Mineração de tópicos e aspectos em microblogs sobre Dengue, Chikungunya, Zika e Microcefalia

Mateus Tarcinalli Machado[1,3,4], **Jéssica Caroline Alves Nunes Temporal**[5], **Thiago Alexandre Salgueiro Pardo**[2,4] , **Evandro Eduardo Seron Ruiz**[3,4],

[1] Programa de Pós-Graduação em Computação Aplicada

[2]Departamento de Ciências de Computação, ICMC

[3]Departamento de Computação e Matemática, FFCLRP

[4]Universidade de São Paulo (USP), Brasil

[5]Data Science Brigade, https://datasciencebr.com

jessicatemporal@gmail.com,{mateusmachado,evandro}@usp.br

taspardo@icmc.usp.br

Abstract. *The proper analysis of opinion texts, as those posted on microblogs and social networks, includes the identification of the topic that is commented by the author of the text. The analysis of the topics may be carried out by a set of techniques for finding what we call 'aspect terms'. This paper shows how the identification of aspect terms in microblogs written in Portuguese may be achieved with frequency-based methods and word vector representation approaches (word2vec). We have obtained a list of n-grams that we believe are adequate indicators of the topics commented by the users. We focus our work on texts related to Dengue, Chikungunya and Zika diseases, as well as Microcephaly, which represent serious threatens nowadays.*

Resumo. *A correta análise de textos opinativos, incluindo aqueles postados em microblogs e redes sociais, passa pela identificação do tópico comentado pelo autor do texto. A análise dos tópicos pode ser realizada por um conjunto de técnicas para a identificação do que chamamos de 'termos de aspectos'. Neste artigo, mostramos como a identificação de termos de aspectos em microblogs em Português pode ser alcançada por métodos baseados em frequência e pela representação vetorial de palavras (word2vec). Obtivemos uma lista de n-gramas que acreditamos que sejam indicadores adequados dos tópicos comentados. Focamos nosso trabalho em textos sobre Dengue, Chikungunya e Zika, assim como Microcefalia, que atualmente são sérias ameaças à saúde.*

1. Introdução

Atualmente, as redes sociais apresentam-se como ferramentas de amplo acesso para a produção de conteúdo opinativo sobre assuntos variados [Cataldi et al. 2010]. Consequentemente, a análise do conteúdo destes textos sobre os mais diversos temas, como economia [Bollen et al. 2011], política [Tumasjan et al. 2010] e saúde [Moorhead et al. 2013, Almansa et al. 2014], tem sido objeto de interesse da área

de Processamento de Linguagem Natural (PLN), mais especificamente das áreas de Análise de Sentimentos (AS) e mineração de opiniões.

A análise textual tem sido utilizada para determinar quais aspectos de produtos ou serviços os usuários estão elogiando ou desaprovando. Como produtos e serviços populares podem receber centenas de comentários, tanto nos sites especializados como também nas redes sociais, a análise automatizada dos sentimentos e da opinião expressa nestes textos tem se tornado um tópico de interesse crescente na pesquisa acadêmica [Hu and Liu 2004a, Hu and Liu 2004b].

A literatura científica documenta o uso de métodos de processamento de textos para analisar, principalmente, duas questões fundamentais em textos opinativos, que são a identificação do tópico comentado e do sentimento associado a este tópico.

A identificação do tópico (ou assunto) comentado é um problema reconhecidamente importante na área de PLN, que pode ser tratado pela tarefa de "extração de aspecto" (EA) [Cataldi et al. 2010]. Cabe ressaltar que o problema de EA transcende a variabilidade de mídias. Em artigo recente, Gandomi e Haide [Gandomi and Haider 2015] citam a importância da pesquisa em EA para facilitar a localização, em textos médicos, de referências a medicamentos, drogas, doenças, entre outros.

O presente artigo procura analisar alguns dos tópicos de saúde que são mais comentados na plataforma de microblogs Twitter no Brasil. Mais especificamente, analisaremos quais os tópicos comentados sobre as doenças Dengue, Chikungunya e Zika, assim como Microcefalia, que são grandes ameaças à saúde atualmente. Acreditamos que os termos utilizados para se referir aos aspectos relacionados a esses problemas de saúde constituem tópicos relevantes nos textos. Por exemplo, no *tweet "essa dengue ta me matando"*, esperamos identificar o termo 'dengue' como o aspecto de referência neste comentário. Tal iniciativa, se aplicada em larga escala sobre os comentários publicados na web e com precisão suficiente, pode permitir a identificação e rastreamento de epidemias emergentes e a reação apropriada de órgãos de saúde, sendo uma contribuição real para a sociedade.

2. Trabalhos relacionados

Em trabalho anterior deste grupo [Almansa et al. 2014], analisamos os textos postados na plataforma de microblogs Twitter, *tweets*, durante a Campanha Nacional de Vacinação contra a Gripe em 2014. Neste estudo, categorizamos os *tweets* em três eixos distintos, ou seja, sobre o conteúdo, sobre a qualificação e os *tweets* que apontavam para outras URLs. No entanto, não pudemos discernir nestes comentários quais aspectos eram preponderantes nas postagens, por exemplo, se eram comentários sobre a vacina ou sobre o processo de vacinação, ou mesmo se os comentários recaiam sobre a doença em si (a gripe, neste caso).

Um dos trabalhos mais relevantes da literatura que aborda esta divisão entre os tópicos de EA e de análise de sentimento é o trabalho de Pavlopoulos [Pavlopoulos and Androutsopoulos 2014]. Neste trabalho, o autor compara os resultados obtidos por quatro métodos automatizados de EA com o objetivo de determinar os assuntos mais comentados por usuários. São os métodos: 1) `FREQ` Baseline; 2) o Método de Hu&Liu; 3) o W2V, ou *word2vec* [Mikolov et al. 2013a]; e, finalmente, 4) o conhecido

Latent Dirichlet Allocation, ou LDA [Blei et al. 2003]. Esses quatro métodos foram utilizados sozinhos ou em pares sobre três conjuntos de dados em Inglês: a) um *dataset* de análise de serviços de restaurantes; b) um conjunto de comentários sobre hotéis contendo 3.600 frases; c) e um conjunto de 3.085 frases de análise de laptops. Entre outros resultados, o autor mostra como a metodologia empregando W2V foi adequada na detecção de aspectos para a avaliação de sentimentos nos vários conjuntos de dados testados.

Notamos que a multiplicação da quantidade de textos opinativos trouxe como consequência o aumento da quantidade de termos de aspecto [Liu 2012]. Esta quantidade de textos opinativos certamente contribuiu para a expansão da área de análise automatizada de sentimentos em textos. A identificação do sentimento no texto funde-se com a extração de termos de aspecto, pois os sentimentos quase sempre estão vinculados a um aspecto particular do objeto comentado. Como exemplo, no excerto "a vacina é dolorida" o adjetivo qualifica o aspecto 'vacina', enquanto que, no fragmento "o período de vacinação foi curto", a qualificação refere-se ao termo composto de aspecto 'período de vacinação'. Num nível mais refinado, o da análise de sentimento baseada em aspecto, *Aspect-Based Sentiment Analysis* (ABSA), a classificação de sentimentos não ocorre para uma frase como um todo, mas para cada aspecto desta, e oferece assim um maior grau de refinamento ao processo. Neste contexto, em outro trabalho relevante, Hui Lek e sua equipe [Lek and Poo 2013] discutem a análise de sentimentos baseada em aspectos do conteúdo exclusivo de *tweets*. Um passo muito importante nessa linha é justamente a EA, que busca listar os aspectos mencionados em um dado *tweet* ou num conjunto destes.

Conrado [da Silva Conrado et al. 2013] e sua equipe trabalharam na extração automática de termos. O seu trabalho focou na utilização de métodos de aprendizado de máquina e atributos de enriquecimento sobre textos em Inglês e mostrou um incremento na taxa de recuperação de termos com o apoio de atributos linguísticos e estatísticos. Trabalhos como esse, de extração de termos, são relevantes porque, em análise de sentimentos, os aspectos comentados normalmente são termos.

O diferencial deste artigo em relação aos demais está na busca pela contextualização dos termos de aspecto e na aplicação em microblogs na língua Portuguesa. Para cada termo de aspecto em potencial, procuramos analisar se o contexto deste termo está relacionado à área de saúde através da abordagem W2V [Mikolov et al. 2013a] aplicada a termos vizinhos. Veremos mais detalhes desta análise de contexto na Seção 5.

3. Objetivos

Dado um conjunto de comentários, em Português, sobre doenças, na plataforma de microblogs Twitter, o objetivo principal deste trabalho é explorar métodos para encontrar uma lista dos tópicos mais citados que estão contextualizados na área de saúde. Para atingirmos este fim, restringimos o trabalho à comentários sobre as doenças transmitidas pelo mosquito *Aedes aegypti*, ou seja, Dengue, Chikungunya e Zika, assim como a condição relacionada de Microcefalia. Detalhamos este objetivo nas seguintes tarefas (ver Figura 1):

1. Recuperar um conjunto de dados brasileiros, em Português, do Twitter, sobre os seguintes temas da área de saúde: Dengue, Chikungunya, Zika e Microcefalia;
2. Realizar a rotulação/classificação morfossintática por meio de um etiquetador morfossintático (POS *tagger*);

3. Pré-processar os textos destes *tweets*, removendo as *stopwords*, termos menores que quatro caracteres e termos formados por caracteres especiais ou numerais;
4. Fazer uma poda frequencial dos termos de aspecto pelo método `FREQ` Baseline;
5. Contextualizar os termos de aspecto simples e compostos (bigramas e trigramas) com a abordagem W2V.

Em síntese, o objetivo deste trabalho é obter uma lista de termos em Português relacionados aos aspectos comentados nos *tweets* e contextualizados na área de saúde. Estes termos poderão ajudar a identificar quais os temas ou os assuntos prevalentes estão sendo comentados nesta plataforma de microblogs. A lista de termos obtida poderá contribuir para (i) conhecermos um conjunto de termos de aspectos da área da saúde sobre os problemas de saúde citados e (ii), futuramente, subsidiar o rastreamento na web de epidemias emergentes, propiciando uma reação mais rápida dos órgãos de saúde.

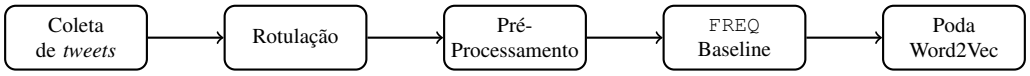

Figura 1. Etapas de processamento neste trabalho.

4. Materiais e métodos

O Laboratório de Sistemas Computacionais Complexos (LSCC) do Departamento de Computação e Matemática da Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, USP, coleta, desde novembro de 2014, *tweets* sobre temas relacionados à saúde a partir de uma lista de termos. Entre estes termos, estão os seguintes: H1N1, Zika, Microcefalia, Gripe, Vacina, Vacinação, Febre, Dor, Alergia, Dengue, Chikungunya, Diarréia, Aids, HIV e doação de sangue. Atualmente, esta base de dados conta com mais de 43 milhões de *tweets*. Aproximadamente 10% destes *tweets* (3.097.294) foram processados neste trabalho.

A classificação morfossintática dos constituintes de uma frase é o reconhecimento das classes gramaticais de suas palavras. São exemplos destas classes: adjetivos, artigos, verbos e substantivos, entre outras [Bird et al. 2009]. O toolkit NLTK [Bird et al. 2009] permite definir vários rotuladores para esta função. Esses rotuladores são formados através do treinamento sobre um conjunto de textos já anotados, os quais chamamos de *corpus*. Da base de dados do NLTK, usamos um rotulador treinado no *corpus* Floresta Sintática [Freitas et al. 2008]. Este *corpus* é composto de textos em português do Brasil e de Portugal previamente rotulados.

O pré-processamento tem como objetivo eliminar os termos muito frequentes ou que podem não agregar contexto ao documento. Para tanto, nos valemos de tarefas simples para esta fase: remoção de *stopwords*; remoção de termos menores que quatro caracteres e termos formados por caracteres especiais ou numerais.

O foco da extração de termos de aspecto é a identificação dos termos de interesse no texto. Existem dois tipos de aspectos, os explícitos e os implícitos. Os primeiros são aqueles que estão sendo explicitamente mencionados na frase. Por exemplo, na oração composta *"Peguei dengue e estou com febre alta"*, o termo 'febre' é um aspecto explícito. Os aspectos implícitos são aqueles que não são diretamente mencionados na frase. Por

exemplo, na oração *"A vacina contra dengue é cara"*, temos como aspecto implícito o termo 'preço', que é inferido pelo qualificador 'cara' [Liu 2012]. O cerne deste trabalho está na identificação dos aspectos explícitos.

Para a EA, iniciamos pela aplicação do método `FREQ` Baseline, que é um método que retorna os substantivos mais frequentes (sem repetição). Os termos são dispostos numa lista ordenada de forma decrescente pela frequência. Este é um método simples, mas que tem demonstrado efetividade em vários trabalhos de pesquisa [Hu and Liu 2004a, Wei et al. 2010, Liu 2012].

O método `FREQ` Baseline foi estendido com uma etapa de poda que utiliza representações dos termos em espaço vetorial contínuo, segundo o modelo W2V [Mikolov et al. 2013b]. O objetivo é remover termos que, mesmo tendo frequências relativamente altas nos textos analisados, não tenham relação com o contexto estudado, ou seja, não pertençam ao contexto dos *tweets* das doenças em questão. Essa representação vetorial dos termos pode ser produzida, por exemplo, através do treinamento de um modelo de linguagem que retorne termos similares a um termo alvo, ou também a partir de uma lista de termos que retorne outro termo similar. Cada termo é representado por um vetor denso dentro de um espaço vetorial contínuo. Cabe enfatizar que, neste mapeamento pelo W2V, termos com semântica similar são mapeados como representações vetoriais próximas umas das outras. Assim, dizemos que os termos semanticamente similares são *embedded nearby each other*. O modelo obtido através deste treinamento é representado por um conjunto de vetores, ao qual chamamos de modelo de semântica distribucional, ou *word embeddings*, em inglês. Este modelo pode ser obtido através do processamento de grandes conjuntos de dados, geralmente compostos por bilhões de palavras, através de treinamento em rede neural. Comumente, em trabalhos semelhantes, são utilizados os dados da *Wikipedia* para a criação desses modelos. Para o Português, foi disponibilizado um modelo treinado pela equipe de João Rodrigues [Rodrigues et al. 2003], que utilizou, além dos dados da *Wikipedia*, outros *corpora* que contêm artigos científicos e jornalísticos, palestras, legendas de filmes e textos técnicos, entre outros documentos, com um total próximo a 1,7 bilhões de palavras.

5. Resultados

Do conjunto de *tweets* descrito anteriormente, recuperamos todos aqueles que apresentam os termos Dengue, Chikungunya, Zika e Microcefalia, referentes aos anos de 2015 e 2016. Escolhemos estes termos devido à grande repercussão que esses problemas de saúde tiveram no cenário brasileiro nos últimos anos. A Tabela 1 mostra o número de mensagens recuperadas em cada ano e os totais parciais por classe de mensagem.

Tabela 1. Número de *tweets* recuperados e analisados para os termos em foco.

Termo	**2015**	**2016**	**Total**
Dengue	922.720	784.330	1.707.050
Chikungunya	44.786	129.186	173.972
Zika	36.130	1.331.898	1.368.028
Microcefalia	5.282	192.992	198.274
Total	**1.008.918**	**2.438.406**	**3.447.324**

Nos *tweets* recuperados, foi aplicado um processo de separação de termos, ou

tokenização, do pacote NLTK [Bird et al. 2009], conhecido como o `TweetTokenizer`. Em seguida, foi aplicado o rotulador morfossintático treinado no modo *bigram backoff* [Freitas et al. 2008], o qual obteve o melhor desempenho em trabalho recente deste mesmo grupo [Temporal 2016].

5.1. Desvendando os unigramas

As etapas de processamento para encontrar os unigramas são comentados abaixo.

Pré-processamento Após a identificação dos substantivos através da rotulação, foram removidas as *stopwords*. Também foram removidos *tokens* formados por números ou caracteres especiais e termos com menos de quatro caracteres. As quantidades de potenciais termos de aspecto para cada doença estão mostrados na Tabela 2 sob o rótulo 'Pré-Proc'. É notável o crescimento no número de termos no ano de 2016 para todas as doenças, exceto a Dengue. Este número crescente de termos acompanha o número crescente de *tweets* recuperados e mostrados na Tabela 1.

`FREQ` Baseline Um forte indicativo de que o substantivo determinado é relevante dentro do contexto estudado é a sua frequência no texto. Termos pouco frequentes foram eliminados nessa etapa, pois muito provavelmente não tem relação com o contexto ou possuem pouca relevância. Filtramos todos os termos com frequência $\leq 0,1\%$ para cada ano estudado. Veja na Tabela 2 o resultado desta poda. Repare que a quantidade de termos é equiparável em ordem de grandeza para todas as doenças e nos dois anos.

`FREQ` Baseline + Poda com W2V Empregamos os vetores de contexto como representativos ora dos textos analisados, *tweets*, ora dos textos próximos à linguagem comum. Para realizarmos essa poda, foram criados dois vetores:

1. O vetor de contexto está relacionado aos 10 termos mais frequentes do conjunto de *tweets* referentes aos problemas de saúde analisados. Este limiar foi encontrado com base nos resultados do método `FREQ` Baseline;
2. O vetor comum, pois foi calculado sobre os 20 termos mais frequentes no *corpus* das obras completas do Machado de Assis no NLTK.

Cada termo de cada um desses *corpora* foi convertido para um vetor obtido através do modelo distribucional da equipe de João Rodrigues [Rodrigues et al. 2003], formando assim duas matrizes de vetores. Para cada uma dessas matrizes, foi calculado seu centroide (média de cada um de seus elementos), tendo como resultado um único vetor para cada *corpus*, ou seja, cada *corpus* representa um contexto e cada contexto é representado por um vetor. Depois, cada termo obtido após a aplicação do método `FREQ` Baseline foi então comparado com esses vetores através do cálculo de similaridade por cosseno. Termos mais próximos do vetor comum foram descartados, enquanto os termos mais próximos do vetor de contexto foram mantidos, pois foram considerados como termos de aspecto contextualizado para os problemas de saúde em foco. Finalmente, os número de termos contextualizados com a abordagem W2V para cada problema são mostrados na última coluna da Tabela 2. Nota-se a grande redução em relação às quantidades obtidas depois da aplicação dos passos anteriores.

5.2. Bigramas e trigramas

A abordagem W2V também foi usada para encontrar termos de aspecto compostos, ou seja, os bigramas e os trigramas.

Tabela 2. Número de termos encontrados nos *tweets*.

Tema dos Tweets	Ano	Pré-Proc.	Freq	W2V
Dengue	2015	69.908	690	84
	2016	67.459	673	84
Chikungunya	2015	7.951	575	80
	2016	17.952	626	110
Zika	2015	8.696	502	66
	2016	105.996	661	73
Microcefalia	2015	2.139	455	58
	2016	18.321	705	104

O processo de identificação dos bigramas e trigramas foi similar ao processo de detecção de aspectos simples (unigramas), descartando-se apenas a etapa de rotulação. Após o pré-processamento, foram identificados todos os bigramas de cada *tweet*, e os mesmos foram ordenados de forma decrescente de acordo com suas frequências. Na aplicação do `FREQ` Baseline, foram removidos os bigramas com frequência $\leq 0,1\%$ dos *tweets* do conjunto de interesse. Por último foi realizada uma poda utilizando a abordagem W2V, em que cada termo foi analisado individualmente para encontrar o seu contexto. Assim, só foram mantidos os bigramas em que ambos os termos estivessem relacionadas ao contexto de saúde. A Tabela 3 mostra os resultados parciais após cada etapa do processamento de EA.

Tabela 3. Número de bigramas encontrados nos *tweets*.

Tema dos Tweets	Ano	Bigramas			Trigramas		
		Pré-Proc.	Freq	W2V	Pré-Proc.	Freq	W2V
Dengue	2015	755.976	630	79	1.133.998	299	25
	2016	716.364	720	77	1.050.112	478	40
Chikungunya	2015	40.716	1.196	114	51.092	1.201	105
	2016	114.164	1.160	142	145.246	1.076	120
Zika	2015	43.568	888	97	52.964	918	92
	2016	1.117.732	638	75	1.590.669	433	42
Microcefalia	2015	7.667	877	95	9.055	865	60
	2016	125.449	1.286	148	167.958	1.157	107

Para a identificação dos trigramas contextualizados, a abordagem usada foi análoga à abordagem de extração de bigramas. Veja na tabela acima os resultados obtidos.

5.3. Listas parciais dos termos de aspecto

Como ilustração dos resultados desse trabalho, o gráfico da Figura 2 mostra a quantidade de alguns dos bigramas ao longo de 24 meses de coleta. Observa-se que o número de mensagens relacionadas à dengue aumentou nos períodos de fevereiro a maio de 2015 e dezembro de 2015 a abril de 2016, épocas em que ocorreram aumentos dos números de casos da doença. Os termos zika, chikungunya e microcefalia foram mais mencionadas no ano de 2016, quando houve grande surto desse problemas, ano em que também se

observa aumento nas menções ao mosquito da dengue. Um ponto importante a ser observado são as menções ao combate à dengue, que também acompanharam a evolução da quantidade de casos da doença, indicando uma possível necessidade de se divulgar com mais antecedência as campanhas preventivas.

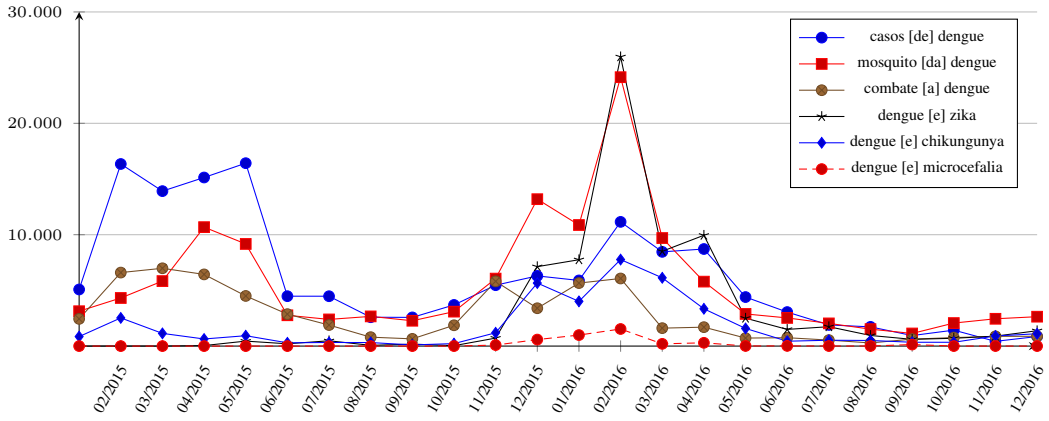

Figura 2. Bigramas por mês.

A seguir, apresentamos listagens de termos extraídos. As listas são apresentadas da seguinte forma: os termos sem formatação apareceram somente nos *tweets* de 2015; os termos sublinhados pertencem aos *tweets* de 2016; os termos em negrito pertencem a ambos os anos. Os termos entre chaves indicam as *stopwords* retiradas no pré-processamento.

Como exemplo de unigramas extraídos, seguem os 20 unigramas mais frequentes em relação à Dengue: (**dengue**) (**casos**) (**mosquito**) (zika) (**saúde**) (**chikungunya**) (**combate**) (**epidemia**) (**vacina**) (**vírus**) (**aedes**) (número) (estado) (**doença**) (**secretaria**) (**focos**) (**mutirão**) (água) (mortes) (**sintomas**)

Seguem os 20 bigramas mais frequentes em relação à Dengue: (**casos [de] dengue**) (**mosquito [da] dengue**) (**combate [a] dengue**) (dengue [e] zika) (**epidemia [de] dengue**) (**dengue [e] chikungunya**) (zika [e] dengue) (**aedes aegypti**) (zika [e] chikungunya) (**combater [a] dengue**) (**secretaria [de/da] saúde**) (número [de] casos) (**casos confirmados**) (zika vírus) (**sintomas [de/da] dengue**) (**focos [de/da] dengue**) (**dengue hemorrágica**) (chikungunya [e] zika) (**confirmados [de] dengue**) (**surto [de] dengue**)

Seguem os 20 trigramas mais frequentes em relação à Dengue: (dengue [,] zika [e] chikungunya) (número [de] casos [de] dengue) (**casos confirmados [de] dengue**) (dengue [,] chikungunya [e] zika) (**casos [de] dengue confirmados**) (zika [,] dengue [e] chikungunya) (**combate [ao] mosquito [da] dengue**) (dengue [e] zika vírus) (dengue [e] zika virus) (**combater [o] mosquito [da] dengue**) (zika virus [e] chikungunya) (chikungunya [e] zika vírus) (zika [,] chicungunha [e] dengue) (**mosquito transmissor [da] dengue**) (sintomas [da] dengue [e] zika) (**focos [do] mosquito [da] dengue**) (**mosquito aedes aegypti**) (risco [de] epidemia [de] dengue) (zika vírus [e] chikungunya) (combater [a] dengue [e] chikungunya)

6. Considerações finais

Neste trabalho, analisamos uma grande quantidade de *tweets* (cerca de 3,4 milhões), buscando extrair os principais termos de aspectos discutidos na rede social *Twitter* sobre Dengue, Chikungunya, Zika e Microcefalia, que são problemas de saúde que tiveram grande repercussão no cenário brasileiro nos últimos anos. No pré-processamento, eliminamos alguns ruídos, através da supressão das *stopwords*, termos menores que quatro caracteres e termos formados por símbolos ou números. Realizamos uma análise morfossintática das postagens, buscando os substantivos mais relevantes com o método `FREQ` Baseline, que retornou uma lista com as frequências de ocorrência desses termos. Também foram criadas listas frequenciais dos bigramas e trigramas encontrados nos *tweets*. E, por último, eliminamos termos não relacionados ao contexto das doenças, utilizando um modelo de semântica distribucional: o *word2vec*.

Apesar da simplicidade do método `FREQ` Baseline na Extração de Aspectos, pudemos observar que, com a adição de uma etapa de poda utilizando o *word2vec*, houve uma significativa melhora na qualidade dos resultados. Termos não relacionados aos contextos estudados são frequentes em textos obtidos a partir de redes sociais, devido à liberdade que os usuários têm de escrever aquilo que pensam. O *word2vec* possibilitou a eliminação desses termos indesejados, permitindo, assim, a construção de listas de aspectos realmente relacionadas aos temas em estudo.

Os métodos explorados mostraram-se satisfatórios para a tarefa, podendo ser aplicados na análise de outras doenças e condições de saúde. Como resultado, tivemos uma visão geral daquilo que foi mencionado na rede social *Twitter*, observando a evolução dos problemas, além de uma percepção sobre a efetividade e o alcance das campanhas preventivas. Espera-se que, futuramente, pesquisas como essa possam, com base em análise de textos da web, subsidiar o acompanhamento de epidemias em tempo real, de forma a se ter mecanismos mais eficazes de intervenção na saúde pelos órgãos competentes.

Agradecimentos

À FAPESP, pelo apoio a este trabalho.

Referências

Almansa, L. F., Machado, M. T., Bosco, G. G., Merlo, E. M., and Ruiz, E. E. S. (2014). Information Learned from Monitoring Microblogs during the 2014 Seasonal Flu Vaccination in Brazil. In *e-Science (e-Science), 2014 IEEE 10th International Conference on*, volume 2, pages 65–66. IEEE.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM.

da Silva Conrado, M., Pardo, T. A. S., and Rezende, S. O. (2013). A machine learning approach to automatic term extraction using a rich feature set. In *HLT-NAACL*, pages 16–23.

Freitas, C., Rocha, P., and Bick, E. (2008). Um mundo novo na Floresta Sintá (c) tica – o treebank do Português. *Calidoscópio*, 6(3):142–148.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

Hu, M. and Liu, B. (2004a). Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM.

Hu, M. and Liu, B. (2004b). Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Lek, H. H. and Poo, D. C. (2013). Aspect-based Twitter sentiment classification. In *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, pages 366–373. IEEE.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Corrado, G., Chen, K., and Dean, J. (2013b). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., and Hoving, C. (2013). A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research*, 15(4):e85.

Pavlopoulos, J. and Androutsopoulos, I. (2014). Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. *Proceedings of LASMEACL*, pages 44–52.

Rodrigues, J., Branco, A., Neale, S., and Silva, J. (2003). LX-DSemVectors: Distributional Semantics Models for Portuguese. *6th International Workshop PROPOR'2003, Faro, Portugal, June 2003*, 8775(2721):214–219.

Temporal, J. C. A. N. (2016). Identificação de entidades mecionadas para análise de sentimentos em microblogs. Monografia (Bacharel em Informática Biomédica), FFCLRP, Universidade de São Paulo, Brazil.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178–185.

Wei, C.-P., Chen, Y.-M., Yang, C.-S., and Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*, 8(2):149–167.