UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

EXPLORANDO REDES SOCIAIS COMO FERRAMENTA DE DISSEMINAÇÃO DE INFORMAÇÕES: UMA ANÁLISE ESPAÇO TEMPORAL EM CASOS DE EPIDEMIA

Liriam Michi Enamoto

Orientadores

Adriana Cesário de Faria Alvim

Vânia Maria Félix Dias

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO de 2016

EXPLORANDO REDES SOCIAIS COMO FERRAMENTA DE DISSEMINAÇÃO DE INFORMAÇÕES: UMA ANÁLISE ESPAÇO TEMPORAL EM CASOS DE EPIDEMIA

Liriam Michi Enamoto

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Prof. Adriana Cesário de Faria Alvim, D.Sc.- UNIRIO

Prof. Vânia Maria Félix Dias, D.Sc. - UNIRIO

Prof. Asterio Kiyoshi Tanaka, D.Sc. - UNIRIO

Prof. Mariano Pimentel, D.Sc. - UNIRIO

Prof. Maristela Terto de Holanda, D.Sc. - UnB

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO de 2016

Enamoto, Liriam Michi
E56 Explorando redes sociais como ferramenta de disseminação de informações: uma análise espaço temporal em casos de epidemia / Liriam Michi Enamoto, 2016.
115f.; 30 cm.

Orientadora: Adriana Cesário de Faria Alvim.
Coorientadora: Vânia Maria Félix Dias.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2016.

1. Redes sociais on-line. 2. Teoria dos grafos. 3. Twitter (Rede social on-line). 4. Ebola (Doença). 5. Saúde pública. I. Alvim, Adriana Cesário de Faria. II. Dias, Vânia Maria Félix. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD - 025.04

A minha família e a meu pai (*in memorian*).

Agradecimentos

Primeiramente eu agradeço a Deus por ter me dado a oportunidade de fazer este curso de mestrado.

Agradeço aos meus pais e a minha família por acreditarem em mim e por todo o apoio e dedicação.

Agradeço as minhas orientadoras, Prof. Adriana Alvim e Prof. Vânia Felix, que estiveram sempre presentes e disponíveis, compartilhando conhecimento e me guiando nesta curta jornada que me proporcionou tantos aprendizados.

Além das orientadoras, eu deixo um agradecimento aos demais professores do PPGI, por todo o aprendizado proporcionado durante o curso e aos colegas de classe.

Enamoto, Liriam Michi. **EXPLORANDO REDES SOCIAIS COMO FERRAMENTA DE DISSEMINAÇÃO DE INFORMAÇÕES: UMA ANÁLISE ESPAÇO TEMPORAL EM CASOS DE EPIDEMIA**. UNIRIO, 2016. 115 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

As redes sociais têm sido utilizadas por pessoas de diversos países como uma ferramenta de comunicação, gerando conteúdo sobre tópicos variados e permitindo o compartilhamento de informações. Em situações de desastre e saúde pública, como, por exemplo, a recente epidemia provocada pelo vírus do Ebola, as redes sociais têm sido utilizadas para fornecer informações atualizadas de fontes oficiais e não-oficiais, incluindo informações educativas à população. A análise destas informações, quando combinadas com suas respectivas localizações geográficas, permite extrair informações valiosas. O objetivo deste trabalho é investigar o uso do Twitter relacionado à epidemia do Ebola ocorrida em 2013, 2014 e 2015 combinando três tipos de análises: análise por meio da Teoria de Grafos; análise espacial utilizando banco de dados geográficos; e análise textual. Para tal, foram coletados comentários do Twitter diariamente durante seis meses, no período de 01/11/2014 a 30/04/2015. A análise dos comentários por meio da Teoria de Grafos permite identificar a presença de vértices que auxiliam na disseminação de informações e seu comportamento ao longo de seis meses. A análise espacial permite visualizar no mapa, países onde foram registrados comentários do Twitter sobre o Ebola e a localização dos vértices que facilitam a disseminação de informações. A análise textual dos comentários de usuários da África permite verificar a relevância do conteúdo postado e a disseminação destas informações para usuários de outros continentes. A primeira contribuição técnica desta pesquisa refere-se ao algoritmo de identificação de país desenvolvido para possibilitar a análise espacial. Como resultado, o algoritmo proposto fornece informações de localização de dados do Twitter, permitindo efetuar análises espaciais em diversas áreas, não se limitando à área de saúde pública. A segunda contribuição técnica é o corpus de categorização de comentários do Twitter para situações de epidemia e saúde pública resultante do algoritmo supervisionado de análise textual. Este corpus permite servir de base

para a categorização de novos comentários do Twitter em inglês relacionados ao Ebola ou a outros tipos de epidemia. Em situações de calamidade e emergência como epidemias, o fornecimento de informações corretas e tempestivas à população são importantes medidas a serem adotadas pelas autoridades de saúde. Estas medidas auxiliam a minimizar o crescimento do surto e diminuir as incertezas e a ansiedade das pessoas. A estratégia de análise de redes sociais como um todo apresentada nesta pesquisa pretende contribuir para aprofundar o conhecimento sobre a utilização das redes sociais como meio de comunicação visando disseminar informações relevantes em casos de saúde pública em nível mundial.

Palavras-chave: Análise de redes sociais, Teoria de Grafos, localização geográfica, epidemia, análise textual.

ABSTRACT

Social network have been used as a communication tool by people in many countries creating content about a variety of topics and sharing information. When disasters and public health situations occurs, such as the recent Ebola virus epidemic, social networks have been frequently used, providing up-to-date information from official and unofficial sources, including educational information to the population. Valuable information can be extracted analyzing the information generated by social network along with its geographic location. This study aims to investigate the Twitter usage related to 2013, 2014 and 2015 Ebola epidemic combining three types of analysis of Twitter posts: Graph Theory analysis, spatial analysis using geographic database and textual analysis. The dataset used in this study was daily crawled from Twitter for a period of six months starting from 01/11/2014 to 30/04/2015. The analysis of Twitter posts using Graph Theory allow identify vertex that help information dissemination and their behavior throughout six months. The spatial analysis allow visualize on map countries with Twitter posts related to Ebola and vertex location that facilitate information dissemination. The textual analysis of African users posts allow study the relevance of Twitter posts and the information sharing to users of other continents. The first technical contribution of this work is the country identification algorithm developed to perform spatial analysis. As a result, the proposed algorithm provide Twitter posts location that can be used to perform spatial analysis in many areas, not limited to public health area. The second technical contribution is the corpus used to categorize Twitter posts related to epidemic and public health situation. This corpus can be used to categorize new English Twitter posts related to Ebola or other kinds of epidemics. In case of risk and emergency situation like epidemics, provide accurate and on time information for the population is one of the important measures to be adopted by public health authorities. This measure might avoid epidemic spread and reduce people anxiety and uncertainty. The overall social network analysis strategy presented in this work aims to contribute improving our knowledge about the social network usage as a communication tool to disseminate relevant information in cases of worldwide public health situation.

Keywords: Social network analysis, Graph Theory, outbreak, geolocation, textual analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Nomenclaturas

GPS	Global Positioning System
ISO	International Organization for Standardization
URL	Uniform Resource Locator
UTF-8	8-bit Unicode Transformation Format

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. Segundo a Organização Mundial de Saúde [18], a epidemia do vírus Influenza H1N1, ocorrido em 2009, foi responsável por 12.220 mortes registradas em 208 países, afetando principalmente grupos de alto risco (idosos e pessoas com doenças crônicas). A transmissão se dá pelo ar, por meio de tosse e espirros de pessoas já infectadas pelo vírus, sugerindo que a mobilidade das pessoas contribuiu para aumentar a exposição do vírus ao redor do mundo.

Mais recentemente, a epidemia provocada pelo vírus do Ebola contabilizou 11.316 mortes, segundo dados divulgados pela Organização Mundial de Saúde em fevereiro de 2016 [10]. O contágio do vírus se dá através do contato com fluidos corpóreos de um paciente que já apresenta os sintomas do Ebola. 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% [65]. Diversas vacinas estão em fase de testes em animais e seres humanos, porém até o presente momento, nenhuma vacina foi aprovada pela Organização Mundial de Saúde 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 [41]. A última epidemia do Ebola iniciou-se na Guiné, na África Ocidental, em dezembro de 2013. No dia 25/08/2014, a Organização Mundial de Saúde 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. 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 [65]. No ritual de enterro tradicional praticado nos países africanos, o corpo é preparado pelo tio paternal ou a mulher mais velha dos parentes, por parte de pai. Após remover as roupas, o corpo é lavado e vestido. Durante a cerimônia de funeral, todos os membros da família lavam as mãos na mesma água e tocam no rosto do morto, como demonstração de amor e respeito. Esta cerimônia tradicional de funeral apresenta possíveis riscos de transmissão por meio de fluidos corpóreos, caso o corpo esteja infectado pelo vírus [37]. Para possibilitar um funeral digno e seguro e evitar a resistência das comunidades, foram criados protocolos em que se busca a compreensão dos familiares em relação a necessidade de práticas seguras de funeral, respeitando seus direitos pessoais e religiosos. Entre alguns exemplos de tais práticas estão: a preparação do funeral por pessoas treinadas e equipadas com máscaras, luvas e roupas de proteção, o acondicionamento do corpo em sacos plásticos e desinfecção do local da cerimônia e dos participantes após o funeral. Apesar dos esforços das organizações governamentais e de saúde, a epidemia continuou a avançar nos países da África Ocidental, registrando um total 10.889 mortes e 26.277 infectados em 29/04/2015. 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 [10]. A partir dos dados divulgados pela Organização Mundial de Saúde foi criado o gráfico da Figura 1.1. O gráfico mostra um aumento acentuado de casos da epidemia de agosto de 2014 a abril de 2015 e uma gradual estabilização até fevereiro de 2016.

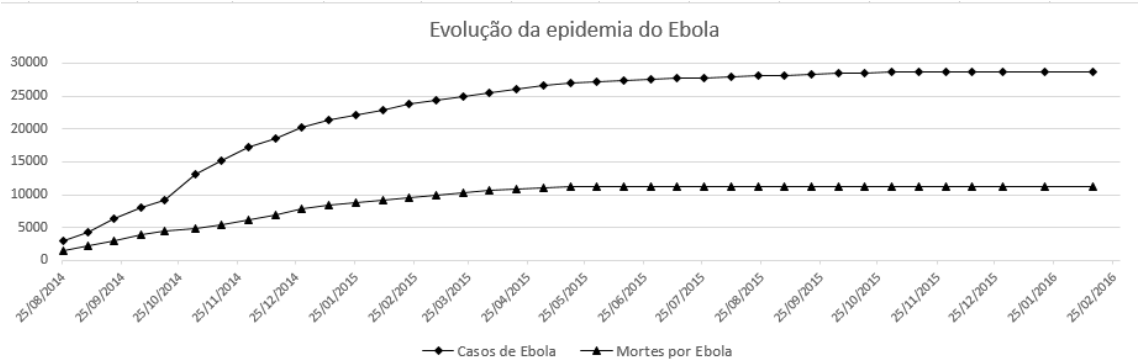


Figura 1.1: Gráfico da evolução da epidemia do Ebola do período agosto/2014 a fevereiro/2016.

Guiné foi declarada sem novos casos de transmissão do vírus no dia 29/12/2015, entrando no período de monitoramento de 90 dias. Durante o período de monitoramento, a população é orientada a reportar às autoridades qualquer caso de pessoas com sintomas do Ebola ou caso de morte suspeita por Ebola. São efetuadas coletas de sangue destas pessoas

e caso seja confirmada a contaminação pelo vírus, os contatos familiares e sociais desta pessoa são investigados para evitar o desencadeamento de novas contaminações. A Libéria terminou o período de monitoramento de 90 dias no dia 14/01/2016, sem registro de novos casos de transmissão. Serra Leoa havia sido declarado sem novos casos de transmissão do vírus no dia 07/11/2015, porém no dia 14/01/2016, durante os 90 dias do período de monitoramento, foi registrada a prática de um enterro não-seguro, cuja morte havia sido causada pelo vírus. O monitoramento de 150 pessoas que participaram do enterro permitiu detectar a transmissão do vírus a um parente da vítima. Apesar da estabilização da epidemia, a Organização Mundial de Saúde divulgou, no dia 24/02/2016, um relatório sobre os cuidados necessários para os sobreviventes do Ebola. Tais cuidados envolvem o tratamento clínico adequado, cuidados com relação ao possível risco de transmissão do vírus por meio do sêmen e do leite materno e monitoramento dos sobreviventes da epidemia, indicando a importância da divulgação de informações educativas atualizadas para evitar uma nova epidemia.

Dados divulgados pela Organização Mundial de Saúde em 02/11/2014 indicaram 310 mortes de profissionais da área de saúde e 546 infectados pelo vírus. Em 04/11/2015, este número aumentou para 513 mortes e 881 infectados. Este fato indica que a epidemia não afetou somente as pessoas que vivem em comunidades, mas também afetou pessoas que trabalham na área de saúde e que teriam mais acesso às orientações de prevenção da epidemia. O estudo de Jacobsen et al. [38] identificou algumas lições aprendidas na epidemia do Ebola de 2013, 2014 e 2015. Entre alguns exemplos, o estudo identificou que a população das áreas mais afetadas recebeu informações confusas e contraditórias sobre as práticas de funerais, proibições de viagens e remédios caseiros; enfermeiros e médicos foram expostos ao vírus por falta de informações corretas sobre procedimentos de remoção de dejetos contaminados, transportes de pacientes e treinamento adequado. O mesmo estudo ainda sugere que, na epidemia do Ebola, a falta de informações corretas de prevenção e forma de transmissão dificultaram a conter os avanços da epidemia nos países africanos.

1.1 Definição do Problema

Em casos de risco à saúde pública, como a epidemia do Ebola, a disseminação de informações corretas de formas de prevenção do vírus e formas de tratamento são importantes medidas a serem adotadas pelas autoridades de saúde. Estas medidas auxiliam a minimizar o contágio do vírus, diminuir as incertezas e ansiedades das pessoas causadas pela ameaça da epidemia e diminuir os investimentos em medidas reativas à epidemia, como

a criação de centros especializados de tratamento, treinamento de equipes, tratamento e monitoramento da doença.

De acordo com Simon et al. [61], 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. Durante a epidemia do vírus da Influenza H1N1 ocorrido em 2009, o departamento de saúde do estado de Virgínia, nos Estados Unidos, utilizou as redes sociais para orientar a população sobre locais de vacinações disponíveis [48]. No trabalho de Lazard et al. [44], 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 [54] 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.

1.2 Objetivos

O objetivo geral desta pesquisa é analisar o uso do Twitter [16] como ferramenta de disseminação de informações na epidemia do Ebola ocorrida em 2013, 2014 e 2015 com ênfase em três abordagens: análise por meio da Teoria de Grafos; análise espacial utilizando banco de dados geográficos; e análise textual do conteúdo disseminado.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral foram definidos os seguintes objetivos específicos:

1. Coletar comentários do Twitter sobre o Ebola, por um período de seis meses;
2. Considerando-se os dados coletados no item anterior, armazenar os dados como grafos, com conjuntos de vértices e arestas;
3. Identificar os países onde Twitter foi utilizado para postar comentários sobre o Ebola, considerando-se os dados produzidos nos itens anteriores;
4. Utilizar métricas da Teoria de Grafos para analisar os grafos subjacentes aos dados coletados e verificar se eles auxiliam na disseminação de informações;

5. Verificar se, em especial, os usuários do Twitter no continente africano participaram na disseminação de informações sobre a epidemia;

6. Por fim, verificar se as informações disseminadas são relevantes.

1.3 Organização do Texto

Esta dissertação está organizada em cinco capítulos. O Capítulo 1 apresenta a introdução, com a definição do problema e a descrição dos objetivos da pesquisa. O Capítulo 2 apresenta conceitos básicos sobre redes sociais, Teoria de Grafos, representação de dados geográficos e apresenta uma revisão bibliográfica sobre redes sociais em situações de epidemia, localização geográfica de dados do Twitter e análise textual de comentários do Twitter. O Capítulo 3 descreve a metodologia utilizada, as ferramentas e dados utilizados, os algoritmos desenvolvidos para compor o fluxo das etapas desta pesquisa. No Capítulo 4, são apresentadas as análises por meio da Teoria de Grafos, a análise espacial e a análise textual dos dados. Por fim, o Capítulo 5 apresenta as contribuições desta pesquisa, discussões relacionadas aos resultados obtidos e as propostas de trabalhos futuros identificados ao longo da pesquisa.

2. Revisão Bibliográfica

Este capítulo está dividido em seis seções. A primeira seção define o conceito de redes sociais segundo Wasserman [69]. Na Seção 2.2 apresentam-se conceitos básicos da Teoria de Grafos utilizados nesta dissertação para representar matematicamente as redes sociais. A Seção 2.3 faz uma revisão bibliográfica sobre o uso de redes sociais em situações de calamidade e epidemia. A Seção 2.4 apresenta conceitos básicos de representação de dados geográficos e a Seção 2.5 apresenta alguns trabalhos relacionados a localização geográfica de dados do Twitter disponíveis na literatura. Por último, a Seção 2.6 faz uma revisão bibliográfica sobre a análise textual de comentários do Twitter relacionados à epidemia.

2.1 Análise de Redes Sociais

De acordo com Wasserman [69], as redes sociais consistem em um conjunto finito de atores e relacionamentos entre eles. Analisar as redes sociais permite estudar o relacionamento de estruturas sociais dentro de um determinado contexto como, por exemplo, o estudo para descobrir membros de rede de terroristas [70], analisar a rede de citações entre pesquisadores para verificar o nível de colaboração entre eles [53], o estudo de redes sociais formadas na internet para verificar as características de redes em grande escala [49], entre outros. Para analisar uma rede social é importante definir a fronteira do conjunto de atores e identificar o *modeling unit*, ou a população a ser estudada, que pode ser um ator, pares de atores, subconjunto de atores ou a rede como um todo. Uma vez identificado o *modeling unit*, define-se a unidade de observação, que é a entidade sob a qual a medição será efetuada e o tipo de relacionamento entre os atores que pode ser não-direcionado ou direcionado. Um relacionamento é direcionado se a relação entre dois atores possui uma origem e um destino, isto é, a relação é orientada de um ator para outro. Um relacionamento é não-direcionado se não houver uma direção na relação entre os atores. [69]

Existem várias formas de representar redes sociais matematicamente. As mais utiliza-

das são por meio da Teoria de Grafos, sociogramas e notação algébrica [69].

2.2 Teoria de Grafos

Uma das formas de representar e analisar as redes sociais é por meio de Grafos, em que os usuários ou atores representam os vértices e os relacionamentos entre os atores representam as arestas. As métricas relacionadas aos vértices permitem analisar a influência dos atores na rede social formada [69].

A seguir serão apresentados alguns conceitos importantes sobre a Teoria de Grafos que serão utilizados ao longo desta pesquisa.

Um *grafo G(V,E)* é um conjunto finito não vazio V de n nodos e um conjunto E de m pares não ordenados de elementos distintos de V [69]. Os elementos de V são os *vértices* e os de E são *arestas* de G, respetivamente. Um vértice é *incidente* a uma aresta e a aresta é incidente a um vértice se o vértice é um dos pares não ordenados de vértices que compõem a aresta. Dois vértices v_1 e v_2 são *adjacentes* quando existe uma aresta $e=(v_1,v_1)$, tal que $e \in E$. Um grafo G_s é um *subgrafo* de G se o conjunto de vértices de G_s é um subconjunto dos vértices de G, e o conjunto de arestas de G_s é um subconjunto das arestas de G. *Díades* representam pares de vértices que podem ou não estar unidos por uma aresta. *Tríades* são subgrafos formados por três vértices os quais podem ou não estarem unidos por arestas.

O *grau* de um vértice é representado pela quantidade de vértices adjacentes que possui. Um *caminho* em um grafo é uma sequência de vértices e arestas, iniciando e terminando em vértices de forma que cada vértice é incidente à aresta anterior e posterior. O *comprimento* do caminho é a quantidade de arestas contidas no caminho.

Um grafo é *conexo* se existe um caminho entre qualquer par de vértices no grafo. Se não existir um caminho em pelo menos um par de vértices, o grafo é dito *desconexo*. Os vértices de um grafo desconexo podem ser particionados em dois ou mais subconjuntos de modo que não existam caminhos entre os vértices em diferentes subconjuntos. O subgrafo conexo é chamado de *componente*. Assim, *componente conexa* é o subgrafo conexo maximial de G. Se existir apenas uma componente conexa no grafo, o grafo é *conexo*. Agrupamento, ou *cluster* em inglês, surge da necessidade de classificar ou organizar elementos em grupos coerentes segundo algum critério de semelhança previamente definido [42]. Em grafos, vértices podem ser agrupados em *clusters* de maneira que exista o máximo de arestas em cada *cluster* e relativamente poucas arestas entre *clusters* diferentes [58].

Um *grafo direcionado* consiste em um conjunto de vértices e um conjunto de arestas

entre pares de vértices representado por um relacionamento direcionado. A métrica *grau de entrada* de um vértice *v* corresponde ao número de arestas que terminam em *v*. A métrica *grau de saída* de um vértice *v* corresponde ao número de arestas que originam de *v*. Em redes sociais, o *grau de entrada* mede a popularidade de um ator e o *grau de saída* mede a expansividade de um ator.

A métrica *centralidade de proximidade* (em inglês, *closeness centrality*) de um vértice *v* representa a distância entre ele e todos os outros vértices do subgrafo do qual *v* faz parte, considerando-se como comprimento, os seus caminhos mais curtos. Esta métrica permite identificar os vértices que conseguem se comunicar de forma mais rápida na rede formada.

Seja $d(v_i,v_j)$ a distância entre os vértices v_i e v_j, a Equação 2.1 calcula a centralidade de proximidade C_c do vértice v_i.

$$C_c(v_i) = \left[\sum_{j=1}^{n} d(v_i, v_j)\right]^{-1} \tag{2.1}$$

O valor máximo desta métrica depende da quantidade de nodos *n* do grafo. Desta forma, para efetuar a comparação desta métrica entre grafos de tamanhos diferentes, o valor deve ser normalizado utilizando a Equação 2.2.

$$C_c'(v_i) = (n-1)C_c(v_i) \tag{2.2}$$

A métrica *centralidade de intermediação* (em inglês, *betweeness centrality*) 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.

Seja g_{jk} a quantidade de caminhos mais curtos entre os vértices *j* e *k*, e seja $g_{jk}(v_i)$ a quantidade de caminhos mais curtos entre os vértices *j* e *k* que passam pelo vértice v_i, a Equação 2.3 calcula a *centralidade de intermediação* C_b de um vértice v_i [69].

$$C_b(v_i) = \sum_{j<K} g_{jk}(v_i)/g_{jk} \tag{2.3}$$

Por fim, aplicando-se a Teoria de Grafos à análise de redes sociais, diz-se que um relacionamento é forte quando existe um caminho direcionado entre os vértices v_i e v_j e um caminho direcionado entre os vértices v_j e v_i. O caminho entre v_i e v_j pode conter

vértices e arestas diferentes do caminho entre v_j a v_i. Um relacionamento é fraco quando existe um caminho direcionado de v_i a v_j ou de v_j a v_i [69].

2.3 Redes Sociais e situações de calamidade e epidemia

Na literatura é possível encontrar alguns trabalhos relacionados com a utilização de redes sociais em situações de calamidade e epidemia, como os apresentados a seguir.

O estudo apresentado por Corley et al. [29], 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. Odlum et al. [51] analisaram a utilização do Twitter na epidemia do Ebola durante o período de 24/07/2014 a 01/08/2014, e verificou-se um acréscimo na quantidade de comentários na Nigéria antes do anúncio oficial sobre o primeiro caso no país. Por meio de análise espaço-temporal das informações postadas no Twitter, Sakaki et al. [57] monitoraram a ocorrência de terremotos no Japão em tempo real. No Brasil, Antunes et al. [21] 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 [16] 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. Ofoghi et al. [52] estudaram a possibilidade de detectar ameaças à saúde pública a partir da variação do conteúdo dos comentários postados no Twitter. Baseado em dois casos de contaminação do vírus Ebola reportados no Reino Unido em dezembro de 2014 e janeiro de 2015, foram efetuadas comparações do conteúdo dos comentários do Twitter sobre o Ebola sete dias antes e sete dias depois dos dois incidentes. Verificou-se que, após os incidentes, os usuários demonstraram preocupação sobre a ameaça da epidemia em seu país, o que não ocorreu nos comentários postados antes dos incidentes.

2.4 Banco de Dados Geográficos

De acordo com Câmara et al. [24], banco de dados geográficos permite que os dados sejam representados por meio de estruturas vetoriais baseadas em três formas: pontos, linhas e polígonos, conforme mostrado na Figura 2.1. Um ponto é um par ordenado (x, y) de coordenadas espaciais e pode ser utilizado para identificar localizações no espaço como, por exemplo, localizações de crimes e ocorrências de doenças. Uma linha é um

conjunto de pontos conectados. Um polígono, ou área, é a região do plano limitada por uma ou mais linhas poligonais conectadas de tal forma que o último ponto de uma linha seja idêntico ao primeiro ponto da próxima linha. Polígonos podem ser utilizados para representar unidades de dados espaciais como, por exemplo, distritos, zonas e municípios. Nesta pesquisa, a estrutura vetorial polígono foi utilizada para representar o país.

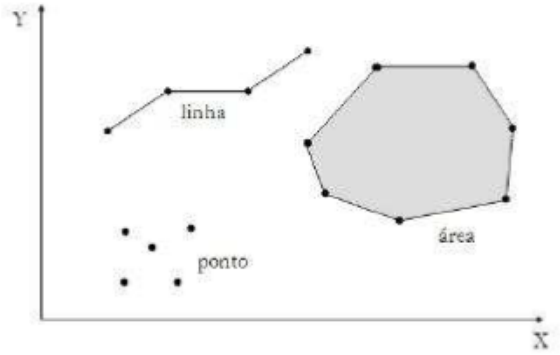


Figura 2.1: Representações de estruturas vetoriais. Fonte: Camara et al. [24].

2.5 Localização geográfica de dados do Twitter

Associar os comentários ou *tweets* sobre diversos assuntos ao local de origem dos usuários do Twitter possibilita efetuar análises espaciais interessantes. No Twitter, a informação da localidade pode ser inserida manualmente no perfil do usuário ou por meio da habilitação do dispositivo GPS, o qual fornece as coordenadas geográficas latitude e longitude. Entretanto, segundo Chandra et al. [25], a habilitação do GPS não tem sido adotada pela maioria dos usuários. Por outro lado, cerca de 75% dos comentários do Twitter possuem alguma informação de localidade informada manualmente. Os formatos desta informação variam desde endereços específicos, coordenadas geográficas, nome de cidade, nome de país, nome de estado, nome de províncias ou palavras sem relação com a localidade [64].

Na literatura, diferentes abordagens vem sendo utilizadas com o objetivo de identificar a localidade do usuário do Twitter, como por exemplo: o uso de modelos probabilísticos baseados no conteúdo dos comentários permite extrair nomes de localidades ou frases associadas às localidades [26, 25]; uma outra maneira de extrair a localidade a partir dos comentários é por meio da utilização de ontologia geoespacial, em que as palavras são classificadas em sujeito, predicado e objeto, e o conteúdo do objeto permite extrair nomes de localidades [30]; a ontologia também pode ser utilizada para tratar as ambiguidades dos nomes das localidades contidas nos comentários [68]; Compton et al. [28] mostraram que

é possível identificar a localidade aplicando-se uma heurística à rede de amigos do Twitter; a localidade também pode ser extraída do texto do comentário utilizando-se como fonte de referência a base de dados geográfica GeoNames [6, 36, 23, 47].

2.6 Análise Textual

A análise textual dos comentários em redes sociais tem sido objeto de diversos estudos. No contexto relacionado à epidemia, Chew et al. [27] analisaram manualmente 5.395 comentários do Twitter em inglês sobre o vírus H1N1. Os comentários foram categorizados em: noticiários sobre a epidemia, experiências pessoais, opiniões pessoais, piadas e anúncios publicitários. Verificou-se que a maioria dos comentários (52%) refere-se a noticiários. Além desta categorização, utilizou-se *emoticons* para classificar os comentários em: sarcasmo, alívio, preocupação e frustração. *Emoticons* são figuras de feições faciais que expressam sentimentos, como alegria e tristeza, criados a partir de combinação de caracteres. O estudo de Ji et al. [40] mostrou que em situações de saúde pública como, por exemplo, influenza, H1N1, sarampo, meningite e tuberculose, cerca de 30% dos comentários do Twitter são noticiários sobre a doença e não expressam a opinião pessoal de usuários. Neste estudo foi adotada a estratégia de separar os comentários em *Tweet* pessoal e *Tweet* de noticiário. *Tweet* pessoal expressa o estado pessoal do autor como sentimento, opinião, emoção, fatos observados pelo autor e que não podem ser verificados em uma observação objetiva. *Tweet* de noticiário relata sobre um fato ocorrido ou que irá ocorrer. A utilização de *emoticons* como classificadores dos comentários do Twitter também foi observada nos trabalhos de Ji et al. [40] e Pak et al. [55], cujo objetivo foi formar o corpus de sentimento positivo e sentimento negativo.

Estudos recentes [50, 56] mostram que, além do uso de *emoticons*, usuários de redes sociais vem adotando *emojis* como forma de expressão. *Emoji* são símbolos gráficos cujo uso iniciou-se em aparelhos celulares no Japão no final dos anos 90, tornando-se mundialmente popular com o aumento do uso de *smartphones*. Ao contrário de *emoticons*, que são criados por uma sequência de caracteres, *emojis* são símbolos gráficos que possuem uma expressividade maior. Além de representarem expressões faciais, são utilizados para representar alimentos, meios de transporte, tempo, atividades como correr e dançar, entre outros. Novak et al. [50] verificaram que o nível de concordância entre vários entrevistados para efetuar a classificação manual dos comentários do Twitter é maior com a presença de *emoji* nos comentários.

De acordo com Liu et al. [45], a análise de sentimentos, também chamada de mineração

de opinião, é um campo de estudo que analisa opiniões, sentimentos, atitudes e emoções das pessoas sobre um determinado produto, serviço, organização, indivíduo, evento ou tópico. A análise de sentimentos pode ser aplicada em diferentes níveis de granularidade como, por exemplo, documentos, sentenças e palavras [45]. A análise de sentimentos no nível de palavras é o mais simples e classifica a palavra em polaridade positiva, negativa ou neutra. Na análise de sentimentos no nível de sentença, além de verificar a polaridade de cada palavra, leva-se em consideração o relacionamento entre as palavras e a sua função gramatical. O resultado desta análise compõe o sentimento da sentença. A análise no nível de documento baseia-se no contexto do documento como um todo e verifica o relacionamento entre as sentenças.

Para efeitos da presente pesquisa, buscou-se um método simples para classificar os comentários pessoais dos usuários do Twitter. Optou-se por efetuar a análise de sentimentos no nível de granulaidade de palavras utilizando o dicionário léxico SentiWordNet [15]. Este dicionário foi proposto no estudo de Baccianella et al. [22] e baseia-se no corpus WordNet com 155 mil palavras em inglês associadas a uma pontuação numérica que varia de -1 a 0 para palavras com sentimento negativo, e de 0 a +1 para palavras com sentimento positivo. Esta pontuação não foi criada para um domínio específico e representa o valor atribuído à palavra em seu uso geral. Lu et al. [46] utilizaram o SentiWordNet para analisar os comentários do Twitter sobre o Ebola e mostrar no mapa dos Estados Unidos as regiões onde foram registrados comentários positivos e negativos. Neste estudo, observou-se a tendência do SentiWordNet em rotular positivamente os comentários coletados sobre a epidemia.

3. Metodologia

Este capítulo descreve a metodologia utilizada nesta pesquisa e está dividido em três seções. A primeira seção apresenta uma visão geral das etapas que compõe a pesquisa e descreve as ferramentas utilizadas em cada etapa. A Seção 3.2 apresenta os critérios utilizados na coleta de dados do Twitter e o período de análise de dados. A Seção 3.3 apresenta os três algoritmos desenvolvidos nesta pesquisa: o algoritmo de identificação de país, o algoritmo de identificação de idioma e o algoritmo supervisionado de análise textual.

3.1 Visão Geral

Esta seção apresenta uma visão geral das etapas que compõe esta pesquisa. A metodologia utilizada nesta pesquisa é quantitativa, baseada no cálculo de métricas da Teoria de Grafos, quantidade de comentários por país e quantidade de comentários categorizados por meio da análise textual. A Figura 3.1 mostra o fluxo das etapas, juntamente com as ferramentas utilizadas. Inicialmente, os comentários do Twitter [16] foram coletados por meio da ferramenta NodeXL [9]. Em seguida, o NodeXL foi utilizado para o cálculo das métricas e para a visualização de grafos. A descrição da ferramenta NodeXL encontra-se na Seção 3.1.1.2. A seguir, para possibilitar a análise espacial e textual dos dados nas etapas posteriores, os comentários do Twitter foram armazenados no banco de dados PostgresSQL [13]. Na etapa seguinte, o algoritmo de identificação de país foi executado para determinar o país de origem dos usuários do Twitter. Este algoritmo foi codificado na linguagem PL/pgSQL [11], cujos detalhes estão descritos na Seção 3.3.1. Após esta etapa, o algoritmo de identificação de idioma, codificado na linguagem Java [7], foi executado para determinar o idioma utilizado nos comentários do Twitter. Os detalhes deste algoritmo estão descritos na Seção 3.3.2. Na sequência, o algoritmo supervisionado de análise textual foi executado utilizando-se a ferramenta Full Text Search [5]. As princi-

pais funcionalidades do Full Text Search estão descritas na Seção 3.1.1.4. O algoritmo supervisionado de análise textual está descrito na Seção3.3.3. Por último, os dados foram visualizados no mapa utilizando-se a ferramenta QGIS [14].

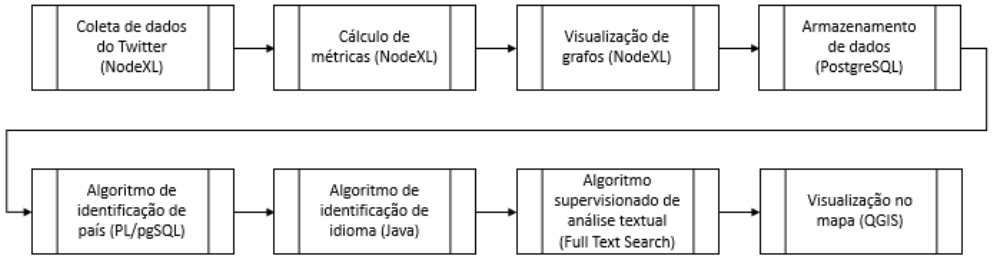


Figura 3.1: Visão geral das etapas que compõe esta pesquisa.

3.1.1 Ferramentas de coleta, armazenamento, visualização e análise de dados de redes sociais

A seguir, apresentam-se as ferramentas utilizadas nesta pesquisa para coleta, armazenamento, visualização e análise de redes sociais.

3.1.1.1Twitter

O Twitter [16] é uma das redes sociais mais utilizadas com cerca de 500 milhões de usuários ao redor do mundo [34]. As principais motivações para o uso desta ferramenta são: postar comentários sobre o que as pessoas estão fazendo no momento, responder os comentários de outros usuários, compartilhar informações contendo URL e comentar sobre noticiários recentes [39]. Uma situação de calamidade em que o Twitter revelou o seu potencial foi no terremoto no Japão, ocorrido em 2010. O Twitter mostrou-se como um importante meio de comunicação logo após o terremoto, permitindo disseminar avisos sobre tsunamis, risco nuclear e facilitando a busca por informações de sobreviventes de forma rápida e fácil [20].

Comparando-se o Twitter com o Facebook [3], outra ferramenta de rede social de grande alcance, é possível enumerar as seguintes vantagens: a possibilidade de postar comentários curtos (*Tweets*) de até 140 caracteres, utilizar palavras-chaves (*Hashtag*) para facilitar a busca por um determinado termo, responder a um comentário (*Reply*) e a possibilidade de um usuário (*Follower*) seguir outros usuários (*Followed*) sem a necessidade de reciprocidade. Neste caso, os usuários seguidores (*Followers*) recebem todos os comentários (*Tweets*) dos usuários que está seguindo (*Followed*). O Twitter permite também que o usuário encaminhe (*Retweet*) um determinado comentário para os seus usuários seguidores, permitindo a disseminação de informações além do alcance do *Tweet* original. Este

mecanismo de relacionamento entre os usuários (*Follower* e *Followed*), acrescido do encaminhamento de comentários (*Retweet*) torna o uso do Twitter uma ferramenta poderosa para a disseminação de informações [43].

De acordo com Shi et al. [60], a probabilidade de um usuário compartilhar uma informação em redes sociais depende de como este avalia a informação como nova. No Facebook [3], o usuário A necessita do consentimento do usuário B para fazer parte da lista de amigos e, após obter o consentimento, o relacionamento torna-se forte. Em relacionamentos fortes, os usuários costumam participar dos mesmos grupos de interesse e existe uma tendência de compartilharem informações similares. Assim, o conteúdo disseminado por um já pode ser de conhecimento do outro. No Twitter, para o usuário A seguir o usuário B, não necessita de consentimento do usuário B, caracterizando o relacionamento como fraco. Em relacionamentos fracos, usuários costumam ter conhecimentos e experiências diferentes e podem atribuir valores diferentes ao mesmo conteúdo de informação. O estudo de Shi et al. [60] sugere que o relacionamento fraco do Twitter é mais propenso ao compartilhamento de informações, comparando-se com os relacionamentos fortes como no Facebook.

3.1.1.2NodeXL

O NodeXL [9] - Network Overview, Discovery and Exploration - é uma extensão da planilha Microsoft Excel [62] e permite coletar dados das principais redes sociais como Twitter, Facebook [3], Flickr [4] e YouTube [19]. 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. O NodeXL permite agrupar os vértices por *cluster*, por componentes conexos e por atributos de vértices. Uma vez agrupados, a ferramenta permite visualizar os grafos, onde o leiaute de apresentação dos vértices e das arestas são configuráveis por meio de atributos de cores, formato e tamanho. 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.

Dependendo do contexto dos dados coletados e da rede social utilizada, o grafo inicial formado pelo NodeXL tende a ser denso e complexo, dificultando sua visualização. Para facilitar a análise do grafo, o NodeXL permite filtrar e tornar visíveis somente os vértices relevantes por meio da seleção de determinadas métricas e permite também destacar estes vértices atribuindo cores e tamanhos diferentes.

A seguir, serão apresentadas algumas situações de uso do Twitter como *Tweet*, *Reply*, *Retweet*, *Mentions* e o resultado do grafo gerado pelo NodeXL. Considere um exemplo com três usuários conectados onde @usuarioC é seguidor do @usuarioB que é seguidor do @usuarioA. Assim, todo comentário postado pelo @usuarioA torna-se visível ao @usuarioB. Por sua vez, todo comentário novo postado pelo @usuarioB e comentários postados pelo @usuarioA que forem encaminhados (*Retweet*) pelo @usuarioB tornam-se visíveis ao @usuarioC. Desta forma, o *Retweet* permite disseminar um comentário para um novo público, que não é o mesmo público do comentário original. A Figura 3.2 mostra a interface visual do NodeXL onde os comentários coletados do Twitter estão armazenados na planilha Excel. As colunas *Vertex1* e *Vertex2* representam os pares de vértices que formam uma aresta, a coluna *Relationship* representa a situação de uso do Twitter e a coluna *Tweet* representa o conteúdo do comentário.

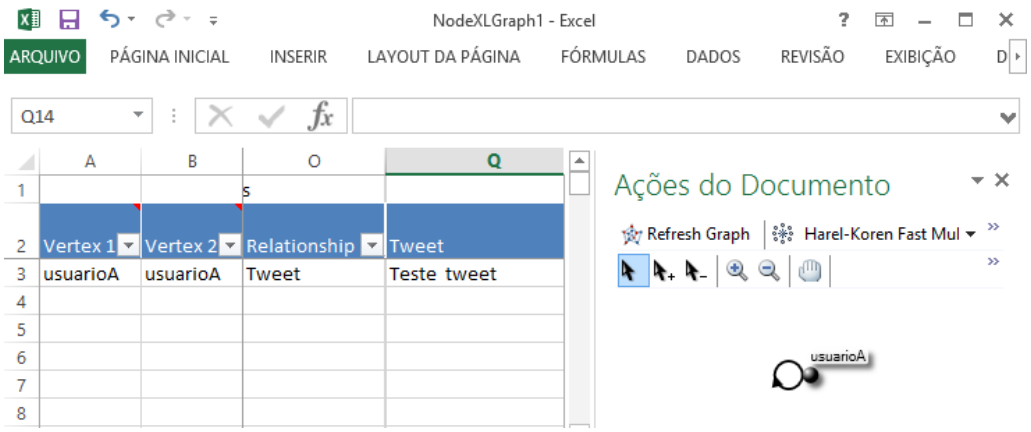
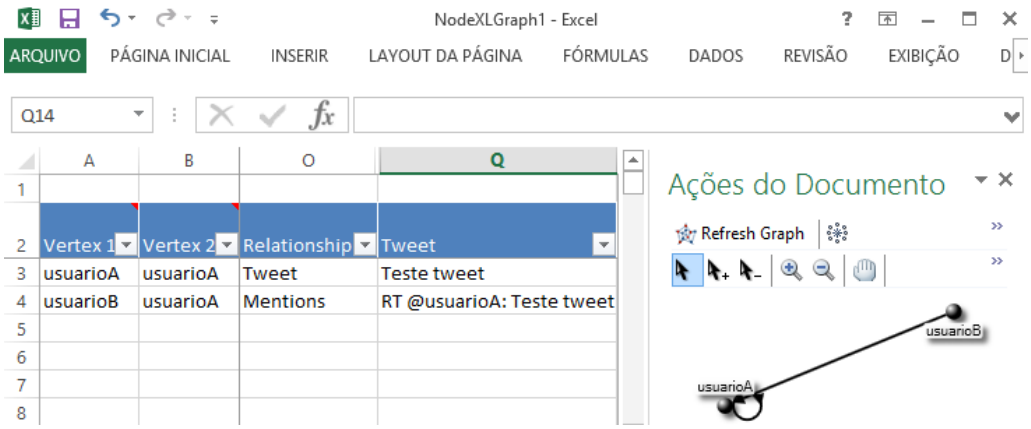


Figura 3.2: Aresta do tipo *self-loop* representando o *Tweet* do usuarioA.

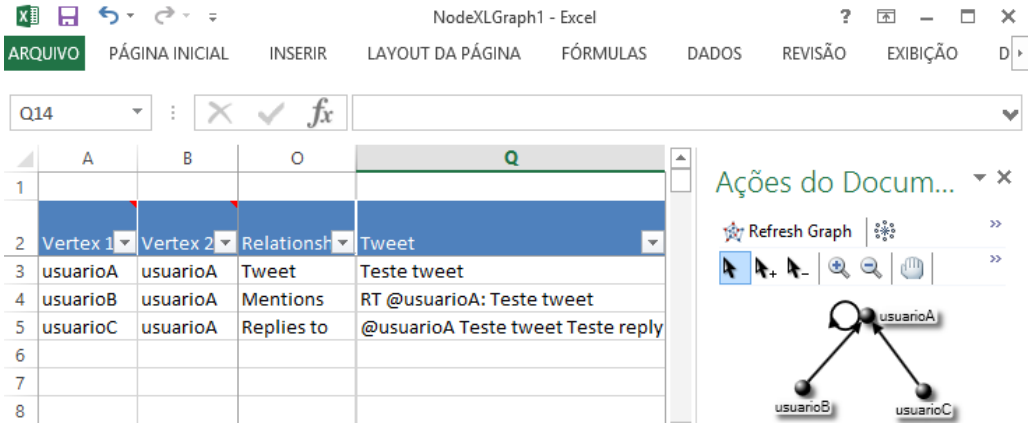
Considere também a situação em que um usuário cria um novo comentário (*Tweet*), conforme ilustrado na Figura 3.2. O @usuarioA posta o seguinte comentário: *Teste tweet*. O comentário postado torna-se visível no Twitter para os seus seguidores. O NodeXL cria uma aresta do tipo laço (*self-loop*) entre @usuarioA e @usuarioA e não cria nenhuma aresta unindo o @usuarioA e os seus seguidores. A seguir, o @usuarioB, que é seguidor do @usuarioA, visualiza o comentário *Teste tweet* e o encaminha (*Retweet*) para seus próprios seguidores. A Figura 3.3 apresenta uma segunda aresta direcionada do @usuarioB para o @usuarioA representando a menção feita ao comentário *Teste tweet*. Este comentário encaminhado (*Retweet*) agora inicia-se com as letras *RT* seguido do nome do usuário mencionado @usuarioA e o conteúdo do comentário original. O conjunto destas simbologias representam a situação de *Retweet*. Neste caso também não é criada nenhuma aresta entre o @usuarioB e o seu seguidor @usuarioC. Na sequência, o @usuarioC recebe todos os comentários (*Tweet*) e encaminhamentos (*Retweet*) do @usuarioB por ser seu seguidor. O Twitter permite responder (*Reply*) o comentário de qualquer usuário, mesmo

que não seja seu seguidor. Desta forma, o @usuarioC visualiza o comentário do @usuarioA e responde (*Replies to*) o @usuarioA adicionando o seguinte comentário: *Teste reply*. O NodeXL cria uma terceira aresta do @usuarioC direcionado ao @usuarioA, conforme mostra a Figura 3.4.



	A	B	O	Q
2	Vertex 1	Vertex 2	Relationship	Tweet
3	usuarioA	usuarioA	Tweet	Teste tweet
4	usuarioB	usuarioA	Mentions	RT @usuarioA: Teste tweet

Figura 3.3: Aresta direcionada do @usuarioB para @usuarioA representando o *Retweet*.



	A	B	O	Q
2	Vertex 1	Vertex 2	Relationsh	Tweet
3	usuarioA	usuarioA	Tweet	Teste tweet
4	usuarioB	usuarioA	Mentions	RT @usuarioA: Teste tweet
5	usuarioC	usuarioA	Replies to	@usuarioA Teste tweet Teste reply

Figura 3.4: Aresta direcionada do @usuarioC para o @usuarioA representando o *Reply*.

O Twitter permite também enviar comentários diretamente a qualquer usuário sem a necessidade de ser uma resposta a um comentário já postado por alguém, por meio da inclusão do nome do usuário no início do comentário. Esta situação é representada da mesma forma que o *Reply*. Para exemplificar, considere as duas situações descritas a seguir e ilustradas na Figura 3.5. O @usuarioA posta o seguinte comentário: *@usuarioB Teste conversa direta*. Neste caso, é criada uma aresta do @usuarioA para o @usuarioB representando o *Reply*. Outra situação possível acontece quando o usuário do Twitter quer postar um novo comentário mencionando um outro usuário na frase. Por exemplo, o @usuarioA posta o seguinte comentário: *Teste mentions @usuarioC*. O NodeXL cria uma aresta do @usuarioA para o @usuarioC do tipo *Mentions*, cuja representação é similar ao *Retweet*, porém sem a adição das letras *RT* no início do comentário.

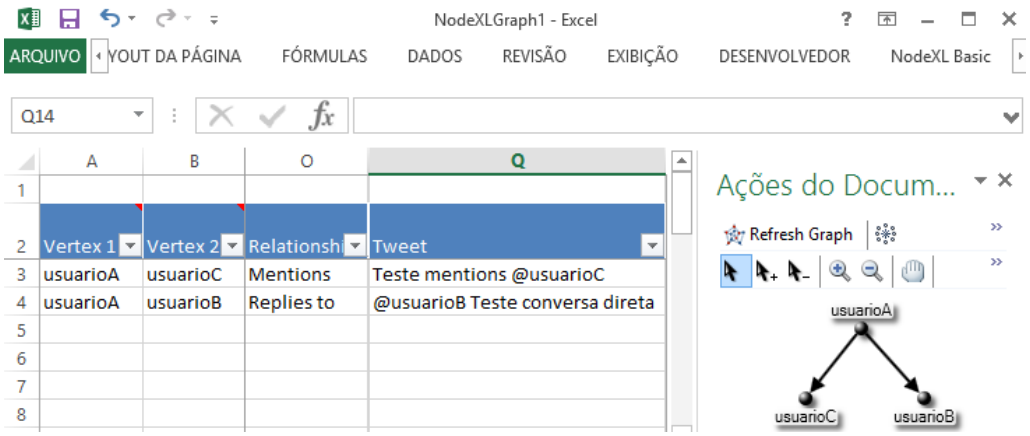


Figura 3.5: Variações de comentários do tipo *Reply* e *Mention*.

Esta seção mostrou como a ferramenta NodeXL permite coletar dados do Twitter, representar relacionamentos de diversos usuários na forma de grafos e analisá-los dentro de um contexto definido.

3.1.1.3 PostgreSQL e PostGIS

O PostgreSQL [63] é um sistema gerenciador de banco de dados objeto-relacional, gratuito e de código aberto, desenvolvido a partir do projeto Postgres em 1986. Entre as principais características do PostgreSQL, está o seu potencial de extensibilidade, o que possibilitou o desenvolvimento de uma extensão geográfica chamada PostGIS [12]. A extensão PostGIS permite o armazenamento e a manipulação de dados espaciais no banco PostgreSQL. A Figura 3.6 mostra os tipos de dados espaciais fornecidos por esta extensão. Nesta pesquisa, o tipo de dado espacial polígono foi utilizado para representar os países afetados pela epidemia. O Full Text Search é outra ferramenta do PostgreSQL utilizada nesta pesquisa, sendo descrita na seção seguinte.

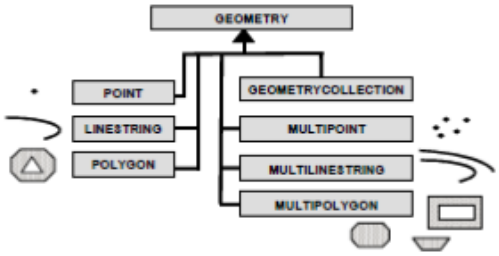


Figura 3.6: Tipos de dados espaciais fornecidos pelo PostGIS. Fonte: Camara et al. [24].

3.1.1.4 Full Text Search

Full Text Search é uma ferramenta do PostgreSQL cujo objetivo é efetuar a busca textual de documentos no banco de dados [5]. Documento é uma unidade de busca como, por exemplo, o texto de um artigo ou uma mensagem de e-mail. Ao analisar um documento, inicialmente a ferramenta separa as frases em *tokens*, que podem ser números, palavras, URL, e-mails, entre outros. Em seguida, os *tokens* são normalizados e transformados em lexemas. A normalização envolve a remoção de sufixos, conversão para letras minúsculas, remoção de palavras de parada e conversão da palavra para o singular. Palavras de parada são termos comuns que aparecem com frequência e podem ser desconsiderados na busca textual como, por exemplo, artigos, preposições, entre outros. Lexemas são palavras normalizadas e tratadas para auxiliar na busca textual. A configuração da busca textual é efetuada por meio de dicionário, em que é possível definir o idioma, as palavras de parada e o dicionário morfológico a ser utilizados no processo de busca. O dicionário morfológico permite normalizar diferentes formas linguísticas para um mesmo lexema. Por exemplo, o dicionário English Ispell reconhece as variações do termo *bank*, como *banking*, *banks* e *bank's* facilitando a busca textual.

O exemplo abaixo utiliza a função *to_tsquery* do Full Text Search, em que a frase *New cases of Ebola in Liberia* é analisada utilizando-se o dicionário em inglês.

```
SELECT to_tsquery('english',
      'New & cases & of & Ebola & in & Liberia');
```

A execução do comando acima gera os seguintes lexemas. Observa-se que as palavas de parada *of* e *in* foram eliminadas.

```
'new' & 'case' & 'ebola' & 'liberia'
```

O PostgreSQL fornece dicionários predefinidos para diversos idiomas como inglês, francês, alemão, português, espanhol e permite criar novos dicionários. Nesta pesquisa foi utilizado o dicionário inglês.

3.1.2 Diagrama de arquitetura geral da solução

A Figura 3.7 exibe um diagrama com o resumo das ferramentas utilizadas. No Apêndice B, detalha-se o roteiro de instalação destas mesmas ferramentas. Os retângulos em azul representam as ferramentas prontas adquiridas de forma gratuita, os retângulos em

rosa são os algoritmos desenvolvidos durante a pesquisa, os retângulos em amarelo são *plug-ins* do banco de dados PostgreSQL e os retângulos em branco representam as informações.

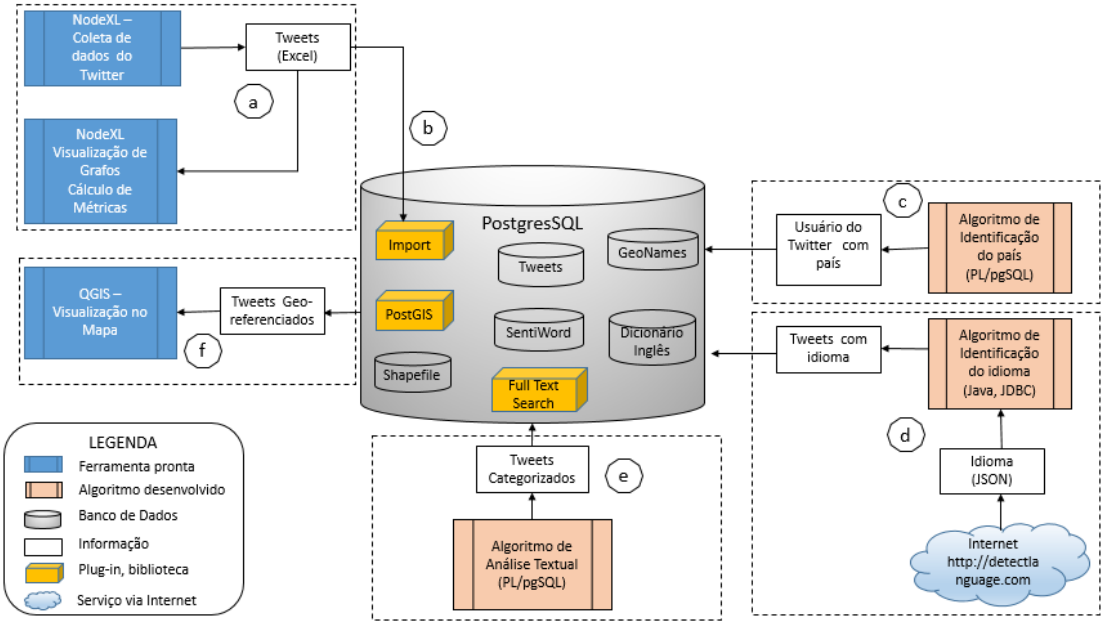


Figura 3.7: Diagrama de arquitetura geral da solução.

O fluxo das informações está descrito abaixo. As letras de "a" à "f" correspondem às letras no diagrama:

a. Os comentários do Twitter são coletados por meio da ferramenta NodeXL e armazenados em arquivos Excel. O NodeXL permite o cálculo das métricas e a visualização de grafos.

b. As planilhas Excel contendo os comentários do Twitter são armazenadas no banco de dados PostgresSQL por meio da ferramenta Import do próprio banco utilizando o formato csv.

c. O algoritmo de identificação de país é executado para determinar o país de origem dos usuários do Twitter.

d. O algoritmo de identificação de idioma é executado para determinar o idioma utilizado nos comentários do Twitter. O algoritmo utiliza o serviço web Language Detection que analisa a frase e retorna o idioma.

e. O algoritmo supervisionado de análise textual é executado para categorizar os comentários em comentários pessoais e não pessoais.

f. A ferramenta QGIS é utilizada para a visualização dos dados do Twitter no mapa.

3.2 Descrição de Dados

Esta seção descreve os critérios utilizados na coleta de dados e o período de análise de cada etapa desta pesquisa.

3.2.1 Coleta de Dados

A coleta dos comentários do Twitter sobre o termo "Ebola" 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*. O NodeXL permite configurar alguns filtros no momento da coleta. Além do termo de busca, é possível configurar, por exemplo, o idioma e a quantidade de comentários, cujo limite máximo é de 18.000 por coleta. Nesta pesquisa, 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. Nos dias iniciais de novembro, em que o termo de busca para esta pesquisa ainda não estava definido, a coleta foi limitada a 5.000 comentários. Nos outros dias, 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.

3.2.2 Período de análise de dados

Esta seção descreve o intervalo de datas utilizado na análise por meio da Teoria de Grafos, na análise espacial e na análise textual. A Figura 3.8 mostra o gráfico representando a evolução da epidemia do Ebola de agosto de 2014 a fevereiro de 2016, segundo dados divulgados pela Organização Mundial de Saúde. O período de novembro de 2014 a abril de 2015, demarcados pela linha tracejada na Figura 1.1, foi utilizado como período de análise desta pesquisa, por representar os meses em que houve a transmissão intensa do vírus nos países da África Ocidental.

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 sele-

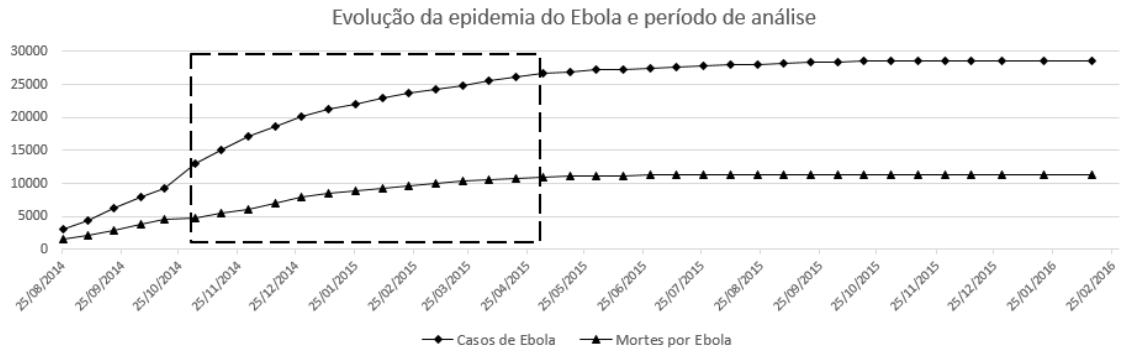


Figura 3.8: Gráfico de evolução da epidemia do Ebola do período agosto/2014 a fevereiro/2016 e o período de análise.

cionados determinados intervalos de datas para possibilitar uma análise longitudinal ao longo dos seis meses.

Para a análise por meio da Teoria de 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 os 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 Processamento de Dados

Esta seção está dividida em três subseções e apresenta os três algoritmos desenvolvidos nesta pesquisa. A Subseção 3.3.1 apresenta o algoritmo de identificação de país. O objetivo deste algoritmo é extrair o país de origem dos usuários do Twitter e possibilitar a análise espacial dos dados do Twitter. A Subseção 3.3.2 apresenta o algoritmo de identificação de idioma, que possibilita a seleção de comentários em inglês. A Subseção 3.3.3 descreve o algoritmo supervisionado de análise textual, cujo objetivo é verificar o conteúdo dos comentários dos usuários dos países da África, onde concentram-se a transmissão intensa da epidemia.

3.3.1 Algoritmo de identificação de país

Esta seção descreve o algoritmo desenvolvido para a identificação do país de origem dos usuários do Twitter. Os trabalhos apresentados por Graham et al. [35], Valkanas [67] e Chandra et al. [25] indicam que menos de 1,06% dos usuários informam as coordenadas geográficas habilitando o GPS. Com base nos dados desta pesquisa com 349.159 usuários, 2.326 (1,06%) informaram as coordenadas geográficas. Apesar deste número representar um percentual pequeno, a presença da latitude e longitude indica a real localização do usuário no momento em que enviou o comentário e esta localidade pode ser diferente da localidade informada em seu perfil, sugerindo usuários em deslocamento ou usuários que não atualizam o seu perfil com frequência [35].

Para fins desta pesquisa, será dada prioridade às coordenadas geográficas, caso seja informada. Caso contrário, a localização geográfica será extraída por meio da comparação do campo "Location" do perfil do usuário com as informações do GeoNames [32, 67] - base de dados geográfica gratuita disponibilizada através da licença Creative Commons License [6]. Desta forma, será assumido que a localidade extraída do Twitter refere-se ao local onde o usuário se encontrava no momento em que enviou o comentário ou o local onde este normalmente permanece.

A seguir, descreve-se o algoritmo de identificação de país. O objetivo deste algoritmo é obter a sigla de país (ISO2) a partir das coordenadas geográficas ou por meio da informação fornecida no campo "Location". ISO2 refere-se a sigla de país de duas letras ISO 3166-1 alfa-2 definida pela Organização Mundial de Padronização e permite que as informações em análise possam ser armazenadas em banco de dados geográficos e representadas em mapas por meio de ferramentas de análise espacial. O algoritmo está dividido em três partes conforme ilustrado na Figura 3.9.

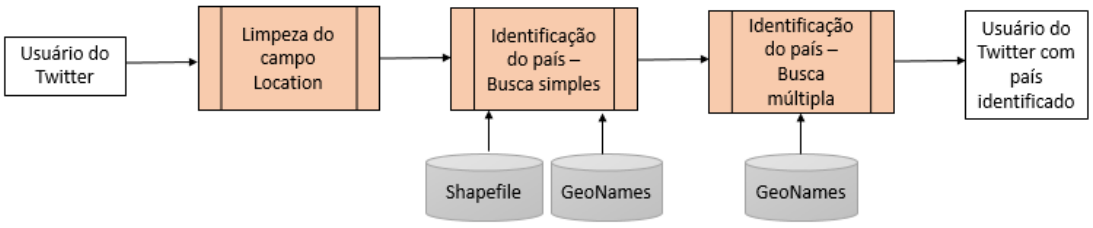


Figura 3.9: Algoritmo de identificação de país.

3.3.1.1 Limpeza do campo "Location"

O objetivo desta primeira etapa do algoritmo é remover do campo "Location" caracteres que dificultam a identificação da localidade como, por exemplo, *emoticons*, caracteres especiais (*carriage return*), ponto ao final da cadeia de caracteres, além de descartar usuários que informaram valores numéricos, e-mails ou URL neste campo.

3.3.1.2 Identificação de país - Busca simples

O objetivo da segunda etapa do algoritmo é obter a sigla do país (ISO2) a partir das coordenadas geográficas ou da informação fornecida no campo "Location". Devido à variedade do conteúdo informado no campo "Location" como, por exemplo, nome de país, nome de cidade, sigla de país, nome de estado, entre outros, foram criadas regras de prioridade para possibilitar a identificação do país.

Descrevem-se, a seguir, as regras adotadas nesta pesquisa para identificação de país, em ordem decrescente de prioridade.

1. Tratamento da latitude e longitude. A sigla do país (ISO2) é obtida por meio das coordenadas geográficas latitude e longitude informadas com a habilitação do GPS. Em uma análise inicial desta pesquisa com 36.527 usuários, 244 (1,04%) habilitaram o GPS. Em alguns casos, foram encontrados usuários com o GPS habilitado, sem a informação da localidade no perfil, como o usuário @mexiconoavanza que enviou comentários cujas coordenadas geográficas pertencem a países diferentes (México e Alemanha), sugerindo um usuário do Twitter em deslocamento. Nestes casos, optou-se pelo país referente as coordenadas geográficas registradas no último comentário enviado pelo usuário. Para determinar a sigla do país a partir da latitude e longitude, utilizou-se as informações do *shapefile* "tm_world_borders-0.3". *Shapefile* é um tipo de arquivo de especificação aberta desenvolvido pela ESRI que permite a interoperabilidade de dados geoespaciais [2].

2. Tratamento de nomes alternativos adicionados manualmente. A sigla do país (ISO2) é obtida por meio de nomes alternativos informados no campo "Location". Verificou-se que alguns usuários costumam informar localidades cujos nomes não são nomes oficiais de país ou cidade, como por exemplo, "England", "Scotland", "New York, NY". O GeoNames possui uma extensa lista de nomes alternativos de localidades cadastradas (cerca de 600 mil), e devido a essa grande quantidade verificou-se que existe duplicidade de nomes que referem-se a locais diferentes, não permitindo a correta extração do país. Por exemplo, "Aebura" é um nome alternativo de locali-

dade utilizado em dez países diferentes. Na Espanha, "Aebura" é o nome histórico atribuído à cidade cujo nome oficial é "La Albuera". Na Itália, "Aebura" é o nome histórico de região administrativa, cujo nome oficial é "Avola". Verificou-se também que a lista de nomes alternativos do GeoNames contém URL e códigos postais que necessitariam de tratamento. Para contornar essa situação, foi criada uma lista manual contendo nomes alternativos extraídos dos dados coletados do Twitter. Esta lista contém, não apenas nomes de cidades, mas também nomes de países, siglas de países e nomes separados por vírgulas. A Tabela A.16 no Apêndice mostra alguns exemplos de nomes alternativos adicionadas manualmente nesta pesquisa.

3. Tratamento de nome de país. A sigla do país é obtida por meio da comparação do campo "Location" com o nome do país em inglês, francês, espanhol e idioma local obtida do *shapefile* "Countries". Verificou-se que existem várias cidades nos Estados Unidos cujo nome é também o nome de um país como, por exemplo, Bolívia, Cuba, Denmark, China, Brazil, Egypt, Mali, entre outros. A cidade Brazil cuja população é de 7.212 habitantes localiza-se no interior do estado de Indiana nos Estados Unidos. O mesmo ocorre com os outros exemplos, isto é, são cidades pequenas com poucos habitantes. Nestes casos, apesar da possibilidade da localidade informada ser um país ou uma cidade, optou-se por dar preferência pelo nome do país, já que existe uma maior probabilidade do comentário do Twitter ter originado de um país do que de uma cidade pouco habitada.

4. Tratamento de sigla de país ISO3. Alguns usuários dos Estados Unidos preferem informar a sigla do país de três letras ISO3 (por exemplo, USA), ao invés do nome do país por extenso. Assim a sigla do país (ISO2) é obtida por meio da comparação do campo "Location" com a sigla do país ISO3 obtida do *shapefile* "Countries".

5. Tratamento de nomes de cidades duplicados. Esta pesquisa baseia-se na lista de localidades (Gazetter) fornecida pelo GeoNames. Inicialmente foi utilizada a lista de 144.329 cidades com mais de 1.000 habitantes para permitir a identificação de um maior número possível de localidades dos usuários do Twitter. Observou-se que, nesta lista, existem 7.025 nomes de cidades iguais localizadas no mesmo país, sendo 2.116 cidades (30,16%) localizadas na China. Neste caso, não houve dificuldade na identificação do país pois a duplicidade ocorria em um mesmo país. Porém, verificou-se que, nesta lista de 144.329 cidades, existem 2.953 cidades com nomes iguais localizadas em países diferentes, o que dificultaria a identificação correta do país. Então optou-se por descartar usuários do Twitter que informaram cidades duplicadas em países diferentes. Com esta abordagem, observou-se que diversos usuários do Twitter de cidades como Madrid, Roma, London, Barcelona, entre ou-

tros, estavam sendo descartados. Então, optou-se por utilizar uma lista de cidades menor do GeoNames com mais de 15.000 habitantes, totalizando 23.444 cidades. Isto diminuiu a quantidade de cidades duplicadas localizadas em países diferentes para 427 cidades. Dentre estas cidades, 35 são capitais de países. Assim, decidiu-se dar prioridade para as capitais baseando-se na premissa de que, caso o nome da cidade esteja duplicada em países diferentes, existe uma maior possibilidade do comentário do Twitter ter originado da capital de um país. Entretanto, cidades como Barcelona e Sevilla, que não são capitais, estavam na lista das 427 cidades duplicadas em países diferentes, o que descartaria diversos usuários do Twitter da Espanha onde havia sido registrado um paciente com os sintomas do vírus do Ebola. A estratégia utilizada para contornar esse problema foi a criação de uma lista das cidades duplicadas em países diferentes, porém priorizada com a cidade de maior população, cuja informação também foi obtida do GeoNames. Esta estratégia baseia-se na premissa de que, caso o nome de uma cidade esteja duplicada em mais um país, existe uma possibilidade maior do comentário do Twitter ter originado da cidade de maior população. Assim, o tratamento da duplicidade do nome da cidade foi efetuado por meio da seguinte ordem de prioridade: capital de país, cidades duplicadas no mesmo país, e cidade duplicada em países diferentes priorizada pelo da maior população.

6. Tratamento de nome de cidade (não capital) utilizando lista de cidades do GeoNames com mais de 15.000 habitantes. A sigla do país (ISO2) é obtida por meio da comparação do campo "Location" com o nome da cidade do GeoNames que não seja capital de país. Neste processo, foi utilizada a lista de cidades do GeoNames com mais de 15.000 habitantes.

7. Tratamento de região administrativa. Alguns usuários preferem informar o nome do estado ou da província, ao invés do nome da cidade ou país. Assim, este processo obtém a sigla do país (ISO2) por meio da comparação do campo "Location" com o nome da região administrativa fornecida pelo GeoNames. Usuários que informaram nomes de regiões administrativas duplicadas em países diferentes foram ignoradas neste processo.

8. Tratamento de nome de cidade (não capital) utilizando lista de cidades do GeoNames com mais de 1.000 habitantes. Verificou-se que usuários de cidades com menos de 15.000 habitantes, como "Java" na Indonésia e "Isla Mujeres" no México, não estavam sendo identificados. Como as capitais de países e cidades com nomes duplicados mais populosos (acima de 15.000 habitantes) já haviam sido tratadas nos processos anteriores, foi incluída a rotina de identificação do GeoNames lista com-

pleta, que abrange cidades com mais de 1.000 habitantes. Neste processo foram descartados nomes de cidades duplicadas em países diferentes.

9. Tratamento de nomes alternativos do GeoNames. A sigla do país (ISO2) é obtida por meio da comparação do campo "Location" com a lista de nomes alternativos fornecida pelo GeoNames. Este último passo da busca simples possibilita a identificação dos usuários que informam nomes alternativos de cidades. Foi efetuado tratamento descartando nomes alternativos duplicados em países diferentes.

3.3.1.3 Identificação de país - Busca múltipla

O objetivo da terceira etapa do algoritmo é obter a sigla do país (ISO2) a partir de múltiplas palavras informadas no campo "Location" como, por exemplo, "Sidney, Australia", "Lagos – Nigeria", "Okinawa / Japan". O algoritmo lê a cadeia de caracteres e quebra em pedaços ou *tokens* utilizando-se os separadores vírgula (,), ponto (.), hífen (-), barra deitada (/), pipe (|) e sinal tironiano (&). Após a separação da cadeia de caracteres em localidades simples, foi adotada a seguinte ordem de prioridade para obter a sigla do país (ISO2):

1. nome de país;
2. sigla de país (ISO3)
3. nome de capital de país;
4. cidades duplicadas no mesmo país;
5. cidades duplicadas em países diferentes;
6. nome de cidade (não capital) utilizando lista de cidades do GeoNames com mais de 15.000 habitantes.

Desta forma, inicialmente buscou-se a identificação de nome de país em cada palavra ou *token*. Por exemplo, se a localidade informada for a cidade e o país "Kampala , Uganda", o algoritmo identificará o país "Uganda". Nos casos em que dois países diferentes forem informados, como, por exemplo, "Brazil / Breland", o país identificado será o primeiro encontrado, neste caso, "Brazil".

3.3.1.4 Resultados

Foram processados 349.159 usuários e 550.779 comentários do Twitter referentes a 51 dias de coleta dos meses de novembro de 2014 a abril de 2015 utilizando o termo de busca “Ebola”. A Tabela 3.1 mostra os dados relativos ao resultado da execução do algoritmo. Do total de 349.159 usuários, 217.469 (62,28%) informaram o campo “Location” e 2.326 (1,06%) habilitaram o GPS, totalizando 219.795 usuários com alguma informação de localidade. Os percentuais mostrados na Tabela 3.1 foram calculados em relação a estes 219.795 usuários. Do total de usuários que informaram o campo “Location”, 133.471 usuários (60,73%) tiveram a sigla do país (ISO2) identificadas por meio deste algoritmo e não foi possível determinar o país de 86.324 usuários (39,27%). Observa-se que 0,83% dos usuários foram descartados por informarem e-mails, números ou palavras que não identificam a localidade. A maioria dos usuários (60,49%) informou apenas uma localidade, permitindo a identificação do país através da busca simples e 52.723 (39,50%) usuários informaram mais de uma palavra na localidade utilizando um separador. O tipo de local mais utilizado nesta amostra de dados foi o nome de cidade (30,88%), incluindo capital e não capital. O segundo tipo de local mais utilizado foi o país (19,82%) incluindo nome e sigla de país. 1,06% dos usuários habilitaram o GPS permitindo a identificação do país onde o usuário se encontrava no momento em que enviou o comentário.

A Tabela 3.2 mostra os dez nomes de localidades mais encontrados no campo “Location”. A coluna “Percentual” exibe o percentual referente a 133.471 usuários com país identificado. O nome da localidade mais encontrado nesta amostra de dados foi “London” (1,55%) identificado pela regra capital de país.

A Tabela 3.3 mostra os dez países com as maiores quantidades de usuários identificados por este algoritmo. A coluna “Perc.” exibe o percentual referente a 133.471 usuários com país identificado. As colunas “Cidade”, “País”, “Lat.long”, “Nome altern.” e “Reg. admin.” exibem, respectivamente, o nome da cidade, do país, a latitude e longitude, o nome alternativo e a região administrativa relativos ao tipo de local utilizado pelo algoritmo na identificação. Os valores numéricos apresentados nestas colunas referem-se ao percentual relativo ao total de usuários de cada país. Os Estados Unidos estão em primeiro lugar com 43.310 usuários (32,45%), sendo que 62,04% destes usuários informaram nomes de cidades. O Reino Unido está em segundo lugar com 10.716 usuários (8,03%) e destes, 76,83% informaram nomes de cidades. A Nigéria está em sexto lugar com 4.184 usuários e o tipo de local mais utilizado foi nome e sigla do país (60,76%). O Brasil aparece em sétimo lugar com 4.150 usuários, sendo que destes usuários, 4,07% ativaram o GPS, estando acima do percentual geral de 1,06%.

Tabela 3.1: Resultado da execução do algoritmo de identificação de país.

Tipo da regra de identificação	Quantidade	Percentual
Descartados	1.827	0,83%
Busca simples - latitude e longitude	2.326	1,06%
Busca simples - nome alternativo adicionado manualmente	9.433	4,29%
Busca simples - nome de país	19.307	8,78%
Busca simples - sigla de país (ISO3)	1.488	0,68%
Busca simples - nome de capital de país	8.720	3,97%
Busca simples - cidades duplicadas no mesmo país	4.108	1,87%
Busca simples - cidades duplicadas em países diferentes	6.055	2,75%
Busca simples - nome da cidade (não-capital) – GeoNames 15.000	16.821	7,65%
Busca simples - nome de região administrativa	6.144	2,80%
Busca simples - nome de cidade (não-capital) – GeoNames 1.000	2.210	1,01%
Busca simples - nome alternativo do GeoNames	4.136	1,88%
Busca múltipla - nome de país	19.714	8,97%
Busca múltipla - sigla de país (ISO3)	3.049	1,39%
Busca múltipla - nome de capital de país	3.711	1,69%
Busca múltipla - cidades duplicadas no mesmo país	8.287	3,77%
Busca múltipla - cidades duplicadas em países diferentes	5.078	2,31%
Busca múltipla - nome de cidade (não-capital) – GeoNames 15.000	12.884	5,86%
Total de localidades identificadas	133.471	60,73%
Localidades não identificadas	86.324	39,27%

3.3.1.5 Validação dos resultados

Para efetuar a validação dos resultados obtidos na execução do algoritmo de identificação de país, foi utilizada a ferramenta Batch Geocoding [1], disponível na internet de forma gratuita. A ferramenta foi utilizada no trabalho de Feingold et al. [31] para obter a localização geográfica dos pacientes cadastrados na lista de espera de transplante de coração nos Estados Unidos. O Batch Geocoding efetua a conversão de coordenadas geográficas (latitude e longitude) em endereços completos (logradouro, bairro, cidade, província, país) e efetua também a conversão inversa, isto é, transforma endereços completos ou endereços incompletos, como bairro e cidade, em coordenadas geográficas e país.

Foi utilizada uma amostra de 804 localidades (2,23%) dos cerca de 36.000 localidades diferentes informadas no perfil do usuário do Twitter. O resultado da comparação entre o algoritmo de identificação de país e o Batch Geocoding indica que, nesta amos-

Tabela 3.2: Dez nomes de localidades mais informadas no campo “Location”.

Location	Quantidade	Percentual
london	2.068	1,55%
usa	1.344	1,01%
new york	1.268	0,95%
nigeria	1.193	0,89%
venezuela	1.071	0,80%
argentina	989	0,74%
madrid	879	0,66%
washington, dc	826	0,62%
paris	824	0,62%
uk	800	0,60%

Tabela 3.3: Dez países com a maior quantidade de usuários identificados.

País	Qtd. usuários	Perc.	Cidade	País	Lat. lon.	Nome altern.	Reg. admin.
Estados Unidos	43.310	32,45%	62,04%	10,89%	2,37%	13,43%	11,28%
Reino Unido	10.716	8,03%	76,83%	4,93%	1,75%	15,60%	0,90%
México	7.068	5,30%	34,24%	32,65%	0,41%	27,74%	4,95%
Espanha	6.209	4,65%	61,19%	28,67%	1,16%	7,54%	1,45%
Argentina	4.700	3,52%	42,98%	49,21%	4,26%	3,23%	0,32%
Nigéria	4.184	3,13%	32,05%	60,76%	0,86%	5,86%	0,48%
Brasil	4.150	3,11%	56,55%	31,57%	4,07%	5,57%	2,24%
Canadá	3.994	2,99%	50,20%	45,67%	1,58%	1,73%	0,83%
França	3.790	2,84%	64,01%	31,50%	1,42%	2,72%	0,34%
Índia	2.993	2,24%	46,91%	48,55%	0,50%	2,41%	1,64%

tra de dados, 678 localidades (84,33%) resultaram na identificação do mesmo país e 126 localidades (15,67%) resultaram na identificação de países diferentes ou falhas na identificação. Analisando-se o resultado, verificou-se que as divergências ocorrem em casos de múltiplas localidades separadas por caracteres, como vírgula (,), traço (-) e barra (/). Por exemplo, no caso da localidade “Basedworld, USA”, o algoritmo desta pesquisa separou a localidade em duas partes: “Basedworld” e “USA” e por meio da regra de priorização, reconheceu primeiro “USA” como sigla de país e retornou os Estados Unidos. A ferramenta Batch Geocoding não reconheceu “Basedworld” como nome de localidade e não retornou nenhum país. No caso da localidade “Southampton/Hastings”, o algoritmo desta pesquisa identificou “Southampton” como uma cidade do Reino Unido e a ferramenta Batch Geocoding identificou “Southhampton” como nome de logradouro da cidade de “Hastings” na Nova Zelândia. Este resultado sugere o uso de regras diferentes entre as duas ferramentas. A abordagem do algoritmo desta pesquisa procura identificar primeiro a localidade no nível de país, para posteriormente, procurar no nível de cidade. O Batch Geocoding procura uma combinação das múltiplas localidades informadas para identificar o nível mais deta-

lhado, isto é, o logradouro de uma cidade, para posteriormente, procurar no nível de país. A Tabela 3.4 mostra outros exemplos em que as duas abordagens retornaram países diferentes. A coluna "Location" refere-se à localidade informada pelo usuário em seu perfil no Twitter, a coluna "País" refere-se ao país identificado pelo algoritmo desta pesquisa e a coluna "Resultado do Batch Geocoding" refere-se ao endereço completo ou incompleto retornado pela ferramenta Batch Geocoding.

Tabela 3.4: Exemplos de divergências nos resultados entre o algoritmo de identificação de país e a ferramenta Batch Geocoding.

Location	País	Resultado do Batch Geocoding
New Orleans / ATL / Chicago	United States	Chicago Ave & New Orleans Cres, Maroubra NSW 2035, Australia
USA-CANADA	Canada	Canada, KY 41519, USA
Los Angeles/San Francisco	United States	San Francisco, Primera Amp Gral Felipe Ángeles, Gral Felipe Ángeles, Hgo., Mexico
Buenos Aires, Argentina.24/09	Argentina	Argentina 24, Loma Bonita, 60983 Buenos Aires, Mich., Mexico
England/France	France	England, UK
Maputo/Cape Town/NYC	Mozambique	New York, NY, USA
South Africa | Brooklyn NY	South Africa	Brooklyn, NY, USA
Tokyo/Sapporo	Japan	Str. Privata di Sapporo e Tokyo, 18036 San Biagio della cima IM, Italy

Por ser um campo de digitação livre do usuário do Twitter, a escolha da abordagem correta para a identificação do país, a partir do campo "Location", torna-se bastante difícil. Por exemplo, quando o usuário informa "USA-CANADA", ele pode estar se referindo aos Estados Unidos, ou ao Canadá, ou pode ser um usuário em movimento. Assim, uma vez definidas as regras de priorização utilizadas no algoritmo desta pesquisa e identificadas as situações em que os resultados divergem nas duas abordagens, o algoritmo de identificação de país foi utilizado para efetuar a análise espacial desta pesquisa.

3.3.2 Algoritmo de identificação de idioma

Esta seção descreve o algoritmo desenvolvido para a identificação de idioma dos comentários do Twitter. Foram coletados 26.448 comentários dos usuários da África durante o período de seis meses. Analisando-se uma amostra dos comentários, verificou-se que além do inglês, os comentários foram escritos em outros idiomas como francês e espanhol. Desta forma, antes de efetuar a análise textual em inglês, foi necessário identificar o idioma dos comentários dos usuários da África para evitar classificações errôneas das

palavras.

A seguir, apresenta-se o algoritmo de identificação de idioma, em que foi utilizada a ferramenta gratuita Language Detection API [8, 59]. Esta ferramenta permite a identificação de 160 idiomas diferentes por meio de serviço web. Além deste serviço, a ferramenta fornece também bibliotecas para a programação nas linguagens Java, C#, Ruby, Python, PHP e Crystal. A Figura 3.10 mostra o fluxo utilizado no algoritmo que é composto pelas duas etapas descritas a seguir.

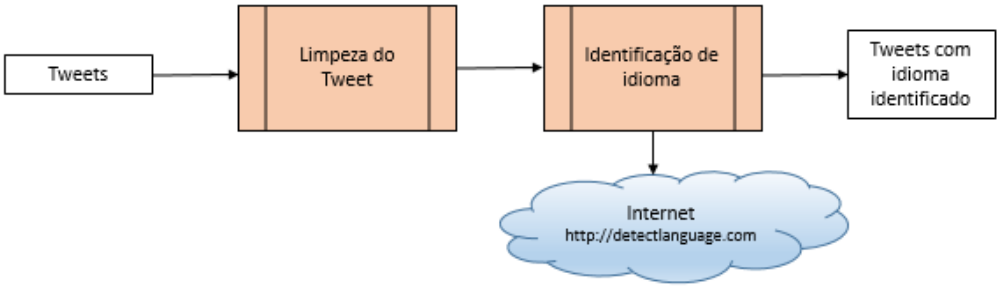


Figura 3.10: Algoritmo de identificação de idioma.

3.3.2.1 Limpeza do *Tweet*

O objetivo desta etapa é remover caracteres especiais (*carriage return*) e nomes de usuários (palavras que iniciam com @ como, por exemplo, @ebolaoutbreak) que dificultam a identificação do idioma e a posterior etapa de análise textual dos comentários.

3.3.2.2 Identificação de idioma

Nesta segunda etapa, o algoritmo utiliza o serviço web da ferramenta Language Detetion API. Envia a cadeia de caracteres do *Tweet* como parâmetro de entrada e recebe o idioma identificado como retorno da requisição.

3.3.2.3 Resultados

Foram processados 2.452 comentários dos usuários da África, referentes aos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015. A Tabela 3.5 mostra os três idiomas mais encontrados nos comentários da África. O idioma mais utilizado foi o inglês com 2.263 comentários (92,29%). O segundo idioma mais utilizado foi o francês, com 72 comentários (2,94%), e o espanhol foi o terceiro idioma mais utilizado com oito comentários (0,33%). Outros idiomas como holandês, italiano e português foram identificados em menor quantidade representando 1,14%. A ferramenta Language Detection API não conseguiu identificar o idioma de 81 comentários (3,30%) devido à presença de palavras

abreviadas, gírias, frases com mais de um idioma, URL, entre outros.

Tabela 3.5: Três idiomas mais utilizados nos comentários da África.

Idioma	Quantidade	Percentual
Inglês	2.263	92,29%
Francês	72	2,94%
Espanhol	8	0,33%
Outros	28	1,14%
Não identificado	81	3,30%
Total	2.452	

3.3.2.4 Validação dos resultados

A validação dos resultados do algoritmo de identificação de idioma foi efetuada por meio de conferência manual de 2.263 comentários em inglês da África. Verificou-se que 99,82% dos comentários foram identificados corretamente e quatro comentários (0,18%) que utilizavam palavras em inglês e francês na frase foram identificados incorretamente como inglês. A identificação correta do idioma é importante para não ocasionar resultados errôneos na análise textual. Dado o baixo índice (0,18%) de erro da ferramenta Language Detection API, considerou-se que o resultado deste algoritmo não ocasionará desvios significativos na pesquisa.

3.3.3 Algoritmo supervisionado de análise textual

Esta seção descreve o algoritmo desenvolvido nesta pesquisa para a análise textual dos comentários do Twitter. Foram selecionados somente comentários de usuários da África por representarem os países onde ocorreram a transmissão intensa da epidemia. Esta pesquisa baseia-se na abordagem de Ji et al. [40], em que os comentários foram divididos em *Tweets* pessoais e *Tweets* de noticiário ou não-pessoais. Após esta classificação, uma abordagem simplificada da análise de sentimentos foi aplicada sobre os *Tweets* pessoais, já que os não-pessoais se referem a notícias que relatam fatos sobre a epidemia e não expressam a opinião pessoal ou o sentimento de um indivíduo. A Tabela 3.6 mostra o quadro-resumo da classificação dos comentários utilizada nesta pesquisa. A automatização da análise textual de comentários em redes sociais apresenta algumas dificuldades devido à informalidade da linguagem, como o uso de gírias, figuras, expressões de ironia e a limitação de 140 caracteres dos comentários. Esta limitação de tamanho restringe a possibilidade de expressão dos usuários, levando-os ao uso de abreviações, URL, figuras, entre outros [46]. Nesta pesquisa optou-se por uma abordagem supervisionada em que amostras de comentários foram analisadas manualmente para efetuar a categorização e a criação do corpus de *Tweets* não-pessoais que tratam sobre a epidemia do Ebola. Após esta etapa, os *Tweets*

pessoais foram classificados automaticamente por meio de uma abordagem simplificada de análise de sentimentos. A Figura 3.11 representa o fluxo utilizado no algoritmo.

Tabela 3.6: Quadro-resumo da classificação dos comentários.

Tipo 1	*Tweets* não-pessoais
0	Relato de casos
1	Solução de contenção
2	Relato de impactos
3	Notícias negativas
4	Necessidade de preparo
5	Outros idiomas
6	Outros
Tipo 2	*Tweets* pessoais
0	Identificação de *emoticon* e *emoji*
1	Análise de sentimentos: sentimento positivo sentimento neutro sentimento negativo
Tipo 3	Outros
0	Outros idiomas
1	Outros

3.3.3.1 Preparação

O objetivo desta etapa é identificar manualmente a presença de URL nos comentários para compor duas listas de URL: a lista de sítios de noticiários, órgãos de saúde, agências não-governamentais e governamentais, cujos exemplos encontram-se na Tabela 3.7; e a lista de sítios pessoais como blogs, sítios de entretenimento, cujos exemplos encontram-se na Tabela 3.8.

3.3.3.2 Identificação de *Tweets* não-pessoais

Após a criação das duas listas de URL, os comentários que contém URL da lista de sítios de noticiários foram identificados como *Tweets* não-pessoais. Estes comentários foram utilizados como conjunto de treinamento para a criação do corpus de noticiários e categorizados conforme as regras descritas na Seção 3.3.3.3. Analogamente, os comentários que contém URL da lista de sítios pessoais foram tratados como *Tweets* pessoais e categorizados conforme descrito na Seção 3.3.3.4.

3.3.3.3 Categorização de *Tweets* não-pessoais

A Tabela 3.9 mostra os critérios de categorização de *Tweets* não-pessoais utilizados nesta pesquisa. Descreve-se a seguir, os procedimentos executados, nesta ordem, para

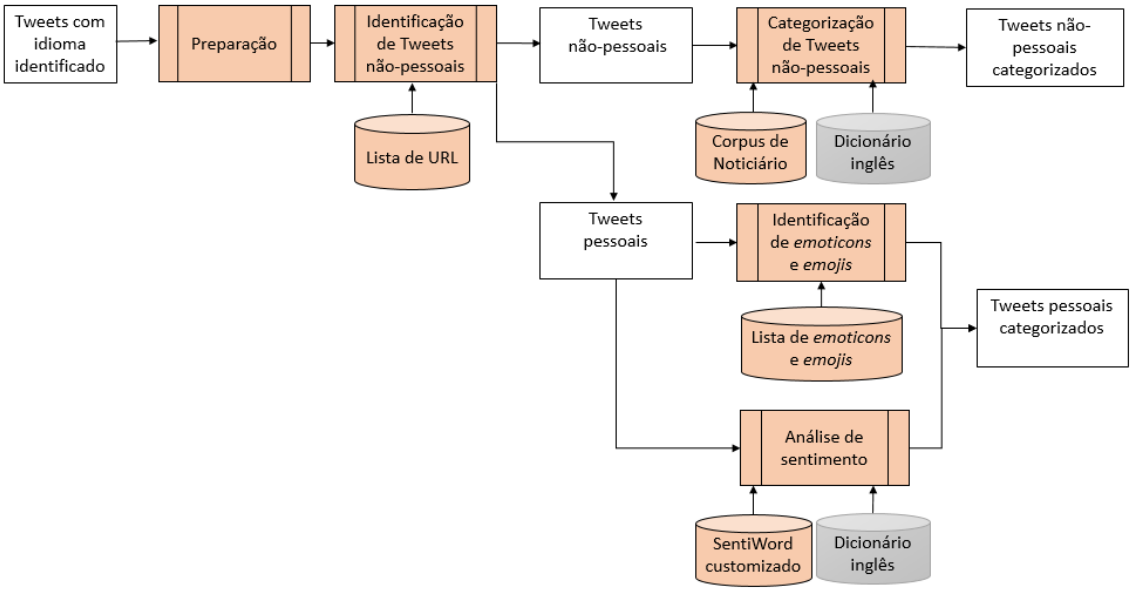


Figura 3.11: Algoritmo supervisionado de análise textual.

efetuar a categorização dos *Tweets* não-pessoais.

1. Outros idiomas. Os comentários de sítios de noticiários em idioma diferente do inglês foram classificados como código 5 (outros idiomas), conforme a Tabela 3.9. Os comentários de sítios de noticiários cujo idioma não foi possível identificar por meio do algoritmo de identificação de idioma foram classificados como código 6 (outros). Este procedimento foi necessário para evitar categorizações errôneas de comentários nas fases seguintes.

2. *Tweets* em inglês de sítios de noticiários. A seguir, *Tweets* da África dos dias 08/11/2014 e 19/11/2014 cuja URL se encontra na lista de sítios de noticiários foram utilizados como conjunto de treinamento e classificados manualmente em cinco categorias: informes da situação, soluções de contenção, impactos causados pelo Ebola, notícias negativas e necessidades de preparo, conforme o conteúdo descrito na Tabela 3.9. O resultado da categorização gerou o corpus de noticiários contendo palavras que auxiliaram na categorização supervisionada dos *Tweets* da África referentes aos dias 23/12/2014 e 30/04/2015 por meio da ferramenta Full Text Search. Foram identificadas novas palavras nos comentários dos dias 23/12/2014 e 30/04/2015, as quais foram acrescentadas ao corpus de noticiários já existente. A Tabela 3.10 mostra alguns exemplos de corpus para cada categoria. Os códigos 5 (outros idiomas) e 6 (outros) não estão presentes nesta tabela porque a identificação de comentários nestas categorias não é efetuada por meio de corpus de noticiários. Os comentários do código 5 são identificados por meio do algoritmo de identificação de idioma, conforme descrito no primeiro passo deste algoritmo. Os comentários

Tabela 3.7: Exemplos de URL de sítios de noticiários, órgãos de saúde, agências não-governamentais e governamentais.

URL	Descrição
bbc.in	BBC News
wp.me	World Press
cnn.it	CNN International
reuters.com	Reuters
ap.org	Associated Press
allafrica.com	AllAfrica
vanguardngr.com	Vanguard Media Limited, Nigeria
n24.cm	News 24 South Africa
enca.com	eNews Channel Africa
premiumtimesng.com	The Premium Times Nigeria
undp.org	Programa de Desenvolvimento das Nações Unidas
who.int	Organização Mundial de Saúde
peaceau.org	União Africana de Segurança e Paz
un.org	Nações Unidas
usa.gov	Governo dos Estados Unidos
unicef.org	Fundo das Nações Unidas para a Infância
worldbank.org	Banco Mundial
cdc.gov	Centro de Controle e Prevenção de Doenças

Tabela 3.8: Exemplos de URL de sítios pessoais como blogs, sítios de entretenimento.

URL	Descrição
youtube.com	Sítio para disponibilização de vídeos
blogspot.com	Sítio de blog
vine.co	Sítio para disponibilização de vídeos
instagram.com	Sítio para disponibilização de fotos e vídeos
9gag.tv	Sítio de entretenimento
vibe.com	Sítio de entretenimento
michelleniniblog.com	Sítio de blog

do código 6 (outros) ou são identificados pelo algoritmo de identificação de idioma, quando não for possível determinar o idioma, ou são processados pelo último passo deste algoritmo quando o idioma for o inglês.

3. *Tweets* em inglês com qualquer URL. Uma vez criado o corpus de noticiários, os comentários que ainda não foram categorizados e que fazem referência a qualquer URL foram analisados automaticamente por meio da ferramenta Full Text Search utilizando o corpus de cada categoria. O objetivo desta etapa é identificar comentários que são copiados de sítios de noticiários cujo conteúdo relatam fatos relacionados à epidemia, independente da URL a que faz referência.

4. *Retweets* de sítios de noticiários. No Twitter, usuários podem encaminhar (*Retweet*) o comentário postado por outro usuário, caso tenha interesse em disseminar a infor-

mação para seus seguidores. Neste caso, o comentário pode conter ou não uma URL. O objetivo desta etapa é identificar os comentários de usuários de sítios de noticiários e órgãos não-governamentais que foram encaminhados por seus seguidores. A identificação deste tipo de comentário foi efetuada através da presença do nome do usuário, por exemplo @unicef, no texto do comentário. Neste caso, o usuário @unicef representa o usuário cujo comentário foi encaminhado por seus seguidores. Uma vez identificados estes comentários, foram categorizados por meio da ferramenta Full Text Search utilizando-se o corpus de noticiários criado anteriormente.

5. Outros. Por fim, os comentários cuja URL está na lista de sítios de noticiários, porém, cujo conteúdo não foi possível classificar nas categorias descritas na Tabela 3.9, foram classificados como código 6 (outros). Nesta etapa, não foram considerados os comentários que continham palavras que expressam sentimento como, por exemplo, *crappy*, em português "porcaria". Tais comentários foram tratados em etapa posterior, como *Tweets* pessoais.

3.3.3.4 Categorização de *Tweets* pessoais e outros

Os comentários que não foram identificados como *Tweets* não-pessoais foram tratados de acordo com as regras descritas nesta seção. Para efeitos da presente pesquisa, buscou-se um método simples para classificar os *Tweets* pessoais. Optou-se por efetuar a análise de sentimentos no nível de granularidade de palavras utilizando o dicionário léxico SentiWordNet. Este dicionário foi proposto no estudo de Baccianella et al. [22] e utiliza o corpus WordNet, em que cada palavra possui uma pontuação numérica, denominada polaridade, que varia de 0 à 1 para o sentimento positivo e de 0 à -1 para o sentimento negativo. Descrevem-se, a seguir, os procedimentos executados, nesta ordem, para efetuar a categorização dos *Tweets* pessoais.

1. Identificação de *emoticon* e *emoji*. Os comentários que contém *emoticon* ou *emoji* foram classificados em sentimento positivo, com a polaridade 0.7, ou negativo, com a polaridade -0.7. Nesta pesquisa, assumiu-se que a presença de *emoticon* ou *emoji* no *Tweet* representa o sentimento do comentário como um todo [55]. Por exemplo, a presença do *emoticon* ":-)" no final do comentário *"$5.7m for ebola :-)"* sugere sentimento positivo do usuário pela doação de 5,7 milhões de dólares para auxiliar na luta contra o Ebola. Desta forma, o comentário foi pontuado com a polaridade 0.7. O comentário *"hunger kils more than ebola, but it is not considered a significant problem, since the rich can't die of it.. #smh #reminisce.. :("* foi pontuado

Tabela 3.9: Critérios para a categorização de *Tweets* não-pessoais.

Código	Categoria	Descrição
0	Informes da situação	Relatos de números oficiais da epidemia, relatos de países com a epidemia já controlada, fechamento de centros de tratamento, etc.
1	Soluções de contenção	Pesquisas de vacinas e diagnóstico da doença, doações e campanhas para ajuda, soluções de isolamento adotada pelas autoridades, hospitais que recebem pacientes, informações educativas de prevenção e formas de contaminação, soluções de tecnologia para disseminação de informação, discussões e negociações entre governos, etc.
2	Impactos causados pelo Ebola	Impactos sociais, econômicos, fechamento de escolas, aumento de gravidez na adolescência, desemprego, crise econômica nas áreas afetadas, aumento de órfãos na região, etc.
3	Notícias negativas	Críticas aos governos, racismo, discriminação, revolta, etc.
4	Necessidades de preparo	Falta de hospitais, doações de sangue, alimentos, saco plástico para funerais seguros, procedimentos para Ebola nos hospitais, alerta sobre a diminuição da noção de risco, etc.
5	Outros idiomas	Quando o idioma identificado não for o inglês.
6	Outros	Quando não for possível identificar o idioma, comentários de sítios de noticiários cuja classificação não foi possível.

com a polaridade -0.7 devido à presença do *emoticon* ":(" que sugere sentimento negativo. A Tabela 3.11 mostra a lista de *emoticons* utilizados nesta pesquisa [55]. A Figura 3.12 mostra a lista de *emojis* que representam sentimento positivo e a Figura 3.13 mostra a lista de *emojis* que representam sentimento negativo, utilizados nesta pesquisa [50].

Figura 3.12: Lista de *emojis* que expressam sentimento positivo. Fonte:Novak et al. [50].

2. Outros. Os comentários em inglês sem relação com a epidemia ou cujo idioma não foi possível identificar foram classificados como Tipo 3 - outros - código 1 (outros), conforme a classificação mostrada na Tabela 3.6. Os comentários em idiomas diferente do inglês que não foram tratados como *Tweets* não-pessoais foram classificados como Tipo 3 - outros - código 0 (outros idiomas). Este procedimento foi

Tabela 3.10: Exemplos de corpus de noticiários para a categorização de *Tweets* não-pessoais.

Código	Categoria	Exemplos de corpus
0	Informes da situação	alert over ebola, ebola symptoms, new ebola cases, situation report, test positive, rising infections rates, confirmed case, number ebola cases, test negative, declare free virus, ebola update, zero cases.
1	Soluções de contenção	frequently asked questions, raise funds, hire doctor, survival guide, information center, educational video, drug clinical trials, treatment research, safe burial, preventative measures, charity ebola, ebola volunteer, construct ebola centre, track contact, wear protective suits, lessons outbreak, revise guideline, technology against ebola, ebola research.
2	Impactos causados pelo Ebola	hit economy, devastating impact, education fall, pregnant women, depreciate currency, education fall, affect trade, collapse, catastrophe, poor economy.
3	Notícias negativas	scare, ebola hysteria, throw stone, blame ebola spread, health crisis, mass panic, children bullied, prejudice, racism, fear black, corruption, anger.
4	Necessidades de preparo	need guidance, need training, need information, food problem, allocate resources, recovery effort, out body bag, no surgeon, lack vaccine, more medics, confuse guideline, need diagnostic test.

necessário neste momento para evitar categorizações errôneas de comentários nas fases seguintes.

3. Análise de sentimentos simplificada. Os comentários do tipo *Tweet*, *Reply* e *Mentions* que ainda não foram classificados até esta etapa foram tratados conforme as regras definidas a seguir.

   (a) *Tweet*. Os comentários em inglês do tipo *Tweet* que não contém URL ou a URL informada não faz parte da lista de sítios de noticiários foram assumidos como comentários pessoais.

   (b) *Reply* e *Mentions*. Os comentários em inglês do tipo *Reply* e *Mentions* que não contém URL ou a URL informada faz parte da lista de sítios pessoais foram assumidos como comentários pessoais.

Inicialmente o comentário foi processado por meio da ferramenta Full Text Search,

Tabela 3.11: *Emoticons* representando sentimentos positivo e negativo utilizados nos comentários. Fonte:Pak et al. [55].

Descrição	Lista de *emoticons*
Sentimento positivo	:-), :), :D, :o), :], :3, :c), :>, =], 8), =), :}, :^:))
Sentimento negativo	:-(, =(, ;(, >:[, :-(, :(, :-c, :c, :-<, :□C, :<, :-[, :[, :{, :-‖, :@, >:(

Figura 3.13: Lista de *emojis* que expressam sentimento negativo. Fonte:Novak et al. [50].

que efetuou o tratamento de eliminação de palavras de parada e retornou somente substantivos, advérbios e adjetivos. Palavras de parada são termos ignorados em mecanismos de busca textual como artigo e preposição. As palavras retornadas pela ferramenta Full Text Search foram comparadas com a base de dados do SentiWordNet, obtendo-se, assim, a polaridade de cada palavra separadamente. Neste processo, verificou-se que a pontuação de algumas palavras no SentiWordNet necessitavam ser revistas para se adequarem ao contexto de calamidade pública e epidemia. Por exemplo, a pontuação original da palavra *"positive"* é 0,672 por representar o sentimento em seu uso genérico. Porém, para se adequarem ao contexto de epidemia, em que a palavra é empregada no sentido de resultado positivo ao vírus do Ebola, a pontuação foi ajustada para -0,2. A Tabela 3.12 mostra outros exemplos de palavras cuja pontuação foi alterada. No comentário *"#Lahoud: People refuse to allow you into places because the first thing they think is that you have Ebola."* as palavras *"people"*, *"first"*, *"think"* e *"place"* possuem pontuações positivas no SentiWordNet e foram alteradas para zero a fim de expressarem um valor neutro. Verificou-se que, na amostra de dados analisada, a presença destas palavras não transmitem opiniões positivas no contexto de epidemia. Por outro lado, no comentário *"If an #Ebola patient died in the #US and has led to spread of infection to one nurse; does this provide a certificate of failure !!"* a palavra *"spread"* com pontuação original 0,147 foi utilizada no sentido de espalhar a epidemia e a pontuação foi alterada para -0,3. Ao final, o somatório da polaridade das palavras foi dividido pela quantidade de palavras com polaridade diferente de zero para obter-se a polaridade do comentário como um todo. Após analisar manualmente a amostra de 100 comentários, observou-se que os comentários cuja polaridade variam entre -0,10 e +0,10 expressam fatos e foram categorizados como sentimento neutro.

Tabela 3.12: Exemplos de palavras cuja pontuação no SentiWordNet foi alterada.

Palavra	Pontuação original	Pontuação alterada	Exemplos de *Tweets*
affected	0,061	-0,061	the countries which are affected wit Ebola, are being punished for the wrong doings
campaign	-0,113	0,200	Overcoming fear, myths and public mistrust to conduct a large vaccination campaign in Guinea
fight	-0,365	0,200	Save Catholic clinics in Africa – they may be the only ones to fight Ebola :: Catholic News Agency (CNA)
hospital	-0,370	0,000	But the last thing that threatened our security and came across the border was #Ebola. So why are we not building hospitals?
joke	0,720	-0,010	Is America done with Ebola jokes?
people	0,175	0,000	#Lahoud: People refuse to allow you into places because the first thing they think is that you have Ebola.
stop	-0,068	0,100	We cannot stop Ebola coming to Zim, but we must know of the very first case so we dont have secondary cases emerging says MoH. #FightEbola
spread	0,143	-0,300	If an #Ebola patient died in the #US and has led to spread of infection to one nurse; does this provide a certificate of failure !!
risk	0,043	-0,100	nobody is safe, the risk is just higher in poor hygiene area's and area's where #ebola already is #hygiene #knowyourfacts
rumour	0,000	-0,150	The reaction to rumours of Ebola can have an economically decapitating effect.

4. Outros. Por fim, os comentários cujo conteúdo não foi possível classificar através das regras citadas anteriormente foram categorizados como Tipo 3 - outros - código 1 (outros).

3.3.3.5 Resultados

Foram processados 2.452 comentários dos usuários da África referentes aos dados coletados nos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015. A Tabela 3.13 mostra o resultado da análise textual. Uma quantidade de 1.476 comentários, que representam 60,20%, referem-se a *Tweets* não-pessoais e 829 comentários, que representam 33,81%, referem-se a *Tweets* pessoais. Destes 829 comentários, 292 (11,91%) foram classificados como negativos, 197 (8,03%) foram classificados como positivos e 340 (13,87%) foram classificados como neutros. 147 comentários (6,00%) referem-se a outros idiomas ou comentários cujo idioma não foi possível identificar ou cujo conteúdo não está relacionado ao Ebola.

Tabela 3.13: Resultado da classificação dos comentários da África.

Tipo 1	*Tweets* não-pessoais	Qtde	Percentual
0	Relato de casos	321	13,09%
1	Solução de contenção	767	31,28%
2	Relato de impactos	129	5,26%
3	Notícias negativas	99	4,04%
4	Necessidade de preparo	35	1,43%
5	Outros idiomas	31	1,26%
6	Outros	94	3,83%
	Subtotal	1.476	60,20%
Tipo 2	*Tweets* pessoais		
0	Identificação de *emoticons* e *emoji*	26	1,06%
1	Análise de sentimentos:	803	32,75%
	sentimento positivo	197	8,03%
	sentimento neutro	340	13,87%
	sentimento negativo	292	11,91%
	Subtotal	829	33,81%
Tipo 3	Outros		
0	Outros idiomas	72	2,94%
1	Outros	75	3,06%
	Subtotal	147	6,00%
	Total	2.452	

Uma das dificuldades encontradas na abordagem de análise textual utilizada foi a grande variedade de URL identificada nos comentários, gerando a necessidade de validação manual de cada URL para criar a lista de sítios de notícias e sítios pessoais. Os 2.452 comentários referentes ao período de análise continham 521 URL diferentes. Outra dificuldade encontrada foi a criação do corpus para a categorização dos *Tweets* não-pessoais. A grande variedade de palavras e expressões utilizadas nos comentários fez com que o corpus necessitasse de revisão a cada nova carga de dados, tornando o processo de classificação demorado e dificultando a análise de uma quantidade maior de comentários. A abordagem de análise de sentimentos utilizada nesta pesquisa é uma técnica simples que baseia-se na polaridade individual de cada palavra. Esta técnica não permite tratar comentários complexos contendo expressões de sarcasmo ou ironia. Por exemplo, o comentário "*I would love to hear PatRobertsons take on Ebola in the US*", que sugere ironia por estar sendo utilizado no contexto de epidemia, foi classificado como *Tweet* pessoal com sentimento positivo e pontuação 0,13. Da mesma forma, comentários envolvendo expressões de negação e perguntas não foram tratados nesta pesquisa.

3.3.3.6 Validação dos resultados

A validação dos resultados da execução do algoritmo supervisionado de análise textual relativo aos *Tweets* pessoais foi efetuada por meio da comparação com a análise manual dos comentários. Foi selecionada uma amostra de 100 comentários da África, representando 4,08% do total de 2.452 comentários processados pelo algoritmo. A análise manual foi efetuada por um colaborador independente, o qual não faz parte dos envolvidos nesta pesquisa. Na análise manual, foram atribuídos valores entre 0 e 1, para comentários com sentimento positivo e valores entre 0 e -1, para comentários com sentimento negativo, não sendo efetuada a classificação de sentimentos neutros. O resultado da comparação entre a análise manual e o algoritmo supervisionado de análise textual mostrou que 66% ou 66 comentários obtiveram o mesmo resultado entre as duas abordagens, isto é, classificaram o mesmo comentário como positivo ou negativo, e 34% ou 34 comentários obtiveram resultados diferentes. Analisando-se estes 34 comentários, verificou-se que o algoritmo classificou 11 comentários como positivo e 23 como negativo. Na análise manual, verificou-se o contrário: 23 foram classificados como positivo e 11 como negativo. Este resultado sugere a tendência do algoritmo classificar negativamente os comentários, provavelmente devido ao tema da análise estar ligado à doença e as palavras utilizadas nos comentários possuírem a pontuação negativa no SentiWordNet. O resultado sugere também a tendência do colaborador interpretar positivamente os comentários. Por exemplo, no comentário *"The fight is on! we will beat ebola"*, o colaborador interpretou a frase como um comentário otimista, no sentido de que a epidemia será eliminada. Na opinião do colaborador, a presença do ponto de exclamação (!) enfatizou o otimismo da frase e foi pontuado com 0,90. O algoritmo pontuou o mesmo comentário com -0,082 por considerar somente a polaridade de cada palavra e não verificar o contexto como um todo, além de não interpretar a presença do ponto de exclamação. A Tabela 3.14 mostra outros exemplos em que ocorreram divergências no resultado da comparação das duas abordagens. A coluna "Análise manual" mostra o resultado da pontuação do colaborador e a coluna "Algoritmo" mostra o resultado da pontuação do algoritmo desenvolvido nesta pesquisa. A comparação do resultado destes 34 comentários não permite verificar qual abordagem é correta, pois a análise de sentimentos é subjetiva e depende de fatores como, experiência pessoal, conhecimento sobre o assunto, familiaridade com frases curtas de comentários (*Tweets*) e opiniões pessoais sobre o assunto comentado. O resultado mostra também a importância de tratar comentários envolvendo expressões de negação e exclamação que podem mudar a classificação do resultado da análise textual. Apesar de 34% dos comentários apresentarem resultados divergentes na validação utilizada nesta amostra de 100 comentários, o percentual de 66% permite verificar a tendência dos comentários da África na análise

textual descrita na Seção 4.3.

Tabela 3.14: Exemplos de *Tweets* com resultados diferentes obtidos pela análise manual e pela análise do algoritmo supervisionado de análise textual.

Tweets	Análise manual	Algoritmo
Despite our differences politically... the threat of ebola... ”a divided house will not stand”.	0,80	-0,14
Europe no dust no mosquito no malaria no ebola no cholera no corruption good democracy but no-backway !!	0,40	-0,26
We don’t have ebola now, but we need to prepare lets have the requirements in place.	0,80	-0,02
The fight is on! we will beat ebola.	0,90	-0,08
It may get worse, but we can defeat ebola.	0,50	-0,16

4. Análise dos Dados

Este capítulo está dividido em quatro seções. A Seção 4.1 utiliza os conceitos da Teoria de Grafos para efetuar uma análise dos dados coletados do Twitter sobre o Ebola. O objetivo desta análise é verificar a presença de vértices que auxiliam a disseminação de informações sobre a epidemia e seu comportamento ao longo de seis meses. A Seção 4.2 apresenta uma análise espacial cujo objetivo é mostrar, 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 Seção 4.3 apresenta uma análise textual dos comentários no Twitter dos usuários da África, em que verifica-se a tendência de variação dos *Tweets* pessoais e não-pessoais durante o período de análise. Por fim, a Seção 4.4 apresenta o resultado da análise desta pesquisa.

4.1 Análise por meio da Teoria de Grafos

Esta seção apresenta uma análise dos dados coletados do Twitter sobre o termo "Ebola" utilizando a Teoria de Grafos. A ferramenta NodeXL foi utilizada para a coleta dos dados, cálculo das métricas em grafos e para a geração dos grafos. Os dados foram coletados durante seis meses, no período de 01/11/2014 a 30/04/2015. 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. A seguir, foram gerados os grafos referentes ao dia 19/11/2014, que faz parte do período de transmissão intensa do vírus do Ebola. A análise baseou-se nos dois maiores componentes conexos, identificados como Grafo G1 e Grafo G2. As métricas grau de entrada, grau de saída, centralidade de intermediação e centralidade de proximidade foram utilizadas para selecionar os vértices chaves dos dados coletados no dia 19/11/2014. A seguir, foi efetuada uma análise temporal das métricas selecionadas para verificar as tendências de comportamento dos

vértices chaves ao longo de seis meses. Por fim, foi efetuada uma análise dos usuários que permaneceram por mais tempo entre os maiores valores das métricas selecionadas durante o período de análise.

4.1.1 Análise

Inicialmente foram analisados os dados coletados no dia 19/11/2014, que contém 9.226 vértices e 10.371 arestas, sendo 8.338 arestas únicas e 2.033 arestas duplicadas. Os dados foram agrupados em 5.263 componentes conexos, sendo 4.148 componentes com vértices únicos formado por laços (*self-loop*), 1.075 componentes conexos com de dois a dez vértices, 31 componentes conexos com de 11 a 30 vértices e nove 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 4.1 foi gerada por meio do NodeXL selecionando-se o algoritmo de Fruchterman-Reingo, o qual permite posicionar os vértices adjacentes próximos uns aos outros mantendo o comprimento das arestas uniforme [33]. A Figura 4.1 está dividida em 112 retângulos. Cada retângulo representa um componente conexo. Nesta figura, os 112 componentes conexos com mais de seis vértices estão ordenados de forma decrescente pela quantidade de vértices de cada componente. Os 5.151 componentes menores, com até cinco vértices, foram ocultos para melhor visualização.

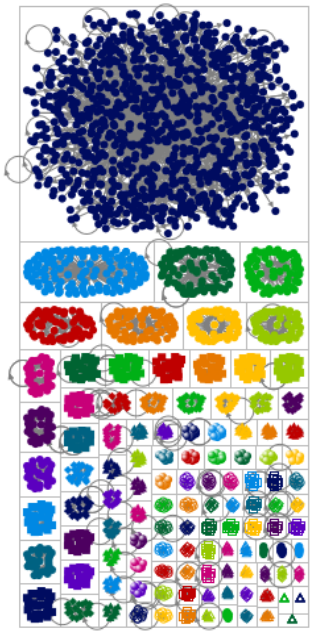


Figura 4.1: Exemplo de 112 componentes conexos com mais de seis vértices.

A seguir, apresentam-se os dois maiores componentes conexos desta coleta. A Figura 4.2 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 4.3 refere-se ao segundo maior componente conexo, identificado por Grafo G2. Este grafo apresenta 118 vértices e 120 arestas, sendo 120 arestas únicas. Nos dois grafos, foi selecionado o algoritmo Fruchterman-Reingo para gerar o leiaute do grafo. 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.

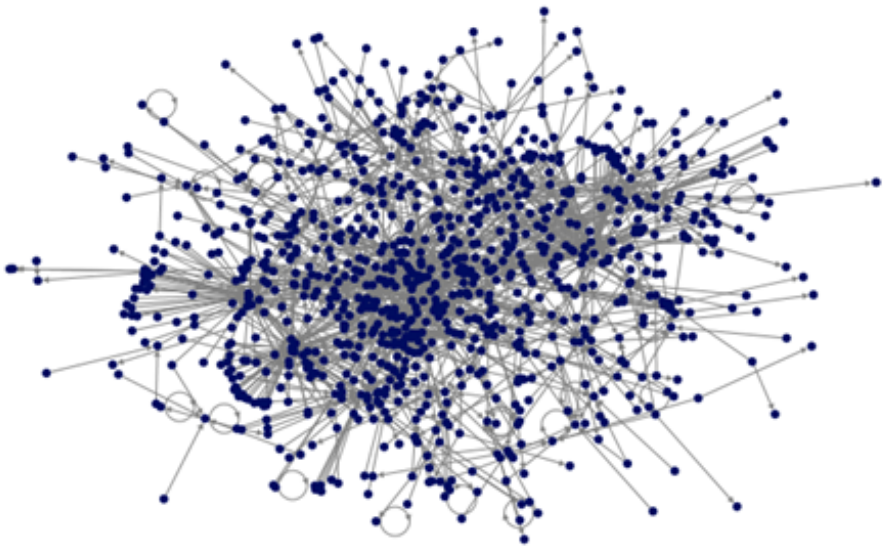


Figura 4.2: Grafo G1 com 987 vértices e 1.582 arestas, sendo 1.252 arestas únicas e 330 arestas duplicadas.

Na sequência, foram analisados vértices com os dez maiores valores das métricas centralidade de intermediação, centralidade de proximidade, grau de entrada e grau de saída referentes a todos os dados coletados no dia 19/11/2014. A Tabela 4.1 apresenta os usuários com os dez maiores valores de centralidade de intermediação, os quais pertencem todos ao Grafo G1, e a Tabela 4.2 apresenta os usuários com os dez maiores valores de grau de saída. As duas tabelas permitem verificar que o usuário @ebolaphone apresenta simultaneamente a maior centralidade de intermediação e o maior grau de saída, sugerindo um usuário ativo que posta comentários (*Tweet*), responde aos usuários (*Reply*) e encaminha comentários (*Retweet*) de outros usuários a seus seguidores, além de atuar como intermediário entre grupos diferentes, facilitando assim a disseminação de informações. Por meio dos perfis disponíveis no Twitter, verificou-se um aspecto interessante sobre os usuários: dentre os usuários listados na Tabela 4.1 com os dez maiores valores de centralidade de intermediação, somente @ebolaphone, @heidi5969 e @yemanish são usuários individuais e os 70% restantes são usuários que representam agências de notícias, jornais ou organi-

Figura 4.3: Grafo G2 com 118 vértices e 120 arestas, sendo 120 arestas únicas.

zações não-governamentais. Verificou-se o inverso com relação aos dez maiores valores de grau de saída, em que 80% dos usuários da tabela Tabela 4.2 são usuários individuais e somente @totalngonews e @unicefdrc são organizações ou agências de notícias.

A Tablela 4.3 apresenta os usuários com os dez maiores valores de grau de entrada. É possível verificar que o usuário @thedouch3 possui o maior grau de entrada, sugerindo que os comentários postados por este usuário chamaram a atenção de seus seguidores sendo encaminhados (*Retweet*) e mencionados (*Mentions*) por outros usuários. Analisando os comentários deste usuário, o comentário "*The US government worrying about Ebola but Gucci Mane still locked up we need to start worrying about the real issues*" resultou em 97 *Retweets* e *Mentions* no dia 19/11/2014. Gucci Mane refere-se a um cantor norte americano detido na prisão e o teor do comentário sugere a insatisfação do usuário @thedouch3 pelo fato das autoridades dos Estados Unidos estarem mais preocupadas com o vírus do Ebola do que com a prisão do cantor. O usuário @thedouch3 é o vértice central do maior subgrafo do Grafo G2, apresentado na Figura 4.3. Apesar de possuir o maior grau de entrada, o usuário @thedouch3 possui um valor inexpressivo de centralidade de intermediação, não constando nos dez maiores valores desta métrica. Por outro lado, o usuário @telegraph, que refere-se ao jornal The Telegraph do Reino Unido, possui o segundo

Tabela 4.1: Usuários com os dez maiores valores de centralidade de intermediação.

Usuário	Centralidade de intermediação	Grau de entrada	Grau de saída	Componente
ebolaphone	526.821,09	12	33	G1
telegraph	317.765,79	83	0	G1
bbcafrica	281.702,50	64	1	G1
reutersafrica	190.567,75	17	4	G1
msf	128.187,38	16	0	G1
ajenglish	127.022,75	10	0	G1
heidi5969	123.930,45	3	18	G1
yemanish	123.846,19	0	3	G1
ap	122.119,24	17	0	G1
guardian	119.455,16	8	0	G1

Tabela 4.2: Usuários com os dez maiores valores de grau de saída.

Usuário	Centralidade de intermediação	Grau de entrada	Grau de saída	Componente
ebolaphone	526.821,09	12	33	G1
ebolaoutbreakus	65.624,00	2	21	G1
heidi5969	123.930,45	3	18	G1
totalngonews	14.384,42	0	13	G1
racecarodds	108.393,49	2	10	G1
tz_uchay	262,00	0	10	G17
dondraper77	50.955,03	1	9	G1
dondraper76	50.955,03	1	9	G1
robinsnewswire	58.827,44	1	9	G1
unicefdrc	9.410,27	5	8	G1

maior valor de centralidade de intermediação e grau de entrada, sugerindo um usuário cujos comentários são encaminhados e mencionados, mas também comporta-se como um importante disseminador de informações entre grupos diferentes. O jornal The Telegraph publicou um artigo sobre a não participação da cantora inglesa Adele na campanha de ajuda ao Ebola, o que desencadeou comentários mencionando o usuário @telegraph. A Tabela 4.4 mostra alguns exemplos de comentários em que o usuário @telegraph foi mencionado. Analisando-se os perfis dos usuários com maior grau de entrada, verificou-se que 40% são de usuários individuais e 60% representam agências de notícias, jornais e organizações não-governamentais.

A seguir, utilizando o NodeXL, aplicou-se um filtro dinâmico no Grafo G1, tornando visíveis somente os vértices cujo valor da métrica centralidade de intermediação é maior do que 80.000. O resultado é exibido na Figura 4.4. O tamanho do vértice é proporcional ao valor desta métrica, e percebe-se que todos os usuários com os dez maiores valores de centralidade de intermediação estão presentes neste grafo, sendo o @ebolaphone o vértice

Tabela 4.3: Usuários com os dez maiores valores de grau de entrada.

Usuário	Centralidade de intermediação	Grau de entrada	Grau de saída	Componente
thedouch3	13.185,00	97	0	G2
telegraph	317.765,79	83	0	G1
bbcafrica	281.702,50	64	1	G1
msf_italia	3.362,00	50	0	G4
yusnaby	2.945,00	48	0	G6
ajenews	92.282,00	44	0	G1
unicefkorea	1.720,00	43	1	G8
debosospechar	1.718,00	40	0	G7
foxnews	4.657,00	39	1	G3
myxphilippines	930,00	32	1	G9

Tabela 4.4: Exemplos de *Tweets* em que o usuário @telegraph foi mencionado.

Tweets
Why Adele was right to ignore Bob Geldof and Band Aid via @Telegraph.
RT @ajsomer: Bob Geldof needs a lesson in PR and maybe a look in the mirror! Great article @Telegraph.
Worst article of the day so far - Why Adele was right to ignore Bob Geldof and Band Aid | via @Telegraph.
Excellent article: Why Adele was right to ignore Bob Geldof and Band Aid - via @Telegraph
Ebola diary: no touching, and a constant stink of chlorine | via @Telegraph

de maior tamanho.

Para analisar os valores de centralidade de poximidade, grafos com baixo número de vértices foram ignorados e foram considerados somente os vértices pertencentes aos dois maiores grafos G1 e G2, por estes possuírem uma conectividade maior. Os valores da métrica centralidade de proximidade foram normalizados para permitir a comparação dos grafos G1 e G2, os quais possuem tamanhos diferentes. Verificou-se que os dez maiores valores desta métrica pertencem a vértices do grafo G2, sendo que o usuário @thedouch3 apresentou o maior valor (0,759798), conforme mostrado na Tabela 4.5. Este resultado sugere que o usuário @thedouch3 consegue interagir rapidamente com todos os usuários do grafo G2 pela sua posição estratégica no grafo. A Tabela 4.5 mostra também que os outros usuários apresentam valores inexpressivos de grau de entrada, grau de saída e centralidade de intermediação, não sendo considerados para fins da análise desta métrica. Desta forma, os usuários @ebolaphone, @thedouch3 e @telegraph serão considerados vértices chaves dos dados coletados no dia 19/11/2014.

A seguir, foi efetuada uma análise temporal das métricas centralidade de intermediação, grau de entrada e grau de saída para verificar as tendências dos vértices chaves ao

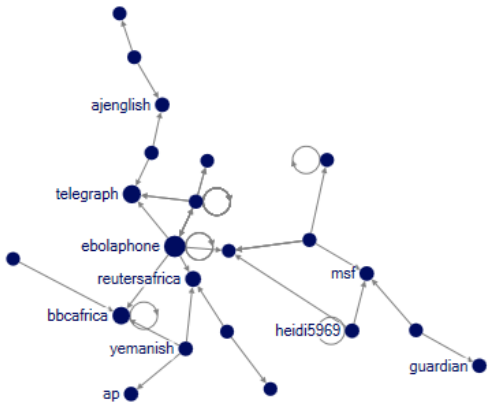


Figura 4.4: Grafo G1 - dez vértices com os maiores valores de centralidade de intermediação.

Tabela 4.5: Usuários com os dez maiores valores de centralidade de proximidade.

Usuário	Centralidade de proximidade	Centralidade de intermediação	Grau de entrada	Grau de saída	Componente
thedouch3	0,759798	13.185,00	97	0	G2
swaggernoswipin	0,500058	1.764,00	0	2	G2
mrweedaddict	0,500058	1.764,00	0	2	G2
kamarprincess	0,436527	232,00	0	2	G2
dakota_mariee	0,436527	232,00	0	2	G2
zaddyez_	0,434889	0,00	1	1	G2
badgal_jade	0,434889	0,00	0	2	G2
wolf_drew7	0,434889	0,00	0	2	G2
karen_txo	0,434889	0,00	1	1	G2
vishwinvlo	0,433368	0,00	0	1	G2

longo de seis meses. As métricas selecionadas foram calculadas para 51 dias distribuídos entre os meses de novembro de 2014 a abril de 2015. O cálculo foi feito em separado para cada dia, representando 28% do total de dias desses seis meses. Esse percentual de 28% não representa todo o universo dos dados postados no Twitter sobre o Ebola nesses seis meses, porém permite analisar a tendência do comportamento dos vértices chaves @ebolaphone, @thedouch3 e @telegraph por meio da variação das métricas em estudo ao longo dos 51 dias.

Analisando as métricas do usuário @ebolaphone neste período, verificou-se que ele está entre os 20 vértices com a maior centralidade de intermediação em quatro dias e entre os 20 vértices com o maior valor de grau de saída em seis dias. Além disso, foram registradas utilização do Twitter deste usuário em 31 dias distribuídos em todos os seis meses de análise, como postagem de comentários (*Tweet*), mencionado por seus seguidores (*Mentions*), repostas a outros usuários (*Reply*), encaminhamento de comentários (*Retweet*) a seus seguidores apresentando valores nas métricas grau de entrada, grau de saída e centralidade

de intermediação. Este fato sugere que o interesse deste usuário em disseminar informações e engajar em discussões não foi limitado a determinados dias, mas permaneceu ao longo de todo o período de análise.

Por outro lado, apesar de registrar o uso do Twitter em 25 dias ao longo dos seis meses, o usuário @telegraph está entre os 20 vértices com a maior centralidade de intermediação e grau de entrada somente no dia 19/11/2014, mostrando valores menos expressivos nos outros dias. Analisando-se o grau de saída nos dias em que a métrica foi calculada, não foi registrado nenhum dia com valor maior que zero. Isto sugere que o usuário @telegraph tem a tendência de postar poucos comentários (*Tweet*) e não responder ou não engajar em discussões com outros usuários. O usuário @thedouch3 está entre os 20 vértices com maior grau de entrada em seis dias e registrou atividades em nove dias ao longo dos seis meses, sugerindo também um perfil de usuário que posta pouco, porém seus comentários são encaminhados (*Retweet*) e o usuário é mencionado (*Mention*) com maior frequência em relação ao usuário @telegraph.

Em seguida, baseado no cálculo das métricas dos 51 dias, foi efetuada uma análise dos usuários como um todo, não sendo limitado aos três usuários chaves selecionados. A Tabela 4.6 exibe a identificação dos usuários que permaneceram por mais tempo entre os 20 vértices com maior centralidade de intermediação durante o período. O usuário @who, que refere-se à Organização Mundial de Saúde, é o que aparece com maior frequência (66,67% ou 34 dias) entre os 20 maiores valores diários. Este percentual representa os 34 dias em relação ao total de 51 dias em que a métrica foi calculada. O usuário @flutematamol aparece em segundo lugar (54,90%), conforme apresentado na Tabela 4.6. Analisando os perfis destes usuários no Twitter que permanceram mais de 49% do tempo entre os 20 maiores valores de centralidade de intermediação, observou-se que usuários representando organizações ficaram em primeiro (@who) e terceiro lugar (@unicef) e usuários individuais ficaram em segundo (@flutematamol) e quarto lugar (@fluffator), não se confirmando o resultado obtido na Tabela 4.1, em que 70% dos usuários representam agências de notícias e organizações.

Os usuários @unicef (56,86%) e @youtube (52,94%) são os que apareceram com maior frequência entre os 20 maiores valores de cada dia da métrica grau de entrada, conforme apresentado na Tabela 4.7. É interessante notar que o usuário chave @thedouch3 está presente nesta tabela confirmando a tendência de ter seus comentários encaminhados (*Retweet*) e de ser mencionado (*Mentions*) por outros usuários com uma quantidade significativa (11,76%) ao longo do período. Analisando o perfil dos usuários no Twitter, verificou-se que os usuários que permaneceram mais de 49% do tempo entre os 20 maiores valores de grau de entrada são contas oficiais de organizações (@unicef e @who),

Tabela 4.6: Usuários que permaneceram por mais tempo com maior centralidade de intermediação.

Usuários	Qtd. de Dias	Percentual
who	34	66,67%
flutematamol	28	54,90%
unicef	27	52,94%
fluffator	25	49,02%
ebola_rt	24	47,06%
ebolaalert	21	41,18%
ebolaoutbreakus	20	39,22%
un, nilimajumder	18	35,29%
nytimes	17	33,33%
unmeer	16	31,37%

sugerindo a relevância das informações postadas sobre o Ebola por estes usuários e o site de mídia digital (@youtube), sugerindo o grande interesse dos usuários em informações sobre o Ebola na forma de vídeo, os quais são compartilhados e encaminhados a outros usuários.

Tabela 4.7: Usuários que permaneceram por mais tempo com maior grau de entrada.

Usuários	Qtd. de Dias	Percentual
unicef	29	56,86%
youtube	27	52,94%
who	25	49,02%
un, fuckingquote_es	16	31,37%
unmeer	15	29,41%
ebolaalert	14	27,45%
bbcafrica e nytimes	12	23,53%
unicefsl, undp	9	17,65%
guardian, oxfam	7	13,73%
conspiracyimage, ebolaoutbreakus, ecorepublicano, fact, louievree, unicef_liberia, tengomalostuits, thedouch3	6	11,76%

Com relação ao grau de saída, os usuários @nilimajumder e @ebolaoutbreakus apareceram com a maior frequência (62,85%) entre os 20 maiores valores de cada dia, conforme apresentado na Tabela 4.8. Nesta tabela é possível verificar a presença do usuário chave @ebolaphone (11,76%), confirmando a tendência do usuário em disseminar informações e engajar em discussões. Analisando o perfil destes usuários no Twittter, verificou-se que todos são usuários individuais, sugerindo que estes costumam encaminhar comentários (*Retweet*), responder aos comentários (*Reply*), mencionar outros usuários (*Mentions*), postar comentários (*Tweet*) e engajar em discussões com outros usuários, ao contrário dos usuários que representam agências de notícias, organizações e jornais.

Tabela 4.8: Usuários que permaneceram por mais tempo com maior grau de saída.

Usuário	Qtd. de Dias	Percentual
ebolaoutbreakus, nilimajumder	32	62,75%
ebola_rt, flutematamol	31	60,78%
fluffator	21	41,18%
ebola_youta, heidi5969	9	17,65%
healthyworld24	8	15,69%
baronianconsult, ebolart, totalngonews	7	13,73%
ebolaphone, ebola_chan666	6	11,76%
beulahkirke, chronik_afrik, jazzyatheart, knityarn	5	9,80%
ebolaalert, infoebola, indiabonita_11, unitedliberians, lateam224, magarya, theee_waviest	4	7,84%
_yung_ebola, bel_grammy, delaney46, drrbai, ebola_2017, ebola_girl, ebola_is_bad, ebola_time, ebolanewsviews, ebolaupdate, healthandcents, jurylady5, mackayim, operationafrica, raggakaas, robinsnewswire, shawnstormmuzic, welapmsimanga, widescopecsr	3	5,88%

4.1.2 Resumo da Análise

A análise dos dados coletados no dia 19/11/2014 permitiu a identificação de três vértices chaves: @ebolaphone, @thedouch3 e @telegraph. 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 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 dos 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 pre-

sença 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

Esta seção apresenta uma análise dos dados por meio de visualização geográfica. Primeiramente, os dados divulgados pela Organização Mundial de Saúde sobre os casos de infecção e mortes por Ebola referentes ao último mês do período de análise, abril de 2015, foram mostrados no mapa. Em seguida, foi efetuada uma análise geográfica dos comentários do Twitter para os meses de novembro de 2014 a abril de 2015. 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.

4.2.1 Análise

A ferramenta QGIS [14] foi utilizada para criar o mapa apresentado na Figura 4.5. Este mapa do mundo destaca em preto o contorno dos países onde foram registrados casos de Ebola de acordo com os dados divulgados pela Organização Mundial de Saúde [10]. Até o dia 26/04/2015, foram registrados 26.312 casos de Ebola e 10.899 mortes por Ebola. O mapa permite observar que existem ocorrências isoladas de casos de Ebola nos Estados Unidos, Reino Unido, Espanha e transmissão intensa da doença na África Ocidental.

A Figura 4.6 permite observar os detalhes dos países da África Ocidental: Serra Leoa é o país com o maior número de casos de Ebola, com 12.371 (47,02%) ocorrências, e a Libéria é o país com o maior número de mortes por Ebola, com 4.608 (42,28%) ocorrências. Neste mapa, a graduação de cores representa a quantidade de casos de Ebola por país

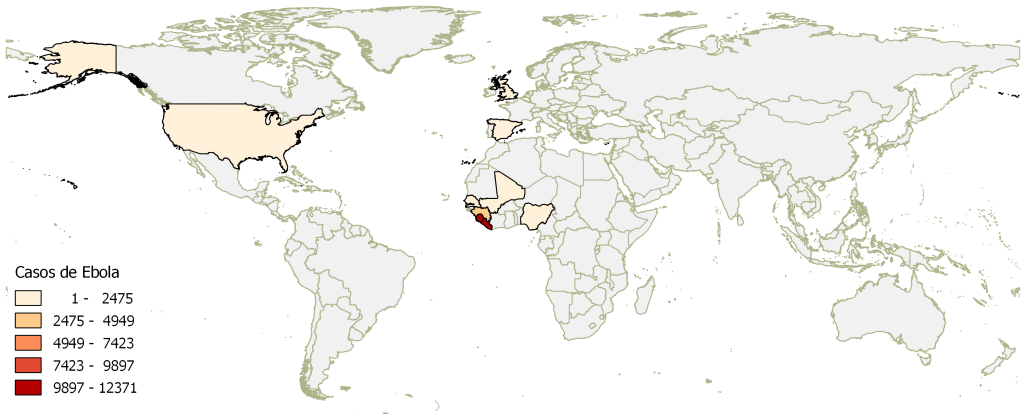


Figura 4.5: Mapa do mundo destacando em preto o contorno dos países onde foram registrados casos de Ebola até 26/04/2015.

e o número ao lado do nome do país representa a quantidade de mortes por Ebola. A Tabela 4.9 mostra os dados oficiais sobre a epidemia do Ebola divulgados pela Organização Mundial de Saúde em 26/04/2015.

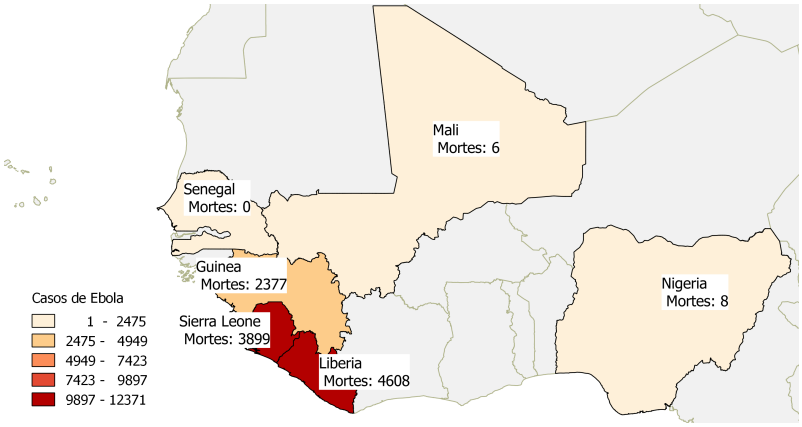


Figura 4.6: Mapa da África Ocidental onde foram registrados casos de Ebola e mortes por Ebola até 26/04/2015.

A fim de possibilitar a análise espacial sobre os dados do Twitter, o algoritmo de identificação de país desenvolvido nesta pesquisa foi executado para os dados referentes a 51 dias distribuídos entre os meses de novembro de 2014 a abril de 2015. Estas datas são as mesmas datas para as quais as métricas centralidade de intermediação, centralidade de proximidade, grau de entrada e grau de saída foram calculadas. Após a execução do algoritmo, verificou-se que 62,28% dos usuários informaram o campo ”Location”, e, dentre estes, foi possível identificar o país de origem de 60,73% dos usuários. Desta forma, esta análise geográfica permite visualizar uma amostra dos dados postados no Twitter ao longo dos seis meses cuja extração da localização foi possível por meio do algoritmo de identificação de país.

As Tabelas 4.10 e 4.11 mostram o percentual de comentários postados no Twitter por

Tabela 4.9: Casos registrados de Ebola e mortes por Ebola até 26/04/2015.

Continente	País	Casos de Ebola	Percentual	Mortes por Ebola	Percentual
África	Serra Leoa	12.371	47,02%	3.899	35,77%
África	Libéria	10.322	39,23%	4.608	42,28%
África	Guiné	35.84	13,62%	2.377	21,81%
África	Nigéria	20	0,08%	8	0,07%
África	Mali	8	0,03%	6	0,06%
África	Senegal	1	0,00%	0	0,00%
Europa	Espanha	1	0,00%	0	0,00%
Europa	Reino Unido	1	0,00%	0	0,00%
América do Norte	Estados Unidos	4	0,02%	1	0,01%
Total		26.312		10.899	

continente referentes, respectivamente, aos meses de novembro de 2014 a janeiro de 2015, e 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* (33,43% a 47,78%), seguida da Europa (16,45% a 26,39%). A África, onde concentra-se os casos de Ebola, está em terceiro lugar (10,14% a 14,79%). A Tabela A.17, do Apêndice A, mostra os 90 países com as maiores quantidades de comentários postados no período de novembro de 2014 a abril de 2015. A coluna "Perc." exibe o percentual de comentários em relação ao total de 212.690 comentários coletados no período.

Tabela 4.10: 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	3.537	14,79%	5.022	11,83%	4.550	11,77%
América do Norte	8.234	34,43%	19.082	44,94%	18.396	47,59%
América do Sul	1.982	8,29%	4.672	11,00%	4305	11,14%
Antártica	1	0,00%	0	0,00%	4	0,01%
Ásia	3.458	14,64%	4.268	10,05%	4.135	10,70%
Europa	5.893	24,64%	8.770	20,66%	6.358	16,45%
Oceania	809	3,38%	643	1,51%	910	2,35%
Total	23.914		42.457		38.658	

A seguir, 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 4.12 mostra os vértices chaves geolocalizados com os 20 maiores valores de grau de entrada referente ao mês de novembro de 2014. A coluna "Location" refere-se ao local informado manualmente pelo usuário em seu perfil no Twitter, a coluna "País" refere-se ao país obtido pelo algoritmo de identi-

Tabela 4.11: 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	3.744	12,28%	3.815	10,14%	5.780	14,60%
América do Norte	13.627	44,71%	17.966	47,78%	15.690	39,65%
América do Sul	3.715	12,19%	3.766	10,01%	3.555	8,98%
Antártica	3	0,01%	0	0,00%	1	0,00%
Ásia	3.402	11,16%	3.676	9,78%	3.482	8,80%
Europa	5.077	16,66%	7.585	20,17%	10.446	26,39%
Oceania	912	2,99%	797	2,12%	622	1,57%
Total	30.480		37.605		39.576	

ficação de país e a coluna "Tipo de local" refere-se ao tipo de local utilizado pelo algoritmo de identificação de país para extrair o país a partir do campo "Location". Observa-se que o usuário @nytimes aparece duas vezes na tabela por ter apresentado os maiores valores de grau de entrada nos dias primeiro e oito de novembro. É possível verificar que o usuário @billgates dos Estados Unidos apresenta o maior valor de grau de entrada (630) sugerindo um usuário que posta comentários relevantes que são encaminhados por seus seguidores e são mencionados por outros usuários. A Tabela 4.13 mostra alguns exemplos de comentários deste usuário. A Tabela 4.14 mostra os vértices chaves geolocalizados com os 20 maiores valores de grau de saída referente ao mês de novembro de 2014. Observa-se a presença de oito usuários do Brasil. O conteúdo dos comentários destes usuários referem-se a rumores ou notícias falsas sobre o Ebola. Nota-se também a presença do usuário @tz_uchay da Nigéria, sugerindo o interesse deste usuário em postar ou responder comentários sobre a doença, apesar da baixa quantidade de comentários na África como um todo (14,79%). A Tabela 4.15 mostra os 20 maiores valores da métrica centralidade de intermediação. Esta tabela possibilita verificar novamente a presença do usuário @billgates, desta vez como um vértice que facilita a disseminação de informações entre grupos diferentes com o maior valor de centralidade de intermediação (4.757.055,81). É possível verificar também a presença de dois usuários da África: @bodacious_lyn da Zâmbia e @benaskay de Gana, sugerindo usuários que auxiliam a disseminar informações.

Foram gerados seis mapas para representar os comentários postados no Twitter em cada mês, referentes a novembro de 2014 até abril de 2015. No mesmo mapa, 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 4.7 mostra os mapas referentes aos meses de novembro de 2014 a janeiro de

Tabela 4.12: 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	Falestīn / 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

2015. A graduação de cores nos países representa a quantidade de comentários do Twitter e o mapa de calor representa a presença de vértices chaves geolocalizados. O mapa de calor é formado por círculos cuja graduação de cores é calculada por meio de um atributo numérico previamente selecionado e a intensidade da cor representa o "calor" deste atributo [66]. No caso dos mapas em análise, a graduação de cor laranja dos círculos são calculados utilizando-se a quantidade de vértices geolocalizados como atributo numérico. Analisando-se o mapa do mês de novembro de 2014, observa-se que os Estados Unidos concentram a maior quantidade de *Tweets* (25,05%) e também apresentam a maior quantidade de vértices chaves geolocalizados. No entanto, nos países em que o uso do Twitter foi menor, como Reino Unido (8,51%), Brasil (3,40%) e Nigéria (5,17%), também pode-se encontrar vértices chaves geolocalizados, sugerindo que o uso do Twitter e a disseminação de informações sobre o Ebola abrangeu vários continentes neste mês. A Tabela 4.16 mostra os dez países com as maiores quantidades de *Tweets* no mês de novembro de 2014. A coluna "Percentual" exibe os valores referentes ao total de 23.914 *Tweets* coletados no mesmo mês.

Tabela 4.13: Exemplos de *Tweets* do usuário @billgates encaminhados e mencionados por outros usuários.

Tweets
@BillGates: The next epidemic could be 1,000x worse than Ebola. Here's what I hope we do : http://t.co/5UljXbr9z3
@BillGates: Watch this short, informative video on why Ebola spread in West Africa: http://t.co/DHAevfNhM3
@BillGates: What did we learn from Ebola? We need to be better WHEN the next epidemic hits
$5.7 million pledged for Ebola blood plasma trials by @gatesfoundation @BillGates @melindagates
@BillGates: The Ebola epidemic is tragic. But the next epidemic could be devastating.
@BillGates Help make it happen for African Ladies NYC Running to Support Ebola Orphans!

Os mapas dos meses de dezembro de 2014 e janeiro de 2015, apresentados na Figura 4.7, e os mapas dos meses de fevereiro de 2015 a abril de 2015, apresentados na Figura 4.8, mostram também a presença de vértices chaves geolocalizados nos países onde foram registrados casos isolados de Ebola como Estados Unidos, Inglaterra e Espanha, além dos países com transmissão intensa do vírus como Serra Leoa, Libéria e Nigéria. As tabelas contendo os 20 maiores valores das métricas grau de entrada, grau de saída e centralidade de intermediação referentes aos meses de dezembro de 2014 a abril de 2015 estão descritas no Apêndice A. Entre os vértices chaves selecionados na Seção 4.1, @ebolaphone, cujo usuário é de origem dos Países Baixos, aparece como vértice chave geolocalizado nos meses de novembro de 2014, março e abril de 2015. O usuário @telegraph do Reino Unido aparece como vértice chave geolocalizado somente em novembro de 2014, quando o jornal The Telegraph publicou um artigo sobre a não participação da cantora inglesa Adele na campanha de ajuda ao Ebola, o que desencadeou comentários do tipo *Retweet* mencionando o usuário @telegraph. O usuário @thedouch3 não aparece em nenhum mês como vértice chave geolocalizado por não ter sido possível extrair sua localização geográfica através dos critérios definidos no algoritmo de localização de país. Da mesma forma, não foi possível determinar o país de origem de diversos outros usuários que permaneceram por vários dias entre os maiores valores das métricas analisadas na Seção 4.1, detalhados nas Tabelas 4.6, 4.8 e 4.7, como por exemplo, @unicef, @flutematamol e @ebolaoutbreakus. Apesar disso, a análise espacial descrita nesta seção sugere que, a partir da amostra de dados extraídos do Twitter sobre o termo "Ebola" durante seis meses, e sobre os quais foi possível identificar o país de origem, a América do Norte possui os maiores percentuais de comentários (33,43% a 47,78%). No entanto, ao analisar a presença de vértices chaves em cada país, observou-se que estes estão distribuídos em vá-

Tabela 4.14: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - novembro/2014.

Usuário	País	Grau de saída	Location	Tipo de local
andreinanews	Colômbia	34	Colombia	Nome de país
ebolaphone	Países Baixos	33	Utrecht, The Netherlands	Nome de cidade
recitouciumes	Brasil	23	Brasil	Nome de país
relatoamor	Brasil	23	Brasil	Nome de país
purasmaidades	Brasil	23	Brasil	Nome de país
fatoswpp	Brasil	23	Brasil	Nome de país
docefatos	Brasil	23	Brasil	Nome de país
indiscretoamor	Brasil	23	Brasil	Nome de país
cortesvingativo	Brasil	23	Santa Maria/RS	Nome de cidade
humidiota	Brasil	23	Brasil	Nome de país
blackphysicists	Estados Unidos	19	Arlington, VA USA	Nome de cidade
followingebola	Reino Unido	17	London, UK	Nome de capital
totalngonews	Reino Unido	13	london	Nome de capital
followingebola	Reino Unido	13	London, UK	Nome de capital
robinsnewswire	Estados Unidos	12	Bothell, WA	Nome de cidade
lokamark	Itália	10	Palermo, Italy	Nome de país
tz_uchay	Nigéria	10	Lagos	Nome de cidade
racecarodds	Estados Unidos	10	Las Vegas	Nome de cidade
robinsnewswire	Estados Unidos	9	Bothell, WA	Nome de cidade
dineitemponews	Brasil	9	Brasil	Nome de país

rios continentes, inclusive na África, onde a quantidade de comentários é menor (10,14% a 14,79%). Este fato sugere o interesse dos usuários do Twitter ao redor do mundo em comentar e disseminar informações sobre o avanço da epidemia ao longo dos seis meses.

4.2.2 Resumo da Análise

A análise espacial no mapa dos dados divulgados pela Organização Mundial de Saúde 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 (33,43% a 47,78%), seguida da Europa (16,45% a 26,39%). A África, onde concentra-se os casos de Ebola, está em terceiro lugar (10,14% a 14,79%).

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

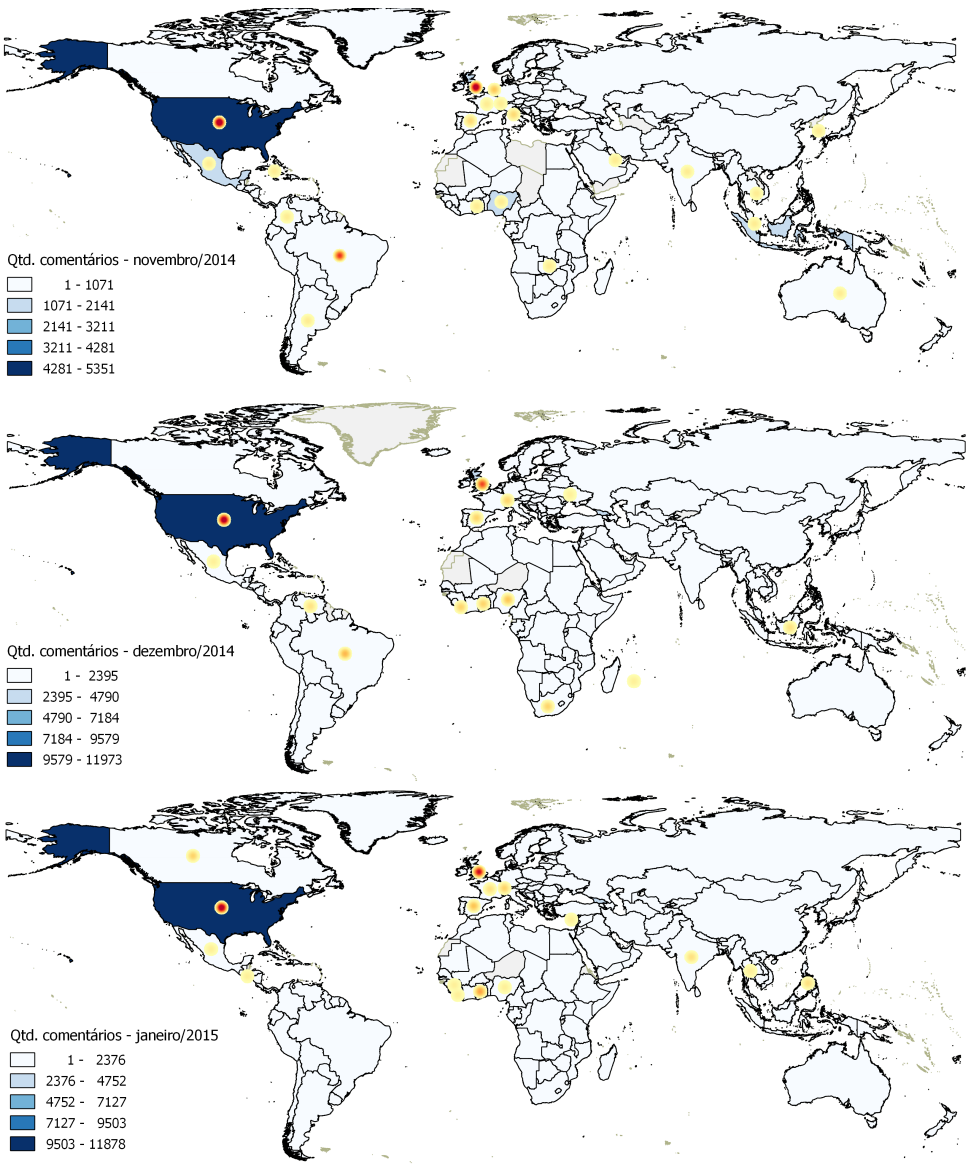


Figura 4.7: Quantidade de *Tweets* por país e vértices chaves geolocalizados dos meses de novembro/2014, dezembro/2014 e janeiro/2015.

em relação à América do Norte e Europa, verificou-se a presença de três vértices geolocalizados da África no mês de novembro de 2014: @tz_uchay da Nigéria, @bodacious_lyn da Zâmbia e @benaskay de Gana.

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.

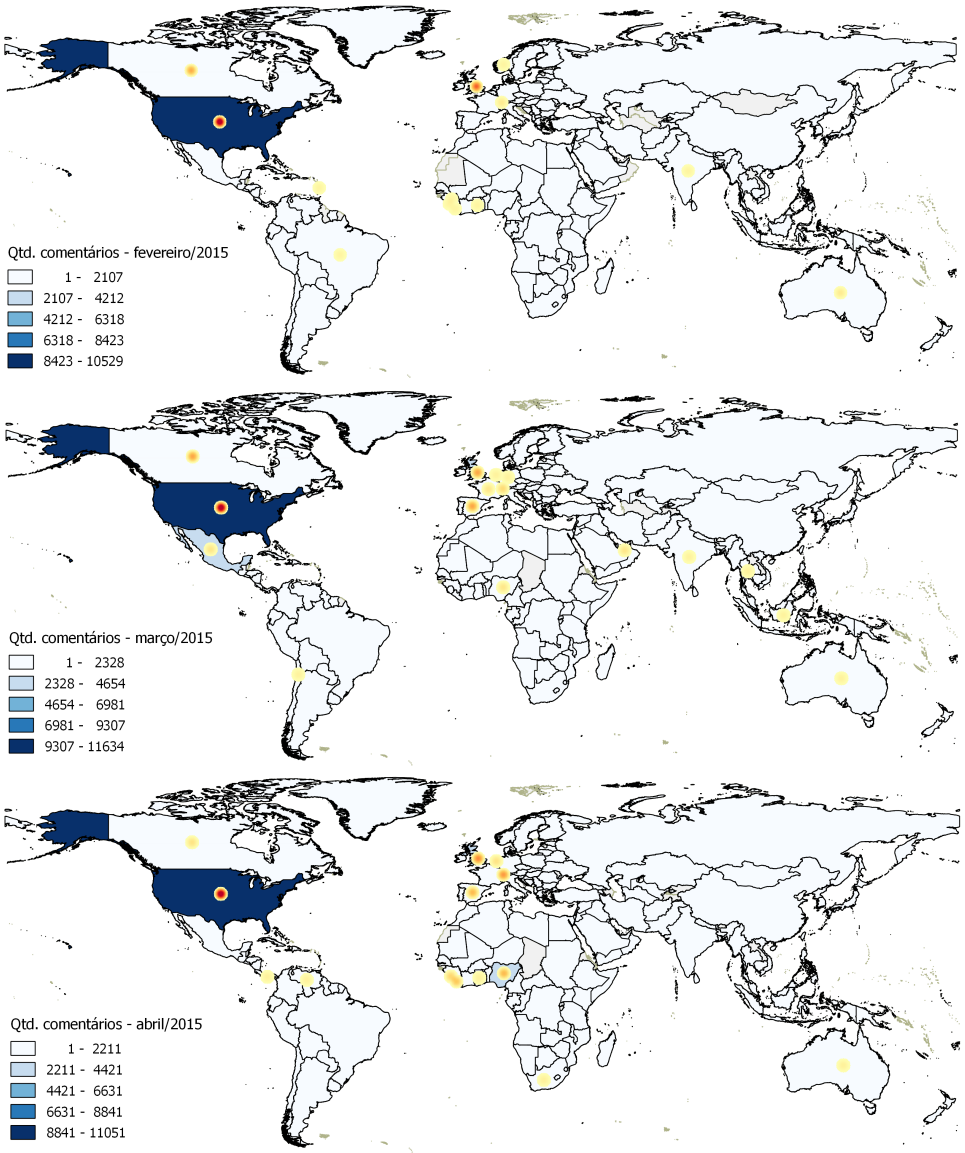


Figura 4.8: Quantidade de *Tweets* por país e vértices chaves geolocalizados dos meses de fevereiro, março e abril/2015.

4.3 Análise Textual

Esta seção apresenta uma análise textual dos comentários do Twitter sobre o Ebola. 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 apresentado na Seção 3.3.3. 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. Os 2.452 comentários referentes aos quatro dias de coleta de *Tweets* da África não são suficientes para efetuar uma análise profunda sobre o conteúdo postado. Porém, devido às dificuldades envolvidas no tratamento dos dados como, por exemplo, a grande quantidade de URL encontradas nos comentários para a identificação de *Tweets* não-pessoais, a variedade de palavras que foram identificadas a cada nova carga de dados e acrescentadas ao corpus de *Tweets* não-pessoais, estes 2.452 comentários foram utilizados como uma amostra para efetuar um estudo inicial sobre a tendência do conteúdo postado. 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.

4.3.1 Análise

A Tabela 4.17 mostra o resultado da classificação dos comentários da África ao longo dos quatro dias em que a análise textual foi executada. Nota-se que o percentual de *Tweets* pessoais nos dias 08/11/2014 e 19/11/2014, que correspondem ao período inicial da epidemia, variou entre 42,78% e 44,69% e diminuiu ao longo do tempo, registrando 27,29% no dia 23/12/2014 e 25,04% no dia 30/04/2015. Por outro lado, o percentual de *Tweets* não-pessoais nos dias 08/11/2014 e 19/11/2014 variou entre 50,13% e 47,23% e aumentou ao longo do tempo, registrando 67,51% no dia 23/12/2014 e 70,77% no dia 30/04/2015. 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. É possível verificar também que o percentual de *Tweets* pessoais foi maior no dia 19/11/2014 (44,69%) em relação aos outros três dias, e uso de *emoticons* e *emojis* também acompanhou esse aumento, registrando 1,64% no dia 19/11/2014, comparativamente aos outros dias em que foram inferiores à 1%.

O gráfico da Figura 4.9 permite visualizar a variação dos *Tweets* não-pessoais durante o período de análise. Observa-se a 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 Tabela 4.18 mostra exemplos de *Tweets* não-pessoais referentes aos quatro dias em que a análise textual foi executada.

A Tabela 4.19 mostra os resultados referentes a *Tweets* pessoais, detalhando a variação

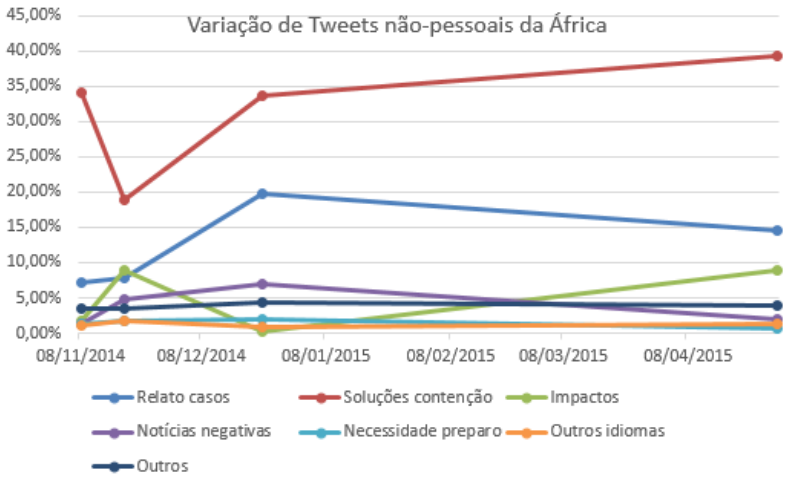


Figura 4.9: Gráfico referente à variação dos *Tweets* não-pessoais da África.

dos *Tweets* pessoais ao longo dos quatro dias de análise. Observa-se que o percentual de *Tweets* pessoais com sentimento negativo diminuiu de 17,59%, no dia 08/11/2014, para 8,83%, no dia 30/04/2015, porém manteve-se superior ao percentual de *Tweets* com sentimento positivo, exceto no dia 23/12/2014.

A seguir, foi efetuada uma análise da disseminação dos comentários dos vértices chaves @ebolaphone, @thedouch3 e @telegraph sobre os usuários na África. Os critérios de escolha destes vértices chaves estão descritos na Seção 4.1. O país de origem do usuário @ebolaphone são os Países Baixos e foi o vértice com o maior grau de saída e centralidade de intermediação do dia 19/11/2014. Seus comentários foram mencionados por oito usuários da África e o usuário @ebolaphone encaminhou *(Retweet)* o comentário de três usuários da África para seus seguidores. O conteúdo dos comentários disseminados pelo @ebolaphone são relacionados a divulgação de mapa dos países afetados na África, ações de contenção da epidemia, sugerindo o interesse em disseminar informações de esclarecimento sobre a epidemia. A Tabela 4.20 mostra exemplos de comentários disseminados pelo @ebolaphone. Não há registros de comentários do vértice chave @thedouch3 disseminados entre os usuários da África nos dados coletados no período de análise. Apenas um comentário do usuário @telegraph foi encaminhado por um usuário da África, classificado como outros por não ter sido possível identificar a relação do seu conteúdo com a epidemia.

Em seguida, foram analisados os comentários dos vértices chaves geolocalizados da África, descritos na Seção 4.1. O vértice @ebolaalaert é um usuário da Nigéria que aparece entre os vértices com 20 maiores valores de centralidade de intermediação nos meses de dezembro/2014, março/2015 e abril/2015, permanecendo por 21 dias ou 41,18% dos 51 dias em que a métrica foi calculada entre os 20 maiores valores desta métrica. Os dez

comentários do usuário @ebolaalert são relacionados a relatos da epidemia, solução de contenção e necessidade de preparo veiculado por agências de notícias. Seus comentários foram encaminhados por 26 usuários de diversos países, como Reino Unido, Venezuela, Nigéira e Índia, sugerindo o interesse destes usuários em disseminar aos seus seguidores os comentários do usuário @ebolaalert. A Tabela 4.20 mostra exemplos de comentários disseminados pelo @ebolaalert. O usuário @unicef_liberia, cujo país de origem é a Libéria, um dos três países da África com transmissão intensa do Ebola, aparece entre os vértices com 20 maiores valores de grau de entrada nos meses de dezembro/2014, janeiro/2015 e fevereiro/2015, permanecendo por seis dias ou 11,76% dos 51 dias em que a métrica foi calculada entre os 20 maiores valores desta métrica. Este fato indica que os comentários disseminados por este usuário é mencionado diversas vezes durante o período de análise. Usuários dos países afetados pela epidemia como, Nigéria, Serra Leoa, Libéria, Estados Unidos, Reino Unido, Espanha e usuários de outros 21 países sem registros da epidemia como, Singapura, Índia, México, Tailândia e Brasil encaminharam o comentário do usuário @unicef_liberia, sugerindo a relevância do seu conteúdo, cujo exemplo pode ser encontrado na Tabela 4.20. Os outros vértices geolocalizados da África @unmeer e @lateam224 seguiram a mesma tendência de disseminar informações sobre relatos de casos e solução de contenção da epidemia. A Tabela 4.20 mostra exemplos de comentários em que os vértices chaves e vértices chaves geolocalizados da África participaram da disseminação da informação, sendo mencionado *(Mention)*, enviando comentários *(Tweet)* ou encaminhando comentários *(Retweet)*.

A análise dos comentários disseminados na África, pelos vértices chaves de outros continentes e os vértices chaves geolocalizados da África, sugere o interesse dos usuários da África e de outros continentes menos afetados pela epidemia em compartilhar aos seus seguidores, informações sobre o avanço da epidemia e as medidas de contenção que foram tomadas ao longo do período de análise.

4.3.2 Resumo da Análise

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 dos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015.

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. Alguns exemplos de comentários de @ebolaphone podem ser visualizados na Tabela 4.20.

A análise mostrou também que os vértices chaves geolocalizados da África: @ebola-alert da Nigéria, @unicef_liberia da Libéria, @unmeer de Gana e @lateam224 da Guiné participaram da disseminação de informações sobre a epidemia para usuários de outros continentes, sugerindo a relevância do conteúdo disseminado.

4.4 Resultados

A análise temporal dos dados de 51 dias por meio da Teoria de Grafos permitiu a identificação de vértices que permaneceram por mais de 27 dias entre os 20 maiores valores das métricas analisadas. Entre os usuários que permaneceram por mais tempo com maior centralidade de intermediação estão @who e @unicef, mostrados na Tabela 4.6, sugerindo usuários que facilitam a disseminação de informações, atuando como intermediários entre grupos diferentes. @unicef e @youtube estão entre os usuários que permaneceram por mais tempo com maior grau de entrada, mostrados na Tabela 4.7, sugerindo usuários que postam comentários que são encaminhados por outros usuários. Entre os usuários que permaneceram por mais tempo com maior valor de grau de saída, estão @ebolaoutbreakus e @nilimajumder, mostrados na Tabela 4.8, sugerindo usuários que postam comentários e respondem aos comentários de outros usuários. A presença destes vértices atuando por mais de 27 dias ao longo dos seis meses de análise sugerem que, na epidemia do Ebola, as redes sociais formadas no Twitter auxiliam na disseminação de informações.

A análise do perfil dos usuários no Twitter a partir da identificação dos vértices chaves sugere que, quando o tema do comentário é sobre epidemia, os usuários do Twitter costumam encaminhar com maior frequência comentários postados por agências de notícias e organizações. Outro fato observado é que usuários individuais costumam postar comentários com maior frequência, em relação aos usuários que representam agências de notícias ou organizações.

A análise espacial permitiu visualizar por meio de mapas os países onde foram registrados comentários no Twitter sobre a epidemia do Ebola. Verificou-se que a concentração de vértices chaves geolocalizados e comentários do Twitter é maior na América do Norte (33,43% a 47,78%), porém os vértices chaves geolocalizados também estão presentes em outros continentes, inclusive na África, onde o percentual de comentários é menor (10,14% a 14,79%). Este fato sugere que, pela importância do tema, o interesse pelas informações disseminadas no Twitter sobre o Ebola não está restrito aos Estados Unidos, onde concentram-se o maior percentual de comentários (32,84%). Atinge também países da África que, apesar de apresentar um percentual menor de comentários, observou-se a presença de vértices geolocalizados que auxiliam na disseminação de informação na Nigégia (@ebolaalert), Libéria (@unicef_liberia), Guiné(@lateam224) e Serra Leoa(@unicefsl), cujos detalhes estão nas Tabelas A.1, A.2, A.3, A.4, A.5, A.7, A.8, A.10, A.11, A.12, A.13 e Tabela A.15, do Apêndice A.

A análise textual supervisionada dos comentários dos usuários da África mostrou que 51,06% das informações disseminadas são relevantes, sendo 13,09% sobre relatos de casos, 31,28% sobre as soluções de contenção da epidemia, 5,26% sobre os relatos de impactos sociais e econômicos e 1,43% sobre as necessidade de preparo, conforme mostrado na Tabela 3.13.

A análise textual temporal dos comentários dos usuários da África ao longo de quatro dias mostrou que o percentual de *Tweets* pessoais diminuiu de 42,78% para 25,04% e o percentual de *Tweets* não-pessoais aumentou de 50,13% para 70,77%. Estes números sugerem que, no mês de novembro quando se registrou cerca de 1000 casos novos de Ebola por semana, os usuários manifestavam opiniões pessoais com mais frequência. À medida em que os meses foram passando, apesar do avanço menor, porém, constante do vírus no continente observado até abril de 2015, os usuários passaram a disseminar com mais frequência noticiários sobre a epidemia.

A análise de *Tweets* não-pessoais permitiu observar uma tendência de aumento do percentual de comentários sobre soluções de contenção da epidemia e uma diminuição do percentual de comentários sobre relatos de casos sobre a epidemia ao longo dos quatro

dias. Este fato sugere que, à medida que aumentam as medidas de controle da epidemia efetuada pelas autoridades e a diminuição de novos casos no período, observa-se um aumento na disseminação de comentários sobre campanhas de ajuda no combate à epidemia, soluções de isolamento de pacientes e pesquisas de vacinas.

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, porém 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 comentários dos vértices chaves geolocalizados da África, como @ebolaalert, @unicef_liberia e @unmeer, mostrou o interesse dos usuários da África e de outros continentes em informações relacionadas a relatos da epidemia, solução de contenção e necessidade de preparo.

Tabela 4.15: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - novembro/2014.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
billgates	Estados Unidos	4.757.055,817	Seattle, WA	Nome de cidade
kokii_r	Camboja	2.298.972,724	Phnom Penh, Cambodia	Nome de país
i_me_adamxd	Singapura	1.081.428,000	Singapore	Nome de país
bodacious_lyn	Zâmbia	1.037.784,958	Lusaka	Nome de capital
nytimes	Estados Unidos	824.677,394	New York City	Nome de cidade
bbcbreaking	Reino Unido	735.315,482	London, UK	Nome de capital
ebolaphone	Países Baixos	526.821,090	Utrecht, The Netherlands	Nome de cidade
nytimes	Estados Unidos	403.287,410	New York City	Nome de cidade
who	Suíça	368.468,118	Geneva, Switzerland	Nome de país
nytimes	Estados Unidos	346.439,754	New York City	Nome de cidade
qualitick	Reino Unido	338.376,617	Chester, UK	Nome de cidade
thomasfarquhar	Reino Unido	338.376,617	UK	Nome alternativo manual
telegraph	Reino Unido	317.765,798	London, UK	Nome de capital
benaskay	Gana	305.963,381	Accra	Nome de capital
thompsonsaskia	Reino Unido	300.900,000	Banbury, England	Nome de cidade
marielammenciso	México	293.936,589	Cancun,Mexico	Nome de país
konakingohana	Estados Unidos	271.140,000	California	Nome de cidade
vjmahon	Austrália	260.916,000	Aireys Inlet, Australia	Nome de país
vidyakrishnan	Índia	249.137,496	Bhopal–New Delhi	Nome de capital
vicente56juan	Espanha	240.828,000	Elche	Nome de cidade

Tabela 4.16: Dez países com as maiores quantidades de *Tweets* em novembro de 2014.

País	*Tweets*	Percentual
Estados Unidos	5.990	25,05%
Reino Unido	2.035	8,51%
México	1.302	5,44%
Nigéria	1.237	5,17%
Indonésia	1.219	5,10%
Espanha	1.021	4,27%
Brasil	814	3,40%
Austrália	710	2,97%
Índia	678	2,84%
África do Sul	621	2,60%

Tabela 4.17: Resultado da classificação dos comentários da África referentes aos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015.

Tipo 1	*Tweets* não-pessoais	08/11/2014	19/11/2014	23/12/2014	30/04/2015	Total
0	Relato de casos	7,09%	7,92%	19,69%	14,62%	16,76%
1	Solução de contenção	34,12%	18,39%	33,61%	39,36%	27,49%
2	Relato de impactos	1,84%	8,82%	0,14%	8,97%	4,28%
3	Notícias negativas	1,31%	4,78%	6,89%	1,88%	4,04%
4	Necessidade de preparo	1,31%	1,64%	1,97%	0,72%	1,43%
5	Outros idiomas	1,05%	1,79%	0,84%	1,30%	1,26%
6	Outros	3,41%	3,44%	4,36%	3,91%	3,83%
	Subtotal	50,13%	47,23%	67,51%	70,77%	59,09%
Tipo 2	*Tweets* pessoais					
0	Identificação de *emoticons* e *emojis*	0,79%	1,64%	0,98%	0,72%	1,06%
1	Análise de sentimentos	41,99%	43,05%	26,30%	24,31%	33,85%
	Subtotal	42,78%	44,69%	27,29%	25,04%	34,91%
Tipo 3	Outros					
0	Outros idiomas	4,20%	1,94%	3,80%	2,32%	2,94%
1	Outros	2,89%	6,13%	1,41%	1,88%	3,06%
	Subtotal	7,09%	8,07%	5,20%	4,20%	6,00%

Tabela 4.18: Exemplos de *Tweets* não-pessoais da África.

Tweets	Tipo
Ebola: 2 Nigerians Test Positive In Sierra Leone. RT @ReutersAfrica: Seventh SierraLeone doctor killed by Ebola: source. RT @ReutersAfrica: Death toll from Ebola in West Africa rises to 7,518 - WHO.	Relato de casos
African Ebola crisis fund is set up: Top African business leaders establish an emergency. RT @BBCAfrica: The Bill and Melinda Gates Foundation pledges $5.7m towards experimental Ebola treatments. Ebola 0utbreak 2014 : How the Text Message is Fighting the Deadly Disease. EBOLA OUTBREAK: Nigeria To Test Vaccines Next February. RT @SkyNews: Ebola Vaccine Declared Safe For Use In Africa. MSF outreach team in SierraLeone tries to track the contact of Ebola. Guinea: Vaccination Teams Defeat 'Ebola Effect' in Guinea.	Solução de contenção
The @AfDB_Group is critically looking at the social-econ implications of ebola. Education falls prey to Ebola in Sierra Leone.	Relato de impactos
RT @Health24com: An Ebola scare has struck one of the world's most densely populated cities. POLITICS: @IMFNews Policies Blamed for Ebola Spread in West Africa. RT @UN: Ebola is not just a health crisis. It is a humandevelopement crisis.	Notítias negativas
RT @nytimes: How has the food supply been affected by the Ebola outbreak? Need greater recovery efforts in Ebola-hit countries: UN. We need more trained Liberians returning to help turn around various sectors in Liberia.	Necessidade de preparo

Tabela 4.19: Resultado da classificação de *Tweets* pessoais da África referentes aos dias 08/11/2014, 19/11/2014, 23/12/2014 e 30/04/2015.

Tweets pessoais	08/11/2014	19/11/2014	23/12/2014	30/04/2015
Sentimento positivo	4,99%	11,66%	9,70%	4,49%
Sentimento neutro	20,21%	18,68%	8,02%	11,72%
Sentimento negativo	17,59%	14,35%	9,56%	8,83%
Total	42,78%	44,69%	27,29%	25,04%

Tabela 4.20: Exemplos de comentários dos vértices chaves e vértices chaves geolocalizados da África.

Tweets	Usuário
RT @UNMEER: Businesses are invited to sign the @GlobalCompact Action Pledge on Ebola Elimination.	ebolaphone
RT @CapeTalk567: Ebola fears have led to @EbolaPhone creating a map to illustrate which areas in Africa affected.	ebolaphone
Good info RT @davidllo: Ebola and a map of Africa. Courtesy of @EbolaPhone @JuryReporter @jillpruetz	ebolaphone
RT @CapeTalk567: RT @EbolaPhone: Wed Nov 19—Unchanged: Multiple Ebola Cases in Mali.	ebolaphone
WSJ U.S. to Complete Ebola Treatment Units in Liberia by End of December.	ebolaalert
NYTIMES C.D.C. Head Says Fight on Ebola Will Be Long.	ebolaalert
BBC Ebola vaccine 'promising in African populations'.	ebolaalert
WORLDBANK World Bank Supports The Gambia's Ebola Prevention Plan.	ebolaalert
RT @UNICEF: No place is too far. We're raising awareness about Ebola in remote parts of Liberia via @UNICEF_Liberia.	unicef_liberia
A member of the 165member medical team Cuba sent to fight Ebola in Sierra Leone has been diagnosed w/ the disease.	unmeer
@EuropeanUnion officials said tht thousands more health professionals like Switala are needed to eradicate the Ebola epidemic. @MSF @UNMEER	unmeer
RT @UNDP: Autoclaves, 1st in SierraLeone decontaminate equipment in Ebolaresponse	unmeer
RT @DrFriedenCDC: Ebola continues to spread in Sierra Leone and in Guinea's capitol city. No time to relax our grip on the response.	lateam224
RT @France24_en: IMF accused of undermining Ebola response in West Africa.	lateam224

5. Conclusão

Este capítulo está dividido em três seções. A Seção 5.1 apresenta as contribuições desta pesquisa. A Seção 5.2 apresenta algumas discussões relacionadas aos critérios adotados e aos resultados obtidos nesta pesquisa. A Seção 5.3 relaciona os possíveis trabalhos futuros.

5.1 Contribuições

O presente 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. Mais especificamente, o principal objetivo da pesquisa foi o de verificar o uso do Twitter em situações de calamidade e saúde pública, como a epidemia do Ebola. Como principal contribuição, apresentou-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 re-

sultados apresentados por cada uma delas separadamente, bem como descartar resultados que não se mostraram claros ou mesmo foram conflitantes entre as três etapas. Entre os resultados, observou-se que o Twitter foi utilizado como meio de comunicação por usuários de diversos países. Identificou-se vértices chaves que auxiliaram na disseminação de informações relevantes como relatos de novos casos, soluções de contenção da epidemia e necessidade de preparo para conter o avanço da epidemia. Um fato interessante a destacar é que, apesar do uso do Twitter ser bem menor na África em relação à América do Norte e à Europa, identificou-se, dentre os vértices chaves, a presença de usuários de países com transmissão intensa do Ebola como Libéria, Guiné e Serra Leoa cujos comentários foram disseminados para outros continentes. A análise do perfil dos usuários no Twitter, a partir dos vértices chaves, mostrou que as pessoas costumam encaminhar com maior frequência comentários postados por agências de notícias e organizações de saúde sugerindo o potencial do Twitter como uma ferramenta de comunicação.

Para viabilizar tal abordagem combinada, foram necessários o estudo de diversas ferramentas disponíveis na literatura, a adequação destas ferramentas para utilização no contexto da pesquisa e o desenvolvimento de novas ferramentas. O roteiro detalhado de instalação destas ferramentas está descrito no Apêndice B.

Além disso, é possível destacar duas contribuições técnicas desta pesquisa: o algoritmo automatizado de identificação de país e o corpus de categorização de comentários do Twitter para situações de epidemia e saúde pública.

A primeira contribuição técnica refere-se ao algoritmo de identificação de país. O algoritmo foi codificado na linguagem PL/pgSQL e permite a identificação do país dos usuários do Twitter a partir da latitude e longitude, caso o usuário habilite o GPS, ou por meio do campo "Location" informado manualmente no perfil do usuário. Neste último caso, o algoritmo possibilita extrair o país quando o usuário informar uma única localidade ou múltiplas localidades. Os tipos de localidades tratados são: nome de país, sigla de país, nome de cidade, nome de região administrativa e nome alternativo de localidade. Este algoritmo baseou-se no fluxo do algoritmo utilizado no trabalho de Valkanas et al. [67], 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. Observou-se algumas limitações do trabalho de Valkanas et al. [67] 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 foram abordadas na presente pesquisa. A primeira limitação foi tratada por meio de definições de regras envolvendo duplicidade de nomes de cidades. A segunda limitação foi tratada pela configuração da codificação UTF8 (8-bit Unicode Transforma-

tion Format) [17] nas informações armazenadas no banco de dados PostgreSQL e pelo uso da linguagem PL/pgSQL no algoritmo de identificação de país. A codificação a ser utilizada no banco de dados é definida no momento da criação da base de dados. Para fins desta pesquisa, foi selecionada a codificação UTF8 por permitir armazenar e manusear caracteres de qualquer idioma, inclusive japonês e chinês. Por ser uma linguagem nativa do banco PostgreSQL e não necessitar de ferramentas intermediárias de conversão de caracteres, o uso da linguagem PL/pgSQL permitiu que localidades informadas utilizando caracteres orientais fossem processadas pelo algoritmo. Estas configurações permitiram a identificação de 130 usuários que informaram localidades utilizando caracteres em japonês em seu perfil e quatro usuários que informaram localidades em caracteres chinês, cujos exemplos estão na Figura 5.1. A coluna "Usuário" refere-se ao usuário do Twitter, a coluna "Location" refere-se ao local informado no perfil do usuário, a coluna "Tradução" refere-se à tradução dos caracteres em japonês e chinês e a coluna "País" refere-se ao nome do país identificado pelo algoritmo. Como resultado geral, o algoritmo possibilitou a identificação de cerca de 60% dos usuários que informaram o campo "Location" dentre os dados do Twitter utilizados nesta pesquisa. O algoritmo de identificação de país permite a análise espacial, a partir de dados do Twitter, podendo ser aplicado em diversas áreas, não se limitando à área de saúde pública.

Usuário	Location	Tradução	País
truth_cn	中国	China	China
7_iuhan_rn	北京	Pequim	China
useiou	日本	Japão	Japão
hamuhamu9434	東京	Tóquio	Japão

Figura 5.1: Exemplo de localidades escritas em caracteres japonês e chinês.

A segunda contribuição técnica é resultante do algoritmo supervisionado de análise textual. A análise de 2.452 comentários em inglês originários da África possibilitou criar o corpus de *Tweets* não-pessoais. Este corpus permite que, por meio da ferramenta Full-Text-Search, comentários sobre a epidemia sejam categorizados em cinco tipos: relatos sobre a epidemia, solução de contenção, impactos sociais e econômicos, notícias negativas e necessidade de preparo, cujos exemplos encontram-se na Tabela 3.10. Este corpus permite servir de base para a categorização de novos comentários do Twitter em inglês, relacionados ao Ebola ou à outros tipos de epidemia.

5.1.1 Exemplo de aplicação da metodologia

A metodologia apresentada nesta pesquisa utilizando a Teoria de Grafos, análise espacial e análise textual assim como as ferramentas gratuitas e os algoritmos desenvolvidos

poderá ser aplicada em outros estudos. Para exemplificar, descreve-se a seguir as possíveis alterações necessárias na metodologia para analisar o uso do Twitter no caso do vírus Zika. A primeira alteração é utilizar o termo de busca "Zika" para coletar comentários do Twitter por meio da ferramenta NodeXL. A segunda alteração refere-se à etapa de análise textual em que será necessário atualizar a lista de URL de sítios de noticiários e a lista de URL de sítios pessoais incluindo novos endereços URL encontrados nos comentários coletados sobre o vírus Zika. A terceira alteração é referente a atualização do corpus criado nesta pesquisa baseado na epidemia do Ebola para efetuar a categorização de *Tweets* não-pessoais. Este corpus poderá ser utilizado como base para categorizar os comentários coletados, porém por se tratar de um novo tema, as novas palavras identificadas nos comentários sobre o vírus Zika deverão ser incluídas no corpus para automatizar a categorização de *Tweets* não-pessoais. A quarta alteração refere-se à categorização de *Tweets* pessoais, em que o dicionário léxico SentiWordNet deverá ser customizado para adequar a pontuação das palavras ao novo tema em estudo.

5.2 Discussões

Esta seção apresenta algumas discussões relacionadas aos critérios adotados nesta pesquisa e aos resultados obtidos.

O primeiro ponto de discussão é referente ao algoritmo de identificação de país, em que determinadas regras foram adotadas quando a localidade informada pode referir-se a mais de um país. Por exemplo, quando o usuário informa mais de um tipo de localidade como país e cidade, foi adotada a regra de priorizar o país. Caso seja informado mais de um país, foi dada a preferência para o primeiro país informado. Um outro exemplo é a regra utilizada para situações de duplicidade de nome de cidades localizadas em países diferentes, em que a prioridade é da cidade mais populosa. Esses três exemplos de regras e outras descritas na Seção 3.3.1 podem não refletir a localização real do usuário.

O segundo ponto de discussão é com relação a validade dos resultados obtidos pelo algoritmo de identificação de país. A validação foi efetuada comparando-se com os resultados obtidos por meio da ferramenta Batch Geocoding. A validação indicou que 15,67% das localidades comparadas resultaram na identificação de países diferentes, devido ao uso de regras diferentes nas duas abordagens. Enquanto o algoritmo desta pesquisa procura identificar prioritariamente o país, dentre as localidades informadas, a ferramenta Batch Geocoding procura identificar prioritariamente uma combinação de logradouro e cidade. Uma vez que o usuário do Twitter pode informar qualquer valor no campo "Location",

torna-se difícil verificar qual a verdadeira intenção do usuário ao digitar o nome de uma ou mais localidades, assim como torna-se difícil a escolha da abordagem correta para a identificação do país a partir deste campo. Em relação aos dois pontos de discussões apresentados, a adoção das mesmas regras de identificação de país para todo os dados coletados do Twitter durante o período de análise possibilitaram um resultado uniforme para serem utilizados na identificação de vértices geolocalizados e análise textual dos comentários da África.

O terceiro ponto de discussão é em relação a representação de vértices geolocalizados no mapa de calor. Ao efetuar a análise espacial por meio de mapas, observou-se que, dependendo da extensão do país, a adoção do país como unidade de representação não possibilitou uma visualização clara em algumas situações. Por exemplo, a maioria dos comentários são originários dos Estados Unidos, cuja área é 130 vezes maior que a de Serra Leoa. No mapa da Figura 4.7, em que são mostrados os vértices chaves geolocalizados dos meses de novembro/2014, dezembro/2014 e janeiro/2015, os círculos do mapa de calor se concentraram em uma única região dos Estados Unidos, baseados nas coordenadas geográficas latitude e longitude do país. Neste caso, a utilização de estado ou cidade como unidade de representação possibilitaria uma melhor visualização, onde os círculos de calor estariam distribuídos pela área do país.

O quarto ponto de discussão refere-se ao algoritmo supervisionado de análise textual. A abordagem utilizada nesta pesquisa para identificar *Tweets* não-pessoais e pessoais baseou-se na presença de URL nos comentários. A grande variedade de URL encontrada gerou a necessidade de verificação manual de cada URL no navegador web para criar a lista de URL de sítios de notícias e a lista de URL de sítios pessoais. Os 2.452 comentários analisados continham 521 URL diferentes, tornando esta etapa manual bastante trabalhosa. Além disso, a variedade de palavras e expressões utilizadas nos comentários fez com que o corpus de *Tweets* não-pessoais necessitasse de revisão a cada nova carga de dados, tornando o processo de classificação demorado. Estes dois fatores limitaram a quantidade de comentários processados permitindo apenas uma análise preliminar de dados.

O quinto ponto de discussão refere-se ao tratamento de *Tweets* pessoais. Foi utilizada uma abordagem simples por meio do SentiWordNet, que baseia-se na polaridade individual de cada palavra. Porém, esta abordagem não efetua o tratamento da frase como um todo além de não tratar comentários complexos contendo questionamentos, expressões de sarcasmo e expressões de negação possibilitando obter resultados inconsistentes.

5.3 Trabalhos Futuros

Esta seção apresenta possíveis trabalhos futuros identificados ao longo da pesquisa.

A primeira melhoria está relacionada ao algoritmo de identificação de país. Ao efetuar a análise espacial por meio de mapas mostrados na Figura 4.7 observou-se que, dependendo da extensão do país, a representação dos vértices geolocalizados no mapa de calor não possibilitou uma visualização clara em algumas situações. Por exemplo, nos Estados Unidos, os círculos do mapa de calor, baseados nas coordenadas geográficas latitude e longitude do país, se concentraram em uma única região. Neste caso, a adoção de nome de cidade ou estado como tipo de local permitiria uma visualização mais clara do mapa de calor. O resultado da execução do algoritmo de identificação de país descrito na Seção 3.3.1.4 sugere que é possível adotar estratégias diferenciadas de identificação de local de acordo com o país. A Tabela 3.3 mostra a tendência do tipo de local utilizado pelos usuários do Twitter em alguns países. Nos Estados Unidos 62,04% dos usuários informaram nomes de cidades e 11,28% informaram nomes de estados no campo "Location", no Brasil 56,55% informaram nomes de cidades. Como trabalho futuro, o algoritmo de identificação de país permitirá identificar o país e a região administrativa ou cidade com base na informação do campo "Location". Esta melhoria proporcionará efetuar análises espaciais mais detalhadas a partir de dados extraídos do Twitter e se adequar ao contexto da pesquisa.

A segunda melhoria foi identificada na análise textual de *Tweets* pessoais. A abordagem de análise de sentimentos utilizada nesta pesquisa é uma técnica simples que baseia-se na polaridade individual de cada palavra. Esta técnica não permite tratar comentários complexos contendo frases interrogativas ou frases exclamativas. Por exemplo, o comentário *"What is being done to disseminate information to those staying in rural areas where most of our nation lives? #Ebola"* postado pelo usuário @ernestmac54 da República do Zimbábue sugere um questionamento sobre o que têm sido feito pelas autoridades para disseminar informações sobre o Ebola nas áreas rurais onde vive a maioria da população. O algoritmo pontuou o comentário com 0,19 e classificou como positivo por considerar somente a polaridade de cada palavra e não verificar o contexto como um todo, além de não interpretar a presença do ponto de interrogação. Similarmente, o comentário *"The fight is on! we will beat ebola"* encaminhado pelo usuário @fgbelee da Libéria foi interpretado na análise manual como uma mensagem de otimismo pela presença do ponto de exclamação(!) e foi pontuado com 0,90. Em contrapartida, o algoritmo pontuou o mesmo comentário com -0,082 por considerar somente a polaridade de cada palavra e não interpretar a presença do ponto de exclamação. Como trabalho futuro, o algoritmo

supervisionado de análise textual considerará estes aspectos observados nas análises dos comentários e efetuará o tratamento de frases interrogativas e frases exclamativas dos comentários do Twitter.

Para finalizar, esta pesquisa teve como objetivo analisar o uso das redes sociais na situação real de epidemia do Ebola combinando três tipos de análises. Estas análises possibilitaram verificar o uso to Twitter como ferramenta de disseminação de informações e identificar os atores relevantes cujas mensagens foram compartilhadas por usuários de diversos países.

A.Apêndice A

Tabela A.1: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - dezembro/2014.

Usuário	País	Grau de entrada	Location	Tipo de local
mittromney	Estados Unidos	842	Massachusetts	Nome de região administrativa
sonypictures	Estados Unidos	833	Culver City, CA	Nome de cidade
dilmarousselff	Brasil	348	Brasil	Nome de país
trovejou	Brasil	300	Brasil	Nome de país
ecorepublicano	Espanha	260	España	Nome de país
independent	Reino Unido	242	London, United Kingdom	Nome de país
abc	Estados Unidos	190	New York, NY	Nome alternativo manual
unicefusa	Estados Unidos	167	New York, NY	Nome alternativo manual
selenagomez	Estados Unidos	165	Los Angeles	Nome de cidade
joeyfatts	Estados Unidos	142	Long Beach, California	Nome de cidade
unicef_liberia	Libéria	138	Liberia, West Africa	Nome de país
unicef_liberia	Libéria	132	Liberia, West Africa	Nome de país
iamch0pper	Estados Unidos	128	LasVegas, Nevada	Nome de cidade
lindaikeji	Nigéria	128	Lagos	Nome de cidade
bbcandrewh	África do Sul	119	Johannesburg	Nome de cidade
edmundi	México	118	México, D.F.	Nome de país
mittromney	Estados Unidos	100	Massachusetts	Nome de região administrativa
sonypictures	Estados Unidos	100	Culver City, CA	Nome de cidade
cbsnews	Estados Unidos	98	New York, NY	Nome alternativo manual
drogalizado	Brasil	91	Porto Alegre - RS	Nome de cidade

Tabela A.2: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - dezembro/2014.

Usuário	País	Grau de saída	Location	Tipo de local
hmarston	Reino Unido	174	london	Nome de capital
areatwitt	Indonésia	168	Sidoarjo	Nome de cidade
elenamartines44	Espanha	132	Spain	Nome de país
yoeloj	Venezuela	70	Venezuela	Nome de país
helpingafricas	Gana	31	Accra Ghana	Nome alternativo manual
mrschida	África do Sul	24	Johannesburg, South Africa!	Nome de cidade
ebolaalert	Nigéria	23	Nigeria	Nome de país
hu_ebola	Ucrânia	21	Украина	Nome alternativo manual
ebolavaccinenow	Estados Unidos	18	Virginia, USA	Sigla de país
totalngonews	Reino Unido	16	london	Nome de capital
knityarn	Estados Unidos	16	United States	Nome de país
totalngonews	Reino Unido	15	london	Nome de capital
dpumgt	Estados Unidos	14	Chicago, IL USA	Nome de cidade
totalngonews	Reino Unido	13	london	Nome de capital
totalngonews	Reino Unido	12	london	Nome de capital
barack_ebola	Reino Unido	12	London	Nome de capital
beejaipls	Estados Unidos	12	teh nearist bar	Latitude, longitude
rcgp	Reino Unido	12	UK	Nome alternativo manual
emciiflames	Gana	11	Ghana, West Africa .	Nome de país
smokexfzayn	Reunião	11	Reunion	Nome de país

Tabela A.3: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - dezembro/2014.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
areatwitt	Indonésia	4.234.687,977	Sidoarjo	Nome de cidade
ebolaalert	Nigéria	2.335.956,331	Nigeria	Nome de país
independent	Reino Unido	1.573.916,305	London, United Kingdom	Nome de país
who	Suíça	1.538.075,140	Geneva, Switzerland	Nome de país
sciam	Estados Unidos	1.321.987,149	New York City, NY, USA	Sigla de país
ecorepublicano	Espanha	1.262.370,000	España	Nome de país
conaquim	Brasil	1.259.009,300	BRASÍLIA - DF -BRASIL	Nome de país
dineitemponews	Brasil	1.184.818,343	Brasil	Nome de país
hmarston	Reino Unido	1.154.678,731	london	Nome de capital
who	Suíça	1.072.069,023	Geneva, Switzerland	Nome de país
mittromney	Estados Unidos	1.049.809,170	Massachusetts	Nome de região administrativa
sonypictures	Estados Unidos	1.006.086,313	Culver City, CA	Nome de cidade
ebola_rt	Reino Unido	987.343,787	Brighton & Hove, UK	Nome de cidade
who	Suíça	811.281,604	Geneva, Switzerland	Nome de país
ebolaalert	Nigéria	810.535,155	Nigeria	Nome de país
unicef_liberia	Libéria	776.725,070	Liberia, West Africa	Nome de país
whitehouse	Estados Unidos	701.074,960	Washington, DC	Nome alternativo manual
yoeloj	Venezuela	664.725,347	Venezuela	Nome de país
afri_democratic	África do Sul	657.965,447	Johannesburg	Nome de cidade
helpingafricas	Gana	610.343,816	Accra Ghana	Nome alternativo manual

Tabela A.4: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - janeiro/2015.

Usuário	País	Grau de entrada	Location	Tipo de local
robdelaney	Reino Unido	351	London	Nome de capital
tuqueridasocia	México	274	México. DF.	Nome de país
mutualzayn	Filipinas	270	Philippines	Nome de país
nytimes	Estados Unidos	269	New York City	Nome de cidade
_tonydennis	Estados Unidos	244	Atlanta, GA	Nome de cidade
sexywgw	Estados Unidos	170	33.97271014 -84.01309717	Latitude, longitude
jackthejokster	Estados Unidos	161	New York	Nome de região administrativa
unicef_liberia	Libéria	157	Liberia, West Africa	Nome de país
mrmichaelspicer	Reino Unido	142	Kent, London	Nome de capital
unmeer	Gana	136	Accra, Ghana	Nome de país
jackthejokster	Estados Unidos	125	New York	Nome de região administrativa
jimmyurine	Estados Unidos	123	Los Angeles, CA	Nome de cidade
unmeer	Gana	113	Accra, Ghana	Nome de país
springnews_tv	Tailândia	97	Bangkok,Thailand	Nome de país
who	Suíça	92	Geneva, Switzerland	Nome de país
statedept	Estados Unidos	87	Washington, DC	Nome alternativo manual
statedept	Estados Unidos	86	Washington, DC	Nome alternativo manual
unmeer	Gana	77	Accra, Ghana	Nome de país
who	Suíça	76	Geneva, Switzerland	Nome de país
unmeer	Gana	76	Accra, Ghana	Nome de país

Tabela A.5: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - janeiro/2015.

Usuário	País	Grau de saída	Location	Tipo de local
paulbassline	Espanha	72	Madrid	Nome de capital
ebolavaccinenow	Estados Unidos	24	Virginia, USA	Sigla de país
ebola_rt	Reino Unido	22	Brighton & Hove, UK	Nome de cidade
cassie_soucy	Estados Unidos	15	Boston, MA	Nome de cidade
fluffator	Canadá	15	Toronto	Nome de cidade
isolonsebola	França	13	Paris	Nome de capital
lateam224	Guiné	13	Conakry - Guinee	Nome de país
paris_ebola	França	12	Paris	Nome de capital
ufuomaomawumi	Nigéria	12	Minna, Abuja ..Nigeria	Nome de cidade
leniashwenda	Estados Unidos	12	Geneva	Nome de cidade
rajbahuleyan	Índia	12	kerala	Nome de região administrativa
snm_carry	Estados Unidos	11	Long Island, New York	Nome alternativo manual
fluffator	Canadá	11	Toronto	Nome de cidade
yaftab	Reino Unido	10	UK	Nome alternativo manual
infectivia	Reino Unido	10	Washington, DC / Global	Nome de cidade
riwired	Reino Unido	10	Broadway	Nome alternativo Geonames
ebola_rt	Reino Unido	10	Brighton & Hove, UK	Nome de cidade
barack_ebola	Reino Unido	9	London	Nome de capital
ebola_rt	Reino Unido	9	Brighton & Hove, UK	Nome de cidade
nildip2010	Índia	9	Rajkot, Gujarat, India	Nome de país

Tabela A.6: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - janeiro/2015.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
ebolafiles	Estados Unidos	2.066.149,392	Atlanta, Georgia	Nome alternativo manual
robdelaney	Reino Unido	1.699.990,000	London	Nome de capital
deightor	Chipre	1.577.162,330	nicosia, cyprus	Nome de país
mutualzayn	Filipinas	1.475.080,000	Philippines	Nome de país
nytimes	Estados Unidos	1.178.828,061	New York City	Nome de cidade
unmeer	Gana	1.159.426,465	Accra, Ghana	Nome de país
jackthejokster	Estados Unidos	1.065.465,000	New York	Nome de região administrativa
who	Suíça	1.000.734,310	Geneva, Switzerland	Nome de país
msf_espana	Espanha	959.648,983	Barcelona, España	Nome de país
unmeer	Gana	937.792,606	Accra, Ghana	Nome de país
c1tyoffl1nt	Estados Unidos	904.948,000	Flint, Michigan	Nome de cidade
ebola_rt	Reino Unido	838.258,804	Brighton & Hove, UK	Nome de cidade
71mdf	El Salvador	833.786,883	El Salvador, C.A.	Nome de país
charitynewsuk	Reino Unido	744.599,034	UK	Nome alternativo manual
gabrialf2	Espanha	717.942,000	Sevilla - 14	Nome de cidade
mikescomedyy	Reino Unido	704.990,000	London	Nome de capital
jorgemadrid_	Espanha	662.848,419	Madrid	Nome de capital
ebola_rt	Reino Unido	647.163,053	Brighton & Hove, UK	Nome de cidade
fluffator	Canadá	601.098,123	Toronto	Nome de cidade
barackobama	Estados Unidos	559.146,784	Washington, DC	Nome alternativo manual

Tabela A.7: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - fevereiro/2015.

Usuário	País	Grau de entrada	Location	Tipo de local
cp24	Canadá	639	Toronto	Nome de cidade
austan_goolsbee	Estados Unidos	221	Chicago	Nome de cidade
unicef_liberia	Libéria	212	Liberia, West Africa	Nome de país
who	Suíça	190	Geneva, Switzerland	Nome de país
unicefsl	Serra Leoa	159	Sierra Leone	Nome de país
dailymail	Estados Unidos	154	New York	Nome de região administrativa
michaeltoole	Reino Unido	148	Cambridge, MA	Nome de cidade
leerhiannon	Austrália	135	Sydney	Nome de cidade
nytimes	Estados Unidos	112	New York City	Nome de cidade
estadao	Brasil	103	São Paulo	Nome de cidade
markcritch	Canadá	100	St. John's, Newfoundland	Nome de cidade
ac360	Estados Unidos	96	New York, NY	Nome alternativo manual
npr	Estados Unidos	96	Washington, DC	Nome alternativo manual
newsroompostcom	Índia	91	INDIA	Nome de país
citynews	Estados Unidos	80	Toronto, Ontario	Nome de cidade
60minutes	Estados Unidos	57	New York, NY	Nome alternativo manual
clarissegvl	Martinica	56	Martinique	Nome de país
cbsnews	Estados Unidos	56	New York, NY	Nome alternativo manual
vickybeanmr	Estados Unidos	54	Dallas, TX	Nome de cidade
vickybeanmr	Estados Unidos	53	Dallas, TX	Nome de cidade

Tabela A.8: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - fevereiro/2015.

Usuário	País	Grau de saída	Location	Tipo de local
ebola_rt	Reino Unido	22	Brighton & Hove, UK	Nome de cidade
_anacdiaz	Estados Unidos	19	los angeles	Nome de cidade
ebola_rt	Reino Unido	16	Brighton & Hove, UK	Nome de cidade
boscaptn	Estados Unidos	14	Boston, MA, USA	Sigla de país
lateam224	Guiné	14	Conakry - Guinee	Nome de país
ebola_rt	Reino Unido	13	Brighton & Hove, UK	Nome de cidade
thriiive	Estados Unidos	13	California	Nome de cidade
fluffator	Canadá	11	Toronto	Nome de cidade
thriiive	Estados Unidos	11	California	Nome de cidade
audibyrne	Estados Unidos	11	Mobile, Alabama	Nome de cidade
london_network_	Reino Unido	9	London	Nome de capital
jekjek19	Estados Unidos	9	Dallas, TX	Nome de cidade
fluffator	Canadá	9	Toronto	Nome de cidade
deathbyepidemic	Estados Unidos	9	USA	Sigla de país
kenzibit	Gana	9	Accra, Ghana.	Nome de país
ebola_rt	Reino Unido	8	Brighton Hove, UK	Nome de cidade
boscaptn	Estados Unidos	8	Boston, MA, USA	Sigla de país
ebola_rt	Reino Unido	8	Brighton Hove, UK	Nome de cidade
kipe76	Reino Unido	8	London	Nome de capital
kipe76	Reino Unido	8	London	Nome de capital

Tabela A.9: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - fevereiro/2015.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
cp24	Canadá	2.190.120,564	Toronto	Nome de cidade
austan_goolsbee	Estados Unidos	2.171.710,572	Chicago	Nome de cidade
fluffator	Canadá	1.626.041,521	Toronto	Nome de cidade
govchristie	Estados Unidos	1.361.584,243	Trenton, NJ	Nome de cidade
dailymail	Estados Unidos	1.077.393,000	New York	Nome de região administrativa
lmextra	Noruega	1.030.002,000	Son by the Oslofjord, Norway.	Nome de país
rtuknews	Reino Unido	1.024.228,000	UK	Nome alternativo manual
cbcnews	Canadá	935.744,324	Canada	Nome de país
leerhiannon	Austrália	911.827,749	Sydney	Nome de cidade
iddochymes	Estados Unidos	855.953,761	New York, NY	Nome alternativo manual
yayayarndiva	Estados Unidos	818.806,024	Sonoma County, California	Nome de cidade
michaeltoole	Reino Unido	686.778,000	Cambridge, MA	Nome de cidade
bbcnews	Reino Unido	582.425,378	London	Nome de capital
who	Suíça	550.813,631	Geneva, Switzerland	Nome de país
citynews	Estados Unidos	490.957,183	Toronto, Ontario	Nome de cidade
danielletv	Estados Unidos	448.788,616	Los Angeles	Latitude, longitude
worldbank	Estados Unidos	448.277,465	Washington, DC	Nome alternativo manual
guardian	Reino Unido	439.318,670	London	Nome de capital
ciocia	Canadá	425.808,813	London, Ontario, Canada	Nome de país
saidmose	Reino Unido	399.943,130	Washington, D.C. and Garowe	Nome de cidade

Tabela A.10: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - março/2015.

Usuário	País	Grau de entrada	Location	Tipo de local
memmosdubai	Emirados Árabes Unidos	803	Dubai, Uniter Arab Emirates	Nome de cidade
billgates	Estados Unidos	600	Seattle, WA	Nome de cidade
bipartisanism	Estados Unidos	277	Seattle, USA	Sigla de país
npr	Estados Unidos	249	Washington, DC	Nome alternativo manual
msf_espana	Espanha	208	Barcelona, España	Nome de país
cp24	Canadá	190	Toronto	Nome de cidade
jackthejokster	Estados Unidos	181	New York	Nome de região administrativa
nytimes	Estados Unidos	153	New York City	Nome de cidade
skynews	Reino Unido	130	London, UK	Nome de capital
msf_espana	Espanha	124	Barcelona, España	Nome de país
gatesfoundation	Estados Unidos	118	Seattle, Washington	Nome alternativo manual
noranartta	Tailândia	117	Bangkok, Thailand	Nome de país
el_pais	Espanha	104	Madrid	Nome de capital
statedept	Estados Unidos	89	Washington, DC	Nome alternativo manual
flotus	Estados Unidos	89	Washington, DC	Nome alternativo manual
gatesfoundation	Estados Unidos	81	Seattle, Washington	Nome alternativo manual
bbcbreaking	Reino Unido	74	London, UK	Nome de capital
tmz	Estados Unidos	72	Los Angeles, CA	Nome de cidade
washingtonpost	Estados Unidos	69	Washington, D.C.	Nome de capital
sridarwanto	Indonésia	69	Depok	Nome de cidade

Tabela A.11: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - março/2015.

Usuário	País	Grau de saída	Location	Tipo de local
ebolaalert	Nigéria	52	Nigeria	Nome de país
helporphankids	Estados Unidos	51	Atlanta	Nome de cidade
padforg	Estados Unidos	28	Washington, DC	Nome alternativo manual
lunatozia	Estados Unidos	28	New York	Nome de região administrativa
sonyasoni	Índia	21	New Delhi, India	Nome de país
ebola_rt	Reino Unido	19	Brighton & Hove, UK	Nome de cidade
lacrimobrien	França	19	Paris-Türkiye	Nome de capital
ebola_rt	Reino Unido	15	Brighton & Hove, UK	Nome de cidade
fluffator	Canadá	14	Toronto	Nome de cidade
ebolaphone	Países Baixos	13	Utrecht, The Netherlands	Nome de cidade
ebola_rt	Reino Unido	12	Brighton & Hove, UK	Nome de cidade
fluffator	Canadá	12	Toronto	Nome de cidade
ebola_fc	México	11	mexico	Nome de país
ro_juan27	Estados Unidos	11	New Jersey, USA	Sigla de país
robhot1982	Estados Unidos	11	Phoenix , Arizona	Nome de cidade
fluffator	Canadá	11	Toronto	Nome de cidade
ebola_rt	Reino Unido	11	Brighton & Hove, UK	Nome de cidade
meghealthinvest	Estados Unidos	10	Seattle	Nome de cidade
logischfaktisch	Suíça	10	Bern - Switzerland	Nome de país
meyerbjoern	Alemanha	10	UK/Germany	Nome de país

Tabela A.12: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - março/2015.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
billgates	Estados Unidos	2.835.349,856	Seattle, WA	Nome de cidade
ebolaalert	Nigéria	1.067.543,466	Nigeria	Nome de país
msf_espana	Espanha	989.853,149	Barcelona, España	Nome de país
afp	França	980.560,118	France	Nome de país
cp24	Canadá	934.778,952	Toronto	Nome de cidade
fluffator	Canadá	893.078,981	Toronto	Nome de cidade
bipartisanism	Estados Unidos	807.305,738	Seattle, USA	Sigla de país
msf_espana	Espanha	781.841,421	Barcelona, España	Nome de país
washingtonpost	Estados Unidos	717.245,422	Washington, D.C.	Nome de capital
memmosdubai	Emirados Árabes Unidos	642.402,000	Dubai, Uniter Arab Emirates	Nome de cidade
fluffator	Canadá	629.775,934	Toronto	Nome de cidade
unicef_es	Espanha	627.741,333	Madrid	Nome de capital
who	Suíça	620.051,266	Geneva, Switzerland	Nome de país
skynews	Reino Unido	597.905,594	London, UK	Nome de capital
mackayim	Austrália	565.318,544	Brisbane, Australia	Nome de país
who	Suíça	546.777,450	Geneva, Switzerland	Nome de país
gatesfoundation	Estados Unidos	538.967,031	Seattle, Washington	Nome alternativo manual
cnnmex	México	523.792,000	México	Nome de país
inghever	Chile	510.436,374	CHILE	Nome de país
abc_es	Espanha	504.020,000	Madrid	Nome de capital

Tabela A.13: Vértices chaves geolocalizados com os 20 maiores valores de grau de entrada - abril/2015.

Usuário	País	Grau de entrada	Location	Tipo de local
ecorepublicano	Espanha	379	España	Nome de país
ebolaalert	Nigéria	328	Nigeria	Nome de país
frankthedoorman	Estados Unidos	261	Los Angeles	Nome de cidade
nytimes	Estados Unidos	236	New York City	Nome de cidade
who	Suíça	210	Geneva, Switzerland	Nome de país
duncanwhitehead	Estados Unidos	206	Miami, USA	Sigla de país
unicefsl	Serra Leoa	200	Sierra Leone	Nome de país
unicefsl	Serra Leoa	186	Sierra Leone	Nome de país
thezaynieworld	Reino Unido	178	England	Nome alternativo manual
unicefsl	Serra Leoa	175	Sierra Leone	Nome de país
who	Suíça	171	Geneva, Switzerland	Nome de país
duncanwhitehead	Estados Unidos	166	Miami, USA	Sigla de país
who	Suíça	157	Geneva, Switzerland	Nome de país
billgates	Estados Unidos	157	Seattle, WA	Nome de cidade
unicef_es	Espanha	140	Madrid	Nome de capital
xor	Estados Unidos	125	san francisco	Nome de cidade
youtube	Estados Unidos	124	San Bruno, CA	Nome de cidade
yanes68	Espanha	124	Torrelodones	Nome de cidade
nytimes	Estados Unidos	124	New York City	Nome de cidade
zuluking709	África do Sul	120	Durban, South Africa	Nome de país

Tabela A.14: Vértices chaves geolocalizados com os 20 maiores valores de grau de saída - abril/2015.

Usuário	País	Grau de saída	Location	Tipo de local
roxzylok	Estados Unidos	159	New York	Nome de região administrativa
africanlbrunch	Estados Unidos	116	New York, NY	Nome alternativo manual
hackairforce1	Estados Unidos	28	Fort Knox	Nome de cidade
unitedliberians	Libéria	27	liberia	Nome de país
unitedliberians	Libéria	25	liberia	Nome de país
ebola_rt	Reino Unido	22	Brighton & Hove, UK	Nome de cidade
unitedliberians	Libéria	19	liberia	Nome de país
roxzylok	Estados Unidos	19	New York	Nome de região administrativa
ebola_rt	Reino Unido	18	Brighton & Hove, UK	Nome de cidade
ebola_rt	Reino Unido	17	Brighton & Hove, UK	Nome de cidade
ndhopeorg	Estados Unidos	17	Denver	Nome de cidade
unitedliberians	Libéria	17	liberia	Nome de país
msf_milafont	Venezuela	16	Valencia	Nome de cidade
ebolaphone	Países Baixos	16	Utrecht, The Netherlands	Nome de cidade
eboladlsph	Estados Unidos	16	Toronto, Ontario	Nome de cidade
ebolaphone	Países Baixos	16	Utrecht, The Netherlands	Nome de cidade
anjumsultana	Canadá	16	Toronto, Canada	Nome de país
ebola_rt	Reino Unido	16	Brighton & Hove, UK	Nome de cidade
ndhopeorg	Estados Unidos	15	Denver	Nome de cidade
ebola_rt	Reino Unido	14	Brighton & Hove, UK	Nome de cidade

Tabela A.15: Vértices chaves geolocalizados com os 20 maiores valores de centralidade de intermediação - abril/2015.

Usuário	País	Centralidade de intermediação	Location	Tipo de local
ebolaalert	Nigeria	1.951.415,616	Nigeria	Nome de país
nytimes	United States	1.838.676,112	New York City	Nome de cidade
bmj_latest	United Kingdom	1.835.269,379	London, UK	Nome de capital
unmeer	Ghana	1.725.417,584	Accra, Ghana	Nome de país
who	Switzerland	1.635.020,420	Geneva, Switzerland	Nome de país
who	Switzerland	1.510.443,232	Geneva, Switzerland	Nome de país
parafinale	United Kingdom	1.505.746,502	North West England	Nome alternativo manual
who	Switzerland	1.384.126,064	Geneva, Switzerland	Nome de país
ecorepublicano	Spain	1.310.233,000	España	Nome de país
frankthedoorman	United States	1.213.120,000	Los Angeles	Nome de cidade
ebolaalert	Nigeria	1.186.100,000	Nigeria	Nome de país
unicef_es	Spain	1.166.208,000	Madrid	Nome de capital
victhetics	Nigeria	1.134.938,000	Abuja	Nome de capital
maritzelr	Panama	1.126.444,000	Panama	Nome de país
fluffator	Canada	1.092.701,190	Toronto	Nome de cidade
africanlbrunch	United States	1.080.130,067	New York, NY	Nome alternativo manual
mackayim	Australia	1.021.511,296	Brisbane, Australia	Nome de país
ebolaalert	Nigeria	1.017.009,558	Nigeria	Nome de país
who	Switzerland	968.083,217	Geneva, Switzerland	Nome de país
samsteinhp	United States	954.613,000	Washington, D.C.	Nome de capital

Tabela A.16: Exemplos de nomes alternativos adicionados manualmente de localidades extraídos do Twitter.

País	Nome alternativo
Nigéria	abuja nigeria
Gana	accra ghana
México	Amealco de Bonfil Qro
Estados Unidos	atlanta , georgia
Estados Unidos	Austin Texas
Brasil	B R A S I L
México	Bacadéhuachi Sonora
Indonésia	Balikpapan
Bélgica	België
Argentina	Berazategui
Reino Unido	Birmingham UK
Colômbia	Bogota Colombia
França	Bretagne
Reino Unido	Britain
Estados Unidos	Bronx, New York
Argentina	Buenos Aires Argentina
Argentina	Burzaco
Estados Unidos	C A L I F O R N I A
México	Cadereyta de Montes Qro
África do Sul	Cape Town South Africa
Venezuela	Caracas Venezuela
Espanha	Catalunya
Estados Unidos	Central Florida
Argentina	Ciudad de Buenos Aires
México	Ciudad Obregon Sonora
Reino Unido	Cornwall, UK
México	Cozumel
México	D.F.
Estados Unidos	Dallas Texas
Reino Unido	Devon
México	Distrito Federal
Reino Unido	Dublin Ireland
Reino Unido	England
Espanha	Españistán
México	Estado de México
México	Guaimas Sonora
Guiné	Guinée
Zimbábue	Harare Zimbabwe
México	Hermosillo Sonora
Turquia	İZMİR
Indonésia	Jakarta Indonesia
Indonésia	Jakarta Selatan
África do Sul	Johannesburg South Africa
Uganda	Kampala Uganda
Nigéria	Lagos nigeria

País	Nome alternativo
Reino Unido	London Town
Estados Unidos	London UK
Estados Unidos	Long Island, New York
México	Los cabos
Estados Unidos	macon, georgia
Espanha	Madrid (España)
Espanha	Madrid (Spain)
Espanha	Mallorca
Reino Unido	Manchester UK
Estados Unidos	marietta, georgia
Estados Unidos	Maryland
Espanha	Mediaset España
Austrália	Melbourne Australia
México	Mexico D. F.
Estados Unidos	Miami FL
Uruguai	Montevideo Uruguay
México	MX
Nigéria	Naija
Quênia	Nairobi Kenya
Estados Unidos	Nashville TN
Estados Unidos	New York State
Estados Unidos	North Texas
Estados Unidos	Northern Virginia
Estados Unidos	northside , georgia
Austrália	NSW Australia
Estados Unidos	NY, NY
Estados Unidos	NYC, NY
Estados Unidos	Oakland CA
Canadá	Ontario Canada
Estados Unidos	Orange County
França	Paris (France)
México	Pedro Escobedo Qro
Austrália	Perth WA
Argentina	Ramos Mejia
Espanha	Región de Murcia
Brasil	RJ
Estados Unidos	San Francisco Bay Area
Espanha	Santiago Bernabéu
Chile	Santiago Chile
Estados Unidos	savannah, georgia
Suíça	Schweiz
Reino Unido	Scotland
Brasil	SP
Turquia	TÜRKİYE
Estados Unidos	U.S.A

Tabela A.17: 90 países com as maiores quantidades de *Tweets* durante novembro de 2014 a abril de 2015.

País	*Tweets*	Perc.	País	*Tweets*	Perc.
Estados Unidos	69.849	32,84%	Uruguai	475	0,22%
Reino Unido	15.281	7,18%	Panamá	435	0,20%
Nigéria	10.663	5,01%	Porto Rico	431	0,20%
México	10.523	4,95%	Egito	387	0,18%
Espanha	8.116	3,82%	Noruega	364	0,17%
Indonésia	6.533	3,07%	Costa Rica	358	0,17%
Venezuela	6.331	2,98%	Finlândia	358	0,17%
Canadá	5.956	2,80%	Senegal	346	0,16%
Brasil	5.414	2,55%	Arábia Saudita	338	0,16%
França	5.379	2,53%	Jamaica	335	0,16%
Índia	5.165	2,43%	Emirados Árabes Unidos	333	0,16%
Argentina	5.159	2,43%	Etiópia	323	0,15%
Austrália	3.976	1,87%	Guatemala	317	0,15%
África do Sul	3.662	1,72%	Grécia	296	0,14%
Itália	2.634	1,24%	China	294	0,14%
Alemanha	2.622	1,23%	Tanzânia	290	0,14%
Quênia	2.486	1,17%	Austria	284	0,13%
Países Baixos	2.195	1,03%	Costa do Marfim	279	0,13%
Colômbia	2.060	0,97%	Polónia	279	0,13%
Filipinas	1.832	0,86%	Comores	257	0,12%
Cuba	1.751	0,82%	Ucrânia	253	0,12%
Gana	1.516	0,71%	Camarões	253	0,12%
República Dominicana	1.291	0,61%	Dinamarca	245	0,12%
Japão	1.175	0,55%	Geórgia	230	0,11%
Paquistão	1.051	0,49%	Marrocos	217	0,10%
Chile	1.023	0,48%	Coreia do Sul	213	0,10%
Turquia	941	0,44%	Líbano	209	0,10%
El Salvador	901	0,42%	Nicarágua	201	0,09%
Suécia	898	0,42%	Hong Kong	199	0,09%
Malásia	890	0,42%	Paraguai	189	0,09%
Libéria	855	0,40%	Bangladesh	187	0,09%
Bélgica	796	0,37%	Bolívia	172	0,08%
Guiné	788	0,37%	Nepal	171	0,08%
Suíça	757	0,36%	Botsuana	170	0,08%
Portugal	739	0,35%	Ruanda	169	0,08%
Irlanda	737	0,35%	Israel	164	0,08%
Serra Leoa	727	0,34%	Mali	155	0,07%
Uganda	724	0,34%	Romênia	151	0,07%
Rússia	697	0,33%	Argélia	148	0,07%
Singapura	641	0,30%	Honduras	143	0,07%
Zimbábue	627	0,29%	Namíbia	134	0,06%
Nova Zelândia	610	0,29%	Qatar	130	0,06%
Equador	590	0,28%	Bulgária	125	0,06%
Tailândia	512	0,24%	Zâmbia	115	0,05%
Peru	485	0,23%	Seychelles	114	0,05%

B.Apêndice B

ROTEIRO DE INSTALAÇÃO

EXPLORANDO REDES SOCIAIS COMO FERRAMENTA DE DISSEMINAÇÃO DE INFORMAÇÕES: UMA ANÁLISE ESPAÇO TEMPORAL EM CASOS DE EPIDEMIA

Liriam Michi Enamoto

B.1 Introdução

O objetivo deste documento é descrever o roteiro de instalação das ferramentas utilizadas e desenvolvidas na pesquisa EXPLORANDO REDES SOCIAIS COMO FERRAMENTA DE DISSEMINAÇÃO DE INFORMAÇÕES: UMA ANÁLISE ESPAÇO TEMPORAL EM CASOS DE EPIDEMIA

B.2 Diagrama de arquitetura geral da solução

A Figura B.1 mostra as ferramentas utilizadas. Os retângulos em azul representam as ferramentas prontas adquiridas de forma gratuita, os retângulos em rosa são os algoritmos desenvolvidos durante a pesquisa, os retângulos em amarelo são plug-ins do banco de dados PostgreSQL e os retângulos em branco representam as informações.

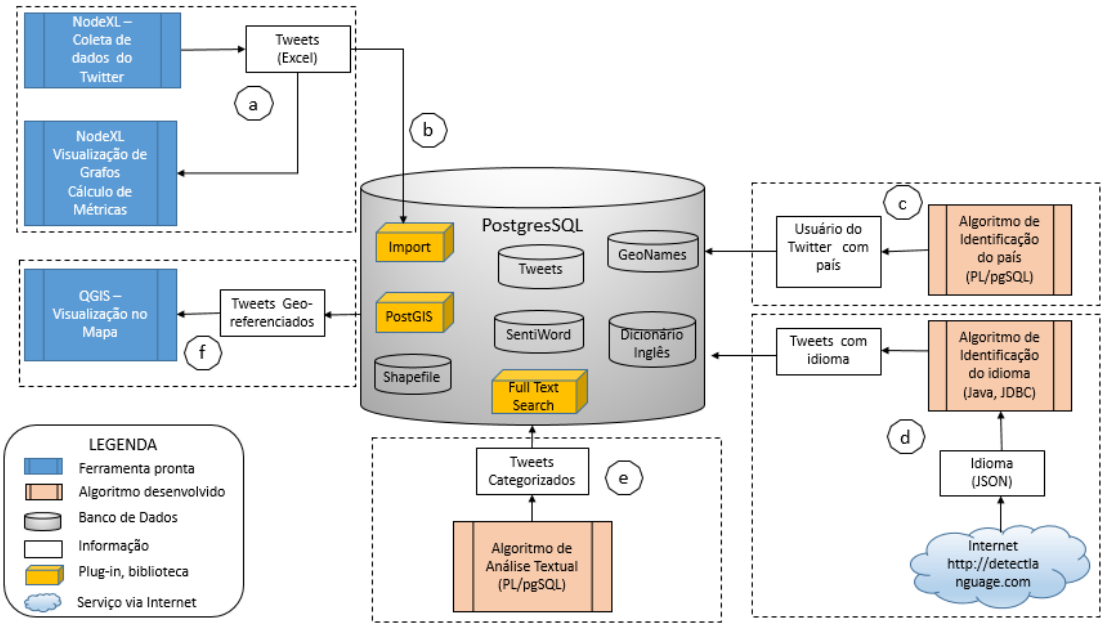


Figura B.1: Diagrama de arquitetura geral da solução.

O fluxo das informações está descrito abaixo. As letras de “a” à “f” correspondem às letras no diagrama:

a. Os comentários do Twitter são coletados por meio da ferramenta NodeXL e armazenados em arquivos Excel. O NodeXL permite o cálculo das métricas e a visualização de grafos.

b. As planilhas Excel contendo os comentários do Twitter são armazenadas no banco

de dados PostgresSQL por meio da ferramenta Import do próprio banco utilizando o formato csv.

c. O algoritmo de identificação de país é executado para determinar o país de origem dos usuários do Twitter.

d. O algoritmo de identificação de idioma é executado para determinar o idioma utilizado nos comentários do Twitter. O algoritmo utiliza o serviço web Language Detection que analisa a frase e retorna o idioma.

e. O algoritmo supervisionado de análise textual é executado para categorizar os comentários em comentários pessoais e não pessoais.

f. A ferramenta QGIS é utilizada para a visualização dos dados do Twitter no mapa.

B.3 Ferramentas adquiridas gratuitamente

B.3.1NodeXL

NodeXL é um template para o Microsoft Excel que permite coletar dados das principais redes sociais (Twitter, Facebook, Flickr e YouTube), criar grafos e calcular métricas de grafos. A ferramenta é disponibilizada na versão básica gratuita, com recursos limitados, e na versão profissional.

Link para download e instalação:

https://nodexl.codeplex.com/

B.3.2PostgreSQL

O PostgreSQL é um sistema gerenciador de banco de dados objeto-relacional, gratuito e de código aberto. A versão utilizada na pesquisa é PostgresSQL 9.4.1-1 Windows x64.

Link para download e instalação:

https://www.postgresql.org/download/

Link para dicas e tutoriais:

http://www.tutorialspoint.com/postgresql/

B.3.3PostGIS

É uma extensão do banco de dados PostgreSQL que permite o armazenamento e a manipulação de dados espaciais. A versão utilizada na pesquisa é PostGIS-pg93x64 ver2.1.5-2.

Link para download e instalação:

http://postgis.net/install/

Link de dicas para instalação:

http://workshops.boundlessgeo.com/postgis-intro/installation.html

Link de tutoriais:

http://www.bostongis.com/?content_name=postgis_tut01

http://workshops.boundlessgeo.com/postgis-intro/

B.3.4 QGIS

QGIS é uma ferramenta gratuita que permite a manipulação e visualização de dados geográficos no mapa.

Link para download e instalação:

http://www.qgis.org

Link de dicas e tutoriais:

http://www.qgistutorials.com/pt_BR/

B.3.5 Language Detection

Esta ferramenta gratuita permite a identificação de 160 idiomas diferentes por meio de serviço web. Além deste serviço, a ferramenta fornece também bibliotecas para a programação nas linguagens Java, C#, Ruby, Python, PHP e Crystal. Para utilizar como serviço web na sua versão gratuita, é necessário efetuar o cadastro no site e obter a chave de liberação.

Link para documentação:

http://detectlanguage.com/

B.3.6 Full Text Search

Full Text Search é uma ferramenta do PostgreSQL cujo objetivo é efetuar a busca textual de documentos no banco de dados. Esta ferramenta é instalada juntamente com o banco PostgreSQL.

Link para documentação:

https://www.postgresql.org/docs/9.4/static/textsearch.html

B.3.7 GeoNames

GeoNames é uma base de dados geográfica disponibilizada gratuitamente por meio da licença Creative Commons License.

Link para documentação:

http://www.geonames.org/

Link para download dos dados:

http://download.geonames.org/export/dump/

Foram utilizados na pesquisa os arquivos de dados "cities1000.zip", "cities15000.zip", "alternateNames.zip" e "admin2Codes.txt" disponíveis no link acima.

B.4 Passos para a instalação do banco de dados PostgreSQL

Os códigos fontes e arquivos necessários para a instalação das ferramentas desenvolvidas poderão ser solicitados por meio do email liriam.enamoto@uniriotec.br. Será disponibilizado o arquivo compactado "codigo fonte.zip". Ao descompactar este arquivo estará disponível a estrutura de diretório mostrada na Figura B.2.

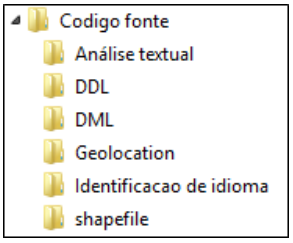


Figura B.2: Estrutura de diretórios que contém os códigos fontes.

B.4.1 Instalação do banco de dados PostgreSQL

O banco de dados PostgreSQL deverá ser instalado conforme as instruções da documentação disponível na Seção B.3.2.

B.4.2 Instalação da extensão PostGIS

O PostGIS deverá ser instalado conforme as instruções da documentação disponível na Seção B.3.3. A seguir serão abordados os principais pontos da instalação:

1. Instalar o PostGIS através da execução do instalador.
2. Criar um banco de dados espacial utilizando a ferramenta pgAdmin do PostgreSQL selecionando "New Database" conforme a Figura B.3. Na tela seguinte, informar o nome do banco de dados a ser criado no campo "name" e "postgres" no campo "owner".

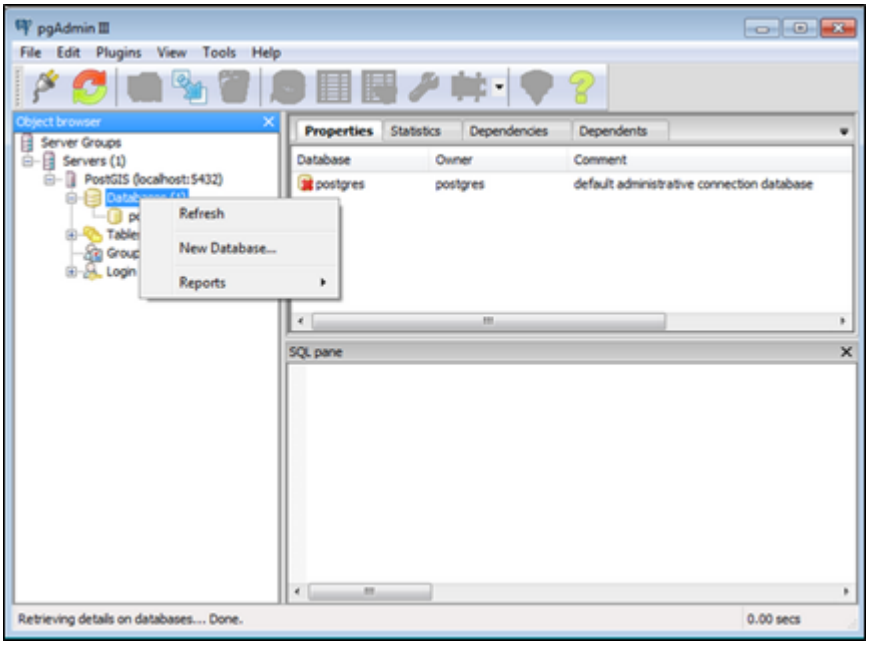


Figura B.3: Criação de um banco de dados espacial utilizando a ferramenta pgAdmin.

3. Selecionar o banco criado e expandir a árvore de objetos. Selecionar o schema "public".
4. No menu "Tools – Query Tool" acionar o "SQL Query". Abrirá uma nova janela onde deverá ser executado o comando abaixo:

   CREATE EXTENSION postgis;

5. A seguir executar o comando abaixo para verificar a versão instalada.

   SELECT postgis_full_version ();

B.4.3 Instalação de shapefiles

O próximo passo será carregar os dados espaciais (shapefiles). Na pesquisa foram utilizados dois shapefiles.

O primeiro shapefile "TM_WORLD_BORDERS" está disponível no link abaixo.

http://thematicmapping.org/downloads/world_borders.php

O shapefile é composto por 4 arquivos:

TM_WORLD_BORDERS-0.3.dbf

TM_WORLD_BORDERS-0.3.prj

TM_WORLD_BORDERS-0.3.shp

TM_WORLD_BORDERS-0.3.shx

O shapefile deverá ser instalado no banco conforme os passos abaixo:

1. Acionar o plug-in "PostGIS Shapefile and DFF Loader" conforme a Figura B.4.

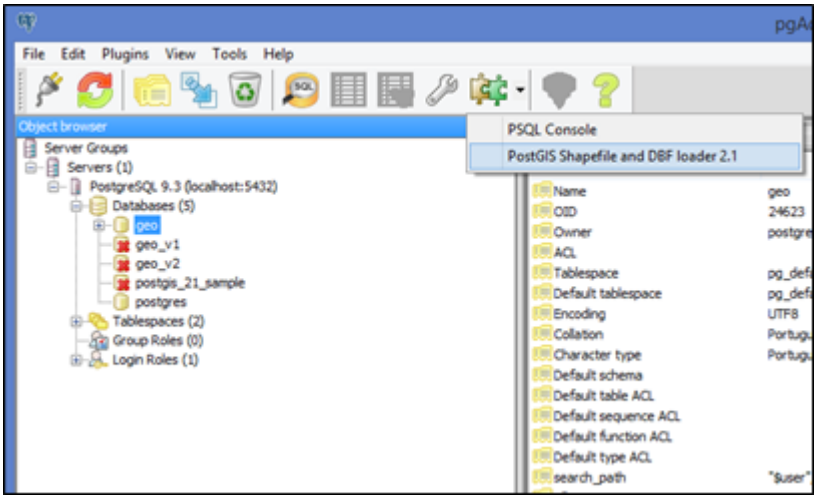


Figura B.4: Acionamento do plug-in "PostGIS Shapefile and DFF Loader".

2. Abrirá a janela mostrada na Figura B.5, em seguida deve-se e clicar em "View connection details" para testar a conexão com o banco de dados criado.

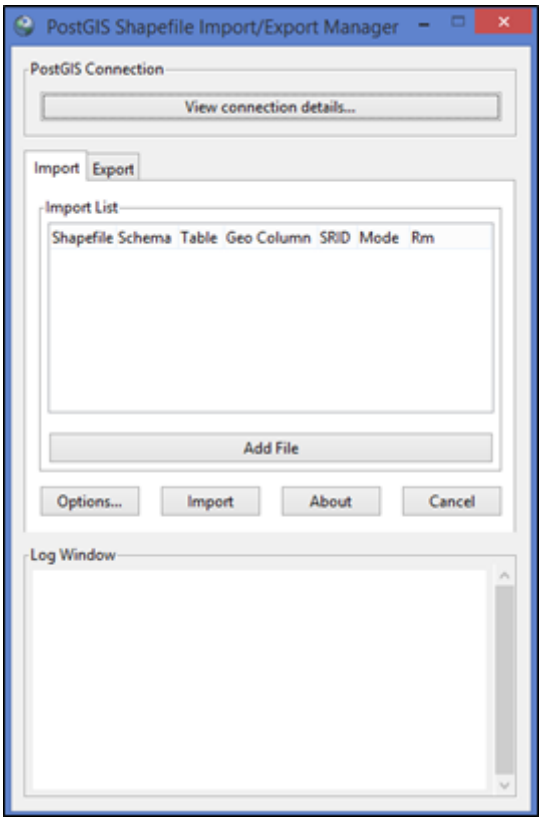


Figura B.5: Instalação do shapefile por meio do PostGIS.

3. Clicar em "Add File" e selecionar o shapefile a ser carregado. Alterar o SRID (Spatial Reference IDentifier) para 4326 e clicar em "Import", conforme a Figura B.6. Finalizado este procedimento, poderá ser verificada a criação da nova tabela "tm_world_borders-0.3" no banco de dados.

O segundo shapefile "Countries" está disponível em:

\\Codigo fonte\shapefile

Este segundo shapefile é composto por 4 arquivos e deverá ser instalado seguindo as mesmas instruções da instalação de "TM_WORLD_BORDERS".

countries.dbf

countries.prj

countries.shp

contries.shx

Após finalizado o procedimento de instalação do segundo shapefile, uma nova tabela "countries" será criada no banco de dados.

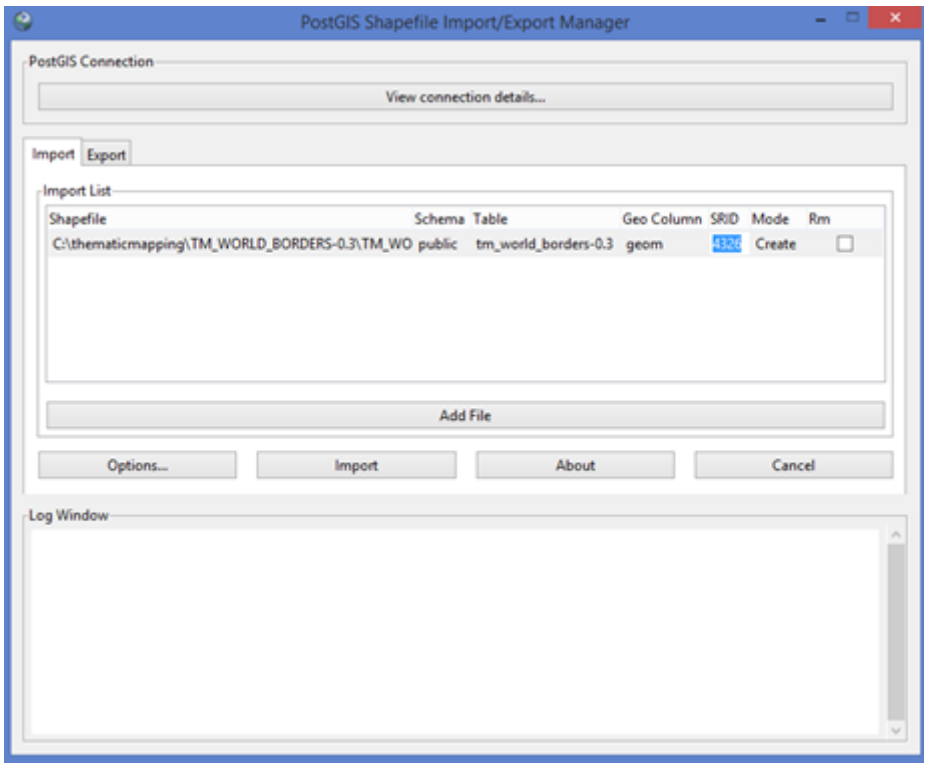


Figura B.6: Confirmação da instalação do shapefile.

B.4.4 Criação de objetos no banco de dados

A Tabela B.1 contém a lista de tabelas do banco de dados e uma breve descrição sobre o seu conteúdo e a origem dos seus dados. As tabelas com o prefixo "g_" são originárias do GeoNames. Os dados para estas tabelas deverão ser obtidos no sítio do GeoNames e carregados no banco de dados. As tabelas "coutries", "spatial_ref_sys" e "tm_world_borders-0.3" são criadas automaticamente no procedimento de instalação do PostGIS ou shapefiles, conforme indicado na descrição de cada tabela.

1. Criar as tabelas executando o script disponível em:

   \\Codigo fonte\DDL

2. Inserir os dados iniciais nas tabelas executando o script disponível em:

   \\Codigo fonte\DML

B.5 Ferramentas desenvolvidas

Foram desenvolvidos três algoritmos nesta pesquisa, cujos detalhes serão apresentados a seguir. Para a execução correta dos algoritmos, é necessário que toda a estrutura

Tabela B.1: Lista de tabelas do banco de dados.

Tabela	Descrição
countries	Tabela criada por meio do procedimento de instalação do shapefile "countries". Utilizada para busca de nome de país em inglês, espanhol, francês e sigla ISO3.
g_admincode1	Efetuar o download do arquivo "admin2Codes.txt" do GeoNames e importar os dados do arquivo para esta tabela. Contém as regiões administrativas da base do GeoNames.
g_alternatename	Efetuar o download do arquivo "alternateNames.zip" do GeoNames e importar os dados do arquivo para esta tabela. Contém os nomes alternativos de localidades da base do GeoNames.
g_geoname	Efetuar o download do arquivo "cities1000.zip" do GeoNames e importar os dados do arquivo para esta tabela. Contém todas as localidades da base do GeoNames.
g_geoname_15000	Efetuar o download do arquivo "cities15000.zip" do GeoNames e importar os dados do arquivo para esta tabela. Contém localidades da base do GeoNames com mais de 15000 habitantes.
spatial_ref_sys	Tabela criada na instalação do PostGIS.
tb_alternate_name	Contém os nomes alternativos inseridos manualmente.
tb_dominio_ex	Contém URL de sítios pessoais.
tb_dominio_tweet	Contém URL de sítios de noticiários.
tb_dup_15000_priorizado	Contém localidades da base do GeoNames com nomes duplicados em países diferentes, priorizados pela cidade mais populosa.
tb_dup_mesmo_pais	Contém localidades da base do GeoNames com nomes duplicados localizados no mesmo país.
tb_dup_paises_dif	Contém localidades duplicadas em países diferentes.
tb_dup_paises_dif_alternate	Contém nomes alternativos duplicados em países diferentes.
tb_pais	Contém informações de países, utilizadas para busca de nome de país.
tb_sentiword_sum	Contém palavras do dicionário SentiWord cujas pontuações foram adaptadas ao tema de epidemia.
tb_tweet	Tabela contendo os comentários do Twitter. Os *Tweets* coletados pelo NodeXL deverão ser carregados no banco no formato csv.
tb_tweet_carga	Tabela temporária utilizada no processamento de algoritmo de identificação de país, idioma e análise textual.
tb_usuario	Tabela contendo as informações de usuário do Twitter. Os usuários do Twitter coletados pelo NodeXL deverão ser carregados no banco no formato csv.
tb_usuario_carga	Tabela temporária utilizada no processamento de algoritmo de identificação de país.
tb_usuario_metrica	Após o cálculo das métricas pelo NodeXL, a planilha Excel contendo as métricas calculadas deverá ser carregada nesta tabela utilizando o formato csv.
tm_world_borders-0.3	Tabela criada por meio do procedimento de instalação do shapefile "tm_world_borders". Utilizada para a identificação do país a partir da latitude e longitude.

de tabelas e procedimentos armazenados já esteja criada no banco de dados PostgreSQL assim como é necessário ter os comentários do Twitter importados na base de dados criada.

A Tabela B.2 contém a lista de todos os procedimentos armazenados desenvolvidos na pesquisa que devem ser criados no banco de dados.

B.5.1 Algoritmo de identificação de país

Desenvolvido na linguagem PL/pgSQL, efetua a identificação do país de origem dos usuários do Twitter, por meio de atualização automática da tabela contendo dados do usuário.

O código fonte dos procedimentos armazenados está disponível em:

\\Codigo fonte\Geolocation

Para executar o algoritmo, executar a rotina principal no pgAdmin:

Select fn_geolocation();

B.5.2 Algoritmo de identificação de idioma

Desenvolvido na linguagem Java, efetua a identificação automática do idioma dos comentários do Twitter previamente armazenados no banco de dados. Utiliza o serviço web Language Detection. Para executar este algoritmo é necessário estar conectado à internet.

O código fonte está disponível em:

\\Codigo fonte\Identificacao de idioma\PostgreSQLJCBC.java

B.5.3 Algoritmo supervisionado de análise textual

Este algoritmo utiliza a ferramenta Full Text Search disponível na instalação do banco de dados PostgreSQL, não sendo necessário um procedimento de instalação específico da ferramenta. É composto por vários comandos em SQL que devem ser executados diretamente no banco de dados PostgreSQL.

O código fonte está disponível em:

\\Codigo fonte\Analise Textual

Tabela B.2: Lista de procedimentos armazenados desenvolvidos.

Função	Descrição	Análise textual	Identificação de país
fn_check_username()	Verifica se o nome do usuário na cadeia de caracteres faz parte da lista de usuários de sítios de noticiários.	X	
fn_clear_word()	Efetua o tratamento do campo local da tabela tb_usuario_carga, descartando valores numéricos, URL, e-mails, emoji.		X
fn_clearemoji()	Remove emoji de uma sequência de caracteres.		X
fn_find_emoji()	Verifica se a cadeia de caracteres contém emoji e caso encontre algum emoji, retorna sua pontuação correspondente.	X	
fn_find_emotion()	Verifica se a cadeia de caracteres contém emotion e caso encontre algum emotion, retorna sua pontuação correspondente.	X	
fn_find_space()	Identifica espaço em uma cadeira de caracteres.	X	
fn_find_username()	Verifica se a cadeia de caracteres contém nome de usuário do Twitter.	X	
fn_geolocation()	Efetua a identificação do país e atualiza a tabela tb_usuario_carga.		X
fn_identifylocation()	Efetua a identificação de país – busca múltipla para cada token informado.		X
fn_isnumeric()	Verifica se o parâmetro de entrada é um valor numérico.		X
fn_multiple_location()	Aciona a função de identificação de país – busca múltipla informando o separador de token.		X
fn_remore_username()	Remove o nome do usuário do Twitter da cadeia de caracteres. Utilizado para remover o nome do usuário do comentário antes de efetuar a identificação de idioma.	X	
fn_single_location()	Efetua a identificação de país – busca simples.		X
fn_tokenizationn_array()	Divide a cadeia de caracteres em tokens a partir de um separador informado.		X

Referências Bibliográficas

[1] Batch Geocoding. http://www.findlatitudeandlongitude.com/batch-geocode/. Último acesso em 24/01/2016.

[2] ESRI Shapefile Technical Description. http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf. Último acesso em 24/01/2016.

[3] Facebook. http://www.facebook.com/. Último acesso em 24/01/2016.

[4] Flickr. http://www.flickr.org/. Último acesso em 24/01/2016.

[5] Full Text Search. http://www.postgresql.org/docs/9.3/static/textsearch.html/. Último acesso em 24/01/2016.

[6] Geonames. http://www.geonames.org/. Último acesso em 24/01/2016.

[7] Java. http://www.oracle.com/technetwork/java/index.html/. Último acesso em 24/01/2016.

[8] Language Detection API. http://detectlanguage.com/. Último acesso em 24/01/2016.

[9] NodeXL. http://nodexl.codeplex.com/. Último acesso em 24/01/2016.

[10] Organização Mundial de Saúde. http://www.who.int/csr/disease/ebola/. Último acesso em 30/04/2016.

[11] PL/pgSQL. http://www.postgresql.org/docs/9.3/static/plpgsql.html/. Último acesso em 24/01/2016.

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

[13] PostgreSQL. http://www.postgresql.org/. Último acesso em 24/01/2016.

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

[15] Sentiwordnet. `http://sentiwordnet.isti.cnr.it`. Último acesso em 24/01/2016.

[16] Twitter. `http://twitter.com/`. Último acesso em 24/01/2016.

[17] UTF8. `https://en.wikipedia.org/wiki/UTF-8/`. Último acesso em 05/04/2016.

[18] WHO Influenza. `http://www.who.int/csr/don/2009_12_30/en/`. Último acesso em 05/04/2016.

[19] Youtube. `http://www.youtube.com/`. Último acesso em 24/01/2016.

[20] ACAR, A., MURAKI, Y. "Twitter for crisis communication: lessons learned from japan's tsunami disaster", *International Journal of Web Based Communities* v. 7, n. 3, pp. 392–402, 2011.

[21] ANTUNES, M. N., SILVA, D., C. H., GUIMARÃES, M. C. S., et al. "Monitoramento de informação em mídias sociais: o e-monitor dengue", *TransInformação* v. 26, n. 1, pp. 9–18, 2014.

[22] BACCIANELLA, S., ESULI, A., SEBASTIANI, F. "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.". In: *Proceedings of Language Resources and Evaluation Conference*, pp. 2200–2204, 2010.

[23] BOUILLOT, F., PONCELET, P., ROCHE, M. "How and why exploit tweet's location information?". In: *Proceedings of AGILE'2012: 15th International Conference on Geographic Information Science*, pp. N–A, 2012.

[24] CÂMARA, G. "Representação computacional de dados geográficos", *CASANOVA, MA et al. Banco de dados geográficos. Curitiba: Mundogeo*, pp. 11–52, 2005.

[25] CHANDRA, S., KHAN, L., MUHAYA, F. B. "Estimating twitter user location using social interactions–a content based approach". In: *Proceedings of Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pp. 838–843, 2011.

[26] CHENG, Z., CAVERLEE, J., LEE, K. "You are where you tweet: a content-based approach to geo-locating twitter users". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768, 2010.

[27] CHEW, C., EYSENBACH, G. "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak", *PloS one* v. 5, n. 11, p. e14118, 2010.

[28] COMPTON, R., JURGENS, D., ALLEN, D. "Geotagging one hundred million twitter accounts with total variation minimization". In: *Proceedings of Big Data (Big Data), 2014 IEEE International Conference on*, pp. 393–401, 2014.

[29] CORLEY, C. D., COOK, D. J., MIKLER, A. R., et al. "Text and structural data mining of influenza mentions in web and social media", *International journal of environmental research and public health* v. 7, n. 2, pp. 596–615, 2010.

[30] DENGA, D.-P., LEMMENSA, R. "An ontology-based approach to incorporate user-generated geo-content into sdi", *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* v. 3825pp. 38–43, 2011.

[31] FEINGOLD, B., PARK, S. Y., COMER, D. M., et al. "Multiple listing for pediatric heart transplantation in the usa: Analysis of optn registry data from 1995 through 2009", *Pediatric transplantation* v. 17, n. 8, pp. 787–793, 2013.

[32] FINK, C., KOPECKY, J., BOS, N., et al. "Mapping the twitterverse in the developing world: An analysis of social media use in nigeria". In: *Proceedings of Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 164–171, 2012.

[33] FRUCHTERMAN, T. M., REINGOLD, E. M. "Graph drawing by force-directed placement", *Softw., Pract. Exper.* v. 21, n. 11, pp. 1129–1164, 1991.

[34] GABIELKOV, M., RAO, A., LEGOUT, A. "Studying social networks at scale: macroscopic anatomy of the twitter social graph". In: *Proceedings of ACM SIGMETRICS Performance Evaluation Review*, pp. 277–288, 2014.

[35] GRAHAM, M., HALE, S. A., GAFFNEY, D. "Where in the world are you? geolocation and language identification in twitter", *The Professional Geographer* v. 66, n. 4, pp. 568–578, 2014.

[36] HAN, B., YEPES, A. J., MACKINLAY, A., et al. "Identifying twitter location mentions". In: *Proceedings of Australasian Language Technology Association Workshop 2014*, p. 157, 2014.

[37] HEWLETT, B. S., AMOLA, R. P. "Cultural contexts of ebola in northern uganda", *Emerging infectious diseases* v. 9, n. 10, pp. 1242–1248, 2003.

[38] JACOBSEN, K. H., AGUIRRE, A. A., BAILEY, C. L., et al. "Lessons from the ebola outbreak: Action items for emerging infectious disease preparedness and response", *EcoHealth*, pp. 1–13, 2016.

[39] JAVA, A., SONG, X., FININ, T., et al. "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*, pp. 56–65, 2007.

[40] JI, X., CHUN, S. A., WEI, Z., et al. "Twitter sentiment classification for measuring public health concerns", *Social Network Analysis and Mining* v. 5, n. 1, pp. 1–25, 2015.

[41] KHAN, A. S., TSHIOKO, F. K., HEYMANN, D. L., et al. "The reemergence of ebola hemorrhagic fever, democratic republic of the congo, 1995", *Journal of Infectious Diseases* v. 179, n. Supplement 1, pp. S76–S86, 1999.

[42] KLEINBERG, J., TARDOS, É., *Algorithm design*. Pearson Education India, 2006.

[43] KWAK, H., LEE, C., PARK, H., et al. "What is twitter, a social network or a news media?". In: *Proceedings of the 19th international conference on World wide web*, pp. 591–600, 2010.

[44] LAZARD, A. J., SCHEINFELD, E., BERNHARDT, J. M., et al. "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* v. 43, n. 10, pp. 1109–1111, 2015.

[45] LIU, B. "Sentiment analysis and opinion mining", *Synthesis lectures on human language technologies* v. 5, n. 1, pp. 1–167, 2012.

[46] LU, Y., HU, X., WANG, F., et al. "Visualizing social media sentiment in disaster scenarios". In: *Proceedings of the 24th international conference on World Wide Web companion*, pp. 1211–1215, 2015.

[47] MACEACHREN, A. M., ROBINSON, A. C., JAISWAL, A., et al. "Geo-twitter analytics: Applications in crisis management". In: *Proceedings of 25th International Cartographic Conference*, pp. 3–8, 2011.

[48] MERCHANT, R. M., ELMER, S., LURIE, N. "Integrating social media into emergency-preparedness efforts", *New England Journal of Medicine* v. 365, n. 4, pp. 289–291, 2011.

[49] MISLOVE, A., MARCON, M., GUMMADI, K. P., et al. "Measurement and analysis of online social networks". In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pp. 29–42, 2007.

[50] NOVAK, P. K., SMAILOVIĆ, J., SLUBAN, B., et al. "Sentiment of emojis", *PloS one* v. 10, n. 12, p. e0144296, 2015.

[51] ODLUM, M., YOON, S. "What can we learn about the ebola outbreak from tweets?", *American journal of infection control* v. 43, n. 6, pp. 563–571, 2015.

[52] OFOGHI, B., MANN, M., VERSPOOR, K. "Towards early discovery of salient health threats: A social media emotion classification technique". In: *Proceedings of Pacific Symposium on Biocomputing (PSB)*, p. 504, 2016.

[53] OTTE, E., ROUSSEAU, R. "Social network analysis: a powerful strategy, also for the information sciences", *Journal of information Science* v. 28, n. 6, pp. 441–453, 2002.

[54] OYEYEMI, S. O., GABARRON, E., WYNN, R. "Ebola, twitter, and misinformation: a dangerous combination?", *British Medical Journal;349:g6178*, , 2014.

[55] PAK, A., PAROUBEK, P. "Twitter as a corpus for sentiment analysis and opinion mining.". In: *Proceedings of Language Resources and Evaluation Conference*, pp. 1320–1326, 2010.

[56] PAVALANATHAN, U., EISENSTEIN, J. "Emoticons vs. emojis on twitter: A causal inference approach", *arXiv preprint arXiv:1510.08480*, , 2015.

[57] SAKAKI, T., OKAZAKI, M., MATSUO, Y. "Earthquake shakes twitter users: real-time event detection by social sensors". In: *Proceedings of the 19th international conference on World wide web*, pp. 851–860, 2010.

[58] SCHAEFFER, S. E. "Graph clustering", *Computer Science Review* v. 1, n. 1, pp. 27–64, 2007.

[59] SENEVIRATNE, S., SENEVIRATNE, A., MOHAPATRA, P., et al. "Your installed apps reveal your gender and more!", *ACM SIGMOBILE Mobile Computing and Communications Review* v. 18, n. 3, pp. 55–61, 2015.

[60] SHI, Z., RUI, H., WHINSTON, A. B. "Content sharing in a social broadcasting environment: evidence from twitter", *Available at SSRN 2341243*, , 2013.

[61] SIMON, T., GOLDBERG, A., ADINI, B. "Socializing in emergencies—a review of the use of social media in emergency situations", *International Journal of Information Management* v. 35, n. 5, pp. 609–619, 2015.

[62] SMITH, M. A., SHNEIDERMAN, B., MILIC-FRAYLING, N., et al. "Analyzing (social media) networks with nodexl". In: *Proceedings of the fourth international conference on Communities and technologies*, pp. 255–264, 2009.

[63] STONEBRAKER, M., ROWE, L. A., *The design of Postgres*. vol. 15. ACM, 1986.

[64] TAKHTEYEV, Y., GRUZD, A., WELLMAN, B. "Geography of twitter networks", *Social networks* v. 34, n. 1, pp. 73–81, 2012.

[65] TEAM, W. E. R., OTHERS. "Ebola virus disease in west africa—the first 9 months of the epidemic and forward projections", *N Engl J Med* v. 371, n. 16, pp. 1481–95, 2014.

[66] TRAME, J., KEßLER, C. "Exploring the lineage of volunteered geographic information with heat maps", *GeoViz: Linking Geovisualization with Spatial Analysis and Modeling, Hamburg, Germany*, pp. 10–12, 2011.

[67] VALKANAS, G., GUNOPULOS, D. "Location extraction from social networks with commodity software and online data". In: *Proceedings of Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pp. 827–834, 2012.

[68] VOLZ, R., KLEB, J., MUELLER, W. "Towards ontology-based disambiguation of geographical identifiers.". In: *Proceedings of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web*, 2007.

[69] WASSERMAN, S., FAUST, K., *Social network analysis: Methods and applications*. vol. 8. Cambridge university press, 1994.

[70] WIIL, U. K., GNIADEK, J., MEMON, N. "Measuring link importance in terrorist networks". In: *Proceedings of Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference*, pp. 225–232, 2010.