Explorando redes sociais como ferramenta de disseminação de informações: uma análise espaço temporal em casos de epidemia

Liriam M. Enamoto
Universidade Federal do Estado do Rio de Janeiro
Av. Pasteur 458, Urca, Rio de Janeiro, RJ, Brazil
liriam.enamoto@uniriotec.br

Adriana C. F. Alvim
Universidade Federal do Estado do Rio de Janeiro
Av. Pasteur 458, Urca, Rio de Janeiro, RJ, Brazil
adriana@uniriotec.br

Vânia M. F. Dias
Universidade Federal do Estado do Rio de Janeiro
Av. Pasteur 458, Urca, Rio de Janeiro, RJ, Brazil
vania@uniriotec.br

Maristela Holanda
Universidade de Brasília
Departamento de Ciência da Computação
mholanda@unb.br

RESUMO

Em situações de desastre e saúde pública as redes sociais têm sido utilizadas com frequência, fornecendo informações atualizadas de fontes oficiais e não-oficiais. O objetivo deste trabalho é investigar o uso do Twitter relacionado à epidemia do Ebola combinando três tipos de análise: (i) por meio de métricas em grafos; (ii) por meio de banco de dados geográficos; (iii) e usando análise textual. A estratégia de análise de redes sociais apresentada neste artigo pretende contribuir, de maneira ainda incipiente, com a utilização destas redes como meio de articular diversos atores para disseminar informações relevantes e intervir em casos de epidemias em nível mundial.

PALAVRAS CHAVE. Redes sociais, epidemias, localização geográfica, Twitter.

Tópicos (OA - Outras aplicações em PO)

ABSTRACT

In situations of disaster and public health, social networks have been frequently used, providing updated information from official and unofficial sources. The objective of this study is to investigate the use of Twitter related to Ebola epidemic combining three types of analysis: using metrics on graphs, using Geographical Database, and textual analysis of comments. The analysis strategy for social networks presented in this paper aims to contribute, in a still incipient way, with the use of these networks as a means of articulating different actors to disseminate relevant information and intervene in cases of epidemics worldwide.

KEYWORDS. Social network, epidemics, geographic location, Twitter.

Paper topics (OA - Other applications in OR)

1. Introdução

O avanço tecnológico das últimas décadas permitiu que as pessoas possam se locomover com facilidade entre países e continentes. Esta vantagem também trouxe vulnerabilidades às pessoas quando se trata de ameaças à saúde pública, como epidemias. Recentemente, a epidemia provocada pelo vírus do Ebola contabilizou 11.301 mortes, segundo dados divulgados pela Organização Mundial de Saúde (OMS) em fevereiro de 2016. O contágio do vírus se dá através do contato com fluidos corpóreos de um paciente que já apresenta os sintomas da doença. Os sintomas mais comuns são febre repentina, vômito, diarreia, dores no estômago, dores musculares e dificuldades de respiração. A taxa de mortalidade da doença chega à 70% [Team et al., 2014]. Diversas vacinas estão em fase de testes em animais e seres humanos, porém, até o presente momento, nenhuma vacina foi aprovada pela OMS para o uso em seres humanos.

O primeiro surto do Ebola ocorreu em 1976 na República Democrática do Congo, onde foram registradas 280 mortes. Após este primeiro surto, foram registradas outras ocorrências do Ebola no Sudão, Gabão e Uganda [Khan et al., 1999]. A última epidemia do Ebola iniciou-se na Guiné, na África Ocidental, em dezembro de 2013. No dia 25/08/2014, a OMS divulgou os primeiros dados sobre a epidemia, registrando 1.546 mortes e 3.052 casos de infecção pelo vírus, com 99% dos casos concentrando-se nos países da África Ocidental: Libéria, Guiné, Serra Leoa e Nigéria. Em outubro de 2014, foram registrados casos isolados de contaminação do vírus nos Estados Unidos e na Espanha. No mês seguinte, em novembro de 2014, registrou-se a primeira morte causada pelo vírus nos Estados Unidos, o que aumentou o estado de atenção ao redor do mundo. Acredita-se que as medidas necessárias para combater o aumento de novos casos são: diagnóstico rápido, isolamento do paciente, controle da infecção, rastreamento dos contatos da vítima e práticas de funeral seguro [Team et al., 2014]. Apesar dos esforços das organizações governamentais e de saúde, a epidemia continuou a avançar nos países da África Ocidental até abril de 2015. Segundo a OMS, os principais motivos na dificuldade de contenção referem-se a aspectos culturais dos países africanos como a resistência ao enterro seguro, desconfiança da população às medidas de quarentena e isolamento do paciente e falta de engajamento da comunidade no combate à doença.

De acordo com Simon et al. [2015], em situações de desastre e saúde pública, as redes sociais têm sido utilizadas com frequência, fornecendo informações atualizadas de fontes oficiais e não-oficiais, incluindo informações educativas à população, informações incorretas e rumores. No trabalho de Lazard et al. [2015], relacionado à epidemia do Ebola, as redes sociais foram analisadas e identificaram que o tema de maior preocupação das pessoas com relação à epidemia é sobre os sintomas causados pelo vírus. Oyeyemi et al. [2014] observaram que as redes sociais foram utilizadas também para disseminar rumores e informações incorretas sobre o Ebola, como a cura por meio de transfusão sanguínea, plantas medicinais e ingestão de água com sal.

O objetivo deste trabalho é colaborar com o estado da arte da área de análise de redes sociais por ocasião da ocorrência de surtos de epidemias que podem vir a se propagar a nível mundial. Especificamente, esta pesquisa teve como foco o caso do surto de Ebola ocorrido no período de 01/11/2014 a 30/04/2015. A rede social escolhida foi o Twitter [Twitter]. Por meio desta pesquisa, pretende-se identificar os países onde o Twitter foi utilizado fortemente para postar comentários sobre o Ebola, estudar os grafos subjacentes às redes sociais observadas a fim de analisar a intensidade com que tais redes auxiliam na disseminação de informações e prescrutar a relevância das mesmas.

Na próxima seção apresentam-se conceitos básicos relacionados à Teoria dos Grafos e os principais trabalhos que relacionam redes sociais e monitoramento de epidemias. Na Seção 3 descreve-se a metodologia utilizada na coleta e armazenamento dos dados. A Seção 4 é dedicada à análise dos dados coletados do Twitter sobre o Ebola e, por fim, na Seção 5, apresentam-se as principais contribuições e possíveis extensões deste trabalho.

2. Referencial Teórico e Trabalhos Relacionados

A seguir, uma breve revisão de alguns conceitos básicos relacionados à Teoria dos Grafos e utilizados neste trabalho. Um *grafo* é um par ordenado $G = (V, E)$ onde V é um conjunto finito não vazio de *vértices*, e E é um conjunto de pares não ordenados de vértices distintos, denominados *arestas* ou *arcos*. Diz-se que G é um *grafo direcionado* se o conjunto E é um conjunto de pares ordenados de vértices. Seja uma aresta $e \in E$, onde $e = (v_i, v_j)$, diz-se que: o vértice v_i é *adjacente* ao vértice v_j, que v_i e v_j são *vizinhos* em G, e que v_i e v_j são *incidentes* a e. Um *laço* é uma aresta onde ambas as extremidades são iguais. Seja G um grafo, o *grau* de um vértice v, denotado por $g(v)$, é igual ao número de vizinhos de v. Se G é um grafo direcionado e v um vértice de G, então o *grau de entrada* de v, denotado por $g_e(v)$, é igual ao número de arcos que terminam em v, e o *grau de saída* de v, denotado por $g_s(v)$, corresponde ao número de arcos com origem em v. Uma sequência de vértices $v_1, ..., v_k$ tal que $(v_j, v_{j+1}) \in E$, $1 \leq j < k$, é denominada *caminho* de v_1 a v_k, onde $k-1$ corresponde ao *comprimento* do caminho entre v_1 e v_k. Denomina-se *distância* entre dois vértices $v, w \in V$, denota-se por $d(v, w)$, o comprimento do menor caminho entre v e w. Um grafo $H = (V, E)$ é um subgrafo de $G = (V, E)$, se $V_H \subseteq V_G$ e $E_H \subseteq E_G$. Um grafo $G = (V, E)$ é *conexo* se existe caminho entre qualquer par de vértices no grafo. Caso contrário, o grafo é dito *desconexo*. Um *componente conexo* é um subgrafo conexo maximal de G. Naturalmente, se o número de componentes conexos de G for igual a um, então G é *conexo*.

A forma mais natural para representação de redes sociais é através da modelagem em grafos, já que os grafos são estruturas subjacentes às redes em geral. Em uma rede social, os usuários correspondem aos vértices e os relacionamentos estabelecidos entre eles correspondem às arestas de um grafo. As métricas relacionadas aos vértices permitem analisar a influência dos atores na rede social examinada. As principais métricas em grafos estritamente relacionadas à análise das redes utilizadas nessa pesquisa são descritas a seguir. A métrica *centralidade de grau* de um vértice v é igual ao valor de $g(v)$. Em grafos direcionados, naturalmente, utilizam-se duas métricas de centralidade de grau, a saber: $g_e(v)$ e $g_s(v)$. Em redes sociais, o grau de entrada é comumente utilizado para medir a popularidade de um ator e o grau de saída costuma estimar a expansividade de um ator. A métrica *centralidade de proximidade* de um vértice v representa a distância entre ele e todos os outros vértices do subgrafo do qual v faz parte. Esta métrica permite identificar os vértices que conseguem se comunicar de forma mais rápida na rede. A métrica *centralidade de intermediação* de um vértice v define o seu grau de participação nos caminhos mais curtos do grafo. Esta métrica permite identificar vértices relevantes que atuam como intermediário entre grupos diferentes, facilitando a disseminação da informação com outros grupos.

Na literatura é possível encontrar alguns trabalhos relacionados com a utilização de redes sociais em situações de epidemia. O estudo apresentado em Corley et al. [2010], sobre o uso de mídias sociais para monitorar as informações sobre o vírus Influenza, mostrou uma correlação entre a frequência das postagens na web relacionadas ao vírus e a quantidade de pacientes reportados pelo Centro de Prevenção e Controle de Desastres dos Estados Unidos. O trabalho apresentado por Mislove et al. [2007] estudou as redes sociais Flickr, YouTube e LiveJournal e analisou a formação de um agrupamento principal fortemente conectado rodeado por agrupamentos menores. Neste contexto, o estudo verificou a importância do agrupamento principal na disseminação de informações na rede como um todo. Por meio de análise espaço-temporal das informações postadas no Twitter, Sakaki et al. [2010] monitoraram a ocorrência de terremotos no Japão em tempo real. No Brasil, Antunes et al. [2014] apresentaram um estudo em andamento sobre o monitoramento da dengue no Brasil por meio da ferramenta e-Monitor Dengue. Neste estudo, observou-se que o número de comentários no Twitter sobre a dengue acompanha o crescimento de números de casos oficiais da doença indicando uma correlação entre os rumores sobre a dengue e o aumento do número de casos notificados.

3. Metodologia

Inicialmente, esta seção descreve um breve resumo de cada uma das principais ferramentas utilizadas neste trabalho. As seções seguintes apresentam as etapas essenciais envolvidas ao longo desta pesquisa. A última seção descreve o fluxo geral das etapas realizadas durante o desenvolvimento do trabalho.

3.1. Principais Ferramentas Utilizadas

O Twitter é uma das redes sociais mais utilizadas no mundo com cerca de 500 milhões de usuários [Gabielkov et al., 2014]. As principais formas de uso do Twitter são: compartilhar acontecimentos do dia-a-dia, responder aos comentários dos amigos, compartilhar informações e comentar sobre noticiários recentes [Java et al., 2007]. Entre as vantagens da utilização do Twitter estão: a possibilidade de postar comentários curtos (*Tweets*) de até 140 caracteres, encaminhar comentários (*Retweet*) de usuários para outros usuários e o mecanismo de relacionamento entre os usuários (*Follower e Followed*). Todas essas vantagens tornam o Twitter uma ferramenta poderosa para a disseminação de informações [Kwak et al., 2010].

O NodeXL - *Network Overview, Discovery and Exploration* - é uma extensão da planilha Microsoft Excel [Smith et al., 2009] que permite coletar dados das principais redes sociais como Twitter, Facebook, Flickr e Youtube. Os dados coletados por meio do NodeXL são armazenados em planilha Excel na forma de grafos com conjunto de vértices e arestas, em que uma aresta unindo dois vértices representa o relacionamento existente entre os mesmos. Com uma interface de fácil uso, o NodeXL possibilita realizar o cálculo das principais métricas em grafos como grau de entrada, grau de saída, coeficiente de clusterização, centralidade de intermediação, centralidade de proximidade, centralidade de autovetor, *pagerank*, além de métricas de grupo.

O PostgreSQL é um sistema gerenciador de banco de dados objeto-relacional, gratuito e de código aberto, desenvolvido a partir do projeto Postgres em 1986 [Stonebraker e Rowe, 1986]. Entre as principais características do PostgreSQL está o seu potencial de extensibilidade, o que possibilitou o desenvolvimento de uma extensão geográfica chamada PosGIS. A extenção PostGIS permite o armazenamento e a manipulação de dados espaciais no banco PostgreSQL. O Full Text Search - *Full Text Search* - é outra ferramenta do PostgreSQL utilizada nesta pesquisa, cujo principal objetivo é efetuar busca textual de documentos no banco de dados.

3.2. Coleta de Dados

A coleta dos comentários do Twitter foi efetuada por meio da ferramenta NodeXL. Os dados foram coletados diariamente durante seis meses, no período de 01/11/2014 a 30/04/2015, totalizando cerca de um milhão de *Tweets*. Foram utilizados o termo de busca 'Ebola' e o filtro da quantidade de comentários. O filtro de idiomas não foi utilizado para não restringir a coleta a determinados países. Utilizou-se o limite de 10.000 comentários diários para efetuar uma análise longitudinal, ao longo de seis meses, ao invés de concentrar a análise em um período restrito de dias ou semanas. Essa quantidade de 10.000 comentários diários não representa todo o universo dos dados postados no Twitter sobre o Ebola neste período. Porém, por meio da análise dos dados de seis meses, é possível estudar a variação e a tendência dos comentários dos usuários nos diversos países.

O período de análise como um todo é referente aos seis meses de coleta, de 01/11/2014 a 30/04/2015. Porém, conforme descrito a seguir, para cada tipo de análise, foram selecionados determinados intervalos de datas para possibilitar uma análise longitudinal ao longo dos seis meses. Para a análise por meio das métricas em grafos, o período de análise refere-se a 51 dias, distribuídos em 27 semanas, ao longo dos seis meses. Em cada semana, foram selecionados dois dias não consecutivos. Para efetuar a análise espacial, o período de análise refere-se aos mesmos 51 dias para os quais foram calculadas as métricas, totalizando 529.772 *Tweets* e 334.500 usuários. Por último, para efetuar a análise textual, foram selecionados quatro dias: 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015. Os primeiros dois dias, por representarem o período inicial da epidemia; o dia 30/04/2015, por representar o final do período de análise desta pesquisa; e o dia 23/12/2014 por permitir acompanhar a tendência dos comentários durante o período de análise.

3.3. Elaboração de Algoritmos

Durante o desenvolvimento desta pesquisa, três algoritmos foram elaborados para dar suporte às análises das redes subjacentes aos dados coletados.

1. *Algoritmo de Identificação de País*: A partir das coordenadas geográficas ou por meio da informação fornecida no campo "Location" dos comentários coletados do Twitter, identifica-se o país de origem. Este algoritmo baseou-se no fluxo do algoritmo utilizado no trabalho de Valkanas e Gunopulos [2012], em que se dividiu a lógica nas fases de limpeza, separação de palavras ou *tokens* e na busca da localidade do usuário do Twitter com o uso das informações do GeoNames - base de dados geográfica gratuita disponibilizada através da licença *Creative Commons License*. Observou-se algumas limitações neste trabalho como, por exemplo, o não tratamento de nomes de localidades iguais que se referem a locais fisicamente diferentes, e o descarte de localidades escritas em caracteres orientais (japonês e chinês). Estas duas limitações são tratadas e superadas no algoritmo desenvolvido nesta etapa.

2. *Algoritmo de Identificação de Idioma*: A motivação para tal deu-se após verificar que, além do inglês, haviam comentários escritos em outros idiomas, como francês e espanhol. Nesta etapa utilizou-se uma ferramenta gratuita que permite a identificação de 160 idiomas diferentes por meio de serviço web: Language Detection API .

3. *Algoritmo Supervisionado de Análise Textual*: Este algoritmo foi desenvolvido para auxiliar na análise textual dos comentários. O algoritmo classifica os comentários em três tipos: *Tweets* não-pessoais (solução de contenção, relato de impactos, notícias de preparo, outros idiomas e outro); *Tweets* pessoais (identificação de *emoticon* e *emoji*, e análise de sentimentos: positivo, neutro ou negativo), e o terceiro tipo Outros.

3.4. Fluxo Geral

A Figura 1 mostra o fluxo das etapas da pesquisa, juntamente com as ferramentas utilizadas. Inicialmente, os comentários do Twitter foram coletados por meio da ferramenta NodeXL. Em seguida, a ferramenta NodeXL foi utilizada para cálculo das métricas e visualização de grafos. A seguir, para possibilitar uma análise espacial e textual dos dados nas etapas posteriores, os comentários do Twitter foram armazenados no banco de dados PostgresSQL. Na etapa seguinte, o algoritmo de identificação de país, codificado na linguagem pgSQL, foi executado para determinar o país de origem dos usuários do Twitter. Após esta etapa, o algoritmo de identificação de idioma, codificado na linguagem Java, foi executado para determinar o idioma utilizado nos comentários do Twitter. Na sequência, o algoritmo supervisionado de análise textual foi executado utilizando-se a ferramenta Full Text Search. Por último, os dados foram visualizados no mapa utilizando-se a ferramenta QGIS.

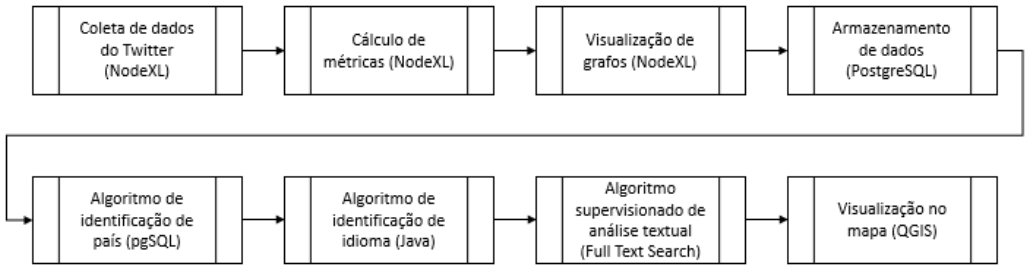

Figura 1: Visão geral das etapas que compõe esta pesquisa.

4. Resultados e Análise

Esta seção efetua uma análise dos dados coletados do Twitter sobre o Ebola e está dividida em três subseções: análise através de grafos, análise espacial e análise textual.

4.1. Análise através de Grafos

Após a coleta, as métricas grau de entrada, grau de saída, centralidade de intermediação e centralidade de proximidade foram calculadas para 51 dias distribuídos entre os seis meses de análise. Entre outros, foram gerados os grafos referentes ao dia 19/11/2014, que faz parte do período de transmissão intensa do vírus do Ebola. Ao todo, o grafo contém 9.226 vértices e 10.371 arestas, sendo 8.338 arestas únicas e 2.033 arestas duplicadas. O grafo possui 5.263 componentes conexos, sendo 4.148 componentes com vértices únicos formado por laços (*self-loop*), 1.075 componentes conexos de dois a dez vértices, 31 componentes conexos de 11 a 30 vértices e 9 componentes conexos com mais de 31 vértices. Os componentes formados por laços representam os usuários cujo comentário (*Tweet*) postado não foi encaminhado (*Retweet*), mencionado (*Mentions*) ou respondido (*Reply*). A Figura 2 apresenta 112 componentes conexos com mais de seis vértices. Os 5.151 demais componentes foram ocultos para melhor visualização. A Figura 3(a) refere-se ao maior componente conexo, identificado por Grafo G1. Este grafo apresenta 987 vértices e 1.582 arestas, sendo 1.252 arestas únicas e 330 arestas duplicadas. A Figura 3(b) refere-se ao segundo maior componente conexo, identificado por Grafo G2. Este grafo apresenta 118 vértices e 120 arestas, sendo 120 arestas únicas. Nota-se que o Grafo G1 é mais denso com várias interconexões entre os vértices e o Grafo G2 é mais esparso e possui dois vértices centrais.

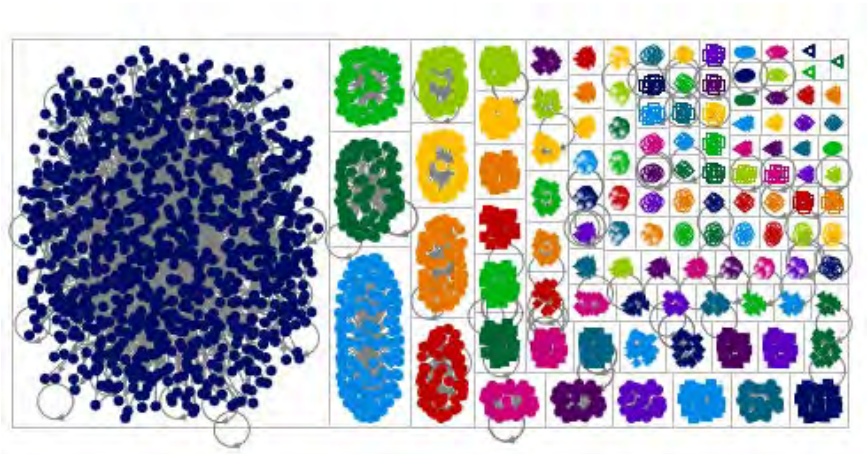

Figura 2: Exemplo de 112 componentes conexos com mais de seis vértices.

(a) Grafo G1 com 987 vértices e 1.582 arestas.

(b) Grafo G2 com 118 vértices e 120 arestas.

Figura 3: Grafos G1 e G2.

Durante esta fase, a análise teve como foco os dois maiores componentes conexos sendo um dos principais objetivos o de observar as tendências de comportamento dos vértices/usuários chaves ao longo de seis meses. Entre tais vértices, o usuário @ebolaphone apresentou os maiores valores de grau de saída e centralidade de intermediação, sugerindo um usuário ativo que posta comentários, responde aos usuários e encaminha comentários de outros usuários, facilitando a disseminação de informações. O usuário @thedouch3 apresentou o maior valor de grau de entrada e centralidade de proximidade, sugerindo um usuário que posta comentários relevantes que são encaminhados por seus seguidores e consegue interagir rapidamente com os outros usuários. O usuário @telegraph apresentou o segundo maior valor de grau de entrada e centralidade de intermediação, sugerindo um usuário que além das características do usuário @thedouch3, comporta-se como um importante disseminador de informações entre grupos diferentes. A análise temporal destes três vértices chaves ao longo de seis meses mostrou que o interesse do usuário @ebolaphone em disseminar informações sobre a epidemia e engajar em discussões permaneceu durante todo o período de análise, o que não ocorreu com o usuário @telegraph que mostrou valores inexpressivos no período. O usuário @thedouch3 mostrou a tendência de postar poucos comentários, porém seus comentários foram encaminhados com maior frequência em relação ao usuário @telegraph. A análise dos vértices que permaneceram por mais tempo ao longo de seis meses entre os 20 maiores valores de centralidade de intermediação e grau de entrada mostrou a presença dos usuários @who e @unicef, que representam respectivamente, a Organização Mundial de Saúde e o Fundo das Nações Unidas para a Infância, sugerindo a relevância das informações postadas por estes usuários. A presença do usuário @youtube entre os vértices que permaneceram por mais tempo entre os 20 maiores valores de grau de entrada sugere o interesse dos usuários em disseminar informações sobre o Ebola na forma de vídeo. A análise dos vértices que permaneceram por mais tempo entre os 20 maiores valores de grau de saída mostrou que são todos usuários individuais, sugerindo que estes costumam postar comentários e responder aos comentários de outros usuários com maior frequência, em relação aos usuários que representam organizações e agências de notícias.

4.2. Análise Espacial

A análise geográfica permite visualizar simultaneamente no mapa os países onde foram registrados comentários sobre o Ebola e os países onde foram identificados vértices com os maiores valores das métricas grau de entrada, grau de saída e centralidade de intermediação. O objetivo desta abordagem é analisar, por meio de visualização no mapa, países onde foram registrados casos de Ebola, países onde foram registrados comentários do Twitter sobre o Ebola, e a localização dos vértices que facilitam a disseminação de informações. A fim de possibilitar a análise espacial sobre os dados do Twitter, o algoritmo de identificação de país foi executado para os dados referentes a 51 dias distribuídos entre os meses de novembro de 2014 a abril de 2015. A Tabela 1 mostra o percentual de comentários postados no Twitter por continente, referente aos meses de novembro de 2014 a janeiro de 2015, e a Tabela 2, referente aos meses de fevereiro de 2015 a abril de 2015. Observa-se que, ao longo dos seis meses, a América do Norte concentra o maior percentual de *Tweets* (34,22% a 47,37%), seguido da Europa (16,37% a 26,22%). A África, onde concentra-se os casos de Ebola, está em terceiro lugar (10,60% a 15,33%).

Foram selecionados os 20 maiores valores das métricas grau de entrada, grau de saída e centralidade de intermediação de cada mês, cujo país de origem do usuário do Twitter foi possível obter por meio do algoritmo de identificação de pais. Estes vértices serão considerados vértices chaves geolocalizados. A Tabela 3 mostra os vértices chaves geolocalizados com os 20 maiores valores de grau de saída referente ao mês de novembro de 2014. Além disso, seis mapas foram gerados para representar os comentários postados no Twitter em cada mês, referentes a novembro de 2014 até abril de 2015. Nesses mapas, estão representados também os usuários que apresentaram os 20 maiores valores das métricas grau de entrada, grau de saída e centralidade de intermediação de cada mês, considerados como vértices chaves geolocalizados. A Figura 4 exibe o mapa referente ao mês de abril de 2015. A graduação de cores nos países representa a quantidade de comentários do Twit-

Tabela 1: Comentários postados no Twitter por continente nos meses de novembro de 2014 a janeiro de 2015.

Continente	nov/2014		dez/2014		jan/2015	
	Tweets	%	Tweets	%	Tweets	%
África	3.688	15,33%	5.203	12,20%	4.725	12,17%
Ásia	3.458	14,37%	4.268	10,01%	4.135	10,65%
Antártica	1	0,00%	0	0,00%	4	0,01%
América do Norte	8.234	34,22%	19.082	44,75%	18396	47,37%
América do Sul	1.982	8,24%	4.672	10,96%	4.305	11,09%
Europa	5.893	24,49%	8.770	20,57%	6.358	16,37%
Oceania	809	3,36%	643	1,51%	910	2,37%
Total	24.065		42.638		38.833	

Tabela 2: Comentários postados no Twitter por continente nos meses de fevereiro de 2015 a abril de 2015.

Continente	fev/2015		mar/2015		abr/2015	
	Tweets	%	Tweets	%	Tweets	%
África	3.902	12,74%	4.005	10,60%	6.043	15,17%
Ásia	3.402	11,10%	3.676	9,73%	3.482	8,74%
Antártica	0	0,00%	0	0,00%	1	0,00%
América do Norte	13.627	44,48%	17.966	47,54%	15.690	39,38%
América do Sul	3.715	12,13%	3.766	9,96%	3.555	8,92%
Europa	5.077	16,57%	7.585	20,07%	10.446	26,22%
Oceania	912	2,98%	797	2,11%	622	1,56%
Total	30.635		37.795		39.839	

ter e o mapa de calor representa os vértices chaves geolocalizados. O mapa de calor é representado por círculos cuja graduação de cores é calculada utilizando-se um atributo numérico previamente selecionado. A intensidade da cor representa o "calor" de um determinado atributo [Trame e Keßler, 2011].

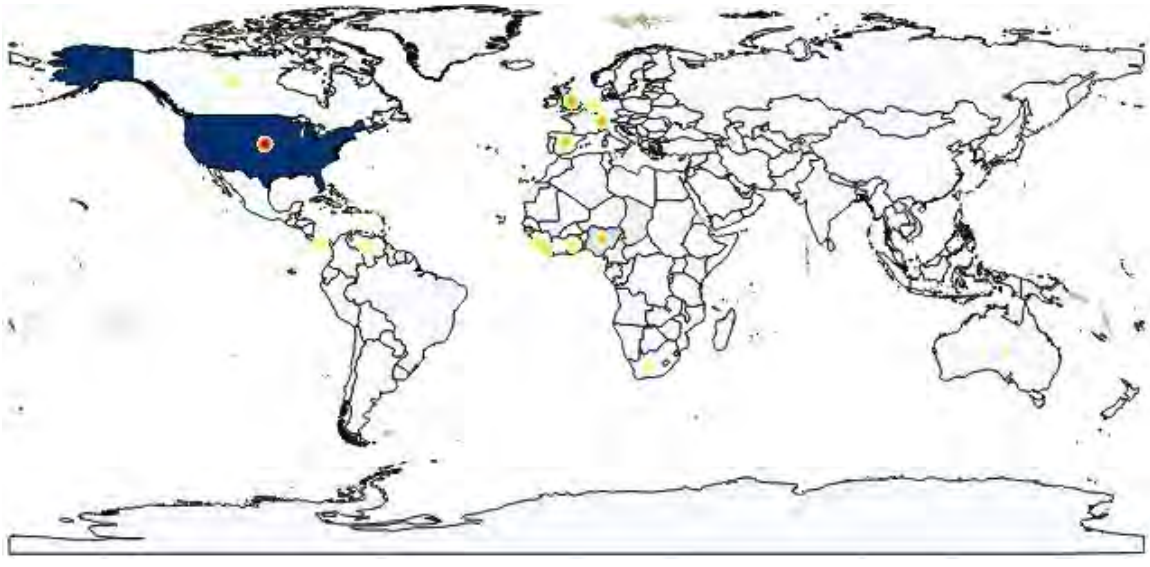

Figura 4: Abril/2015 - Quantidade de tweets por país e vértices chaves geolocalizados.

Tabela 3: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - novembro/2014.

Usuário	País	Grau de entrada	Location	Tipo de local
billgates	Estados Unidos	630	Seattle, WA	Nome de cidade
alexmanns1	Estados Unidos	390	Detroit , Michigan	Nome de cidade
decappeal	Reino Unido	185	London, UK	Nome de capital
savechildrenuk	Reino Unido	181	London, UK	Nome de capital
onedirection	Reino Unido	181	London	Nome de capital
nytimes	Estados Unidos	150	New York City	Nome de cidade
dxxmien	França	133	Falestín / le mans	Nome de cidade
nytimes	Estados Unidos	101	New York City	Nome de cidade
telegraph	Reino Unido	83	London, UK	Nome de capital
nytimes	Estados Unidos	80	New York City	Nome de cidade
bbcbreaking	Reino Unido	74	London, UK	Nome de capital
adrianlb_ok	Argentina	73	Lomas de Zamora, Buenos Aires	Nome de capital
bbcworld	Reino Unido	66	London, UK	Nome de capital
jaimegmora	Espanha	63	Madrid	Nome de capital
ebolaphone	Países Baixos	62	Utrecht, The Netherlands	Nome de cidade
eboladeeply	Estados Unidos	51	New York	Nome de região administrativa
msf_italia	Itália	50	Italia	Nome de país
yusnaby	Cuba	48	La Habana, #Cuba	Nome de país
ajenews	Qatar	44	Doha, Qatar	Nome de país
unicefkorea	Coreia do Sul	43	Seoul, Korea	Nome de capital

A análise espacial no mapa dos dados divulgados pela OMS em 26/04/2015 mostrou casos isolados do Ebola nos Estados Unidos, Reino Unido, Espanha e transmissão intensa nos países da África Ocidental: Libéria, Serra Leoa e Guiné. A análise geográfica dos comentários do Twitter possibilitou verificar que, durante os meses de novembro de 2014 a abril de 2015, a América do Norte apresentou o maior percentual de comentários (34,22% a 47,37%), seguido da Europa (16,37% a 26,22%). A África, onde concentra-se os casos de Ebola, está em terceiro lugar (10,60% a 15,33%). A análise de vértices chaves geolocalizados mostrou os 20 vértices com os maiores valores de centralidade de intermediação, grau de entrada e grau de saída durante os meses de novembro de 2014 a abril de 2015. Apesar do baixo percentual de comentários da África em relação à América do Norte e Europa, verificou-se a presença de quatro vértices geolocalizados da África no mês de novembro de 2014: @tz_uchay da Nigéria, @bodacious_lyn da Zâmbia, @benaskay de Gana e @allafrica da Tunísia. A análise temporal do mapa de calor, referente aos seis meses de análise, mostrou uma concentração de comentários e de vértices geolocalizados nos Estados Unidos, porém verificou-se também a presença de vértices geolocalizados nos países da África Ocidental (Nigéria, Libéria e Serra Leoa) ao longo dos seis meses, sugerindo o interesse dos usuários destes países em comentar e disseminar informações sobre a epidemia.

4.3. Análise Textual

Para esta análise, foram selecionados somente comentários de usuários da África, por representarem os países onde ocorreram a transmissão intensa da epidemia. Foram processados 2.452 comentários referentes aos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015, por meio do algoritmo supervisionado de análise textual. Os dias 08/11/2014 e 19/11/2014 foram selecionados por representarem o período inicial da epidemia, permitindo verificar o que os usuários comentavam

quando o vírus se espalhou de forma intensa na África Ocidental. O dia 30/04/2015 foi selecionado por representar o final do período de análise desta pesquisa e os dados do dia 23/12/2014 permitem acompanhar a tendência dos comentários durante o período de análise. Inicialmente foi analisada a variação dos *Tweets* pessoais e não-pessoais ao longo dos quatro dias selecionados. Em seguida, foi efetuada uma análise sobre a disseminação dos comentários dos vértices chaves na África e dos vértices chaves geolocalizados da África para outros continentes.

A análise textual temporal dos comentários mostrou que o percentual de *Tweets* pessoais diminuiu de 42,78% para 25,04%, entre os dias 08/11/2014 e 30/04/2015 e o percentual de *Tweets* não-pessoais aumentou de 50,13% para 70,77% no mesmo período. Estes números sugerem que, no início da epidemia, quando verificou-se uma rápida expansão do vírus nos países da África Ocidental, os usuários manifestavam opiniões pessoais com mais frequência. À medida em que os meses foram passando, apesar do avanço constante do vírus no continente, os usuários passaram a disseminar com mais frequência noticiários sobre a epidemia. A análise dos *Tweets* não-pessoais mostrou uma tendência de aumento de comentários sobre relatos de casos da epidemia até dezembro/2014 e a diminuição até abril/2015. Com relação aos comentários sobre solução de contenção, ocorre uma queda no dia 19/11/2014 e um aumento gradativo até abril/2015. A análise dos *Tweets* pessoais mostrou uma diminuição do percentual de comentários com sentimento negativo de 17,59% para 8,83%, entre os dias 08/11/2014 e 30/04/2015. Não foi possível observar um padrão de variação dos comentários com sentimento positivo e sentimento neutro a partir dos dados destes quatro dias.

A análise dos vértices chaves mostrou que o usuário @ebolaphone, cujo país de origem são os Países Baixos, encaminhou comentários de esclarecimento sobre a epidemia para usuários da África e também teve seus comentários mencionados por usuários da África, sugerindo a importância deste usuário na disseminação de informações. A análise mostrou também que os vértices chaves geolocalizados da África: @ebolaalert, da Nigéria, @unicef_liberia, da Libéria, @unmeer, de Gana, @lateam224, de Guiné e @allafrica, da Tunísia, participaram da disseminação de informações sobre a epidemia para usuários de outros continentes, sugerindo a relevância do conteúdo disseminado.

5. Conclusão

Este trabalho se insere nos esforços investigativos de pesquisadores que buscam compreender as possibilidades de utilização das redes sociais como meio de articular diversos atores para disseminar informações relevantes e intervir em casos de epidemias em nível mundial. Como contribuição principal, apresenta-se uma estratégia combinada de análise de redes sociais. Após coletar aproximadamente um milhão de *Tweets* ao longo de seis meses, foram selecionados determinados intervalos de datas para possibilitar uma análise longitudinal ao longo deste período. Então, três tipos de análises foram realizadas. Na primeira etapa, obtidos os grafos subjacentes aos dados coletados no Twitter, foram selecionadas as métricas mais relevantes para o estudo em questão. Nesta fase, o foco da análise concentrou-se em 51 dias, distribuídos em 27 semanas, totalizando 529.772 *Tweets* e 334.500 usuários. Em seguida, investigou-se a análise espacial, cujo período de análise refere-se aos mesmos 51 dias. Por fim, foi realizada a análise textual dos comentários dos atores considerados centrais nas duas etapas anteriores. Nesta última etapa, selecionou-se quatro dias: 08/11/2014 e 19/11/2014, por representarem o período inicial; 30/04/2015, por ser o último dia de coleta; e o dia 23/12/2014 por permitir acompanhar a tendência dos comentários durante o período da análise global. O uso da combinação dos três tipos de análise possibilitou reforçar os principais resultados apresentados por cada uma delas separadamente, bem como descartar resultados que não se mostraram claros ou mesmo foram conflitantes entre as três etapas. Além disso, foram desenvolvidos três algoritmos para auxiliar no processo de análise combinada. A saber: *Algoritmo de Identificação de País*, *Algoritmo de Identificação de Idioma* e *Algoritmo Supervisionado de Análise Textual.*

6. Agradecimentos

Os autores agradecem à CAPES e à FAPERJ (projeto E-26/203.446/2015 - BBP) pelo apoio financeiro.

Referências

Antunes, M. N., da SILVA, C. H., Guimarães, M. C. S., e Rabaço, M. H. L. (2014). Monitoramento de informação em mídias sociais: o e-monitor dengue. *TransInformação*, 26(1):9–18.

Corley, C. D., Cook, D. J., Mikler, A. R., e Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health*, 7(2):596–615.

Facebook. `http://www.facebook.com/`. Último acesso em 24/01/2016.

Flickr. `http://www.flickr.org/`. Último acesso em 24/01/2016.

Full Text Search. Full text search. `http://www.postgresql.org/docs/9.3/static/textsearch.html/`. Último acesso em 24/01/2016.

Gabielkov, M., Rao, A., e Legout, A. (2014). Studying social networks at scale: Macroscopic anatomy of the twitter social graph. *arXiv preprint arXiv:1404.1355*.

GeoNames. Geonames. `http://www.geonames.org/`. Último acesso em 24/01/2016.

Java, A., Song, X., Finin, T., e Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, p. 56–65. ACM.

Khan, A. S., Tshioko, F. K., Heymann, D. L., Le Guenno, B., Nabeth, P., Kerstiëns, B., Fleerackers, Y., Kilmarx, P. H., Rodier, G. R., Nkuku, O., et al. (1999). The reemergence of ebola hemorrhagic fever, democratic republic of the congo, 1995. *Journal of Infectious Diseases*, 179(Supplement 1):S76–S86.

Kwak, H., Lee, C., Park, H., e Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, p. 591–600. ACM.

Language Detection API. Language detection api. `http://detectlanguage.com/`. Último acesso em 24/01/2016.

Lazard, A. J., Scheinfeld, E., Bernhardt, J. M., Wilcox, G. B., e Suran, M. (2015). Detecting themes of public concern: A text mining analysis of the centers for disease control and prevention's ebola live twitter chat. *American journal of infection control*, 43(10):1109–1111.

LiveJournal. `http://www.livejournal.com/`. Último acesso em 24/01/2016.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., e Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, p. 29–42. ACM.

NodeXL. `http://nodexl.codeplex.com/`. Último acesso em 24/01/2016.

Oyeyemi, S. O., Gabarron, E., e Wynn, R. (2014). Ebola, twitter, and misinformation: a dangerous combination?

pgSQL. `http://www.postgresql.org/docs/9.3/static/plpgsql.html/`. Último acesso em 24/01/2016.

PosGIS. http:/postgis.net/. Último acesso em 24/01/2016.

QGIS. http://www.qgis.org/. Último acesso em 24/01/2016.

Sakaki, T., Okazaki, M., e Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, p. 851–860. ACM.

Simon, T., Goldberg, A., e Adini, B. (2015). Socializing in emergencies-a review of the use of social media in emergency situations. *International Journal of Information Management*, 35(5): 609–619.

Smith, M. A., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., Capone, T., Perer, A., e Gleave, E. (2009). Analyzing (social media) networks with nodexl. In *Proceedings of the fourth international conference on Communities and technologies*, p. 255–264. ACM.

Stonebraker, M. e Rowe, L. A. (1986). *The design of Postgres*, volume 15. ACM.

Team, W. E. R. et al. (2014). Ebola virus disease in west africa-the first 9 months of the epidemic and forward projections. *N Engl J Med*, 371(16):1481–95.

Trame, J. e Keßler, C. (2011). Exploring the lineage of volunteered geographic information with heat maps. *GeoViz: Hamburg, Germany*.

Twitter. http://twitter.com/. Último acesso em 24/01/2016.

Valkanas, G. e Gunopulos, D. (2012). Location extraction from social networks with commodity software and online data. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, p. 827–834. IEEE.

YouTube. http://www.youtube.com/. Último acesso em 24/01/2016.