Análise Exploratória das Dúvidas sobre a COVID-19 Publicadas no Twitter

Tiago de Melo[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Manaus, AM – Brasil

tmelo@uea.edu.br

Abstract. *People commonly post health questions on social media. In this work, a method based on topic modeling was developed to discover the main issues related to COVID-19 published on Twitter. A model was also developed to identify the main entities in four different categories. The findings may help politicians and health organizations to understand the main questions about COVID-19.*

Resumo. *As pessoas comumente postam dúvidas sobre saúde em redes sociais. Neste trabalho foi desenvolvido um método baseado em modelagem de tópicos para descobrir as principais questões relacionadas à COVID-19 publicadas no Twitter. Também foi desenvolvido um modelo para identificar as principais entidades de diferentes categorias. As descobertas podem ajudar os políticos e organizações de saúde no entendimento das principais dúvidas sobre a doença.*

1. Introdução

Devido a propagação da COVID-19 no mundo, as plataformas de mídias sociais tornaram-se locais onde ocorre uma intensa e contínua troca de informações entre órgãos governamentais, profissionais da área de saúde e o público em geral. Um representativo número de estudos científicos têm mostrado que as mídias sociais podem desempenhar um papel importante como fonte de dados para análise de crises e também para entender o comportamento das pessoas durante uma pandemia.

Com o objetivo de auxiliar o monitoramento da saúde pública e também para apoiar a tomada de decisão de profissionais, diversos sistemas de monitoramento vêm sendo desenvolvidos para classificar grandes quantidades de dados oriundos das mídias sociais. A análise sistemática destes dados pode ajudar os governantes e profissionais da saúde a identificar as questões mais comuns entre os usuários. Dentre as plataformas de mídias sociais, O Twitter é uma das mais populares. Existe aproximadamente 200 milhões de usuários registrados nesta plataforma e que publicam mais de 500 milhões de tuítes diariamente. Portanto, pode-se aproveitar desse alto volume e troca frequente de informações para se conhecer as dúvidas sobre determinadas crises.

Diante deste cenário, foi realizado um estudo exploratório de mineração de opinião das mensagens de usuários do Twitter relacionadas à COVID-19. O estudo focou na análise das perguntas dos usuários, pois assume-se que seja um tipo de mensagem apropriada para se compreender as principais dúvidas das pessoas sobre a atual pandemia. A técnica de modelagem de tópicos foi usada para identificar os principais tópicos discutidos pelas pessoas no Twitter. Ainda foi desenvolvido um modelo de Reconhecimento de Entidades Mencionadas (REM) para identificar as principais menções a um grupo pré-definido de entidades.

O trabalho está organizado da seguinte maneira. Na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 é apresentada a metodologia desenvolvida neste trabalho. Na Seção 4 são apresentados e discutidos os resultados obtidos nos experimentos. Por fim, na Seção 5 são apresentadas as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

Diversos estudos vêm investigando as questões publicadas por usuários em redes sociais. Em [Zhao and Mei 2013] foi apresentado um estudo que analisou as necessidades publicadas por usuários em redes sociais. Para isto, os autores analisaram um representativo volume de questões postadas pelos usuários do Twitter. Paul *et al.* [Paul et al. 2011] conduziram um estudo sobre os tipos de perguntas que os usuários postam no Twitter. Mais recentemente, diversos autores vêm considerando as mensagens publicadas no Twitter como fonte de dados para lidar com graves problemas sociais, tais como desastres e pandemias. Em [Zahra et al. 2017], os autores investigaram o uso das características dos usuários do Twitter baseadas em diferentes localizações durante desastres. Eles examinaram a atividade dos usuários desta plataforma durante os terremotos na Itália e em Myanmar. Sinnenberg *et al.* [Sinnenberg et al. 2017] desenvolveram um estudo sobre o uso do Twitter na saúde pública. Para isso, os autores definiram uma taxonomia para descrever o uso do Twitter e caracterizar o estado atual da plataforma na pesquisa de saúde pública. Alaa *et al.* [Abd-Alrazaq et al. 2020] desenvolveram um estudo para identificar os principais tópicos relacionados à pandemia de COVID-19 nos tuítes.

Apesar dos trabalhos relacionados assumirem a importância de considerar as mídias sociais, nenhum destes trabalhos foca em perguntas sobre a pandemia de COVID-19. Na verdade, para o melhor de nosso conhecimento, esse é o primeiro trabalho que analisou um tipo específico de postagem, qual seja, as perguntas em português do Brasil.

3. Materiais e Métodos

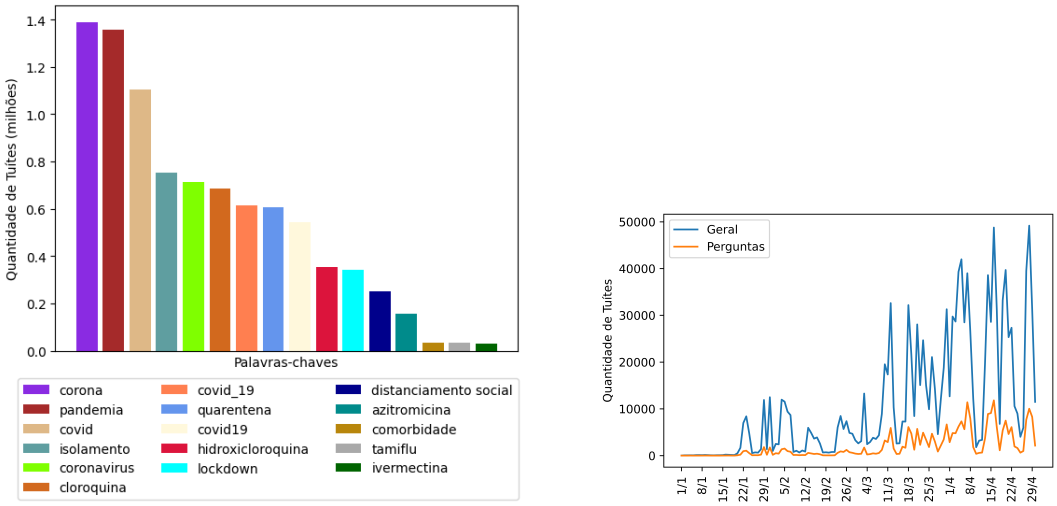
3.1. Coleta de Dados

A biblioteca Twitterscraper[1] de Python foi utilizada para coletar os tuítes relacionados à COVID-19 em português entre janeiro a abril de 2020. A opção `--lang` do Twitterscraper permitiu coletar as mensagens que estavam em português. Foi utilizado um conjunto de palavras-chaves para identificar os tuítes que faziam menção à COVID-19. Foram consideradas as seguintes palavras-chaves: corona, coronavírus, COVID, COVID19, COVID-19, distanciamento social, isolamento, *lockdown*, quarentena, cloroquina, hidroxicloroquina, ivermectina, tamiflu, azitromicina, pandemia e comorbidade.

O conjunto de dados apresentou 2.619.215 tuítes. A Figura 1a mostra a distribuição das mensagens coletadas para cada palavra-chave. Neste gráfico, é possível observar que *corona* e *COVID* são as formas mais comuns de nomear a doença. Além disso, é possível observar que cloroquina é o medicamento mais popular entre os usuários.

Com o objetivo de identificar as perguntas dos usuários nas mensagens, os textos foram segmentados em sentenças e as sentenças que terminavam com o caracter ? foram consideradas perguntas. Após essa etapa, a coleção de dados apresentou 2.313.070 perguntas. Figura 1b mostra a distribuição diária de tuítes coletados. O gráfico mostra que a quantidade de perguntas segue o fluxo da quantidade geral de postagens sobre a doença.

[1] https://pypi.org/project/twitterscraper



(a) Agrupamento por palavras-chaves. **(b)** Número de postagens.

Figura 1. Distribuição do número de postagens.

É possível ainda notar que houve um crescimento sobre a discussão relativa à COVID-19 a partir de meados do mês de março, devido ao registro do primeiro brasileiro morto por causa da doença. O gráfico também mostra, através dos pontos mais baixos, que as pessoas postam menos nos finais de semana ou feriado.

3.2. Pré-Processamento dos Dados

Inicialmente, foi realizada uma limpeza dos textos que inclui a remoção de *hyperlinks*, e-mails, *hashtags*, caracteres que indicam a replicação de tuítes (RT), duplos espaços em branco e pontuação. Então cada pergunta foi convertida em uma lista de palavras e as *stopwords* foram removidas. Dentre os termos gerados, foram mantidos apenas os termos classificados como substantivos, adjetivos, verbos e advérbios.

3.3. Modelagem de Tópicos

O modelo *Latent Dirichlet Allocation* (LDA) foi usado para identificar os tópicos relevantes sobre a COVID-19. LDA permite descobrir tópicos latentes usando a distribuição de probabilidade multinomial dos termos em documentos não estruturados. Nos experimentos, a variação de tópicos foi de 1 a 60, e foi selecionado o modelo com o maior valor de pontuação de coerência (*coherence score*). A geração dos tópicos foi executada para cada mês com o objetivo de identificar a mudança temporal dos tópicos.

O modelo que gerou 20 tópicos foi selecionado com um valor médio de pontuação de coerência entre os quatro meses de 0.674. Este valor é usado como uma métrica que calcula a concordância de um conjunto de pares e subconjunto de palavras e as probabilidades associadas [Röder et al. 2015]. Em geral, os tópicos são interpretados como sendo coerentes se todos os termos, ou a maioria destes, são relacionados.

3.4. Reconhecimento de Entidades Mencionadas

Os métodos de Reconhecimento de Entidades Mencionadas (REM) são baseados principalmente em modelos de aprendizagem de máquina. A extração de entidades mencionadas no Twitter é uma tarefa ainda mais desafiadora, pois os tuítes são curtos e, portanto, são mais difíceis de se interpretar quando comparados com textos mais longos. Os textos

curtos também apresentam muitas variações linguísticas e tendem a ser menos corretos em termos gramaticais. Além disso, a maioria das pesquisas sobre ferramentas de processamento de linguagem natural são voltadas para o idioma inglês [Santos et al. 2015].

Para esta tarefa, foi empregada a ferramenta spaCy [Honnibal 2020] que é baseada em modelos de redes neurais. Foi criado um modelo para reconhecer as seguintes entidades: Medicamentos (DRUG), Doenças (DIS), Pessoas (PER) e Organizações (ORG). A razão para a escolha destas entidades é que as informações relativas a essas categorias são essenciais durante uma crise de pandemia. Apesar de existir bastante interesse sobre o desenvolvimento de modelos REM em Português [Santos et al. 2019], nenhum dos modelos investigados poderia ser usado diretamente para reconhecer as entidades relacionadas à COVID-19 porque não foram treinados para reconhecer tais entidades.

4. Resultados e Discussões

4.1. LDA Aplicado

O modelo final de LDA gerou 20 tópicos utilizando os valores padrão de parâmetros do modelo Gensim LDA MultiCore. Em janeiro de 2020, a notícia de pessoas infectadas com a COVID-19 na China foi recebida com um certo grau de ironia e incredulidade. Muitos usuários faziam uma brincadeira relacionando o nome da doença com o nome de uma famosa marca de cerveja, tal como na pergunta *"Se eu beber cerveja eu pego corona vírus?"*. Neste período, pôde-se observar que a preocupação para muitos usuários do Twitter era saber se a doença poderia atrapalhar a realização do carnaval no Brasil.

Observa-se uma mudança da percepção dos usuários em relação à doença no mês seguinte. As pessoas começaram a fazer perguntas para entender melhor a terminologia da doença. Por exemplo, muitos usuários questionam a diferença entre epidemia e pandemia. Ainda no mês de fevereiro, um questionamento comum foi saber o alcance da doença em países afastados da China.

No mês de março ocorreu uma nítida mudança dos questionamentos dos usuários em face à doença. Termos como "morrer" e "sobreviver" foram bastante utilizados. Isto é devido, especialmente, pelo fato da morte do primeiro brasileiro por COVID-19. Um outro tópico que apareceu com frequência no mês de março foi a preocupação das pessoas com as ações governamentais para minimizar o efeito da pandemia na saúde pública e na economia do País.

No mês de abril, o principal foco dos usuários foi nas ações de governantes, tais como isolamento social e *lockdown*, para lidar com o alastramento da doença. Um outro tópico bastante ativo nos questionamentos foram as discussões políticas. Por exemplo, houve diversos questionamentos sobre a decisão do presidente do Brasil de demitir o Ministro da Saúde durante o período de pandemia. Finalmente, uma outra dúvida bastante comum foi a eficácia dos remédios e outros tratamentos para combater a doença.

4.2. Modelo REM Aplicado

Nesta seção, discutem-se a criação e a performance do modelo REM proposto e também a aplicação do modelo no conjunto de dados coletados. Foram consideradas quatro entidades: Doença (DIS), Medicamento (DRUG), Organização (ORG) e Pessoa (PER). Para cada tipo de entidade, foram usadas as métricas de *Precisão* (P), *Revocação* (R) e *F1 Score* (F_1). Com o objetivo de treinar o novo modelo, foi realizada a anotação manual

dos dados de treino. A anotação ocorreu em 2.000 das perguntas coletadas do Twitter selecionadas aleatoriamente. Este *dataset* anotado foi dividido em 80% para treino e 20% para teste. Tabela 1 apresenta os resultados alcançados pelo modelo para cada categoria.

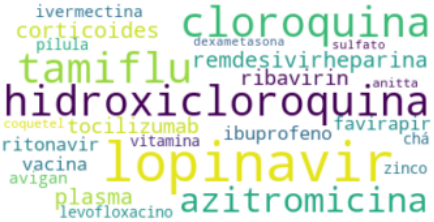
Tabela 1. Resultados alcançados pelo modelo REM.

	Doença	Medicamentos	Organização	Pessoa
Precisão	98,97	87,80	93,40	95,00
Revocação	96,50	88,26	92,92	96,44
F_1 *Score*	97,72	88,04	93,16	95,71

De acordo com os resultados obtidos, a média de F1 entre as quatro entidades foi 93,65. O desempenho mais baixo do modelo foi na categoria de Medicamentos. Isto ocorreu porque o modelo identificou diversas substâncias que têm efeito sobre as pessoas, mas que não foram anotadas como medicamento. É o caso, por exemplo, dos vários tipos de chá que são mencionados pelas pessoas como alternativas para o tratamento da doença.

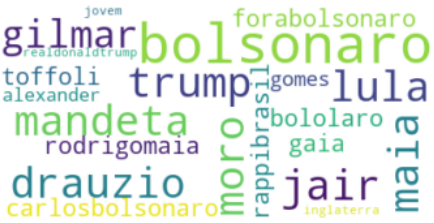
Na Figura 2a, são apresentadas as menções mais frequentes em perguntas relacionadas à doença. É possível observar que a doença é nomeada de diversas maneiras. As pessoas também fizeram muitos questionamentos sobre outras doenças. Comumente, as perguntas sobre as demais doenças fazem uma relação com a COVID-19. Por exemplo, um usuário perguntou *"Alguém sabe a diferença dos sintomas de H1N1 e COVID19?"*.

(a) Doença (DIS)



(b) Medicamento (DRUG)

(c) Organização (ORG)



(d) Pessoa (PER)

Figura 2. Menções às entidades identificadas pelo modelo de REM.

Na Figura 2b, são apresentadas as menções mais frequentes em perguntas relacionadas aos principais medicamentos usados para o tratamento da doença. As dúvidas mais comuns foram sobre o uso e a eficácia dos medicamentos. Interessantemente, o termo chá foi identificado pelo modelo de REM como uma menção a medicamentos. Ao se analisar o contexto do uso da menção de chá nas questões dos usuários, é possível identificar a dúvida comum se alguns tipos de chá poderiam ser usados no combate à COVID-19.

Na Figura 2c, são apresentadas as menções mais frequentes em perguntas relacionadas às organizações diante da pandemia. A menção a Estado costuma aparecer nas perguntas se referindo ao Governo Federal ou à alguma das Unidades da Federação. Ao se

analisar o contexto em que essas menções aparecem, é possível observar que as perguntas costumam ter uma crítica a essas organizações.

Na Figura 2d, são apresentadas as menções mais frequentes às pessoas em perguntas relacionadas a pandemia. Os nomes mais frequentes foram de políticos e pessoas que aparecem na grande mídia. Este é o caso, por exemplo, do médico Dráuzio Varela que ficou em bastante evidência após algumas declarações polêmicas na TV.

5. Conclusões e Trabalhos Futuros

Atualmente, muitas pessoas fazem uso das mídias sociais como o Twitter para expressar diversos tipos de questionamentos sobre a doença. A compreensão das dúvidas comuns dos usuários dessas redes sociais pode ser um ponto de partida para projetar mensagens estratégicas para campanhas de saúde e estabelecer um sistema de comunicação eficaz durante a pandemia para um melhor enfrentamento à doença. O Twitter não divulga dados sobre o perfil de seus usuários, tais como idade ou gênero. Assim, não foi possível realizar uma análise estratificada dos usuários. Como trabalhos futuros, pretende-se buscar obter esses dados e inclui-los na análise. Pretende-se também investigar a aplicação dos métodos desenvolvidos neste trabalho em outras fontes de mídias sociais.

Referências

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., and Shah, Z. (2020). Top concerns of tweeters during the covid-19 pandemic: infoveillance study. *JMIR*, 22(4):e19016.

Honnibal, M. (acessado em 26 de maio de 2020). *spaCy*.

Paul, S. A., Hong, L., and Chi, E. H. (2011). Is twitter a good place for asking questions? a characterization study. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Santos, J., Consoli, B., dos Santos, C., Terra, J., Collonini, S., and Vieira, R. (2019). Assessing the impact of contextual embeddings for portuguese named entity recognition. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 437–442.

Santos, J. T. L., Anastacio, I. M., and Martins, B. E. (2015). Named entity disambiguation over texts written in the portuguese or spanish languages. *IEEE Latin America Transactions*, 13(3):856–862.

Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., and Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. *American journal of public health*, 107(1):e1–e8.

Zahra, K., Ostermann, F. O., and Purves, R. S. (2017). Geographic variability of twitter usage characteristics during disaster events. *Geo-spatial information science*, 20(3):231–240.

Zhao, Z. and Mei, Q. (2013). Questions about questions: An empirical analysis of information needs on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1545–1556.