

Um sistema de alarme para vigilância epidemiológica de rumores utilizando redes sociais

Denise Brito, Janaina Gomide, Walter Santos, Wagner Meira Jr., Adriano Veloso e Virgilio Almeida

Universidade Federal de Minas Gerais

{denise.brito, janaina, walter, meira, adrianov, virgilio}@dcc.ufmg.br

Abstract. Online social networks are used at a daily basis by millions of users around the world. More and more people use these networks to interact, to issue opinions, and to share content about several different topics, such as entertainment, weather, work, family, traffic, and even their health conditions. In summary, social networks became another social place with their own meaning and evolving dynamically. Many events are late perceived and mentioned by traditional media, but may happen in the social networks at real time, so that it is possible to detect them and such detection subsidize the determination of predictive models. The goal of this work is to exploit content available in the social networks to detect the occurrence and predict epidemics surges, in particular those associated with dengue. Starting from the collection and processing of data from social networks, we identify how surges manifest in social networks and exploit these features to build an alarm system, as part of an epidemics surveillance system based on rumours. The results show that we were able to achieve high precisions (more than 99%), as well as to predict surges in advance.

Resumo. As redes sociais online fazem parte do cotidiano de milhões de pessoas do mundo inteiro. Cada vez mais pessoas utilizam essas redes para interagir, opinar e compartilhar conteúdos sobre os mais diversos tópicos, como diversão, clima, trabalho, família, trânsito e mesmo sua condição de saúde. Em suma, as redes sociais se tornaram mais um lugar social com significados próprios, evoluindo dinamicamente. Muitos acontecimentos são tardiamente percebidos e divulgados pelos meios de comunicação tradicionais, mas podem acontecer nas redes sociais em tempo real, sendo passíveis de serem detectados e de subsidiarem a construção de modelos de previsão. O objetivo deste trabalho é explorar conteúdo disponível nas redes sociais para detectar a ocorrência e prever surtos epidemiológicos, em particular de dengue. A partir da coleta e processamento de dados das redes sociais, identificamos as características de surtos e exploramos essas características para a construção de um alarme, constituindo um sistema de vigilância epidemiológica de rumores. Resultados mostram que fomos capazes de atingir altas precisões (mais de 99%), assim como antecipar surtos.

1. INTRODUÇÃO

A *World Wide Web* tem sido cada vez mais utilizada como meio para compartilhamento de informações e interação. Milhões de pessoas utilizam as redes sociais online para publicar opiniões sobre diversos assuntos, assim como relatar eventos que acontecem em suas vidas e seu estado de saúde. Muitos acontecimentos são percebidos e divulgados tardiamente, enquanto podem estar sendo discutidos em tempo real nas redes sociais. Sendo assim, tais eventos podem ser detectados através de mensagens postadas pelas pessoas que os vivenciam e essas mensagens podem também servir para construir modelos de previsão.

Para que as mensagens possam ser utilizadas como fonte de dados para detecção de eventos, elas devem possuir algumas características importantes: devem ser publicadas espontaneamente, contendo uma opinião ou sentimento, e ter localização no espaço e no tempo bem definida. Mais ainda, é necessário que essas redes sociais sejam utilizadas de forma significativa pela sociedade, o conteúdo da mensagem possa ser obtido e essas mensagens também informem a localização do usuário. Em parti-

cular, uma rede que possui essas características é o Twitter, pois permite que os usuários publiquem mensagens de até 140 caracteres, o que faz com que as mensagens sejam objetivas, e, ao mesmo tempo, possuam algum contexto. Os eventos também devem possuir algumas características para que possam ser detectados: causar algum impacto na vida das pessoas, para que elas sejam motivadas a postar informações sobre ele; atingir um grande número de pessoas, para que o número de mensagens seja significativo, e não deve haver nenhum impedimento social, para que ele seja divulgado publicamente.

Neste trabalho, a detecção de eventos através de redes sociais foi aplicada no contexto da dengue. A dengue é uma doença que ocorre em regiões de clima tropical e subtropical, transmitida pela picada do mosquito *Aedes aegypti* infectado. Até o momento, em termos de vigilância epidemiológica, o Ministério da Saúde monitora a quantidade de mosquitos transmissores encontrados e, então, se houver muitos focos do mosquito, as campanhas e medidas de prevenção são intensificadas naquele local. No entanto, a presença de mosquitos transmissores não significa que haja uma incidência alta de dengue, e, quando uma epidemia ocorre, o número de casos registrados é conhecido pelas autoridades com algumas semanas de atraso. Logo, o objetivo deste trabalho é permitir uma detecção precoce de surtos de dengue para tornar mais rápido o planejamento do governo.

Em [Gomide 2012] foi mostrado um método de detecção precoce de surtos de dengue no Brasil através de mensagens postadas no Twitter. A partir deste trabalho desenvolvemos um sistema de alarme, que está sendo utilizado pelo Ministério da Saúde desde Novembro de 2011. Este artigo descreve o sistema de alarme e a sua avaliação, com base nos *tweets* e nos dados oficiais, agrupados por semana, correspondentes às trinta primeiras semanas de 2012.

2. TRABALHOS RELACIONADOS

Nesta seção discutimos alguns trabalhos relacionados à detecção de eventos e sistemas de alarme baseados em redes sociais.

Nos últimos anos, técnicas de mineração de dados têm sido aplicadas em redes sociais para detecção de eventos, como por exemplo em [Sakaki et al. 2010], em que os *tweets* foram utilizados para detectar a ocorrência de terremotos e sua localização. Em [Tumasjan et al. 2010], *tweets* contendo opiniões sobre as eleições alemãs foram utilizados. Os primeiros trabalhos que visavam a detecção de surtos epidemiológicos utilizavam artigos de jornais contendo informações sobre a gripe ([Mawudeku and Blench 2006; Brownstein et al. 2008; Freifeld et al. 2008]), e, mais recentemente, foram utilizados blogs ([Corley et al. 2009]), consultas em sites de busca ([Ginsberg et al. 2009; Chan et al. 2011; Althouse et al. 2011]), e também o Twitter ([Culotta 2010; Lampos and Cristianini 2010; 2011; Lampos et al. 2010; Chen et al. 2010; Achrekar et al. 2011; Chunara et al. 2012]).

Após o desenvolvimento do método de detecção do evento, é importante que haja um sistema de alarme para alertar a população em situações críticas. Em [Sakaki et al. 2010], foi construído um sistema que detecta a ocorrência de um terremoto e envia e-mails para as pessoas cadastradas. Em [Freifeld et al. 2008], as notícias sobre a gripe são visualizadas numa página web e em [Lampos et al. 2010] são disponibilizadas as taxas de incidência da Influenza estimadas com o método proposto em [Lampos and Cristianini 2010]. O Google Dengue Trends (<http://www.google.org/denguetrends/>) permite visualizar a frequência de consultas relacionadas à dengue no Brasil e em mais 9 países. A diferença do alarme proposto neste trabalho para os descritos anteriormente é a especificidade do alarme (em nível de cidade, e não de estados ou países), com atualização semanal e também, além da análise da incidência, é feita uma análise da tendência do número de casos de dengue.

3. ALARME

O ponto de partida do sistema de Alarme é o Observatório da Dengue [Gomide 2012], que coleta e analisa as mensagens postadas no Twitter que referenciam dengue. Em particular, o Observatório é

capaz de determinar aquelas mensagens que representam experiências pessoais. Neste mesmo trabalho foi demonstrado haver uma alta correlação entre o volume de mensagens que expressam experiência pessoal e notificações da doença por localidade e por tempo.

As mensagens são classificadas utilizando o LAC (classificador associativo sob demanda) [Velooso et al. 2006]. O classificador compara as regras presentes nas mensagens do treino, previamente rotuladas manualmente, com as regras encontradas nas novas mensagens coletadas e as classifica em uma das cinco categorias: experiência pessoal, paródia ou ironia, opinião, informação e campanha.

Com base nesses achados, o alarme foi criado para complementar o sistema de vigilância tradicional de detecção e acompanhamento dos surtos de dengue. O alarme é específico para cada cidade com mais de cem mil habitantes e é baseado em dois tipos de análise: a análise de volume e a análise de tendência dos casos de dengue. A análise de volume estima a incidência de dengue no período atual, enquanto que a análise de tendência informa se o número de *tweets* relatando dengue aumentou, diminuiu ou permaneceu constante em relação às duas semanas anteriores.

É importante fornecer as duas informações porque, por exemplo, mesmo que o número de casos previstos seja baixo, se houve um aumento significativo de *tweets*, isso é um fato possivelmente preocupante, que não seria percebido apenas com a análise de volume. E mesmo que o número de *tweets* caia, em comparação com as semanas anteriores, o volume pode estar alto, o que não seria notado apenas com a análise de tendência.

Cabe ressaltar que o período considerado para o alarme é definido pelas características da doença sendo monitorada. No caso da dengue, que tem um ciclo semanal, optou-se por considerar duas semanas, tendo em vista a praxe epidemiológica.

3.1 Análise de volume

A metodologia de análise epidemiológica do Ministério da Saúde se baseia no número de casos por cem mil habitantes e reconhece 3 faixas: de 0 a 100 casos por cem mil habitantes, é dito que há uma baixa incidência de dengue naquela semana; entre 100 e 300 casos é considerada média incidência e a partir de 300 casos a incidência é alta. Para a análise de volume, o número de *tweets* relatando experiência pessoal de cada semana é a entrada de uma função de previsão. Seus parâmetros foram ajustados por regressão linear, assumindo que há uma relação linear entre o número de *tweets* e o número de casos reais de dengue, usando os dados do Twitter e os dados oficiais do fim de 2010 e início de 2011, e são diferentes para cada cidade, pois a porcentagem da população que possui acesso à internet e a incidência de dengue variam de região para região. O alarme então fornece em qual faixa de classificação recai o número de casos previstos.

Por exemplo, aplicando o método à cidade do Rio de Janeiro, com 6355949 habitantes¹, a função de previsão para essa cidade é: $y = 5.826x + 402.31$, onde y é o número de casos previstos e x o número de *tweets*. Assim, é calculado o número de casos previstos para todas as semanas. Os casos previstos e os casos reais são divididos pelo número de habitantes e multiplicados por cem mil, para serem comparados com a classificação do Ministério da Saúde (dada em termos de cem mil habitantes). Então, a classificação do número de casos previstos é comparada com a classificação dos casos reais, para validação. A Tabela I mostra o número de *tweets* para cada semana, os casos previstos, a faixa de classificação dos casos previstos, os casos reais e a faixa de classificação deles.

Para a cidade de Recife, PE, com 1546516 habitantes, a função de previsão é: $y = 2.4596x + 5.8507$. A classificação é feita exatamente como no exemplo anterior, e os resultados estão na Tabela II.

¹http://www.ibge.gov.br/home/presidencia/noticias/noticia_visualiza.php?id_noticia=1961&id_pagina=1

Tabela I. Previsão e validação do número de casos previstos para Rio de Janeiro, RJ

<i>Tweets</i>	Casos previstos	Alerta previsão	Casos reais	Alerta real	Alerta tendência
127	1142.212	baixo	964	baixo	estabilização
108	1031.518	baixo	1107	baixo	estabilização
106	1019.866	baixo	1499	baixo	estabilização
84	891.694	baixo	1667	baixo	estabilização
81	874.216	baixo	1862	baixo	estabilização
109	1037.344	baixo	2280	baixo	estabilização
98	973.258	baixo	2758	baixo	estabilização
176	1427.686	baixo	3317	baixo	estabilização
227	1724.812	baixo	4839	baixo	estabilização
241	1806.376	baixo	5800	baixo	estabilização
211	1631.596	baixo	5513	baixo	estabilização
274	1998.634	baixo	6605	médio	estabilização
270	1975.33	baixo	6764	médio	estabilização
290	2091.85	baixo	6972	médio	estabilização
331	2330.716	baixo	8577	médio	estabilização
530	3490.09	baixo	9269	médio	aumento
453	3041.488	baixo	8734	médio	estabilização
570	3723.13	baixo	8375	médio	estabilização
506	3350.266	baixo	7580	médio	estabilização
354	2464.714	baixo	6860	médio	redução
347	2423.932	baixo	6461	médio	estabilização
326	2301.586	baixo	4932	baixo	estabilização
217	1666.552	baixo	3460	baixo	redução
173	1410.208	baixo	2692	baixo	redução
197	1550.032	baixo	1791	baixo	estabilização
146	1252.906	baixo	1080	baixo	estabilização
135	1188.82	baixo	582	baixo	estabilização
137	1200.472	baixo	289	baixo	estabilização
61	757.696	baixo	88	baixo	redução
5	431.44	baixo	5	baixo	redução

3.2 Análise de tendência

A análise de tendência serve para identificar se o histórico do número de *tweets* sugere alguma tendência do número de casos de dengue. É calculado o Z-score ([Larsen and Marx 1986]) usando a média diária de *tweets* na semana atual (x) e nas duas semanas anteriores (μ), e do desvio padrão do número de *tweets* durante esse período (σ) para cada cidade. Foi utilizado o período de duas semanas anteriores pela natureza do evento. O Z-score é dado pela seguinte fórmula:

$$(x - \mu)/\sigma$$

A tendência é também classificada em três faixas, de acordo com o resultado do Z-score: para o Z-score abaixo de -1 , é dito que houve uma tendência de diminuição, para o Z-score entre -1 e 2 , é dito que a tendência é de estabilização e, acima de 2 , é considerado que houve um aumento.

3.3 Metodologia de avaliação do alarme

O Ministério da Saúde fornece o número de casos confirmados de dengue em cada cidade, por semana. Para avaliar o alarme, foi calculado o coeficiente de correlação cruzada entre as séries temporais do número de *tweets* e dos dados oficiais, para cada cidade, para verificar se as séries possuem boa correlação e descobrir se é possível fazer uma detecção precoce do número de casos de dengue através do Twitter.

A correlação cruzada [Broucke 1996] foi calculada aplicando-se um atraso de d semanas na série dos dados oficiais $o(i)$ em relação à série do Twitter $t(i)$, onde $i = 0, 1, 2, \dots, N - 1$ e N é o número de

Tabela II. Previsão e validação do número de casos previstos para Recife, PE

<i>Tweets</i>	Casos previstos	Alerta previsão	Casos reais	Alerta real	Alerta tendência
81	205.0783	baixo	359	baixo	estabilização
28	74.7195	baixo	290	baixo	estabilização
48	123.9115	baixo	294	baixo	estabilização
53	136.2095	baixo	359	baixo	estabilização
48	123.9115	baixo	461	baixo	estabilização
72	182.9419	baixo	508	baixo	estabilização
60	153.4267	baixo	465	baixo	estabilização
100	251.8107	baixo	574	baixo	estabilização
122	305.9219	baixo	615	baixo	estabilização
84	212.4571	baixo	582	baixo	estabilização
90	227.2147	baixo	635	baixo	estabilização
89	224.7551	baixo	598	baixo	estabilização
75	190.3207	baixo	438	baixo	estabilização
66	168.1843	baixo	416	baixo	estabilização
64	163.2651	baixo	420	baixo	estabilização
52	133.7499	baixo	317	baixo	estabilização
43	111.6135	baixo	226	baixo	estabilização
57	146.0479	baixo	233	baixo	estabilização
43	111.6135	baixo	180	baixo	estabilização
30	79.6387	baixo	147	baixo	estabilização
42	109.1539	baixo	120	baixo	estabilização
28	74.7195	baixo	104	baixo	estabilização
27	72.2599	baixo	77	baixo	estabilização
16	45.2043	baixo	65	baixo	estabilização
17	47.6639	baixo	65	baixo	estabilização
21	57.5023	baixo	55	baixo	estabilização
15	42.7447	baixo	30	baixo	estabilização
14	40.2851	baixo	18	baixo	estabilização
9	27.9871	baixo	29	baixo	estabilização
2	10.7699	baixo	4	baixo	redução

semanas avaliadas, no caso deste experimento, 30. A fórmula da correlação cruzada é:

$$r = \frac{\sum_{i=1}^n (t_i - \bar{t})(o_i - \bar{o})}{\sqrt{\sum_{i=1}^n (t_i - \bar{t})^2} \sqrt{\sum_{i=1}^n (o_i - \bar{o})^2}}$$

sendo \bar{t} e \bar{o} as médias das séries correspondentes.

Os valores que o coeficiente pode assumir estão entre -1 e 1 . Entre 0 e 1 , significa uma correlação direta entre as duas séries, e quanto mais próximo de 1 , indica maior correlação. Analogamente, entre 0 e -1 , significa uma correlação inversa entre as duas séries, e quanto mais próximo de -1 , a correlação inversa é maior. Foram calculados os coeficientes de correlação cruzada considerando um atraso de 0 , 1 e 2 semanas dos dados oficiais em relação aos dados do Twitter. Logo, quando o coeficiente considerando um atraso de 1 ou 2 semanas for mais próximo de 1 do que o coeficiente correspondente a nenhum atraso, isto sugere que há uma detecção precoce do número de casos de dengue pelo Twitter em comparação ao sistema de vigilância tradicional. No entanto, os dados do Twitter foram agrupados com a semana começando numa quinta-feira e terminando na quarta-feira seguinte, enquanto que os dados do Ministério da Saúde provavelmente são agrupados considerando a semana começando no domingo, ou seja, a série do Twitter já está meia semana adiantada. Para compensar essa diferença, consideramos que houve detecção precoce apenas quando o coeficiente correspondente ao cálculo com atraso de 2 semanas é mais próximo de 1 do que os demais coeficientes.

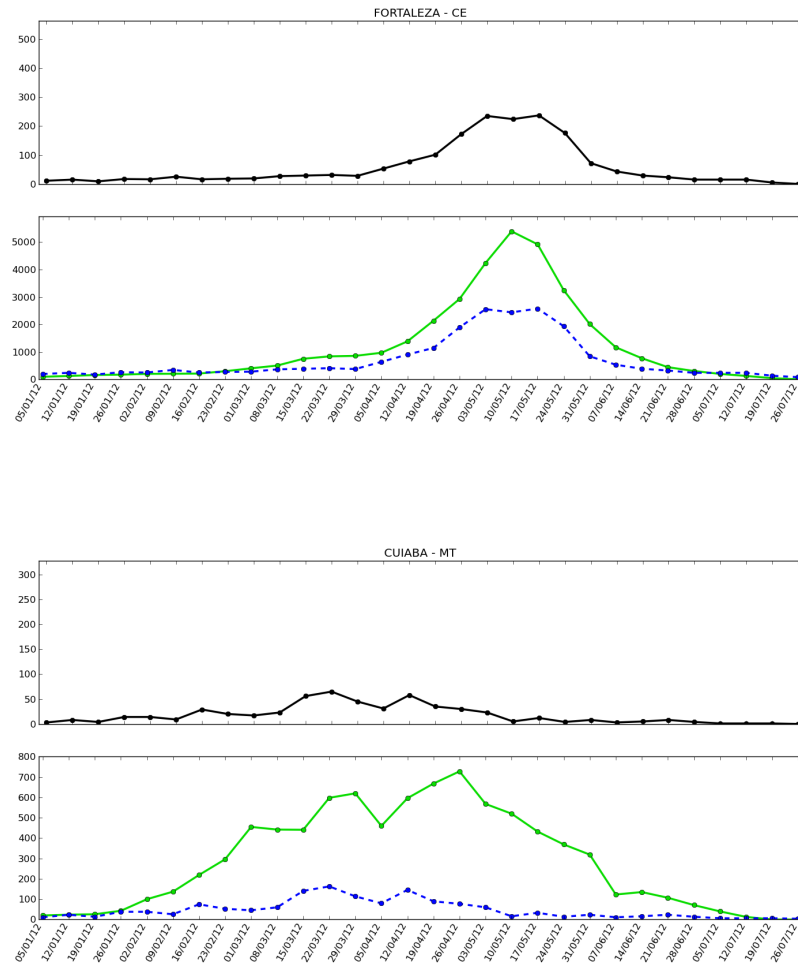


Fig. 1. Acima, número de *tweets*. Abaixo, em verde, número de casos reais e em azul, número de casos previstos.

4. RESULTADOS

Foram avaliados os resultados para as 285 cidades brasileiras com mais de cem mil habitantes. Em 2012, o número de casos de dengue registrados foi significativamente menor do que em 2011. Ainda assim, em 99.12% das previsões, o número de casos previstos através dos dados do Twitter recai na mesma faixa de classificação do número de casos registrados pelo Ministério da Saúde.

Nenhuma das 285 cidades obteve 7 ou mais *tweets* durante todas as semanas analisadas. Optou-se por considerar então todas as cidades com 7 ou mais *tweets* em pelo menos 20 das 30 semanas, e assim o número de cidades que se encaixam neste perfil são 16. Dentre estas 16 cidades, estão Porto Alegre, RS, e Curitiba, PR, que pertencem à região Sul do Brasil e possuem pouquíssimos casos de dengue registrados. O grande número de *tweets* registrados nessas cidades pode ser devido à localização informada pelo usuário do Twitter e será objeto de trabalho futuro.

A Tabela III mostra os resultados de algumas cidades, com o número de casos de dengue e o valor das correlações cruzadas considerando um atraso na série dos dados oficiais de 0, 1 e 2 semanas em relação aos dados do Twitter. Observou-se que em capitais onde há um histórico de surtos de dengue

e que o número de *tweets* é suficiente, o coeficiente de correlação se mostrou melhor do que em outras capitais em que o número de *tweets* é suficiente mas o número de casos de dengue é mais baixo (Tabela III). Em alguns desses casos, houve mais *tweets* sobre dengue do que casos de dengue registrados. Das 16 cidades analisadas, 10 obtiveram ao menos uma das correlações superior a 0.5.

Tabela III. Correlação entre casos de dengue e *tweets*

Cidade	Casos de dengue	Correlação 0	Correlação 1	Correlação 2
RIO DE JANEIRO – RJ	122722	0.901898	0.937201	0.914303
FORTALEZA – CE	35253	0.982306	0.892424	0.677947
RECIFE – PE	8684	0.928756	0.901308	0.876676
NATAL – RN	8577	0.714485	0.792293	0.833409
GOIÂNIA – GO	7045	0.366342	0.348550	0.221251
SALVADOR – BA	4834	0.431263	0.397027	0.197968
JOÃO PESSOA – PB	2726	0.752976	0.680269	0.556181
MANAUS – AM	2175	0.604449	0.618463	0.504184
ARACAJU – SE	2165	0.161315	0.294573	-0.222900
SÃO PAULO – SP	1986	0.429890	0.576528	0.605924
BELÉM – PA	1839	0.707398	0.740384	0.714267
BRASÍLIA – DF	1143	0.610918	0.613832	0.479530
SANTOS – SP	927	0.673293	0.520498	0.366419
BELO HORIZONTE – MG	747	-0.165831	-0.334045	-0.322071
CURITIBA – PR	37	0.438474	0.038517	0.012693
PORTO ALEGRE – RS	24	0.370767	0.450461	0.324811

5. CONCLUSÕES E TRABALHOS FUTUROS

Neste artigo apresentamos um sistema de alarme para realizar vigilância epidemiológica de rumores da epidemia de dengue no Brasil. O ponto de partida do alarme é o número de mensagens postadas na rede social Twitter que expressam experiência pessoal com a doença. O alarme se baseia em dois critérios e foi validado usando os dados oficiais de notificação da dengue nas 30 primeiras semanas de 2012.

Apesar de que em 2012 não houve grandes surtos como no ano anterior, a taxa de acerto (quando o número de casos previstos recai na mesma faixa de classificação que o número de casos reais) é bastante significativa. O volume insuficiente de *tweets* em determinadas cidades é uma limitação inerente ao modelo de previsão, por isso optamos por salientar as cidades em que isso não ocorreu. O principal motivo de falhas, mesmo com altas correlações entre *tweets* e casos reais, parece ser a grande variação no volume de casos de dengue de um ano para outro num mesmo local. Nas cidades onde isso aconteceu, como Rio de Janeiro – RJ (Tabela I), Fortaleza – CE e Cuiabá – MT (Figura 1), a proporção de *tweets* por casos de dengue variou de 2011 para 2012. Para aproximar ainda mais o número de casos previstos do número real, outras funções de previsão podem ser geradas, e se elas estiverem bem ajustadas, o alarme pode até mesmo ajudar em casos de subnotificação dos dados oficiais.

Em algumas cidades, como em Natal, RN, e São Paulo, SP, o coeficiente de correlação mais próximo de 1 é o correspondente ao atraso de 2 semanas na série dos dados oficiais, o que indica uma detecção precoce. Mas mesmo quando o melhor coeficiente é correspondente ao atraso zero, a detecção através de dados do Twitter contribui com a vigilância tradicional, pois na última, os dados demoram algumas semanas para chegar às mãos dos responsáveis pela elaboração de políticas preventivas e de combate à doença.

Em termos de trabalhos futuros, pretendemos concluir a validação utilizando os dados oficiais de 2012 e identificar as causas que justificam os desvios ou erros de previsão. Pretendemos também aprimorar as técnicas de análise de sentimento e georeferenciamento para que tenhamos uma maior

acurácia na detecção dos *tweets* pertinentes por cidade. E, finalmente, pretendemos aperfeiçoar as técnicas de detecção, em particular para lidar com aumentos ou reduções suaves, que não são bem detectados pela estratégia utilizada, reajustando os parâmetros com os dados coletados em 2011 e 2012, sendo que o primeiro ano representa uma época em que houve grande incidência de dengue e o segundo, com a grande maioria das cidades com baixa incidência.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado por CNPq, CAPES, Fapemig e InWeb.

REFERÊNCIAS

- ACHREKAR, H., GANDHE, A., LAZARUS, R., SSU-HSIN YU, AND LIU, B. Predicting flu trends using Twitter data. In *IEEE INFOCOM 2011 - IEEE Conference on Computer Communications Workshops*. IEEE, pp. 702–707, 2011.
- ALTHOUSE, B. M., NG, Y. Y., AND CUMMINGS, D. A. T. Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis* 5 (8): e1258, 08, 2011.
- BROURKE, P. Cross Correlation. <http://paulbourke.net/miscellaneous/correlate/>, 1996.
- BROWNSTEIN, J. S., FREIFELD, C. C., REIS, B. Y., AND MANDL, K. D. Surveillance sans fronti res: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Med* 5 (7): e151, 07, 2008.
- CHAN, E. H., SAHAI, V., CONRAD, C., AND BROWNSTEIN, J. S. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Negl Trop Dis* 5 (5): e1206, 05, 2011.
- CHEN, L., ACHREKAR, H., LIU, B., AND LAZARUS, R. Vision: towards real time epidemic vigilance through online social networks. In *ACM Workshop on Mobile Cloud Computing Services: Social Networks and Beyond*. ACM, pp. 1–5, 2010.
- CHUNARA, R., ANDREWS, J. R., AND BROWNSTEIN, J. S. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Journal of Tropical Medicine and Hygiene* 86 (1): 39–45, 2012.
- CORLEY, C., MIKLER, A. R., SINGH, K. P., AND COOK, D. J. Monitoring influenza trends through mining social media. In *Proceedings of International Conference on Bioinformatics & Computational Biology (BIOCOMP)*. CSREA Press, pp. 340–346, 2009.
- CULOTTA, A. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*. SOMA '10. ACM, New York, NY, USA, pp. 115–122, 2010.
- FREIFELD, C. C., MANDL, K. D., REIS, B. Y., AND BROWNSTEIN, J. S. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association (JAMIA)* 15 (2): 150–157, 2008.
- GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S., AND BRILLIANT, L. Detecting influenza epidemics using search engine query data. *Nature* 457 (7232): 1012–4, 2009.
- GOMIDE, J. S. *Minera o de Redes Sociais para Detec o e Previs o de Eventos Reais*. M.S. thesis, Universidade Federal de Minas Gerais, BR, 2012.
- LAMPOS, V. AND CRISTIANINI, N. Tracking the flu pandemic by monitoring the social web. In *2nd IAPR Workshop on Cognitive Information Processing (CIP 2010)*. IEEE Press, pp. 411–416, 2010.
- LAMPOS, V. AND CRISTIANINI, N. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, September, 2011.
- LAMPOS, V., DE BIE, T., AND CRISTIANINI, N. Flu detector - tracking epidemics on twitter. *Machine Learning and Knowledge* vol. 6323, pp. 599–602, 2010.
- LARSEN, R. AND MARX, M. *An introduction to mathematical statistics and its applications*. Prentice-Hall, 1986.
- MAWUDEKU, A. AND BLENCHE, M. Global public health intelligence network (gphin). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, 2006.
- SAKAKI, T., OKAZAKI, M., AND MATSUO, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*. WWW '10. ACM, New York, NY, USA, pp. 851–860, 2010.
- TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G., AND WELPE, I. M. Predicting elections with twitter : What 140 characters reveal about political sentiment. *Word Journal Of The International Linguistic Association*, 2010.
- VELOSO, A., MEIRA JR., W., AND ZAKI, M. J. Lazy associative classification. In *International Conference on Data Mining*. IEEE Computer Society, pp. 645–654, 2006.