e estrutura gramatical, a ter que lidar também com anáfora e ambiguidades [8].

Sistemas de text mining normalmente apresentam processos semelhantes aos do data mining clássico e podem ser divididos em quatro áreas principais: preprocessing tasks; core mining operations; presentation layer components; refinement techniques [7].

Preprocessing tasks, como o próprio nome diz, são tarefas de pré-processamento, na qual os textos são preparados para os processos seguintes. Apesar de o processo de text mining ser parecido com o de data mining, nessa parte o text mining é bem mais custoso pelo fato já citado de trabalhar com dados não estruturados em forma de texto, pois os dados devem ter tratamentos diferenciados para serem submetidos aos algoritmos de mineração [17].

Core mining operations são as operações mais importantes que incluemos aldorigmos de: descobrimento de padrões; análise de tendência; descoberta de conhecimento [7].

Presentation layer components, incluem a interface gráfica e funcionalidade de navegação padrão bem como o acesso à linguagem de consulta, de forma geral semelhante a parte de visualização e avaliação do data mining, para assim criar ou modificar clusters conceituais a fim de classificar perfis para padrões ou conceitos específicos [7].

Refinement techniques, conhecidas como pós-processamento, são métodos que filtram informações redundantes a fim de suprimir, ordenar, podar, generalizar e agrupar abordagens destinadas a descoberta de otimização.

Devido às características de o text mining ser mais específico que o data mining e trabalhar com dados não estruturados, o mesmo vem sendo muito utilizado em diversas áreas de aplicação, descobrindo ou gerando novos padrões e/ou conhecimento em diversas áreas. Como exemplos de sua utilização valem citar: o seu uso para avaliar estudantes em ambiente online de interação em streaming de vídeo ao vivo, na qual se obteve resultados com padrões interessantes na interação do estudante com outros estudantes e instrutor, entre outras contribuições [9]; mesmo usado combinando com um modelo de escolha simples podem chegar a resultados significantes [18]; uso de text mining para superar dificuldades que envolvem a extração e quantificação de dados que os consumidores geram [19]; análise de opinião de consumidor em redes sociais onlines [12]; entre outros [10], [11].

3. Objetivos

Este trabalho tem por objetivo pesquisar e comparar ferramentas e técnicas de text mining de mensagens do Twitter na qual terá como objetivo recuperar as informações e reconhecer um padrão associado a marcas ou eventos com possível validação de dados providos do LinkedIn, para posterior estudo relacionado ao Marketing e Publicidade

4. Procedimentos metodológicos/Métodos e técnicas

Serão utilizadas ferramentas, técnicas e métodos de text mining e data mining. A princípio serão analisadas as técnicas LSI[7], NLP e non-NLP[8].

5. Cronograma de Execução

.

6. Contribuições e/ou Resultados esperados

Espera-se que este trabalho possa contribuir de forma a apresentar um comparativo que possa a vir ser útil para se saber quais ferramentas, métodos e técnicas podem ser mais eficazes entre as comparadas.

7. Espaço para assinaturas

Londrina, 15/04/2013.

_____ _____

Aluno Orientador.

Referências

[1] J. Heidemann, M. Klier, F. Probst, Online social networks: A survey of a global phenomenon, in: Computer Networks, (2012).

[2] R. D. Groot, Consumers don't play dice, influence of social networks and advertisements, Elsevier B.V. (2005).

[3] U. M. Dholakia, R. P. Bagozzi, L. K. Pearo, A social influence model of consumer participation in network- and small-group-based virtual communities, in: Intern. J. of Research in Marketing, (2003).

[4] S. Pookulangara, K. Koesler, Cultural influence on consumers' usage of social networks and its' impact on online purchase intentions, in: Journal of Retailing and Consumer Services, (2011).

[5] C. Kiss, M. Bichler, Identification of influencers—Measuring influence in customer networks, in: Decision Support Systems, (2008).

[6] I. H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, 2nd ed, Morgan Kaufmann Publishers is an imprint of Elsevier, (2005).

[7] R. Feldman, J. Sanger, The text mining handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press, (2007).

[8] A. Kao, S. R. Poteet, Natural Language Processing and Text Mining, Springer-Verlag London Limited, (2007).

[9] W. He, Examining students' online interaction in a live video streaming environment using data mining and text mining, in: Computers in Human Behavior, (2012).

[10] W. He, Improving user experience with case-based reasoning systems using text mining and Web 2.0, in: Expert Systems with Applications, (2012).

[11] W. He, S. Zha, L. Li, Social media competitive analysis and text mining: A case study in the pizza industry, in: International Journal of Information Management, (2013).

[12] M. M. Mostafa, More than words: Social networks' text mining for consumer brand sentiments, in: Expert Systems with Applications, (2013).

[13] D. M. Boyd, Social Network Sites: Definition, History, and Scholarship, in: Journal of Computer-Mediated Communication, (2008).

[14] K.K. Nusair, A. Bilgihan, F. Okumus, C. Cobanoglu, Generation Y travelers' commitment to online social network websites, in: Tourism Management, (2012).

[15] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, The MIT Press, (2001);

[16] Encyclopædia Britannica, (2012)

[17] I.M. de Oliveira, Estudo de uma Metodologia de Mineração de Textos Científicos em Língua Portuguesa, Dissertação(mestrado), UFRJ/COPPE, (2009).

[18] N. Archak, A. Ghose, P. Ipeirotis, Deriving the Pricing Power of Product Features by Mining Consumer Reviews, in: Management Science, (2011).

[19] O. Netzer, R. Feldman, J. Goldenberg, M. Fresko, Mine Your Own Business: Market-Structure Surveillance Through Text Mining, in Marketing Science, (2012).