Entendendo a Twitteresfera Brasileira

Fabrício Benevenuto
UFOP - Ouro Preto

Diego Silveira
UFOP - Ouro Preto

Thalisson Oliveira
UFOP - Ouro Preto

Leonardo Bombonato
UFOP - Ouro Preto

Reinaldo Fortes
UFOP - Ouro Preto

Álvaro Pereira Jr.
UFOP - Ouro Preto

ABSTRACT

Twitter has been constantly growing as an important system where users discuss about everything, expressing opinions, political view, sexual orientation, and even their mood like happiness or sadness. Social networks are pointed as places where users influence and are influenced by others, being a perfect environment for word-of-mouth marketing, advertisement, and political campaigns. In order to offer a better understand of the use of Twitter in Brazil, this work provides a wide characterization of Brazilian users in Twitter as well as a deep understand of the content posted by Brazilians in Twitter. We correlate Brazilian demographic data with geographic data from the Twitter users to show that some Brazilian states are underestimated in Twitter. Additionally, we characterize the different linguistic patterns adopted, we analyzed the most propagated URLs, and we identified the most influential Brazilian users in Twitter on each Brazilian region.

RESUMO

O Twitter vem constantemente crescendo como um importante sistema onde usuários discutem sobre tudo, expressando opiniões, visão política, orientação sexual e até mesmo humor, como felicidade ou tristeza. Redes sociais são apontadas como locais onde usuários influenciam e são influenciados por outros sendo, portanto, ambientes perfeitos para a realização de marketing boca-a-boca, propagandas e campanhas políticas. Com o intuito de oferecer entendimento sobre o uso do Twitter no Brasil, este trabalho provê uma ampla caracterização dos usuários brasileiros no Twitter e do conteúdo postado por esses usuários. Nós correlacionamos dados demográficos brasileiros com dados da localização dos usuários do Twitter para mostrar que alguns estados brasileiros estão subestimados nesse sistema. Além disso, nós caracterizamos os diferentes padrões linguísticos adotados, analisamos as URLs mais propagadas, e identificamos os usuários brasileiros mais influentes no Twitter em cada região brasileira.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; H.4.m [**Information Systems Applications**]: Miscellaneous; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Experimentation, Measurement

Keywords

Twitter, Microblog, Web 2.0

1. INTRODUÇÃO

Desde seu início a Internet tem sido palco de uma série de novas aplicações incluindo a WWW e email. Atualmente, a Web tem recebido uma nova onda de aplicações associadas ao crescimento e proliferação das redes sociais online. Vários desses sistemas surgiram, incluindo redes de profissionais (ex. LinkedIn), redes de amizade (ex. MySpace, Facebook e Orkut), e redes voltadas para o compartilhamento de algum tipo específico de conteúdo como mensagens curtas (ex. Twitter), diários e blogs (ex. LiveJournal), fotos (ex. Flickr) e vídeos (ex. YouTube).

Em particular, o Twitter surgiu como uma plataforma social na Web onde usuários discutem sobre tudo, expressando opiniões, visão política, orientação sexual e até mesmo conceitos vagos como humor e felicidade. Além disso, o Twitter tem atraído a popularidade de jornais e celebridades, conectados a usuários comuns através de elos de seguidores e seguidos pelos quais usuários podem trocar informações. Estimativas sugerem que, atualmente, os 200 milhões de usuários ativos do Twitter postam cerca de 150 milhões de tweets (mensagens) contendo cerca de 23 milhões de URLs diariamente [5, 31]. Segundo a comScore, o Brasil é o terceiro país onde o Twitter mais cresce atualmente [13].

Com tanta popularidade, o Twitter se tornou uma plataforma para a descoberta de informação em tempo real e vem se mostrando uma importante ferramenta para a disseminação de notícias de última hora, tais como revoluções e desastres [18]. Além disso, Google e Bing já indexam tweets públicos como forma de prover busca por informação em tempo real [29]. Como consequência, o Twitter vem sendo alvo de diversos sistemas relacionados à descoberta de informações em tempo real, como a identificação automática de terremotos a partir da análise de tweets [25] e extração de opniões de eleitores em relação a candidatos a cargos políticos [30]. No Brasil, o Twitter foi recentemente utilizado como fonte de dados para o monitoramento Web das eleições presidenciais de 2010 [26] e como fonte de dados para o monitoramento de casos de dengue no Brasil [15].

Apesar da comprovada aplicabilidade desses estudos e ferramentas, pouco se sabe sobre o uso do Twitter no Brasil. Como exemplo, não sabemos se um determinado estado brasileiro está sub ou super representado no Twitter em relação à sua demografia real, o que poderia ter profundo impacto nos resultados reportados por ferramentas brasileiras que monitoram eventos no Twitter. Pouco se sabe sobre o uso do Twitter no Brasil com relação aos temas discutidos, usuários mais influentes, sites mais divulgados no Twitter, etc. Neste cenário, apesar de haver um grande interesse de empresas brasileiras em divulgar produtos e realizar propaganda boca-a-boca, pouco se sabe sobre quais estratégias realmente seriam efetivas. Nesse contexto, acreditamos que uma ampla análise da *twitteresfera* poderia oferecer respostas para várias dessas ques-

tões.

Este trabalho visa dar um primeiro passo nessa direção de forma a realizar uma ampla análise do conteúdo postado por brasileiros no Twitter. Para isso, utilizamos uma coleção que possui 55 milhões de usuários do Twitter, todos os elos de seguidores e seguidos (grafo com quase 2 bilhões de elos) e todos os tweets postados por esses usuários (1.8 bilhões de tweets). Dado que o Brasil é um país tão extenso e diversificado, a localização geográfica das pessoas pode ter um importante papel nos assuntos discutidos, nos temas compartilhados e até mesmo na felicidade e no humor das pessoas. Sendo assim, identificamos a localização geográfica de usuários do twitter para realizar uma análise sobre seu uso nas diferentes regiões brasileiras.

Nossas análises revelam que alguns estados brasileiros estão subestimados no Twitter em relação a dados demográficos, sendo que características sócio-econômicas desses estados parecem influir no número de usuários do Twitter. Além disso, nossas análises ainda revelam padrões linguísticos associados ao uso de microblogs no Brasil, identificando não só os termos mais frequentes nos tweets postados por brasileiros, mas também aqueles associados a contextos felizes e tristes. Nós ainda analisamos as URLs postadas por brasileiros no Twitter e comparamos a popularidade dos sítios web mais populares no Twitter com a popularidade desse sítios na Web de maneira geral. Por último, foram estudadas diferentes abordagens para identificar os usuários brasileiros mais influentes no Twitter e foi construído um ranking de usuários influentes para cada região do Brasil. Acreditamos que a metodologia empregada nas análises presentes nesse trabalho pode guiar futuros trabalhos e ferramentas que pretendem realizar algum tipo de monitoramento de dados extraídos de redes sociais online, em especial o Twitter.

O restante do trabalho está organizado da seguinte forma. A seção 2 descreve trabalhos relacionados. A seção 3 descreve a estratégia adotada para a coleta de dados da twitteresfera brasileira. A seção 4 apresenta dados da demografia dos usuários brasileiros do Twitter. A seção 5 apresenta os termos mais frequentes utilizados pelos brasileiros, bem como uma análise da polaridade desses termos. Em seguida, a seção 6 analisa os domínios das URLs mais populares propagadas no Twitter. A seção 7 indentifica os usuários brasileiros mais influentes de cada região brasileira. Finalmente, a seção 8 oferece conclusões e direções para trabalhos futuros.

2. TRABALHOS RELACIONADOS

O Twitter tem sido alvo de um grande número de trabalhos relacionados às suas características topológicas e aspectos da interação de seus usuários. Kwak e colaboradores [20] estudaram a topologia do Twitter, encontrando distribuições que seguem leis de potência para o número de seguidores e seguidos, um diâmetro curto e baixa reciprocidade. Eles ainda estudaram abordagens para ranquear usuários baseadas no uso do Pagerank calculado sobre a estrutura de seguidores e seguidos no Twitter. Mais recentemente, Cha e colaboradores mostraram que usuários influentes não são necessariamente os usuários mais seguidos do Twitter, o que significa que a topologia do Twitter pode não ser suficiente para capturar a influência dos usuários. Scellato e colaboradores [27] estudaram como informação propagada no Twitter pode ser explorada para melhorar sistemas de caches de arquivos multimídia em redes de distribuição de conteúdo (*CDNs - Content Distribution Networks*). Como resultados eles mostraram que o número de acertos em cache pode ser aumentado em relação a políticas de caches que não levam em consideração informações geográficas e sociais. Rodrigues e colaboradores [23] provêem uma série de análises sobre os padrões de propagação de informação entre os usuários do Twitter. Além de quantificar o aumento de audiência de uma informação que retweets podem causar, eles identificaram características típicas da estrutura das árvores de propagação de informação nesse sistema. Finalmente, a existência de phishing [12] e spammers no Twitter vem sendo constantemente reportada [21] e alguns esforços recentes propuseram abordagens para a detecção de spammers no Twitter [8, 17].

Alguns trabalhos recentes analisaram a repercussão de eventos específicos no Twitter. Em particular, Sakaki e colaboradores [25] mostraram o poder da informação disponibilizada em tempo real nas redes sociais online propondo um mecanismo para detecção de ocorrências de terremotos baseado em monitoramento do Twitter. A abordagem, que consiste em simplesmente identificar tweets relacionados a terremotos por região, foi capaz de enviar alertas sobre terremotos mais rapidamente do que agências meteorológicas. Mais recentemente, Tumasjan e colaboradores [30] mostraram que opiniões identificadas em tweets relacionados à eleição federal alemã foi capaz de refletir o sentimento político registrado fora das redes sociais. Mais recentemente, uma análise geográfica e temporal sobre dengue no Twitter foi realizada por Gomide e seus colaboradores [15]. No cenário nacional, um exemplo de ferramenta brasileira que estuda a repercussão de um fenômenos na Web é o *Observatório da Web* [26]). O observatório da Web já monitorou as eleições de 2010 na Web em diferentes mídias sociais, além de monitorar a repercurssão dos jogos da copa do mundo de 2010 e comentários sobre dengue na Web. De forma complementar a esses trabalhos, nosso trabalho oferece uma visão geral do uso do Twitter no Brasil e não analisa somente a repercussão de um determinado evento.

3. METODOLOGIA DE MEDIÇÃO

Em um passado recente, redes sociais eram um domínio de sociólogos e antropólogos, que utilizavam pesquisas e entrevistas com pequenos grupos de usuários como ferramentas de coleta de dados [32]. Com o surgimento das redes sociais online, a obtenção de dados reais em larga escala se tornou possível, e pesquisadores de diversas áreas da computação começaram a realizar coletas de dados. Entretanto, a coleta de conteúdo criado por usuários envolve um grande número de desafios. Esta seção apresenta a metodologia adotada para a obtenção dos dados utilizados em nossas medições. Inicialmente, a seção 3.1 descreve a estratégia adotada para coletar dados do Twitter. Em seguida, a seção 3.2 discute a coleta da localização geográfica de usuários do Twitter, de forma a permitir a realização de análises restritas à twitteresfera brasileira. A seção 3.3 descreve o processo utilizado para identificar tweets felizes, tristes ou neutros. Finalmente, a seção 3.4 discute possíveis limitações dos dados obtidos.

3.1 Coleção de dados do Twitter

Esta seção descreve brevemente a estratégia utilizada para construir uma coleção de dados do Twitter. Podemos acessar cada usuário do sistema, visto que o Twitter atribui um identificador (ID) numérico e sequencial para cada usuário cadastrado [7]. Como novos usuários recebem um identificador sequencial, podemos percorrer todos os IDs, sem ter que verificar a lista de amigos desses usuários em busca de novos IDs para coletar.

Recentemente, nós realizamos uma coleta do Twitter seguindo essa estratégia. Foi solicitado aos administradores do Twitter a permissão para realizar uma coleta em larga escala. Em resposta, eles adicionaram os endereços IPs de 58 máquinas sob nosso controle em uma lista branca, com permissão para coletar dados. Cada uma das 58 máquinas, localizadas no *Max Planck Institute for Software*

Systems (MPI-SWS), na Alemanha[1], teve permissão para coletar dados a uma taxa máxima de 20 mil requisições por hora. Utilizando a API do Twitter, nosso coletor investigou todos os 80 milhões de IDs de forma seqüencial, coletando todas as informações públicas sobre esses usuários, bem como seus elos de seguidores e seguidos e todos os seus tweets. Dos 80 milhões de contas inspecionadas, encontramos cerca 55 milhões em uso. Isso acontece porque o Twitter apaga contas inativas por um período maior do que 6 meses. No total, coletamos **54.981.152** milhões de usuários, todos os elos de seguidores e seguidos, que correspondem a **1.963.263.821** elos únicos e todos os tweets postados por esses usuários, em um total de **1.755.925.520** tweets. Ao inspecionar as listas de seguidores e seguidos coletadas, não encontramos nenhum identificador acima dos 80 milhões inspecionados, sugerindo que coletamos todos os usuários do sistema na época. Para uma descrição mais detalhada desses dados e das técnicas empregadas para a realização da coleta desses dados, recomendamos ao leitor as seguintes referências [11, 7].

3.2 Coleta de informações geográficas

Para identificarmos usuários e tweets brasileiros precisamos primeiramente identificar a localização e origem dos tweets nos dados coletados. A localização dos usuários do Twitter coletados aparece na forma de texto livre e frequemente contém localizações inválidas, tais como "Marte" ou "minha casa". Além disso, é muito comum usuários abreviarem localizações como BH para Belo Horizonte, Sampa para São Paulo, e Floripa para Florianópolis. A API do Twitter provê a latitude e a longitude apenas dos usuários que utilizam aparelhos móveis com serviço GPS e permitem o compartilhamento desse tipo de informação. Esses usuários correspondem a uma fração muito pequena dos usuários, especialmente no período em que os dados foram coletados.

Como forma de realizar uma estratégia sistemática de se obter a localização dos usuários, foi utilizada a API do Google Geocoding para traduzir endereços em formato texto em coordenadas geográficas (latitute e longitute). A API do Google Geocoding é utilizada por diversos serviços do Google que utilizam mapas e tem como vantagem identificar corretamente localizações abreviadas, como BH, Sampa e Floripa. Essa API provê um mecanismo direto de se obter as coordenadas geográficas, o país, o estado e a cidade a partir de uma localização fornecida em texto através de uma requisição HTTP [16]. Nos casos em que o usuário forneceu diretamente coordenadas geográficas através do uso de aparelhos móveis com GPS, o Google Geocoding foi utilizado apenas para buscar o nome da cidade e do país desses usuários.

É importante ressaltar que a API do Google Geocoding limita o número máximo de requisições por IP a 200 requisições por hora. Essa taxa torna difícil a obtenção da localização de todos os usuários coletados que correspondem a quase 55 milhões de usuários. Sendo assim, nós restringimos a coleta aos usuários cujo fuso horário fosse um dos adotados no Brasil. Além disso, foi implementado um sistema de cache de localizações, de forma a evitar que duas localizações idênticas fossem requisitas à API do google Geocoding. Dos 923.232 usuários investigados, foram identificados 198.714 usuários com localização detalhada no nível de cidades. No total esses usuários postaram 29.013.854 tweets. Esses usuários e tweets correspondem ao conjunto de dados que analisaremos nas próximas seções.

[1]Esta coleta foi realizada pelo primeiro autor desse trabalho durante uma visita ao MPI-SWS

Categoria	Emoticons	Tweets
Positivo	:-) :D :) =D =)	2.244.628
Negativo	:-(:(D:): T.T =\~ :\~ :'(=[T_T :'-(:,( ]=)-: :,-( ),: :-'(=-[:-,(	468.903
Neutro		26.300.323

Tabela 1: Categorias de tweets de acordo com a presença de emoticons

3.3 Identificação de sentimentos

Emoticons representam uma excelente forma de capturar emoções expressas no texto, pois eles capturam a emoção do escritor que inclui uma expressão fácil utilizando caracteres ASCII. Nós definimos três categorias para classificar tweets de acordo com a presença de emoticons: positivo, negativo ou neutro. A tabela 1 apresenta os emoticons utilizados, o número de tweets em cada categoria. Note que nós classificamos somente os tweets com localização geográfica, que são os que utilizaremos em nossas análises.

3.4 Limitação dos dados

Apesar do conjunto de dados coletado nos permitir a oportunidade única de estudar as características da twitteresfera brasileira, esses dados possuem algumas limitações. Primeiro, nossa análise é restrita ao período coletado do Twitter, que contém dados desde a criação do Twitter em 2006 até Julho de 2009, que é um período no qual o Twitter era muito mais popular nos EUA. Segundo, para examinar a localização geográfica dos usuários, nós consideramos que cada usuário possui uma única localização fixa, desconsiderando que usuários podem postar tweets de diferentes localidades. Finalmente, 21,5% dos usuários investigados tiveram suas localizações identificadas.

4. USUÁRIOS E TWEETS POR REGIÃO

Nesta seção apresentamos uma análise do número de usuários do Twitter com relação à população de cada estado e à proporção da população com acesso à internet, a partir de dados do Instituto Brasileiro de Geografia e Estatística (IBGE) [2]. A Tabela 2 apresenta o número de usuários do Twitter, o número de tweets, a população em 2007 e o percentual da população com acesso à internet (2008), para cada um dos estados brasileiros e agregados por região. As duas últimas colunas apresentam, respectivamente, a média de tweets enviados por usuário para cada estado (usuários ativos: razão entre o número de tweets e o número de usuários), e o percentual da população que usa o Twitter (razão entre o número de usuários e a população de cada estado).

Observando o percentual da população que usa o Twitter, vemos que as unidades da federação que mais usam o Twitter são Distrito Federal, São Paulo e Rio de Janeiro, com 0,21% da população, seguidos pelos estados do Sul (Santa Catarina, Paraná e Rio Grande do Sul), todos com 0,11%. Em geral, o Norte é a região que menos usa o Twitter, seguido pela região Nordeste e logo pela região Centroeste.

Observando a coluna com a média de tweets enviados por usuário, de forma geral vemos que as unidades da federação com mais usuários em proporção à população são também as unidades com os usuários mais ativos. Este dado pode ser consequência do próprio fato do uso ser maior entre as pessoas no estado, uma vez que há influencia do meio social que a pessoa está inserida no comportamento da pessoa (neste caso o comportamento é usar de forma ativa o Twitter como ferramenta social). Apesar dessa observação geral, ela não é uma regra. Há estados como Pernambuco, Ceará, Alagoas, Piauí e Espírito Santo que possuem alto índice de usuários ativos, mas poucos usuários em proporção.

Observamos que **não** há uma correlação entre o número de usuários e a população dos estados. Por exemplo, vários estados da

Estado	Número de usuários	Número de Tweets	População 2007	Percentagem do uso de internet 2008	Usuários ativos	Perc. da população que usa o Twitter
Rondônia	0	0	1.453.756	31,0	0,0	0,000
Tocantins	596	72.500	1.243.627	31,3	121,6	0,047
Amapá	267	31.066	587.311	31,3	116,4	0,045
Pará	1.814	206.048	3.641.395	23,9	113,6	0,049
Amazonas	935	120.375	3.221.939	30,2	128,7	0,029
Roraima	59	1.401	7.065.573	35,3	23,7	0,001
Acre	304	18.139	655.385	29,4	59,7	0,046
Região Norte	**3.975**	**449.529**	**17.868.986**	**30,3**	**113,1**	**0,022**
Alagoas	809	113.172	3.037.103	17,8	139,9	0,026
Pernanbuco	4.569	722.882	8.485.386	25,6	158,2	0,053
Rio Grande do Norte	1.956	263.550	3.013.740	43,2	134,7	0,064
Bahia	4.034	488.144	14.080.654	26,9	121,0	0,028
Paraíba	1.710	223.584	3.641.395	26,4	130,8	0,047
Maranhão	1.010	86.098	6.118.995	20,2	85,2	0,016
Sergipe	716	68.274	1.939.426	29,3	95,4	0,036
Ceará	3.710	554.833	8.185.286	25,8	149,6	0,045
Piauí	838	119.015	3.032.421	20,2	142,0	0,027
Região Nordeste	**19.352**	**263.9552**	**51.534.406**	**26,2**	**136,4**	**0,037**
Santa Catarina	6.527	999.253	5.866.252	40,2	153,1	0,111
Rio Grande do Sul	12.114	1.671.759	10.582.840	36,3	138,0	0,114
Paraná	11.774	1.685.625	10.284.503	40,2	143,2	0,114
Região Sul	30.415	4.356.637	26.733.595	38,9	143,2	0,113
São Paulo	84.584	12.895.459	39.827.570	43,9	152,5	0,212
Rio de Janeiro	33.247	4.953.433	15.420.375	40,9	149,0	0,215
Minas Gerais	1.5118	2.100.818	19.273.506	33,1	139,0	0,078
Espírito Santo	1.714	241.625	3.351.669	35,3	141,0	0,051
Região Sudeste	**134.663**	**20.191.335**	**77.873.120**	**38,3**	**149,9**	**0,172**
Distrito Federal	5.327	867.362	2.455.903	56,1	162,8	0,216
Mato Grosso	824	78.166	2.854.642	33,8	94,9	0,028
Mato Grosso do Sul	817	85.244	2.265.274	39,2	104,3	0,036
Goiás	3.341	346.029	5.647.035	35,2	103,6	0,059
Região Centroeste	**10.309**	**1.376.801**	**13.222.854**	**41,1**	**133,6**	**0,078**
Total	**198.714**	**29.013.854**	**187.232.961**	**32,7**	**146,0**	**0,106**

Tabela 2: Dados de usuários e tweets do Twitter relacionados com dados do IBGE

região norte brasileira praticamente não possuem usuários identificados em nossa base de dados. A Bahia possui a maior população da região nordeste e a quarta população do Brasil, mas é um dos estados que menos usam o Twitter no Brasil. Em contrapartida, o Distrito Federal tem uma baixa população e uma grande proporção de usuários no Twitter, ao passo que São Paulo tem a maior população do Brasil e grande proporção de usuários no Twitter.

Com relação à fração da população com acesso à internet, há alguma relação com o uso do Twitter, mas não é uma regra. Mais de 56% da população do Distrito Federal tem acesso à Internet (índice bem maior que dos demais estados), e é um dos que mais usam o Twitter. Entre os demais estados que mais usam o Twitter (São Paulo, Rio de Janeiro, Santa Catarina, Paraná e Rio Grande do Sul), todos possuem bom índice de pessoas com acesso à internet, comparando com os demais estados brasileiros e com a média nacional (32,7%). Por outro lado, há estados que têm boa taxa de acesso à internet, mas baixo uso do Twitter, como Rio Grande do Norte e Mato Grosso do Sul.

Estudamos ainda se existe relação entre o uso do Twitter e a renda média mensal familiar dos estados e o PIB dos estados, mas nenhuma relação foi observada. Por este motivo não apresentamos estes dados na Tabela 2.

5. TERMOS FREQUENTES

Nesta seção faremos uma análise dos termos mais frequentes encontrados em tweets. A figura 1 apresenta a nuvem dos 150 termos mais frequentes, sem qualquer filtro (inclusive *stop words* foram mantidas).

Figura 1: Nuvem dos termos mais frequentes em tweets

Alguns padrões interessantes foram observados na lista dos termos mais frequentes (considerando os 300 termos mais frequentes). Primeiramente, vemos que há um grande número de substantivos

Figura 2: Nuvem dos termos que ocorrem em tweets com polaridade positiva que não ocorrem com frequência em tweets com polaridade negativa

Figura 3: Nuvem dos termos que ocorrem em tweets com polaridade negativa que não ocorrem com frequência em tweets com polaridade positiva

relacionados a tempo de alguma forma, ou adjuntos adverbiais de tempo, como: dia, agora, hoje, noite, tempo, semana, amanhã, sempre, hora, tarde, ontem, nunca, anos, dias, sexta, ano, horas (nesta ordem de frequência). O Twitter parece ser uma ferramenta usada para descrever fatos ou eventos que aconteceram recentemente, estão acontecendo, ou vão acontecer, o que justifica a forte presença de termos indicadores de tempo.

Observamos também que o texto nos tweets são em geral informais, com a presença de um conjunto de abreviações comuns entre os usuários, como: d (de), q (que), p (para), vc (você), ta (está), pq (porque), gt (gente), tô (estou), hj (hoje), s (sem), tava (estava), rs (risos), eh (é), tb (também), tbm (também) (nesta ordem de frequência). Ainda, há um conjunto de termos bem informais que muitas vezes não fazem parte do dicionário da língua portuguesa, mas são usados com frequência por um grande número de usuários, como: só, cara, aí, hahaha, né, haha, legal, ah, hehehe, hahahaha, hehe, oi, xd. Das formas utilizadas para expressar um sorriso, ou uma risada, vemos que "hahaha" é a mais usuada, e que não há um padrão único, sendo usadas ainda as expressões: haha, hehehe, hahahaha, hehe, rs, xd.

A seguir analisamos os termos relacionados a tweets com polaridade positiva e relacionados a tweets com polaridade negativa. Dois conjuntos com os 1.000 termos mais frequentes foram gerados: um somente dos tweets com polaridade positiva e outro somente dos tweets com polaridade negativa. A diferença destes conjuntos foi usada para gerar duas listas de termos: uma dos termos de tweets com polaridade positiva que não ocorrem com frequência em tweets com polaridade negativa, e outra dos termos de tweets com polaridade negativa que não ocorrem com frequência em tweets com polaridade positiva. As nuvens dos 150 termos mais frequentes destas listas são apresentadas nas figuras 2 e 3, respectivamente.

É interessante observar os termos que ocorrem em tweets com polaridade positiva que não ocorrem com frequência em tweets com polaridade negativa, como: obrigado, adorei, viva, ótima, boas, bacana, bastante, beleza, sucesso, indico, bons, rir, gostar, recomendo, concordo, bonito, divertido, engraçado, especial, parabens, querida, graças. Também é interessante observar os termos que ocorrem em tweets com polaridade negativa que não ocorrem com frequência em tweets com polaridade positiva, como: triste, droga, morreu, tédio, chorar, morrer, infelizmente, doente, sozinha, cansada, raiva.

6. URLS MAIS PROPAGADAS

Mensagens no Twitter podem conter no máximo 140 caracteres. Tal restrição de espaço serviços tem tornado encurtadores de URLs cada vez mais populares. Esses sistemas funcionam da seguinte forma. Eles traduzem uma URL (que pode consistir de centenas de caracteres) em uma nova URL, tipicamente com poucos caracteres que retorna os códigos HTTP 301 ou 302 de redirecionamento para a URL longa original [6]. Por exemplo, o link `http://topics.nytimes.com/top/news/business/companies/twitter/` pode ser encurtado para `http://nyti.ms/1VKbrC` pelo bit.ly [3], que irá redirecionar qualquer requisição para o sítio Web original.

Para estudarmos os principais domínios propagados no Twitter, precisamos obter a versão longa de todas as URLs encurtadas encontradas em nossa base. Para isso, desenvolvemos uma ferramenta capaz de resolver a URL de um tweet, enviando uma requisição ao sistema encurtador e resolvendo a URL. Se o domínio obtido fosse diferente, nós consideramos a URL no tweet ser uma URL encurtada, senão, consideramos a URL como longa. Foram encontrados 30 serviços encurtadores de URL nos dados de usuários brasileiros, sendo que os serviços tinyurl.com [4] e bit.ly [3] são os mais populares com mais de 90% do total de URLs encurtadas.

Twitter	Domínio	Alexa
1	twitpic.com	47
2	blip.fm	2654
3	lolquiz.com	-
4	youtube.com	3
5	plurk.com	5141
6	ff.im	-
7	tumblr.com	34
8	flickr.com	36
9	twitter.com	13
10	meadd.com	695

Tabela 3: Ranking de domínios mais propagadas no Twitter e suas respectivas popularidades no ranking do alexa.com

Com base nesses dados, podemos verificar se as URLs mais compartilhadas por brasileiros no Twitter são URLs de domínios populares na Web brasileira. Como forma de comparação, nós listamos os domínios brasileiros mais populares segundo o sítio Web alexa.com [1], que mede o número de acessos recebidos por diversos sítios web. A tabela 3 apresenta os domínios mais populares das URLs propagadas por brasileiros no Twitter e suas respectivas posições no ranking do alexa.com. De forma similar, a tabela 4 mostra o ranking dos domínios mais populares brasileiros segundo o alexa.com e suas respectivas posições no ranking dos domínios

Alexa	Domínio	Tweets
1	google.com.br	-
2	google.com	78
3	youtube.com	4
4	facebook.com	65
5	uol.com.br	-
6	orkut.com.br	16
7	live.com	-
8	globo.com	11
9	blogspot.com	-
10	yahoo.com	-

Tabela 4: Ranking dos domínios mais populares no alexa.com e suas respectivas popularidades no ranking do Twitter

mais propagados no Twitter.

Podemos fazer várias observações interessantes a partir dessas tabelas. O ranking do alexa.com é bastante diferente do ranking dos domínios propagados no Twitter. Em particular, o domínio mais popular propagado nos tweets de brasileiros é o twitpic.com, um sistema para permite o compartilhamento de fotos. No alexa.com, o sítio web com maior número de acessos é do Google, que aparece muito pouco nos dados do Twitter. Ao observarmos os 100 domínios mais populares no Twitter notamos uma predominância de domínios de sítios web associados a notícias, entretenimento e redes sociais. Apenas um sítio web de compras aparece no ranking, na posição 97. No ranking do alexa.com, além de redes sociais, encontramos uma predominância de sistemas de busca, serviços como o sítio do banco Itaú na posição 23 e ministério da fazenda na posição 41, além de dezessete sistemas de compra online, incluindo Mercado livre na posição 17, buscape na posição 30 e submarino na posição 34.

7. USUÁRIOS INFLUENTES NO BRASIL

Usuários influentes possuem uma função vital em qualquer ecossistema, devido à associação da noção influência à noção de poder de formar a opinião de outras pessoas, afetando funcionalidades da sociedade, como por exemplo, em quem eleitores votam [9] ou como uma determinada moda se espalha [14]. Apesar de amplamente estudada em diferentes áreas, tais como sociologia, marketing e ciência política [24, 19], a noção de influência nunca foi quantificada nesses contextos por se tratar de um conceito vago e difícil de ser quantificado. Entretanto, com a enorme popularidade do Twitter, várias idéias conflitantes e algoritmos para a identificação de usuários influentes no Twitter vêm surgindo.

Várias máquinas de busca, incluindo o Google, estimam a importância dos tweets baseados no Pagerank na tentativa de retornar tweets mais "importantes" como resultado de buscas [28]. O PageRank é um algoritmo interativo que assinala um peso numérico para cada nodo com o propósito de estimar sua importância relativa no grafo. O algoritmo foi inicialmente proposto por Brin e Page [10] para ordenar resultados de busca do protótipo de máquina de busca da Google. A intuição por trás do PageRank é que um usuário do Twitter é importante se existem muitos usuários o seguindo ou existem outros usuários importantes o seguindo. A equação que calcula o PageRank (PR) de um usuário i, $PR(i)$, é definida da seguinte forma:

$$PR(i) = (1 - d) + d \sum_{v \in S(i)} \frac{PR(v)}{N_v} \quad (1)$$

onde $S(i)$ é o conjunto de usuários que seguem i, N_v denomina o número de arestas que saem do nodo v, e o parâmetro d é um fator que pode ter valor entre 0 e 1. Em nossos experimentos, o Pagerank foi computado para o grafo completo do Twitter e não somente para os usuários brasileiros, visto que o uso de uma amostra do grafo para o cálculo do Pagerank pode alterar significativamente os resultados. O valor do parâmetro d utilizado foi o valor padrão de 0,85, definido em [10].

Apesar de sua imensa aplicabilidade, o Pagerank ainda não foi rigorosamente verificado como métrica capaz de capturar os usuários mais relevantes do Twitter. Algumas celebridades, como Barack Obama, são seguidos por muitos usuários, mas também os seguem de volta tornado esses usuários importantes e, possivelmente afetando o uso do Pagerank como medida de influência. Trabalhos recentes mostraram a existência de um grande número de spammers no Twitter que atuam seguindo automaticamente de usuários no Twitter na tentativa de ser seguido de volta e ter sua influência aumentada [8, 22]. Além disso, Cha e seus colaboradores [11] mostraram que usuários que possuem muitos seguidores no Twitter não são necessariamente usuários muito retweetados. O número de retweets funciona como um indicador da habilidade de um usuário gerar conteúdo com valor suficiente para ser passado para a frente. Alguns sítios web com propósito comercial têm ganhado popularidade ao prover o Retweetrank de usuários comuns, como é o caso do `www.retweetrank.com`. Sendo assim, além do Pagerank, vamos considerar como medida de influência o retweetRank, que é o ranking de usuários mais retweetados. Para isso, foi preciso compararmos cada tweet postado pelos usuários brasileiros com os tweets postados posteriormente pelos seus seguidores na tentativa de identificar potenciais retweets. Um retweet ocorre quando um usuário posta uma mensagem com os seguintes padrões `RT @usuário` ou `via @usuário` ao final do tweet.

A tabela 5 mostra os 10 usuários brasileiros mais bem ranqueados de acordo com Pagerank e suas respectivas posições de acordo com o Retweetrank. De maneira semelhante, a tabela 6 mostra o ranking dos usuários de acordo com Retweetrank e suas respectivas posições de acordo com o Pagerank. Podemos fazer interessantes observações sobre essas tabelas. Primeiro, podemos notar que o usuário com maior Pagerank é o usuário *@manomenezes*, que é o perfil do atual técnico da seleção brasileira de futebol, Mano Menezes[2]. Entretanto, o usuário mais retweetado é o jornalista Marcelo Tas (*@marcelotas*), apresentador do programa de televisão CQC. Apesar de muito seguido, o técnico Mano Menezes aparece apenas na posição 64 do ranking dos mais retweetados. De fato, usuários do Twitter podem estar mais interessadas em receber em primeira mão as informações de um técnico de futebol, mas não necessariamente repassá-las aos seus seguidores. Outra importante observação é relacionada ao grande número de artistas brasileiros entre os mais influentes, especialmente aqueles ligados a programas de humor.

Apesar de oferecer uma visão geral sobre os usuários mais influentes no brasil, nossas análises apontam usuários da twitteresfera brasileira com maior potencial de influenciar um grande número de pessoas. Entretanto, influência pode variar de acordo com as diferentes regiões brasileiras, especialmente em um país tão vasto como o Brasil. De fato, empresas ligadas a estratégias de marketing boca-a-boca no Brasil podem estar mais interessadas em identifcar usuários que sejam influentes locais em determinadas regiões do que usuários já famosos na televisão e que são consequentemente populares no Twitter.

Como forma de identificar tais usuários, a Tabela 7 mostra os 10 usuários mais influentes de cada região do Brasil de acordo com o Pagerank e sua posição no ranking geral. De maneira similar,

[2]Na época da coleta de dados, Mano Menezes era o técnico do Corinthians

Pagerank	Usuário	Retweetrank
1	manomenezes	64
2	marcelotas	1
3	DaniloGentili	13
4	marcoluque	26
5	ivetesangalo	91
6	kibeloco	2
7	rodrigovesgo	20
8	christianpior	6
9	OscarFilho	22
10	andreolifelipe	100

Tabela 5: Posição dos 10 usuários com maior Pagerank no ranking do Retweetrank

Retweetrank	Usuário	Pagerank
1	marcelotas	2
2	kibeloco	6
3	rosana	12
4	Cardoso	58
5	melhoresfrases	17
6	christianpior	8
7	millorfernandes	15
8	esoterismo	209
9	abduzeedo	33
10	malvados	55

Tabela 6: Posição dos 10 usuários com maior Retweetrank no ranking do Pagerank

a Tabela 8 mostra os 10 usuários mais influentes de cada região do Brasil de acordo com o Retweetrank e sua posição no ranking geral. Note que além das celebridades, a maioria concentrada na região sudeste, nossas rankings de usuários influentes reveleram a existência de influentes locais em diversas regiões. Como exemplo, o usuário @diariodopara, na região norte aparece na posição 1505 no ranking brasileiro do Pagerank, mas é o décimo usuário mais influente de toda a região Norte, tendo uma atuação ainda mais importante no estado do Pará. Já no caso de retweets, o usuário *@Cardoso* se mostrou mais retweetado até mesmo do que o usuário *@marcelotas* na região nordeste, que aparece em primeiro lugar nos demais estados. Tais observações sugerem que algumas regiões podem possuir influentes locais que, dependendo da região, podem representar uma melhor opção para a realização de uma ação de marketing e propaganda.

8. CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresenta uma ampla análise geográfica dos usuários brasileiros no Twitter e de seus tweets. Nossas análises revelam que alguns estados brasileiros estão subestimados no Twitter em relação a dados demográficos, sendo que características sócio-econômicas desses estados parecem influenciar no número de usuários do Twitter. Além disso, nossas análises ainda revelam padrões linguísticos associados ao uso de microblogs no Brasil, identificando palavras geralmente associadas a contextos felizes e tristes. Nossas análises revelam as URLs mais postadas por brasileiros no Twitter e mostra que essas não são as mais acessadas na web de maneira geral. Por último, identificamos os usuários brasileiros mais influentes no Twitter e mostramos que o ranking de influência varia em diferentes regiões brasileiras. Como última contribuição, cabe ressaltar que a metodologia aqui empregada em diversas análises pode guiar futuras análises relacionadas ao Twitter.

Como trabalhos futuros, pretendemos desenvolver uma ferramenta de busca, de forma a permitir que usuários web possam extrair informações de nossa base de dados. Em outras palavras pretende-se desenvolver as bases de um sistema que permita que interessados possam entender a repercussão e a opinião de usuários do Twitter sobre eventos ocorridos no passado em diferentes localizações geográficas. Exemplos de cenários que esperamos identificar com essa ferramenta são: (i) epidemia de dengue afeta negativamente o humor das pessoas, onde pela nuvem de tags poderemos ver nomes e entidades criticadas; (ii) a vitória do Rio de Janeiro como sede das Olimpíadas de 2016 causou enorme alegria aos brasileiros; vitória de um determinado time de futebol gera alegria em uma região e infelicidade em outra; (iii) quais as regiões receberam positivamente/negativamente o lançamento de um novo produto.

9. AGRADECIMENTOS

O presente trabalho foi realizado com o apoio da Fapemig e do UOL (www.uol.com.br), através do Programa UOL Bolsa Pesquisa, processo número 20110210152501.

10. REFERÊNCIAS

[1] Alexa, the web information company. `http://www.alexa.com`. Accessed in March/2010.

[2] Arquivos do ibge. *http://www.ibge.gov.br/servidor_arquivos_est*.

[3] Bit.ly. *http://www.bit.ly*.

[4] tinyurl. *http://www.tinyurl.com*.

[5] There Are Now 155m Tweets Posted Per Day, Triple the Number a Year Ago. `http://rww.to/gv4VqA`, April 2011.

[6] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ionnadis, E. P. Markatos, and T. Karagiannis. we.b: The web of short URLs. In *Int'l World Wide Web Conference (WWW)*, 2011.

[7] F. Benevenuto, J. Almeida, and A. Silva. Coleta e análise de grandes bases de dados de redes sociais online. In *Jornadas de Atualização em Informática (JAI)*, pages 11–57. 2011.

[8] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[9] J. Berry and E. Keller. *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. Free Press, 2003.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

[11] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[12] S. Chhabra, A. Aggarwal, F. Benevenuto, and P. Kumaraguru. Phi.sh/$ocial: The phishing landscape through short urls. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2011.

[13] comScore. The netherlands ranks number 1 worldwide in penetration for twitter and linkedin. `http://bit.ly/hOpdAb`, 2011.

[14] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, 2002.

[15] J. Gomide, A. Veloso, W. M. Jr., F. Benevenuto, V. Almeida, F. Ferraz, and M. Teixeira. Dengue surveillance based on a

Top Norte	Top Nordeste	Top Centro-oeste	Top Sudeste	Top Sul
DiscoLee (151)	ivetesangalo (5)	bfeneto (18)	manomenezes (1)	abduzeedo (33)
talivale (767)	sigacl (30)	passagensaereas (37)	marcelotas (2)	federicodevito (43)
jerilsonduarte (836)	comunicadores (66)	WelderMM (62)	DaniloGentili (3)	piangers (46)
DEZMinutos (903)	murilogun (96)	oab_brasil (167)	marcoluque (4)	garotasemfio (75)
educhaves (1051)	gilbertogil (124)	consumidor_gov (227)	kibeloco (6)	lucasfresno (82)
asttro (1093)	scheilacarvalho (138)	brazilians (238)	rodrigovesgo (7)	nanypeople (91)
joaopedrosenado (1122)	m_camelo (168)	agenciabrasil (301)	christianpior (8)	bobagento (102)
jenizambiazzi (1255)	LisiSilveira (187)	AlonFeuerwerker (431)	OscarFilho (9)	azaghal (106)
pedrox (1451)	Danosse (230)	jorgeemateus (450)	andreolifelipe (10)	laumundial (102)
diariodopara (1505)	veramartins (248)	comediamm (464)	oceara (11)	anderssauro (164)

Tabela 7: Top influentes de cada região do Brasil de acordo com o Pagerank e sua posição no ranking geral

Top Norte	Top Nordeste	Top Centro-oeste	Top Sudeste	Top Sul
pedrox (306)	Cardoso (4)	marcelotas (1)	marcelotas (1)	marcelotas (1)
Cardoso (4)	marcelotas (1)	kibeloco (2)	rosana (3)	Cardoso (4)
kibeloco (2)	rosana (3)	Cardoso (4)	kibeloco (2)	kibeloco (2)
marcelotas (1)	kibeloco (2)	rosana (3)	Cardoso (4)	rosana (3)
rosana (3)	melhoresfrases (5)	melhoresfrases (5)	melhoresfrases (5)	millorfernandes (7)
diariodopara (838)	christianpior (6)	christianpior (6)	christianpior (6)	melhoresfrases (5)
melhoresfrases (5)	millorfernandes (7)	millorfernandes (7)	millorfernandes (7)	tplayer (25)
christianpior (6)	comunicadores (11)	inagaki (15)	malvados (10)	tiagomx (103)
millorfernandes (7)	malvados (10)	malvados (10)	esoterismo (8)	christianpior (6)
paolelli (1022)	inagaki (15)	samara7days (16)	comunicadores (11)	malvados (10)

Tabela 8: Top influentes de cada região do Brasil de acordo com o Retweetrank e sua posição no ranking geral

computational model of spatio-temporal locality of twitter. In *ACM SIGWEB Web Science Conference (WebSci)*, 2011.

[16] Google Geocoding API. `http://code.google.com/intl/en/apis/maps/documentation/geocoding/`.

[17] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The underground on 140 characters or less. In *ACM conference on Computer and communications security (CCS)*, pages 27–37, 2010.

[18] US confirms it asked Twitter to stay open to help Iran protesters. `http://tinyurl.com/klv36p`.

[19] E. Katz and P. Lazarsfeld. *Personal Influence: The Part Played by People in the Flow of Mass Communications*. The Free Press, New York, 1955.

[20] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Int'l World Wide Web Conference (WWW)*, pages 591–600, 2010.

[21] K. Lee, B. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[22] K. Lee, B. Eoff, and J. Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *AAAI Int'l Conference on Weblogs and Social Media (ICWSM)*, 2011.

[23] T. Rodrigues, F. Benevenuto, M. Cha, K. P. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2009.

[24] E. M. Rogers. *Diffusion of Innovations*. Free Press, 1962.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[26] W. Santos, G. Pappa, W. M. Jr., D. Guedes, A. Veloso, V. Almeida, A. Pereira, P. Guerra, A. Silva, F. Mourão, T. Magalhães, F. Machado, L. Cherchiglia, L. Simões, R. Batista, F. Arcanjo, G. Brunoro, N. Mariano, G. Magno, M. Ribeiro, L. Teixeira, A. Silva, B. Reis, and R. Silva. Observatório da web: Uma plataforma de monitoração, síntese e visualização de eventos massivos em tempo real. In *Anais do XXXVII Seminário Integrado de Hardware e Software*, SEMISH'10, pages 110–120, 2010.

[27] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft. Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades . In *Int'l World Wide Web Conference (WWW)*, 2011.

[28] D. Talbot. How Google Ranks Tweets. `http://www.technologyreview.in/web/24353/`, Jan 2010.

[29] T. N. Y. Times. Google Adds Live Updates to Results, December 2009. `http://nyti.ms/cnszI5`.

[30] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. 2010.

[31] L. Rao, Twitter Seeing 90 Million Tweets Per Day, 25 Percent Contain Links, *TechCrunch*, 2010. `http://tinyurl.com/27x5cay`.

[32] S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 1994.