ANÁLISE EXPLORATÓRIA DE DADOS DO TWITTER: compreendendo as conexões da informação de saúde durante o surto da febre amarela em 2017

Exploratory Analysis of Twitter data: Understanding the health information connections during the Yellow Fever Outbreak in 2017

Gabriela Denise de Araujo (1); Fabricio Landi de Moraes (2); Ivan Torres Pisa (3)

(1) Universidade Federal de São Paulo (Unifesp), Brasil gabrieladenise.a@gmail.com (2) fabricio.landi@gmail.com (3) ivanpisa@gmail.com

Resumo

Este artigo apresenta uma análise exploratória de como as informações de saúde foram compartilhadas e discutidas no Twitter em termos de tópicos de conscientização e posicionamentos durante surto da Febre Amarela de 2017 no Brasil. Para isso, foi utilizada a abordagem de mineração de dados com análise exploratória de grafos. Como principais resultados, foram identificados os picos de mensagens comparados aos picos de casos notificados em algumas regiões do país, uma análise das hashtags vinculadas ao principal assunto e diferentes tópicos oriundos da análise exploratória de grafos como campanha de vacinação, sentimentos, prevenção, rumores, outras doenças vinculadas ao mesmo transmissor entre outros. Este estudo demonstrou que as redes sociais, como o Twitter, oferecem oportunidades únicas para a vigilância participativa, podendo auxiliar no monitoramento de alguns aspectos da saúde pública e oferecer dados adicionais aos gestores de saúde de como as pessoas interagem durante um surto.

Palavras-chave: Redes sociais; Mineração de dados; Febre Amarela; Análise Exploratória de Grafos

Abstract

This paper presents a detailed analysis of how health information was shared and discussed on Twitter in terms of awareness and opinions during the 2017 Yellow Fever outbreak in Brazil. For this, the data mining approach with exploratory graph analysis was performed. As main results, Twitter activity peaks were identified compared to peaks of cases reported in some regions of the country, an analysis of hashtags linked to the main subject and different topics from the exploratory analysis of graphs such as vaccination campaign, feelings, prevention, rumors, other diseases linked to the same transmitter, among others. This study illustrates that social networks, such as Twitter, offer unique opportunities for

Araujo, Gabriela Denise de; Moraes, Fabricio Landi de and Pisa, Ivan Torres. Análise exploratória de dados do Twitter: compreendendo as conexões da informação de saúde durante o surto da febre amarela em 2017. *Brazilian Journal of Information Science: Research trends*, vol.14, no.3, jul.-set. 2020. e020006. https://doi.org/ 10.36311/1940-1640.2020.v14n3.10179

participatory surveillance, which can assist in monitoring some aspects of public health and offer additional data to health managers on how people interact during an outbreak.

Keywords: Social Network; Data Mining; Yellow Fever; Exploratory Graph Analysis

1 Introdução

Entre final de 2016 e 2017, o Brasil vivenciou o surto mais expressivo de Febre Amarela (FA) outrora observado no país que afetou principalmente os estados da região Sudeste, a região mais populosa do país. Este surto foi considerado o maior da série histórica desde 1980 quando o Ministério da Saúde passou a disponibilizar os dados contabilizados dos casos (Boldrini 2017). Segundo o Informe Especial Febre Amarela no Brasil nº 01/2017 divulgado pelo Ministério da Saúde (Brasil 2017a) diversas ações de enfrentamento da FA foram realizadas baseando-se em vigilância integrada, diagnóstico laboratorial e prevenção/imunização da população. Ainda houve ações transversais nas áreas de assistência à saúde, comunicação e financiamento.

No campo da saúde pública, a vigilância é uma área fundamental, pois está diretamente relacionada às práticas de atenção e promoção da saúde dos cidadãos. O sistema de vigilância deve ser composto por estruturas de informação que subsidiam a tomada de decisões, planejamentos e investigações sobre epidemias que ocorrem ao longo do tempo em certos territórios, sendo então capaz de promover ações específicas para contornar situações de risco (Guimarães et al 2017). Todavia, para promover ações eficazes é necessário que os órgãos responsáveis tenham informações atuais e consistentes para a realização do monitoramento da saúde pública.

Desenvolver sistemas de vigilância envolve acessar um conjunto de informações relacionadas à saúde como: morbidade, mortalidade, estado imunitário e nutricional da população. Quanto maior a quantidade de dados de saúde úteis que possam auxiliar os processos de vigilância em saúde, melhor será a identificação de alterações oportunas do padrão epidemiológico de doenças, detecção e monitoramento, provendo respostas às emergências em saúde pública.

As redes sociais têm proporcionado um montante de conteúdos na web que amplia as possibilidades de obtenção de informações também relacionadas à saúde. A mineração de dados de saúde do Twitter, por exemplo, possibilita o rastreamento de relatos médicos ao longo do tempo com geolocalização. A participação dos usuários expondo suas opiniões e experiências sobre saúde nas mídias sociais pode auxiliar no monitoramento de alguns aspectos da saúde pública e colaborar para uma vigilância participativa, oferecendo uma percepção aos gestores de saúde de como as pessoas interagem com temas de saúde na web.

Contudo, a tarefa de automatizar a extração de informação em dados textuais e organizá-los é desafiadora. O grande volume de dados textuais torna as ferramentas automáticas de processamento de texto cada vez mais requeridas. Por isso, investigações de novas metodologias para análise desses dados têm sido impulsionadas com o intuito de monitorar hábitos comunicacionais e medir diferentes características da população, incluindo aspectos de saúde (Araújo et al. 2018).

A proposta deste estudo é investigar as percepções do público em situações de surto a partir de mensagens publicadas no Twitter, utilizando técnicas de mineração de dados textuais com análise exploratória de grafos e identificar relacionamentos entre os achados que possam descrever características da opinião da população brasileira fomentando a produção científica de informações sobre vigilância em saúde utilizando as redes sociais.

2 Referencial Teórico

Diversos estudos (Beykikhoshk et al. 2015; Miller et al. 2017; Stefanidis et al. 2017) presentes na literatura mostram o potencial do uso de abordagens de mineração de texto e redes complexas para detectar eventos e sumarizar informações úteis nas redes sociais. Há evidências de que essas plataformas facilitadoras de compartilhamento de conteúdo podem auxiliar tanto às autoridades de saúde pública, através da identificação dos interesses e preocupações da população, eventos/surtos, quanto a própria população que faz uso desses ambientes para buscar informações opinativas e relatos experiências de outras pessoas.

Sanders-Jackson et al. (2015) apresentaram um método de análise de dados relacionadas ao tabaco e à saúde provenientes de redes sociais. No artigo os autores descrevem uma metodologia linguística de comparação da frequência relativa de termos e apresentam uma análise de rede semântica comparando a força das conexões entre as palavras. A análise mostra vários aspectos da rede que ajudaram a identificar frases-chaves ou conceitos importantes nas conversas online. Esse estudo revelou que há questões sobre tabaco sendo discutidas no Twitter e análise de redes semânticas pode auxiliar na caracterização de conversas online, assim intervenções futuras podem aproveitar as redes sociais e os principais eventos atuais para aumentar a conscientização sobre questões relacionadas ao tabagismo (Sanders-Jackson et al. 2015).

Tangherlini et at. (2016) se dedicaram a entender o que é falado sobre vacinação em dois sites populares de rede social dedicados aos pais/mães e analisaram a estrutura narrativa persuasiva que é discutida nesses sites. Para isso, utilizaram métodos probabilísticos para determinarem os tópicos de discussões, desenvolveram um modelo generativo de narrativa estatística-mecânica para extrair automaticamente as histórias e fragmentos subjacentes de milhões de postagens. Agregaram as histórias em um grafo de estrutura narrativa abrangente. Identificaram uma forte estrutura narrativa relacionada à busca de isenções e uma cultura de desconfiança do governo e das instituições médicas (Tangherlini et al. 2016).

Klein; Guidi Neto and Tezza (2017) realizaram um estudo transversal que analisou a associação entre variáveis obtidas no monitoramento das mídias sociais Facebook, Twitter, Instagram, Flickr, Youtube e Blog e as variáveis obtidas pelos indicadores da Diretoria de Vigilância Epidemiológica de Santa Catarina. Os assuntos monitorados foram: dengue, chikungunya, zika e microcefalia. Esse monitoramento foi restrito ao estado de Santa Catarina e realizado no período de um mês para coincidir com as semanas epidemiológicas da Vigilância Epidemiológica. Para verificar a associação estatística entre os dados foi utilizado o Coeficiente de Correlação de Pearson. As variáveis analisadas apresentaram uma alta correlação indicando que o monitoramento e a mineração de conteúdo compartilhado nas redes sociais podem ser um bom indicador para gestores da área de saúde (Klein; Guidi Neto and Tezza 2017).

Grover et al. (2018) exploraram as discussões no Twitter relacionadas às tecnologias utilizadas na saúde. O estudo apresenta as principais tecnologias no domínio da saúde por meio de uma análise de hashtag e as principais doenças (agudas, crônicas, transmissíveis e não transmissíveis) por meio da análise de palavras e sua associação pela co-ocorrência de palavras nos tweets. A associação mostrou que as tecnologias foram usadas no tratamento, identificação e cura de várias doenças. O estudo também realizou diferentes análises estatísticas e de redes sociais apresentando diversos grafos e matrizes de correlação com o intuito de mostrar um quadro geral de como as várias tecnologias são sendo relacionadas ao domínio da saúde (Grover et al. 2018).

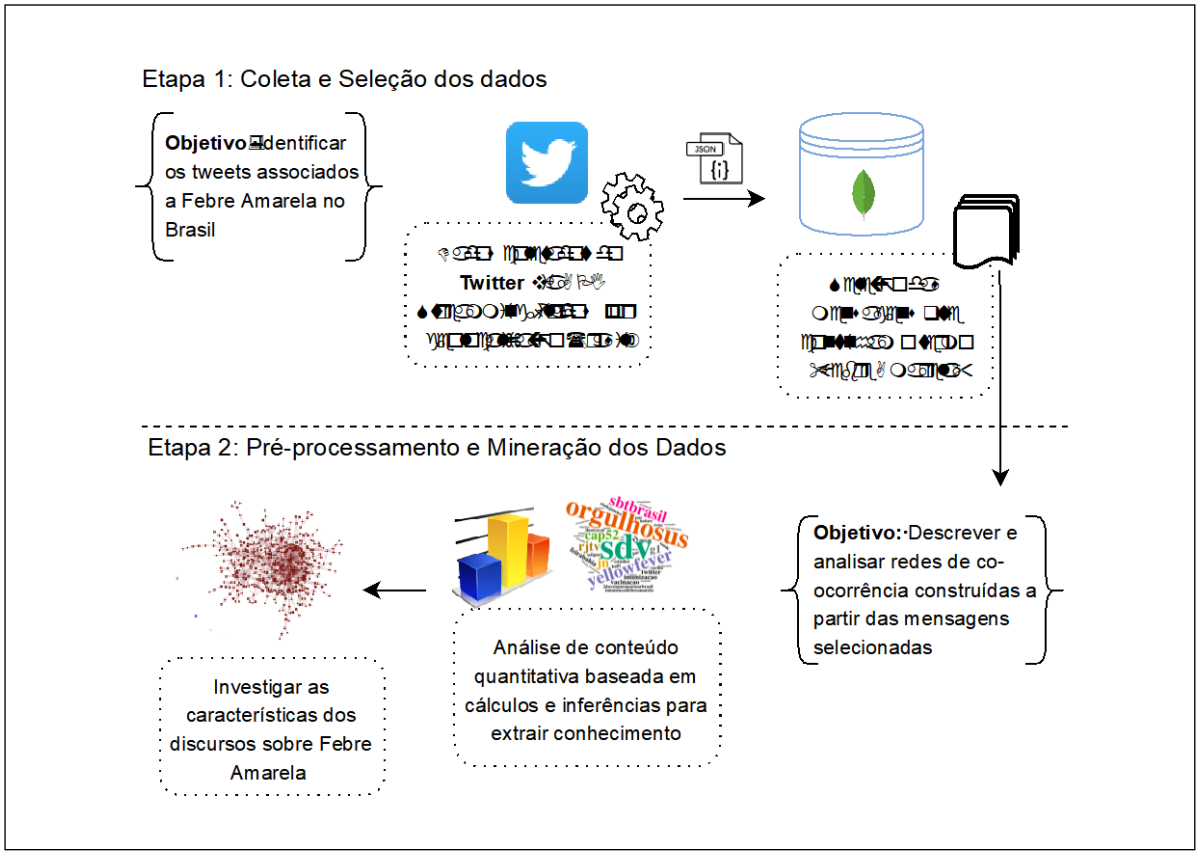
Um estudo (Hoffman et al. 2019) focou no tema vacinação, buscando caracterizar os indivíduos que compartilham mensagens sobre anti-vacinação, as informações que geralmente são publicadas e a disseminação desse conteúdo nas redes sociais. O conjunto de dados analisado era composto por mensagens publicadas no Facebook de 197 usuários em resposta a uma mensagem que promove a vacinação. Realizaram uma análise quantitativa, análise descritiva, análise de redes sociais e uma avaliação qualitativa. A análise das redes sociais descobriu que os tópicos e as pessoas tendiam a se agrupar em quatro subgrupos distintos (“confiança”, “alternativa”, “segurança” e “conspiração”). A identificação de subgrupos distintos alerta que não se pode utilizar uma abordagem geral para campanhas ou programas educacionais sobre vacinação, combater um único tema ou argumento provavelmente não será bem-sucedido com todas as crenças anti-vacina.

3 Metodologia

Este artigo é um estudo exploratório-descritivo (Lakatos 2017) que utiliza técnicas de mineração de dados e análise de grafos para investigar as características dos discursos sobre FA em dados disponíveis publicamente na rede social Twitter. A Figura 1 apresenta uma ilustração das etapas metodológicas propostas para a elaboração deste estudo, incluindo as duas etapas principais: (1) coleta e seleção dos dados e (2) pré-processamento e mineração dos dados. Uma descrição mais abrangente e detalhada de cada etapa é apresentada a seguir na Figura 1.

A etapa 1 contempla a coleta, armazenamento e seleção dos dados. O Twitter, a rede social foco deste estudo, detém uma política de privacidade (https://twitter.com/pt/privacy) que torna público todos os tweets registrados pelos seus usuários, ou seja, visível e pesquisável por qualquer pessoa e disponível inclusive por meio das interfaces de programação de aplicativos (APIs), que suportam as ferramentas de monitoramento e análise dos conteúdos publicados. Sendo assim, neste estudo foi desenvolvida uma aplicação em NodeJS (https://nodejs.org/) que interage com a API do Twitter para coletar todas as mensagens publicadas no Brasil. A API do Twitter (https://developer.twitter.com/) disponibiliza a utilização de filtros de pesquisa e nesse caso foi utilizado o filtro de geolocalização para retornar apenas as mensagens publicadas no Brasil. Foram coletadas e selecionadas para análise 3.061 mensagens do Twitter (tweets) que continham o termo "Febre Amarela" no período de 28/02/2017 a 31/08/2017 (185 dias).

Figura 1 - Ilustração das etapas metodológicas propostas para realização deste estudo englobando as duas etapas



Fonte: autores

Com os dados selecionados, a primeira atividade da etapa 2 foi realizar o pré-processamento ou transformação dos dados. O pré-processamento de dados têm a responsabilidade de melhorar a qualidade dos dados para aumentar a eficiência do processo de mineração de dados. Para isso foi utilizado o software R (https://www.r-project.org/) que possui um grande poder de análise de dados fornecendo ferramentas robustas para manipular e preparar os mesmos. Por meio desse software foi possível realizar a limpeza e transformação dos dados com os seguintes processos: remoção das *stopwords* (artigos, preposições, e etc.), caracteres especiais, pontuação, acentos, espaços duplos, URLs, tags de usuários e padronizando as palavras em caixa baixa. Essa é a atividade mais onerosa, apesar das implementações existentes para grande parte das tarefas de pré-processamento. Essas tarefas consomem boa parte do tempo do processo de extração do conhecimento. Além disso, com o entendimento dos dados alguns cuidados devem ser realizados, como tratamento de exceções e particularidades.

Após o pré-processamento, foi realizada a análise de conteúdo quantitativa (Bardin 2016) baseada na identificação das palavras mais frequentes e suas correlações, e a análise de *hashtag* que são palavras consideradas rótulos de metadados que podem ter relação com o conteúdo da mensagem em que está presente. A rede social Twitter foi um dos principais disseminadores da *hashtag* que posteriormente se tornou um recurso para filtrar os conteúdos postados na rede por assunto.

Por fim, com o intuito de investigar as relações entre as palavras e características dos discursos sobre FA, foi realizada uma análise exploratória de grafo por rede complexa. Uma rede complexa corresponde a um grafo com elementos discretos representados por um conjunto de nós ligados por arestas (Appel 2010). Neste artigo, os nós da rede serão as palavras e as arestas as conexões entre as palavras que representam uma co-ocorrência entre dois termos que estavam em um mesmo tweet. Através dessa representação é possível calcular diferentes métricas para compreensão dos objetos e suas ligações, buscando o entendimento dos dados e o significado de suas interações. As medidas que foram analisadas a partir de métricas da análise de rede são

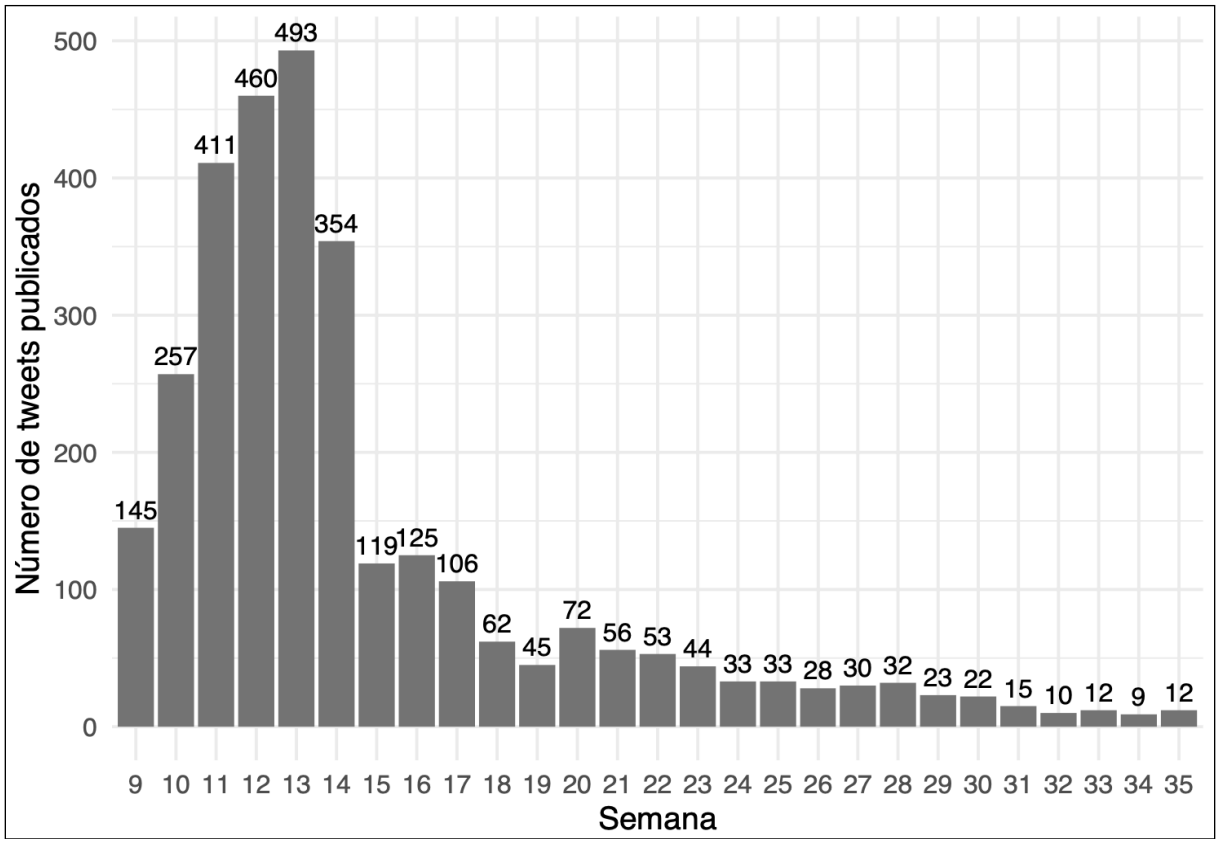
centralidade de grau (*degree*), centralidade de intermediação (*betweenness*) e agrupamentos (*cluster*).

A análise de grafo também foi realizada no software R utilizando um pacote para visualização e manipulação de rede chamado "visNetwork" que faz uso da biblioteca javascript "vis.js" (http://visjs.org).

4 Resultados e Discussão

Na Figura 2 é apresentado um gráfico com o número de mensagens que foram publicadas em cada semana e na Figura 3 um gráfico com a distribuição temporal dos casos suspeitos de FA notificados à SVS/MS até 31/05/2017.

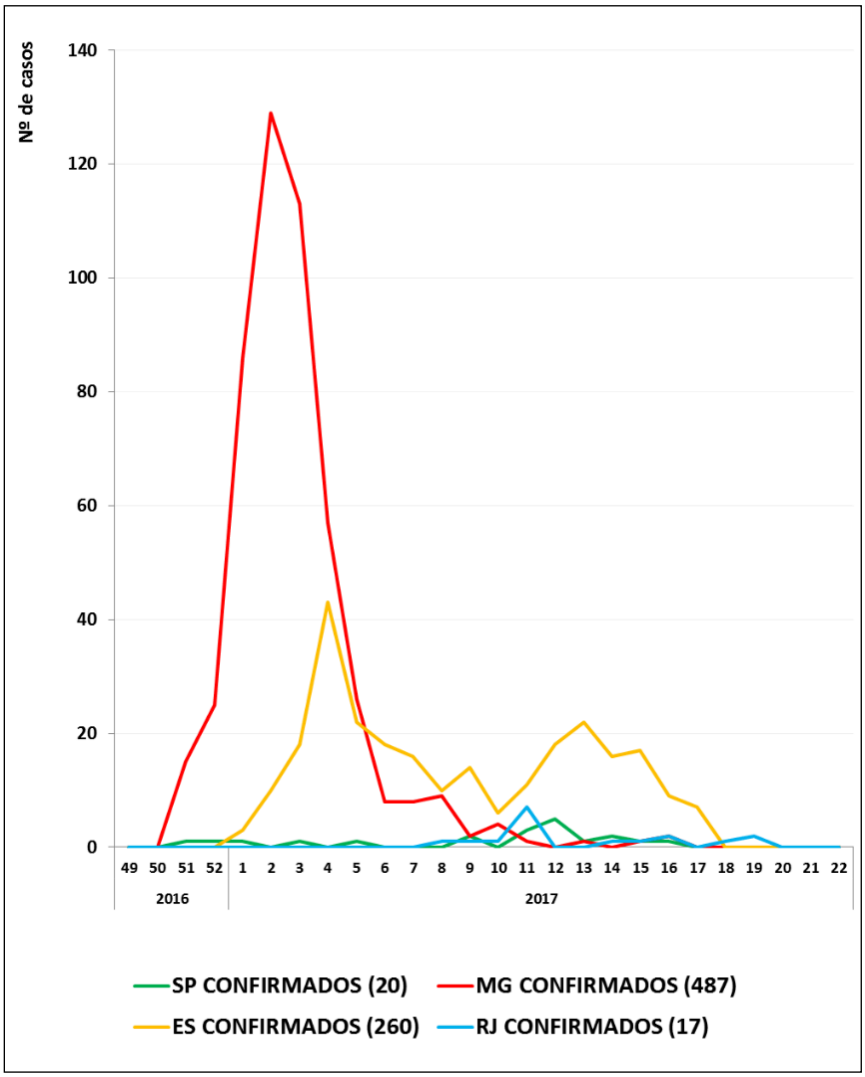
Figura 2 - Número de tweets publicados em cada semana do período da coleta



Fonte: dados da pesquisa

Segundo o Informe do Centro de Operações de Emergências em Saúde Pública sobre Febre Amarela nº43/2017 (Brasil 2017b), o pico do número de casos confirmados e em investigação ocorreu no mês de janeiro de 2017 (entre as semanas 1 a 6), todavia os casos continuaram a aumentar em diversas regiões do país ao longo do tempo, principalmente nos estados de Minas Gerais, Espírito Santo, Rio de Janeiro e São Paulo.

Figura 3- Distribuição temporal dos casos suspeitos de FA notificados à SVS/MS até 31/05/17, com data de IS a partir de 01 dezembro de 2016, por UF do LPI, data de IS e classificação



Fonte Brasil (2017c)

O comportamento do Gráfico 1 (Figura 2) mostrou um pico de mensagens entre as semanas 11 e 14, que também foram as semanas que apresentaram picos de casos confirmados em algumas regiões como Rio de Janeiro, São Paulo e Espírito Santo. Uma limitação deste estudo foi a impossibilidade de utilizar os dados de geolocalização, pois poucos tweets apresentaram latitude e longitude, ou seja, poucos usuários publicaram mensagens de dispositivos móveis com marcação geográfica ativada. No entanto, foi possível identificar em diversas mensagens a citação de algumas cidades e estados que podem ter relação aos eventos.

Após a seleção dos tweets, técnicas de mineração de dados foram aplicadas nas mensagens para extrair regularidades, padrões e tendências dos dados textuais. O intuito foi explorar o conteúdo de forma quantitativa baseando-se na eleição e na classificação de palavras em conceitos e a posterior análise de suas correlações.

4.1 Análise de *hashtags*

Primeiro foi contabilizado o número de mensagens do corpus que apresentavam *hashtags* (#). Foram encontradas 149 *hashtags* únicas em 183 mensagens. As Figuras 4 e 5 apresentam a nuvem de palavras com todas estas *hashtags* e um gráfico ordenado por frequência com as 20 *hashtags* que mais ocorreram nas mensagens, respectivamente.

Através da visualização do gráfico na Figura 5 foi possível identificar assuntos específicos agrupados pelas *hashtags*.

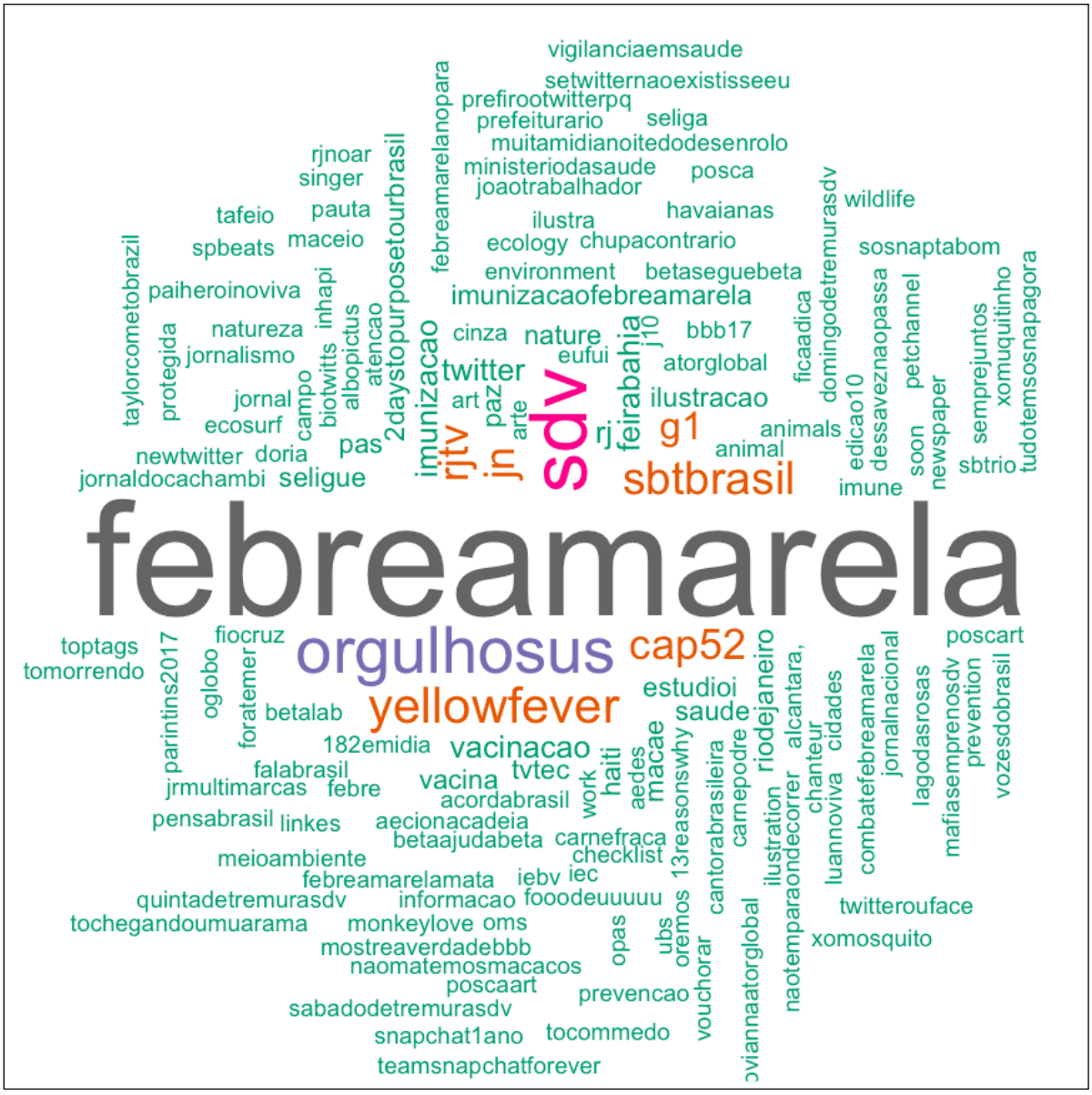
A nuvem de palavras evidencia as palavras importantes em um conjunto de textos, que neste artigo são os tweets. Quanto maior a frequência da palavra, maior o tamanho da fonte, e diferentes cores são aplicadas.

Várias *hashtags* relacionadas a imunização e vacinação foram destaques, analisando alguns dos tweets com essas *hashtags* foi possível notar mensagens sobre campanhas de vacinação, locais e as situações nas instituições de saúde, e também sobre a própria situação de saúde dos usuários que publicaram mensagens relacionadas a estarem ou não imunes e terem ou não tomado a vacina contra a FA.

Além disso, várias *hashtags* relacionadas a noticiários e telejlornais (ex: "rjtv", "jn", "g1", "sbtbrasil") foram identificadas e, entre os tweets que continham essas hashtags, alguns eram mensagens que propagavam informações sobre a situação da doença no país e outros se tratavam de opiniões/questionamentos das pessoas sobre a doença e a situação de surto em algumas regiões.

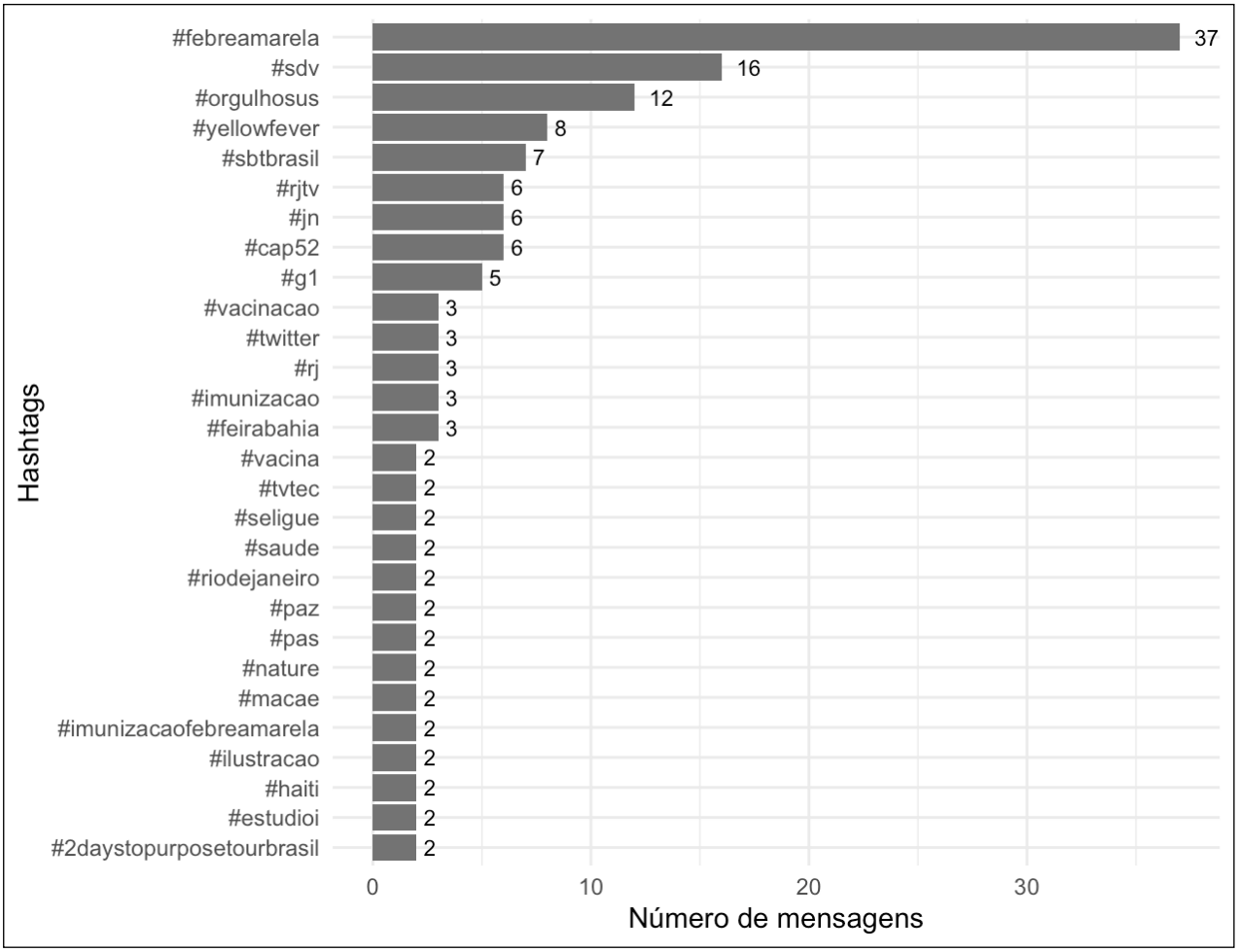
Araujo, Gabriela Denise de; Moraes, Fabricio Landi de and Pisa, Ivan Torres. Análise exploratória de dados do Twitter: compreendendo as conexões da informação de saúde durante o surto da febre amarela em 2017. *Brazilian Journal of Information Science: Research trends*, vol.14, no.3, jul.-set. 2020. e020006. https://doi.org/ 10.36311/1940-1640.2020.v14n3.10179

Figura 4- Nuvem de palavras das hashtags. O tamanho da fonte das palavras é proporcional à frequência da mesma correspondente no conjunto de dados



Fonte: dados da pesquisa

Figura 5 - Ranking das 20 mais frequentes hashtags no conjunto de dados



Fonte: dados da pesquisa

Também foram encontradas com grande frequência as *hashtags:* "orgulhosus" e "cap52", ambas relacionadas a serviços públicos de saúde. A primeira foi criada pelo Conselho Municipal de Saúde que trouxe ao Brasil o Dia do Orgulho SUS e convocou os usuários e profissionais do sistema de saúde a proporem ideias de como melhorar a assistência ao paciente dentro do serviço de saúde. A segunda foi criada pela Coordenadoria Geral de Saúde da AP 5.2 com a intenção de promover informativos, ações e promoção de saúde.

Outras *hashtags* de menor relevância se referiam a eventos, a programas de TV e a cidades. Esses exemplos são uma prova do tipo de informação que pode ser obtida em mensagens do Twitter apenas monitorando *hashtags*, esse processo colabora para o entendimento dos dados revelando alguns dos assuntos discutidos sobre o tema FA.

4.1 Análise Exploratória de Grafo

Após analisar as *hashtags* e descobrir alguns dos assuntos que foram discutidos nos tweets, uma análise exploratória de grafo examinando as relações entre as palavras foi realizada. Essa análise buscou compreender as narrativas das mensagens tanto de maneira ampla, quanto em suas singularidades, identificando vários tópicos de conversas que também fizeram parte do escopo das discussões sobre FA, realçando algumas características e padrões nos discursos.

Para dados textuais, três tipos de redes complexas são predominantes: as redes de co-ocorrência, sintáticas e semântica (Amancio 2013). A rede abordada neste estudo é a de co-ocorrência. Este tipo de rede fundamentalmente conecta palavras adjacentes, capturando a maior quantidade possível de ligações relevantes entre as palavras.

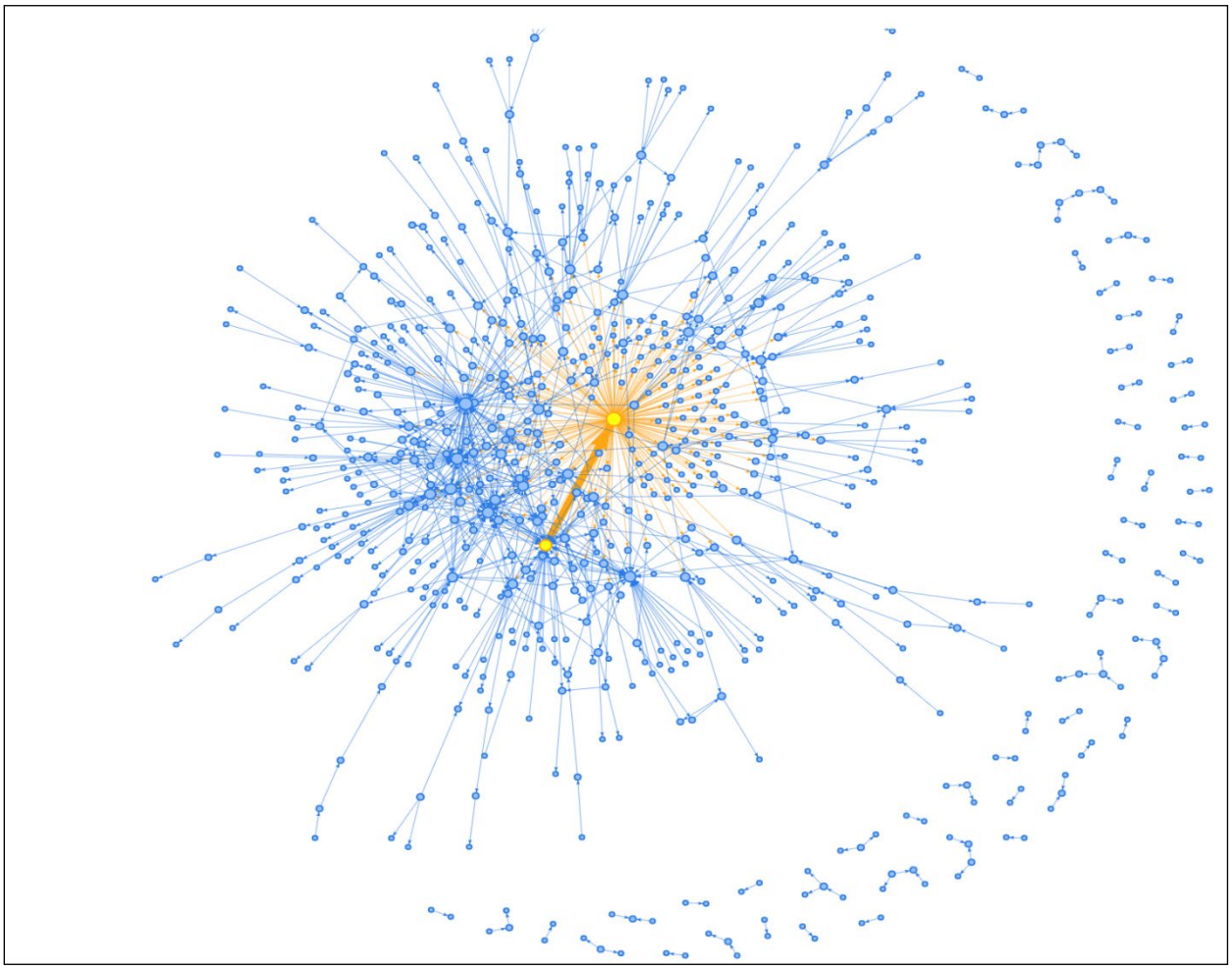
Para uma investigação significativa da rede com a intenção de deixá-la inteligível, apenas as conexões mais expressivas foram contempladas na mesma, ou seja, somente as palavras que tiveram mais de uma conexão, removendo os *loops* que ocorrem quando as palavras estão conectadas a elas mesmas, foram adicionadas à rede. Assim, foi gerada uma rede com 715 nós e 995 arestas[1].

A partir dessa rede (Figura 6) foi possível extrair algumas medidas de centralidade que possibilitam identificar e quantificar a importância de um nó (palavra) ou um grupo de nós. Foram calculadas as medidas de centralidade grau (*degree*), o número de arestas que saem ou entram de um nó, tanto grau de entrada quanto grau de saída pelo fato da rede ser direcionada.

[1]As Figuras 6 e 7 estão em formato digital interativo e podem ser visualizadas nos seguintes links: https://saude360.unifesp.br/projeto/TwitterHealth/febreamarela_grafo_clusters.html e https://saude360.unifesp.br/projeto/TwitterHealth/febreamarela_grafogeral.html. Na figura interativa é possível selecionar os agrupamentos, os nós e visualizar os rótulos dos nós e suas conexões diretas.

Assim, o grau de entrada revela o número de arestas que chegam a um nó e mostra as palavras que receberam mais conexões direcionadas a elas obtendo um maior prestígio na rede. E o grau de saída contabiliza todas as conexões que saem de um nó.

Figura 6 - Visualização da rede de co-ocorrência direcionada de palavras gerada com o conjunto de dados com 715 nós e 995 ligações. Os nós em amarelo representam as palavras “Febre” e “Amarela”



Fonte : dados da pesquisa

Outro conceito de centralidade que foi calculado foi o *betweenness*, essa medida identifica os nós com maior vantagem na rede e influência no controle da distribuição da informação, medindo sua importância relativa na rede (Amancio 2013).

Entre os dez nós com maior grau (*degree*) e valor *betweenness* na rede, excluindo as palavras "Febre" e "Amarela" que compõem o tema foco deste estudo, foram identificados termos chaves das discussões como "Vacina", alguns verbos/advérbios e suas variações que podem indicar que em diversas conversas os usuários estavam expondo o fato de terem sido vacinados ou não ("tomar", "não", "vou", "contra", "hoje").

A palavra "Morte" também foi identificada com um grau alto na rede, assim como a palavra "Rio" que está relacionada a cidade do Rio de Janeiro, esta que foi uma das cidades que tiveram pico de casos entre as semanas 10 e 12. As duas palavras citadas podem ser consideradas focos de grandes discussões nesse conjunto de dados.

Em redes que possuem uma quantidade significativa de nós é possível detectar agrupamentos (*clusters*), ou seja, grupos de nós que são mais conectados entre si do que com o restante da rede. Esses grupos são chamados de comunidade. O coeficiente de agrupamento de uma rede é calculado através da medida de transitividade (Dorow 2000), que tem o objetivo de indicar quão próximo o grafo está de ser um grafo completo, ou seja, de ter todos os nós conectados. Nota-se pela Figura 6 que o grafo não é completo pois há diversos nós que não se conectam, porém é possível notar um componente maior ao centro da rede. O componente maior segue a estrutura do grafo original e inclui a maioria dos nós da rede.

A partir do componente maior, foi utilizado o algoritmo *cluster_edge_betweenness* (https://igraph.org/r/doc/cluster_edge_betweenness.html) do pacote "igraph" do R para identificar os agrupamentos na rede. A Figura 7 mostra a rede com os 12 agrupamentos identificados por diferentes cores.

Em cada agrupamento revelado na Figura 7 é possível analisar as dinâmicas interacionais que ocorrem nos tweets, investigando o contexto nas quais as conversas estavam relacionadas e os diferentes enfoques.

Alguns agrupamentos chamaram a atenção pelo fato de relacionarem diversos conceitos de saúde, como por exemplo, o agrupamento 1 que contemplou diferente doenças ligadas ao nó "Febre" (o nó com maior grau de entrada da rede), como por exemplo: meningite, câncer,

hepatite, dengue e chikungunya. As duas últimas doenças estavam inclusive conectadas entre si e com um outro nó representado pela palavra "zika". É interessante notar como o tema FA foi relacionado a tantas outras doenças, trazendo o fato notório de que o mosquito Aedes Aegypti transmite tanto FA, dengue, chikungunya e zika vírus.

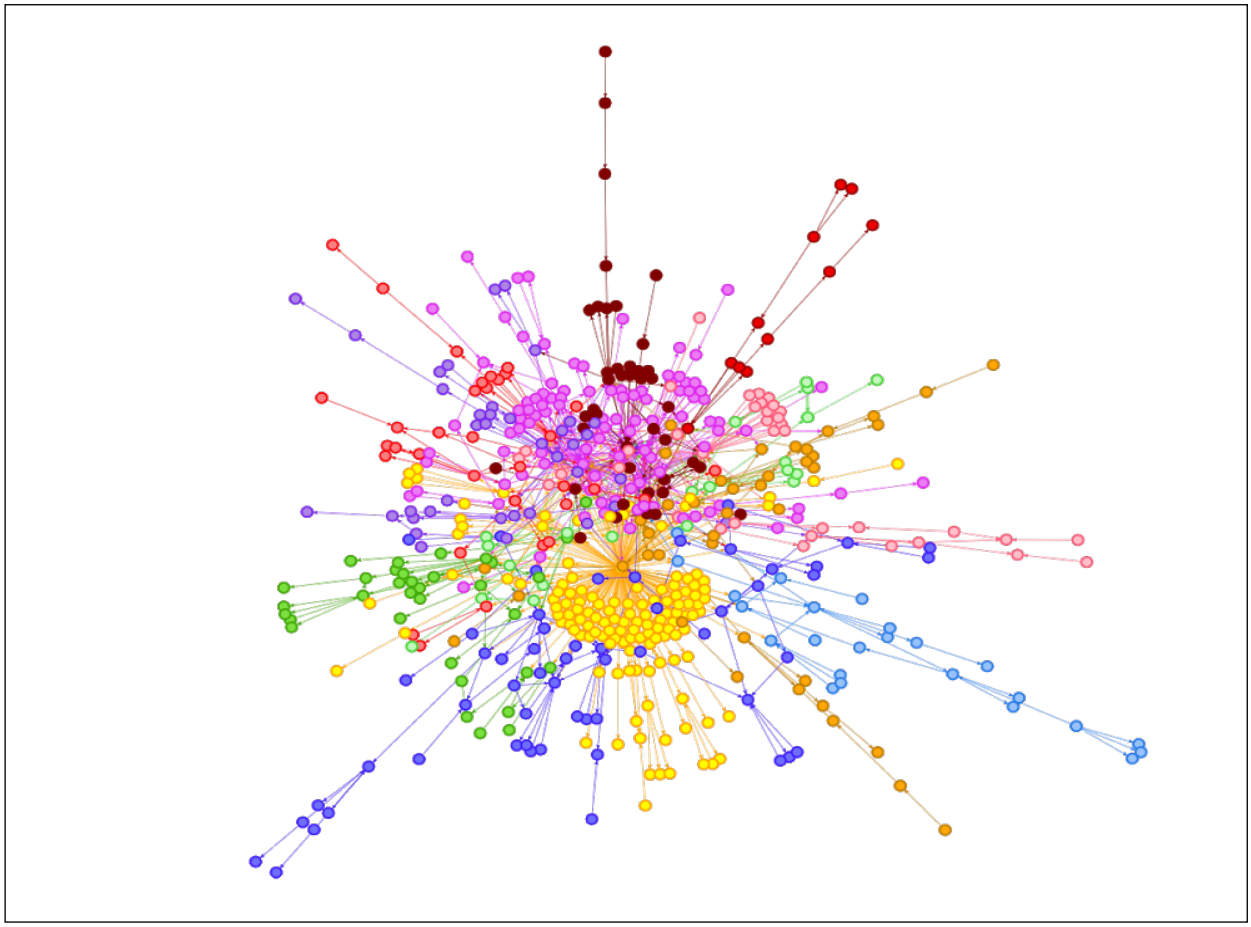


Figura 7 - Visualização do componente maior da rede com os 12 agrupamentos identificados em cores variadas

Fonte: dados da pesquisa

Os agrupamentos 2, 3 e 5 mostraram uma forte relação ao tópico "Vacina", evidenciando algumas palavras que norteiam narrativas sobre campanhas de imunização, prevenção, doses e também incluindo outras doenças como gripe, gripe suína e influenza, doenças que possuem vacinas como método de prevenção.

Em evidência no agrupamento 3, diversos discursos foram identificados sobre tomar ou não a vacina, com os nós "não", "tomar" e "tomei". O fato de que as pessoas publicam no Twitter sobre ter ou não tomado a vacina, pode auxiliar na criação de uma variável de análise de observação e previsão daquelas pessoas que ainda não tomaram a vacina, mais uma ferramenta que pode contribuir no planejamento de campanhas e investigações de possíveis casos.

Neste mesmo agrupamento encontram-se palavras que podem estar relacionadas aos sentimentos dos usuários sobre tomar a vacina, como por exemplo, as palavras "odeio" e "agulha" diretamente conectados e que demonstram uma ênfase de sentimento negativo. Em contrapartida também foram encontrados os nós "não" e "doeu" diretamente conectados e que apresentam a ausência do sentimento de dor no evento de tomar a vacina, este está subentendido. Houve também a verificação de que os usuários também compartilharam outras sensações e sintomas após tomar a vacina, como os nós "braço", "direito", "dormente" e "inchado" encontrados no agrupamento 8.

Assim como o agrupamento 2, o agrupamento 5 também trouxe "Vacinação" como tópico, porém revelou alguns nós com palavras referentes a cuidados/prevenção como; "evitar", "perfumes", "utilizar", "repelentes" e "telas". A palavra "OMS" que se refere à Organização Mundial da Saúde também foi identificada neste agrupamento e diversos links com notícias sobre as instruções da OMS no combate à Febre Amarela foram compartilhados nas mensagens.

Outro agrupamento apresenta associações entre termos que demonstram frustração dos usuários, como a conexão entre os nós "prefiro", "ficar", "doente" e os nós "enfrentar", "fila", "enorme" que indica que alguns estabelecimentos de saúde estavam com filas grandes para aplicar a vacina contra FA.

Nos agrupamentos 10 e 7 foi possível identificar uma discussão importante que houve na época do surto da FA. Um nó com a palavra "macaco" teve ligação com as palavras: "infectados", "encontrados", "mortos", "transmissores" e novamente no agrupamento 12 notam-se nós conectados a "mico-leão-dourado". Na época do surto houve notícias de mortes de macacos infectados por FA e rumores surgiram relacionando o surto com os macacos serem os transmissores da doença. Esses rumores viriam depois a serem contrapostos por diversos

veículos de comunicação, inclusive nas redes sociais. Um exemplo de uma notícia que foi compartilhada em uma das mensagens cujo título era “Febre Amarela provoca caça e morte de macacos” (link compartilhado na mensagem: https://t.co/tTHCCaKYpr), expõe o fato da preocupação das pessoas com relação aos macacos infectados e as denúncias de ataques aos macacos em diversas regiões. É interessante notar, analisando alguns tweets que continham esse tópico, que a maioria dos usuários estava compartilhando informações principalmente com a intenção de negar os rumores e transmitir a informação correta.

Verificou-se também agrupamentos relacionados a notícias que podem ter sido compartilhadas sobre número de casos, mortes, regiões/cidades. Nesses agrupamentos foi possível identificar termos que permitem a localização da narrativa como Minas Gerais, São Paulo, Rio de Janeiro, São Gonçalo, São Fidélis e Vitória, que foram regiões que constataram diversos casos no período.

Há agrupamentos mais amplos com maior quantidade de nós como o agrupamento 6, com o nó em destaque com a palavra "Amarela" (o nó com maior grau rede), se conectando com os mais variados nós incluindo palavras relacionadas a doenças (malária, gastrite e tétano).

Essas associações mostram a variabilidade de discursos que um tema de saúde pode gerar em uma rede social. Identificar esses diferentes tópicos podem auxiliar na sumarização dos elementos narrativos que são discutidos no Twitter, aumentando o conhecimento sobre um determinado assunto e evidenciando características dos discursos que podem ser monitoradas e investigadas a fundo. Apesar da mensagem do Twitter ser um texto curto, a análise exploratória de grafos em um conjunto grande de mensagens é bastante útil para compreensão de uma narrativa ampla, auxiliando na identificação dos assuntos discutidos de forma simplificada.

5 Conclusões

As redes sociais vêm fornecendo cada vez mais dados e informações que podem auxiliar na compreensão da opinião popular e dos eventos que ocorrem ao redor do mundo. Todavia, extrair e analisar esses dados de forma automática e eficiente tem sido um desafio. Essa

dificuldade se dá por diversos fatores entre eles o grande volume de dados envolvidos e a manipulação de dados textuais, ou seja, não estruturados. Apesar disso, diversas técnicas de recuperação de informação e mineração de dados têm sido amplamente utilizadas com resultados satisfatórios.

Este artigo utilizou técnicas de mineração de dados e análise exploratória de grafos para extrair e investigar assuntos que estavam sendo tratados sobre Febre Amarela em diversas mensagens publicadas na rede social Twitter durante o surto de 2017. Essa abordagem auxiliou na seleção e no entendimento dos dados identificando palavras-chaves das discussões e através de suas conexões possibilitou descobrir contextos e padrões nos discursos analisados.

A identificação de mensagens sobre o fato dos usuários compartilharem se tomaram ou não a vacina de Febre Amarela pode servir para criar uma estratégia mais escalável de análise da evolução desses assuntos em tempo real, auxiliando no planejamento de campanhas de conscientização ou melhorias no sistema para atender melhor a população.

Outros elementos observados foram os termos que permitiram a localização da narrativa. Apesar de não ter sido possível realizar uma análise por geolocalização dos tweets, foi possível identificar algumas cidades e estados em referências encontradas em alguns termos publicados nas mensagens que eram justamente regiões com grande número de casos notificados no período.

Em suma, as redes sociais são, cada vez mais, canais de comunicações com grande potencial de informações atuais sobre a sociedade, inclusive sobre saúde. Através de novos dados, novas perguntas podem ser feitas com intuito de adquirir mais informações que podem ser consideradas uma fonte complementar para vigilância em saúde pública. Explorar ferramentas para analisar estes dados de forma automatizada e intuitiva irá beneficiar tanto a área de vigilância em saúde quanto a informática em saúde do consumidor. Esta pesquisa fornece uma estrutura para servir como referência na coleta e análise de dados obtidos em redes sociais online, e com isso pode contribuir metodologicamente com demais pesquisas da área da Ciência da Informação.

Notes

(1) O presente estudo foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES).

Referências

Amancio, Diego Raphael. *Classificação de textos com redes complexas*. 2013. Instituto de Física de São Carlos, Universidade de São Paulo, Tese de doutorado.

Araújo, Gabriela Denise., et al. "Sentiment Analysis of Twitter's Health Messages in Brazilian Portuguese". *J. Health Inform*, vol. 10, no. 1, p. 17-24, 2018.

Appel, Ana Paula. *Métodos para o pré-processamento e mineração de grandes volumes de dados multidimensionais e redes complexas*. 2010. Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Tese de doutorado.

Bardin, Laurence. *Análise de conteúdo.* São Paulo: Edições 70, 2016.

Beykikhoshk, A., et al. "Using Twitter to Learn about the Autism Community". *Soc. Netw. Anal. Min.*, vol. 5, no. 22, 2015. https://doi.org/10.1007/s13278-015-0261-5

Boldrini, Angela. "Surto de febre amarela no Brasil é o maior de série histórica, desde 1980". *Folha de S. Paulo,* São Paulo, 26 jan. 2017.

Brasil, Ministério da Saúde, Secretaria de Vigilância em Saúde. *Informe especial febre amarela no Brasil nº 01/2017,* Saúde, 2017a, portalarquivos.saude.gov.br/images/pdf/2017/marco/18/Informe-especial-COES-FA.pdf

Brasil, Ministério da Saúde, Secretaria de Vigilância em Saúde, Centro de Operações de Emergências em Saúde Pública sobre Febre Amarela. *Informe nº43/2017b*, 2017b, portalarquivos.saude.gov.br/images/pdf/2017/junho/02/COES-FEBRE-AMARELA---INFORME-43---Atualiza----o-em-31maio2017.pdf

Brasil, Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância de Doenças Transmissíveis. *Febre amarela silvestre no Brasil*, 2017c, www.saude.gov.br/images/pdf/2017/junho/22/1.%20b%20-%20FA_CIT_22.06.2017.pdf

Dorow, Beate. *A Graph Model for Words and their Meanings*, 2006. Universität Stuttgart, PhD Dissertation.

Grover, Purva, et al. “Technology enabled Health” – Insights from twitter analytics with a socio-technical perspective. *International Journal of Information Management*, vol. 43, p. 85–97, 2018.

Guimarães, Raphael Mendonça, et al. “Os desafios para a formulação, implantação e implementação da Política Nacional de Vigilância em Saúde”. *Ciênc. saúde coletiva*, Rio de Janeiro, vol. 22, no. 5, p. 1407-1416, 2017.

Hoffman, L. Beth., et al. “It’s not all about autism: The emerging landscape of anti-vaccination sentiment on Facebook”. *Vaccine*, vol. 37, p. 2216-2223, 2019.

Klein, Gisiela Hasse, Guidi Neto, Pedro and Tezza, Rafael. “Big Data e mídias sociais: monitoramento das redes como ferramenta de gestão”. *Saúde soc.*, vol.26, no.1, p.208-217, Mar 2017, ISSN 0104-1290. http://dx.doi.org/10.1590/s0104-12902017164943.

Lakatos, Eva Maria. *Fundamentos de metodologia científica*. 8th ed. São Paulo: Grupo Gen - Atlas, 2017.

Miller, Wendy R., et al. “Word Adjacency Graph Modeling: Separating Signal From Noise in Big Data.” *Western Journal of Nursing Research*, vol. 39, no. 1, pp. 166–185, Jan. 2017, https://doi.org/10.1177/0193945916670363

Sanders-Jackson, A., et al. “Applying linguistic methods to understanding smoking-related conversations on Twitter”. *Tobacco Control*, vol. 24, p. 136-138, 2015.

Stefanidis, A. et al. “Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts”. *JMIR Public Health Surveill,* vol. 3, no. 2, 2017.

Tangherlini, R. T. et al. “Mommy Blogs” and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites. *JMIR Public Health Surveill,* vol. 2, no. 2, p. 166, 2016.

Recived: 12/05/2020 Acepted: 09/07/2020