FREDERICO CHAVES CARVALHO

DIEGO MARIANO

MARCOS MATOS

Painel SimplificaSUS

Proposta apresentada ao Hackathon do Projeto InovaSUS: painel de apresentação de evidências científicas coletadas de artigos acadêmicos e avaliadas por meio de técnicas de *machine learning*, mineração de dados e processamento de linguagem natural.

Belo Horizonte

2022

Resumo

Artigos acadêmicos condensam os principais achados de pesquisas científicas, sendo vital para tomada de decisão principalmente quando se trata de dados de saúde. Entretanto, devido à linguagem técnica utilizada nesse tipo de manuscrito, sua compreensão se torna complexa para profissionais que não tenham maior afinidade com a pesquisa científica. Assim, é de grande importância a construção de estratégias que melhorem a comunicação entre profissionais das áreas da saúde e acadêmicos. Aqui apresentamos uma proposta de painel científico, denominado **SimplificaSUS**, que engloba evidências retiradas de artigos científicos avaliadas por meio de técnicas de *machine learning* e processamento de linguagem natural. Como estudo de caso, apresentamos uma comparação com o "Painel de Evidências Científicas sobre Tratamento Farmacológico e Vacinas – COVID-19". Nossos resultados demonstram como técnicas de aprendizagem de máquina e processamento de linguagem natural podem ser de grande ajuda na compreensão de dados levando em consideração o contexto. Um protótipo do painel está disponível em: http://inovasus.alfahelix.com.br/.

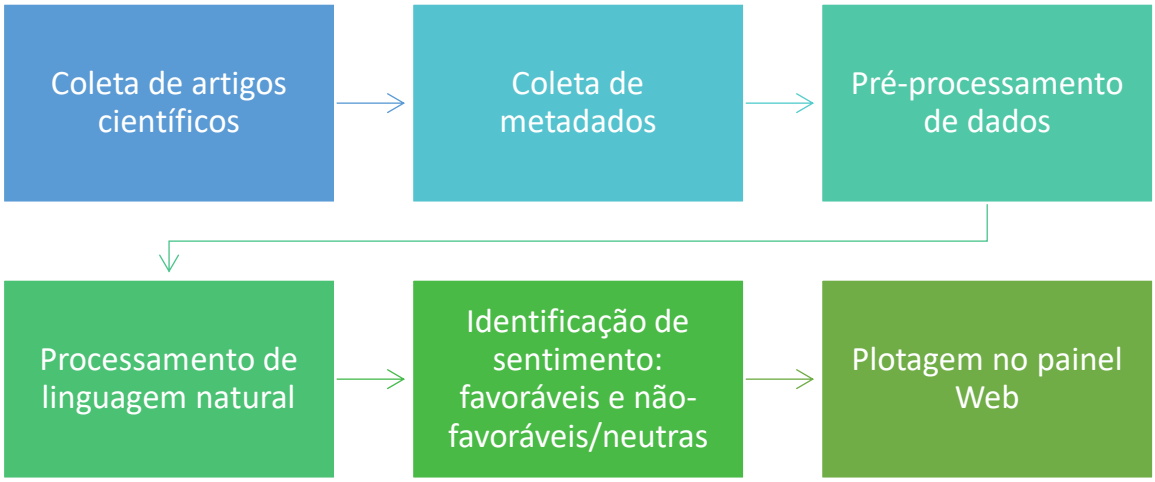
Introdução

Apresentamos uma proposta de painel **SimplificaSUS**. Nossa proposta visa coletar, analisar e apresentar evidências retiradas de artigos científicos avaliadas por meio de técnicas de *machine learning* e processamento de linguagem natural. Nosso objetivo principal é fornecer uma ferramenta amigável que facilite a compreensão de artigos científicos e que forneçam visualizações que complementem as ferramentas já existentes.

As técnicas de *machine learning* (aprendizagem de máquina) permitem utilizar algoritmos e modelos matemáticos para compreensão de padrões nos dados, ajudando na descoberta de conhecimento e tomada de decisão. Processamento de linguagem natural (PLN) são técnicas de aprendizado de máquina que se baseiam na tentativa de compreensão da linguagem humana por meio de software.

Assim, técnicas de *machine learning* e PLN podem ser de grande ajuda na sumarização de informações presentes em artigos científicos. Para isso, estabelecemos a seguinte metodologia:

Figura 1. Passos da metodologia.



Para avaliar nossa proposta, realizamos uma comparação com um painel de divulgação produzido pela equipe do Ministério da Saúde. Detalhes são apresentados a seguir.

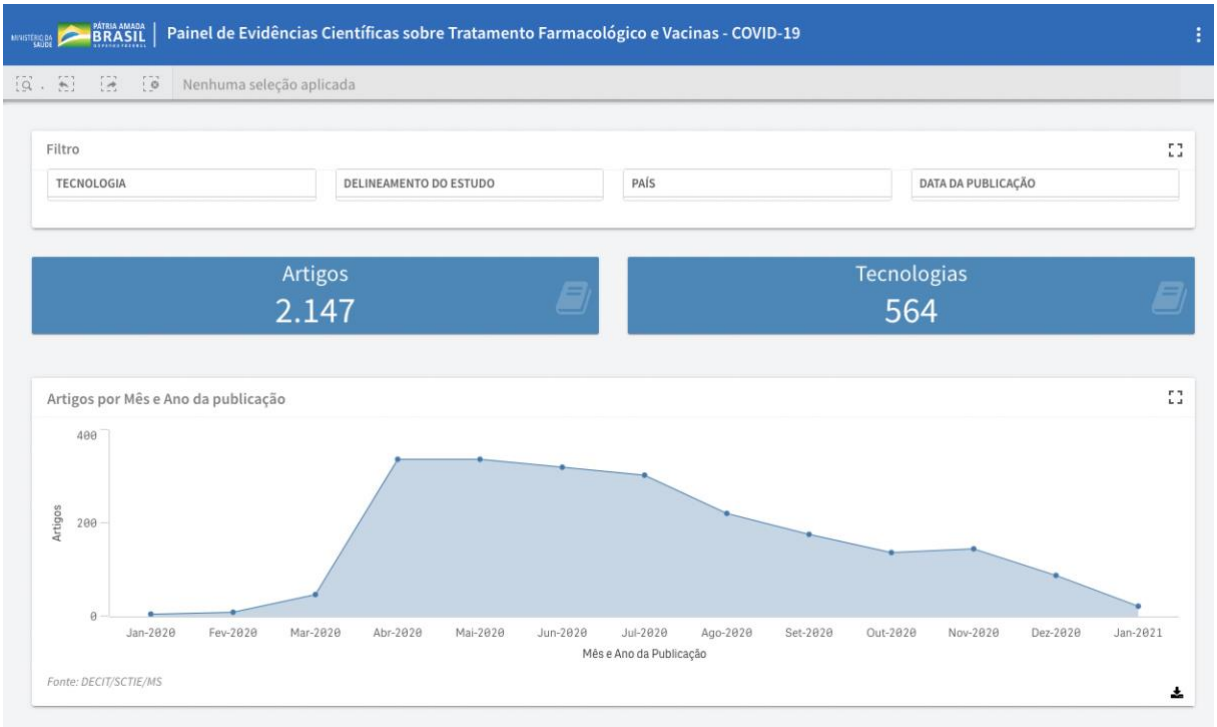
Estudo de caso

O Painel de Evidências Científicas sobre Tratamento Farmacológico e Vacinas - COVID-19[1], visa “reunir em tempo real as informações sobre publicações técnico-científicas de revistas

[1] Disponível em https://infoms.saude.gov.br/extensions/evidencias_covid/evidencias_covid.html. Acesso em 17 de junho de 2022.

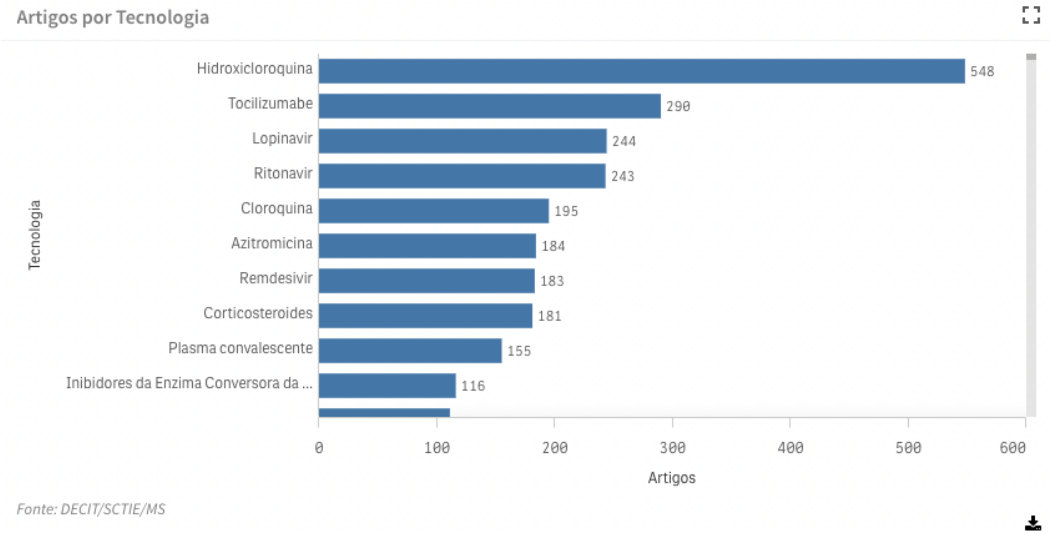
indexadas e em pré-impressão que investigam a eficácia, segurança e efetividade de medicamentos e produtos biológicos usados para tratamento e prevenção da doença provocada pelo novo coronavírus". Até a data avaliada (17/06/2022), o painel condensava informações de 2147 artigos. Entretanto, a grande quantidade de dados pode ser prejudicial para a compreensão do problema, uma vez que a gama de dados pode levar a entendimentos extremamente diversos das informações contidas dificultando uma clara transmissão do conhecimento ali contido.

Figura 2. Visão geral do Painel de Evidências Científicas sobre Tratamento Farmacológico e Vacinas - COVID-19. Fonte: https://infoms.saude.gov.br/extensions/evidencias_covid/evidencias_covid.html.



Por exemplo, o gráfico que apresenta uma sumarização das tecnologias encontradas nos artigos nos dá uma falsa percepção sobre a natureza dos dados.

Figura 3 Tecnologias detectadas pelo Painel de Evidências Científicas sobre Tratamento Farmacológico e Vacinas - COVID-19. Fonte: https://infoms.saude.gov.br/extensions/evidencias_covid/evidencias_covid.html.



Assim, partindo desse exemplo, utilizamos técnicas de processamento de linguagem natural NLTK para remover palavras de baixa semântica. Em seguida, buscamos padrões que permitam classificar as informações presentes em resumos para detectar expressões favoráveis ou não favoráveis a presença de certa tecnologia citada.

Por exemplo, o artigo "*A systematic review and meta-analysis on chloroquine and hydroxychloroquine as monotherapy or combined with azithromycin in COVID-19 treatment*" de Ghazy *et al.* (2020) apresenta o seguinte trecho:

> "*Overall VQR [virologic cure rate], and that at days 4, 10, and 14 among patients exposed to HCQ [Hydroxychloroquine] did not differ significantly from the SC [standard care],* [...] *despite the scarcity of published data of good quality, the effectiveness and safety of either HCQ alone or in combination with AZM [azithromycin] in treating COVID-19 cannot be assured*".

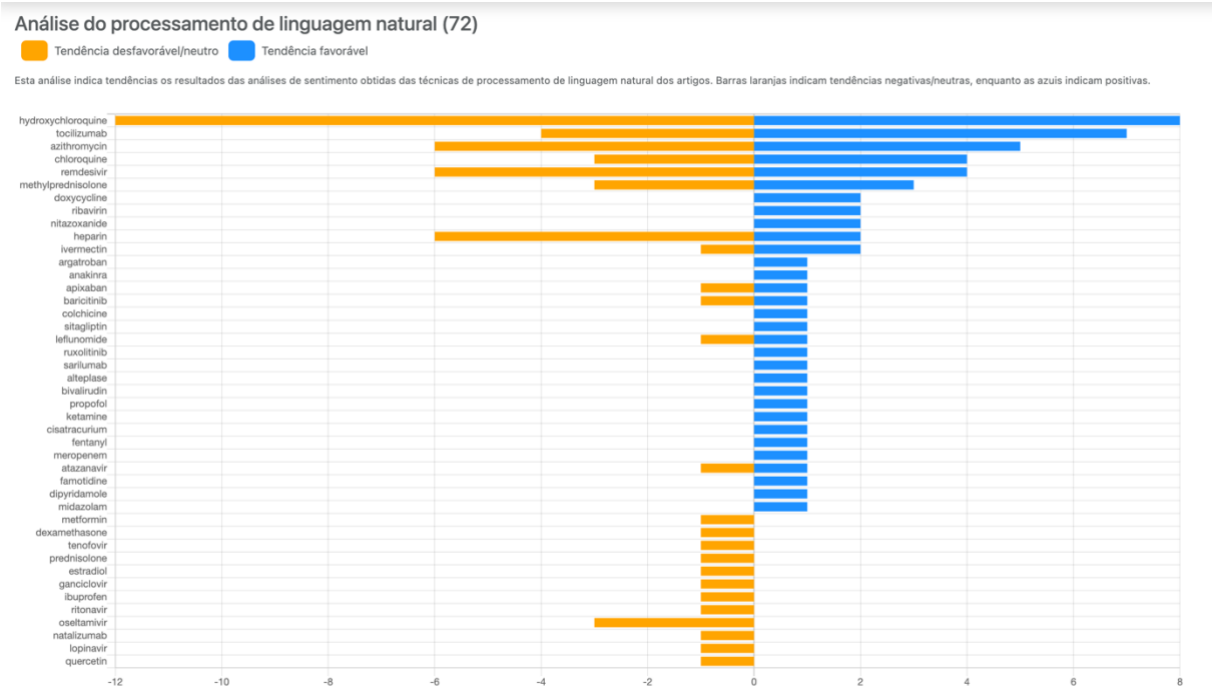
Em tradução livre:

> "A taxa de cura virológica geral, e que nos dias 4, 10 e 14 entre os pacientes expostos a Hidroxicloroquina não diferiu significativamente do cuidado padrão [...] apesar da escassez de dados publicados de boa qualidade, a eficácia e segurança da Hidroxicloroquina sozinha ou em combinação com o azitromicina no tratamento do COVID-19 não pode ser garantido".

Note que a análise padrão apenas detecta a citação ao termo "*hydroxychloroquine*" (Hidroxicloroquina). Entretanto, apenas com uma análise manual aos dados podemos detectar o contexto a qual a citação é apresentada (neste caso, vemos que a eficácia e segurança não puderam ser garantidos).

Técnicas de processamento de linguagem natural permitem essa detecção automática, identificando assim padrões favoráveis ou não. Por exemplo, o gráfico a seguir apresenta o resultado da análise de processamento de linguagem natural realizada com uma amostra de dados obtidos do Painel de Evidências Científicas sobre Tratamento Farmacológico e Vacinas - COVID-19.

Figura 4. Resultados das análises de sentimento usando PLN. Fonte: http://inovasus.alfahelix.com.br/.



Observe como as técnicas foram capazes de descrever com mais detalhes, o contexto a qual a citação foi apresentada. Nesta figura, as barras laranja indicam descrições consideradas “não favoráveis ou neutras”, enquanto as barras azuis são “favoráveis”.

A seguir descrevemos com mais detalhes os principais resultados da nossa proposta.

Resultados

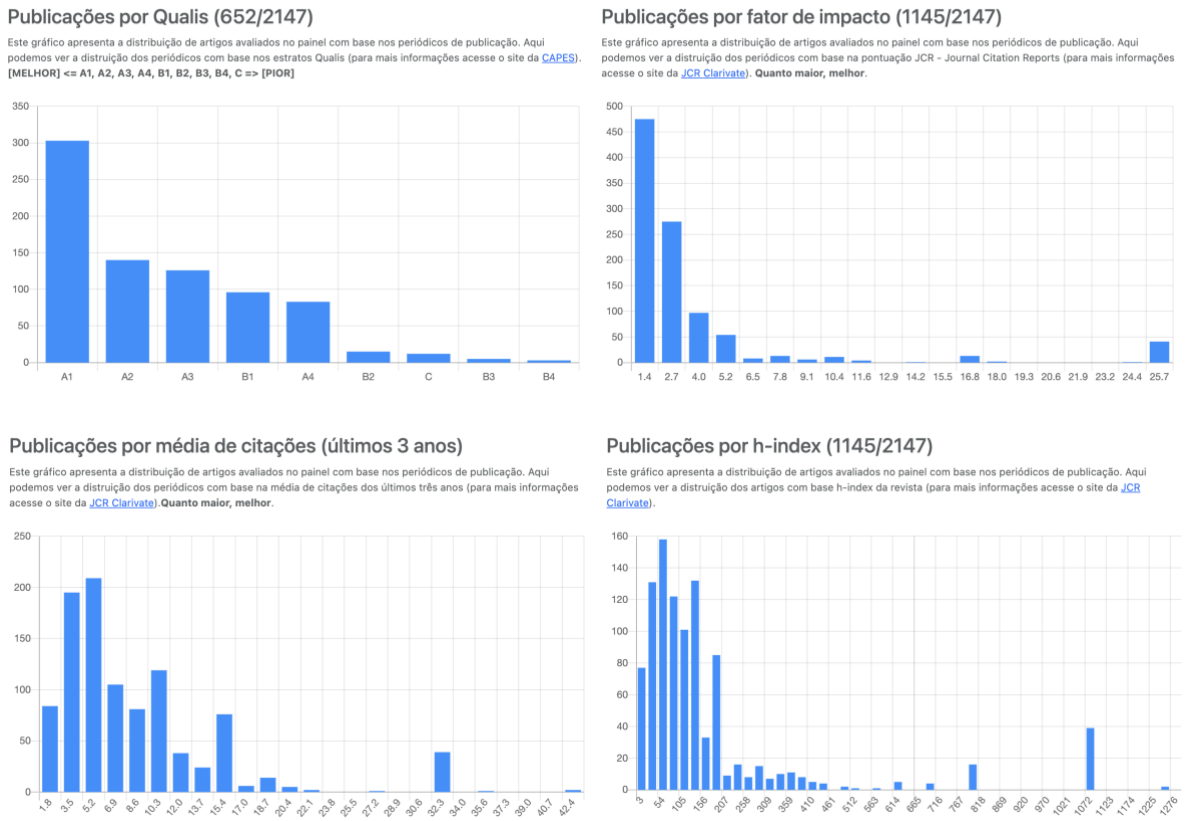
Um protótipo do painel foi disponibilizado em https://inovasus.alfahelix.com.br/. A figura a seguir apresenta uma visão geral da interface:

Figura 5. Visão geral da interface. Aqui vemos os artigos coletados sobre CoViD-19. Fonte: http://inovasus.alfahelix.com.br/.

No gráfico a esquerda podemos ver as publicações por data de publicação. Esse gráfico foi incluído apenas para indicar aos usuários que os dados utilizados são os mesmos disponíveis no painel de evidências científicas. À direita realizamos uma análise inicial das publicações. Nessa análise verificamos a relevância de onde o manuscrito foi publicado. Observamos então, que aproximadamente metade dos artigos foi publicada em revistas sem fator de impacto ou estratos Qualis (métrica usada pela CAPES que divide periódicos em estratos que vão de A a C, sendo os com estrato A de maior relevância acadêmica). Esses índices são métricas usadas para avaliar periódicos e são uma boa indicação de que o artigo em questão passou pelo processo de revisão por pares.

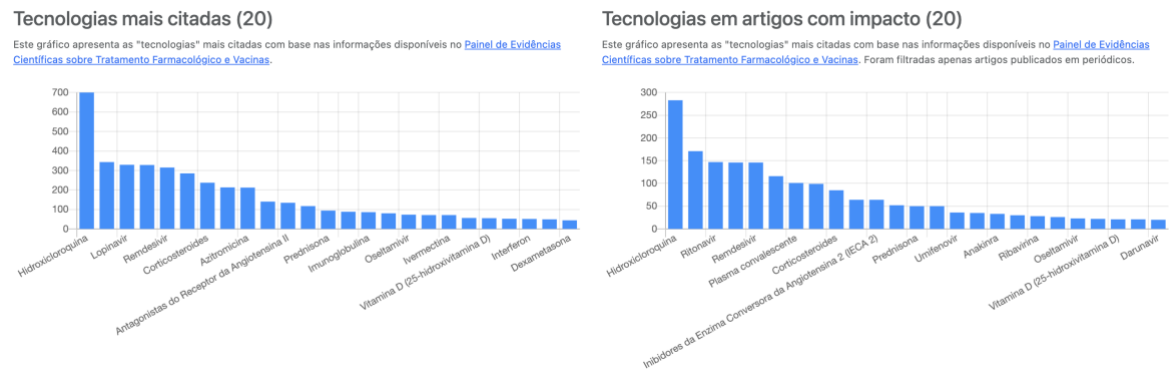
Ao avaliar artigos, é importante analisar onde ele foi publicado, uma vez que revistas de baixo impacto ou servidores de deposito de artigos de pré-impressão (preprint) tendem a ter processos de revisão menos rigorosos (ou até mesmo não ter nenhum processo de revisão). A figura a seguir ilustra algumas das visualizações que indicam a qualidade dos periódicos que publicaram os artigos:

Figura 6. Gráficos de avaliação dos periódicos onde estão as publicações. Fonte: http://inovasus.alfahelix.com.br/.



A seguir, reproduzimos o gráfico de tecnologias detectadas pelo painel de evidências científicas de CoViD-19 (figura a seguir; lado esquerdo). Neste caso, tecnologias indicam medicamentos, vacinas ou outros componentes relacionados ao contexto dos artigos. Entretanto, essa comparação não nos mostrou variação entre publicações sem impacto e publicações com impacto.

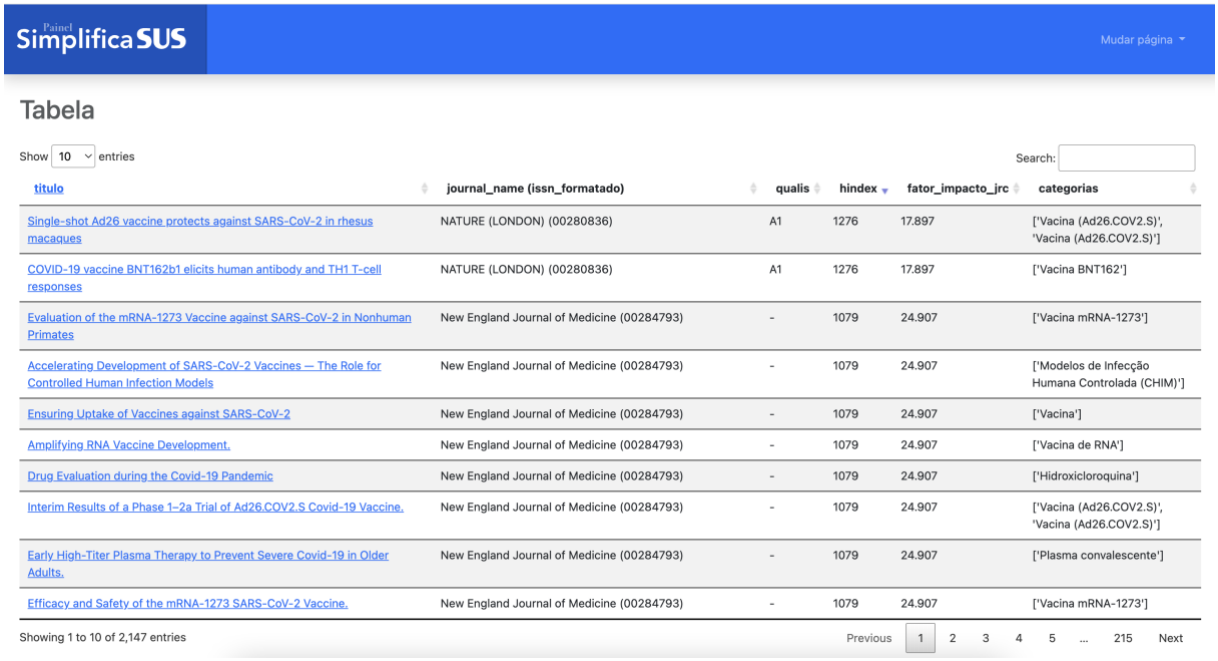
Figura 7. 20 tecnologias mais citadas. No lado direito, filtramos essas tecnologias por artigos com maior fator de impacto. Fonte: http://inovasus.alfahelix.com.br/.



Por fim, apresentamos uma tabela interativa com os dados coletados de artigos. Essa tabela permite a exploração dos artigos, avaliando diversos metadados, títulos, resumos, quais

tecnologias citam, e qual o impacto do periódico a qual foram publicados (por meio do estrato QUALIS, H-Index da revista e do fator de impacto JCR).

Figura 8. Tabela interativa de artigos. Fonte: http://inovasus.alfahelix.com.br/.



Avaliação e validação da interface do painel

O painel foi avaliado em duas etapas:

- **alfa**: avaliações realizadas pelo time de desenvolvedores;
- **beta**: avaliações feitas com três usuários externos que não tinham conhecimento prévio da ferramenta.

Os potenciais usuários entrevistados demonstraram satisfação com o visual apresentado. São aspectos destacados pelos usuários: os gráficos escolhidos, a forma de disposição das informações, a fluidez da plataforma, sobriedade, possibilidade de comparações rápidas, clareza das informações, da disponibilização do conhecimento de forma direta e na integra, abrangência dos dados contidos e da fácil compreensão da avaliação de qualidade do conhecimento disponibilizado. Além dessas colocações, os usuários apresentaram propostas para a melhoria da interface, em geral, relacionadas a responsividade da aplicação (sugestões que foram acatadas e incorporadas ao código). Por fim, os usuários externos demonstraram interesse em acompanhar o avanço da plataforma, bem como sua funcionalidade no futuro e quais impactos trarão quando estiverem em pleno funcionamento.

Discussão sobre os aspectos avaliados na Hackathon

Viabilidade técnica

A viabilidade técnica do projeto se comprova pela sua eficácia em coletar e tratar informações de maneira automatizada e com boa precisão. Além disso, os protótipos apresentados demonstram a capacidade de produção de ferramentas finalísticas com interface interativa, simples e amigável para usuários finais.

Os scripts utilizados foram validados através da aplicação para processamento automático dos dados de publicações científicas referentes a seis doenças de relevância no cenário atual brasileiro: Covid-19, Zika, Dengue, Febre Amarela e Varíola do Macaco (esses quatro últimos ainda em desenvolvimento). Neste ponto, observou-se que os algoritmos possuem bom desempenho e fornecem dados corretamente classificados, o que permite sumarização automática dos dados dos artigos, facilitando a compreensão do contexto.

Aplicabilidade

O painel apresenta soluções pertinentes para o problema de sumarização de conhecimento científico. O painel pode ser acessado em qualquer navegador e por qualquer usuário com acesso à internet.

Criatividade e originalidade

Apesar do projeto se inspirar em ferramentas já existentes, como o Painel de Evidências Científicas de CoViD-19, ele apresenta uma série de fatores originais, que dão uma nova perspectiva na análise desses dados. Por exemplo, apresentamos visualizações que consideram o impacto do artigo na comunidade científica. Além disso, propusemos técnicas de aprendizagem de máquina para detectar o contexto do uso de tecnologias (como vacinas e fármacos) citadas em artigos.

Barreira de entrada

Uma das dificuldades encontradas foi na coleta automatizada dos dados. Há limitações nas APIs públicas usadas para coleta de dados, que não puderam ser contornadas devido a restrição de tempo da Hackathon. Acreditamos que com mais tempo possamos coletar mais dados.

Aderência ao projeto Inova Dados

O projeto em questão apresenta problemas abordados nos seminários. A sumarização de artigos por meio de painéis evidências científicas são de grande valia em vários aspectos e a compreensão correta dos dados, levando em consideração o escopo e contexto, são fundamentais.

Elemento Futuro

Esperamos no futuro a aplicação da estratégia para análises de artigos para outros tipos de doenças virais, como varíola, febre amarela, zika, dengue e HIV.

Metodologia

Dados dos artigos foram coletados do Painel de Evidências Científicas disponível em https://infoms.saude.gov.br/extensions/evidencias_covid/evidencias_covid.html.

Metadados dos artigos foram coletados na API do PubMed, disponível em https://www.ncbi.nlm.nih.gov/home/develop/api/.

Dados foram processados usando scripts Python. Interface foi desenvolvida usando Laravel framework. Interface gerada usando Bootstrap Framework. Gráficos gerados usando a biblioteca Chart.js (https://www.chartjs.org/).

Detalhes da metodologia estão disponíveis em https://github.com/fccarvalho2/InovadadosSUS.

Conclusões

Após desenvolvermos esta proposta do Painel SimplificaSUS. A qual engloba o uso de tecnologias web em combinação com técnicas de aprendizagem de máquina, mineração de dados e processamento de linguagem natural para extrair conhecimentos de artigos científicos e convertê-los para uma linguagem mais compreensível por meio de visualizações de dados. Como diria o ditado: “uma imagem, vale mais do que mil palavras”. Nossa proposta estende funcionalidades de propostas existentes, trazendo uma série de inovações que poderão fazer a diferença em relação a acessibilidade, publicidade, fomentar debates, formação e aplicabilidade deste conhecimento nas tomadas de decisão por profissionais da área de saúde. Atualmente realizamos uma implementação completa com dados de CoViD-19 e planejamos realizar análises de artigos relacionados a outras doenças virais, como varíola, febre amarela, dengue e

Zika. Códigos-fonte das aplicações foram compartilhados *open source* no endereço: https://github.com/fccarvalho2/InovadadosSUS. Um vídeo descrevendo a proposta está disponível em https://youtu.be/rjjsh-I5r_A. Um protótipo funcional da aplicação está disponível em http://inovasus.alfahelix.com.br/.

Referências

Ghazy, R.M., Almaghraby, A., Shaaban, R. et al. A systematic review and meta-analysis on chloroquine and hydroxychloroquine as monotherapy or combined with azithromycin in COVID-19 treatment. Sci Rep 10, 22139 (2020). https://doi.org/10.1038/s41598-020-77748-x